Alzheimer’s Disease Knowledge Graph

Athish Gopal Rajesh

December 17, 2025

Abstract

Alzheimer’s disease (AD) research is marked by a rapidly expanding and fragmented literature spanning biomarkers, clinical phenotypes, therapeutics, and molecular mechanisms. While retrieval augmented generation (RAG) systems can summarize individual documents, they often struggle to capture explicit relationships between biomedical entities, reason across heterogeneous sources, and produce answers that are transparently grounded in evidence.

This project presents an Alzheimer’s Disease centric Knowledge Graph (KG) and a Graph RAG question answering pipeline that produces answers grounded strictly from the retrieved graph facts. The KG integrates authoritative biomedical ontologies like MONDO, HPO, GO, PRO, ChEBI, and HGNC to provide stable identifiers and synonym resolution, together with curated Alzheimer’s resources from Alzforum to capture disease specific biomarkers, phenotypes, drugs, and pathways.

The system follows a reproducible end to end workflow from raw ontology files and curated HTML pages to normalized entities, Neo4j importable graph structures, and intent driven graph retrieval. A Graph RAG layer classifies user intent, executes targeted Cypher queries, constructs ultra compact graph derived context, and constrains a local large language model to answer only from retrieved evidence. By explicitly modeling biomedical relationships at the graph level, this approach enables grounded, interpretable question answering for AD research queries while mitigating hallucination and improving transparency relative to document centric RAG systems.

1. Introduction

Alzheimer’s disease (AD) is one of the most extensively studied neurodegenerative disorders, with thousands of research papers published across diverse subdomains such as biomarkers, genetics, clinical phenotypes, and therapeutic development. For researchers, clinicians, and pharmaceutical stakeholders, keeping pace with this rapidly growing body of work is increasingly challenging. Relevant findings are distributed across heterogeneous sources, expressed using inconsistent terminology, and often require cross referencing multiple studies to understand how biomarkers, drugs, pathways, and phenotypes relate to one another.

Traditional approaches to literature review and evidence synthesis are time consuming and do not scale well as the volume of research grows. More recently, retrieval augmented generation (RAG) systems have been proposed as a way to assist users by retrieving relevant documents and generating natural language summaries. However, standard RAG pipelines are largely document centric since they retrieve text chunks based on similarity and rely on large language models to infer relationships implicitly. In complex biomedical domains such as Alzheimer's disease, this often leads to shallow reasoning, missing cross document connections, and limited transparency about how answers are derived.

Knowledge graphs offer a complementary paradigm by explicitly representing entities like diseases, biomarkers, drugs, genes, pathways and their relationships in a structured, queryable form. When combined with curated biomedical ontologies, knowledge graphs can provide stable identifiers, controlled vocabularies, and a principled way to integrate heterogeneous data sources. However, on their own, knowledge graphs are not easily accessible to non expert users and typically require familiarity with query languages such as Cypher or SPARQL.

This project addresses these challenges by designing and implementing an Alzheimer's Disease centered Knowledge Graph coupled with a Graph RAG question answering pipeline. The core design principle is that ontologies provide the structural backbone and identifier while curated Alzheimer's resources provide domain specific content, and the knowledge graph schema connects them into a coherent disease centric neighborhood. Natural language questions are first classified by intent (e.g., biomarker, drug, phenotype), then answered by executing targeted graph queries, converting the results into compact, interpretable context, and constraining a large language model to generate answers strictly grounded in retrieved graph evidence.

By prioritizing structured graph retrieval over unstructured text inference, this approach aims to support reliable, transparent, and interpretable question answering for key Alzheimer's disease research questions, such as identifying biomarkers that change direction in specific biofluids, understanding the clinical trial landscape of therapeutics, and linking drugs to their molecular targets and pathways.

2. Previous Work

2.1 Document Centric Biomedical Question Answering and RAG

Retrieval augmented generation (RAG) has been proposed as a method to improve factual correctness in large language model (LLM) outputs by retrieving relevant documents prior to generation (Lewis et al., 2020). In biomedical domains, RAG style systems typically retrieve abstracts or passages from large corpora such as PubMed, then synthesize answers using neural language models. Biomedical question answering benchmarks such as PubMedQA have demonstrated the feasibility of answering scientific questions using abstract level evidence (Jin et al., 2019).

However, document centric RAG systems suffer from several limitations in highly technical biomedical domains. Retrieved text passages often contain partial or implicit information, requiring multi step reasoning across multiple documents. As a result, LLMs may hallucinate connections between entities or fail to correctly interpret relationships such as causal direction, experimental context, or biological mechanism (Liang et al., 2025). These limitations are especially pronounced in Alzheimer's research, where questions often require structured reasoning across biomarkers, disease stages, drugs, and molecular pathways rather than surface level text matching.

2.2 Alzheimer's Disease Knowledge Graphs

To address the complexity of Alzheimer's research data, several studies have constructed knowledge graphs that integrate heterogeneous biomedical information. The Alzheimer's Knowledge Base (AlzKB) integrates genes, chemicals, diseases, and anatomical entities from multiple public databases into a unified graph representation, enabling downstream tasks such as drug repurposing and hypothesis generation (Romano et al., 2024). Graph based analytics on AlzKB have demonstrated the value of explicit relational structure for translational research.

Other efforts have focused on ontology driven Alzheimer's knowledge graphs using graph databases such as Neo4j. For example, Spasov et al. (2024) proposed an ontology based AD knowledge graph that integrates clinical and diagnostic information, enabling structured querying and visualization of Alzheimer's related entities. More broadly, recent surveys highlight the growing role of Alzheimer's disease knowledge graphs in organizing complex biomedical knowledge and supporting data driven discovery (Dobreva et al., 2025).

Despite their strengths, most Alzheimer's knowledge graphs are designed primarily for offline analysis, visualization, or graph mining tasks. They do not directly support natural language question answering, nor do they integrate LLMs to translate structured graph facts into user friendly explanations.

2.3 Graph Augmented Question Answering for Alzheimer's Disease

Recent research has begun to explore the combination of knowledge graphs and LLMs for question answering. Graph augmented QA frameworks aim to retrieve structured graph neighborhoods and then use LLMs for explanation or summarization, reducing hallucinations compared to purely text based RAG systems. Li et al. (2024) proposed a dynamic co augmentation framework in which Alzheimer's specific knowledge graphs are iteratively constructed from literature and used to guide LLM reasoning over disease related questions.

More recently, Xu et al. (2025) investigated GraphRAG style approaches for Alzheimer's disease, demonstrating that graph based retrieval improves answer accuracy and interpretability compared to standard RAG baselines. These studies suggest that explicit modeling of entities and relationships is critical for complex biomedical reasoning tasks.

Nevertheless, existing graph augmented approaches often rely on automatically extracted or incomplete graphs, lack standardized biomedical identifiers, or do not clearly separate retrieval logic from generation. As a result, traceability to authoritative sources and consistency across heterogeneous data remain challenging.

2.4 Positioning of This Work

In contrast to prior approaches, this project emphasizes (i) ontology grounded normalization using authoritative biomedical ontologies, (ii) curated Alzheimer's specific resources for high quality domain content, and (iii) an intent aware Graph RAG pipeline that retrieves structured Neo4j subgraphs before generating answers. By grounding responses strictly in retrieved graph context, the system aims to combine the interpretability of knowledge graphs with the flexibility of LLM based explanation.

3. System Design and Implementation

3.1 End-to-End System Architecture

Designed and implemented an end to end Alzheimer's Disease centric Knowledge Graph (KG) and Graph RAG system as shown in Fig. 1., that integrates heterogeneous biomedical resources into a unified, queryable framework. The system architecture follows a layered pipeline that transforms raw biomedical data into structured graph representations and supports grounded question answering through targeted retrieval and controlled language generation.

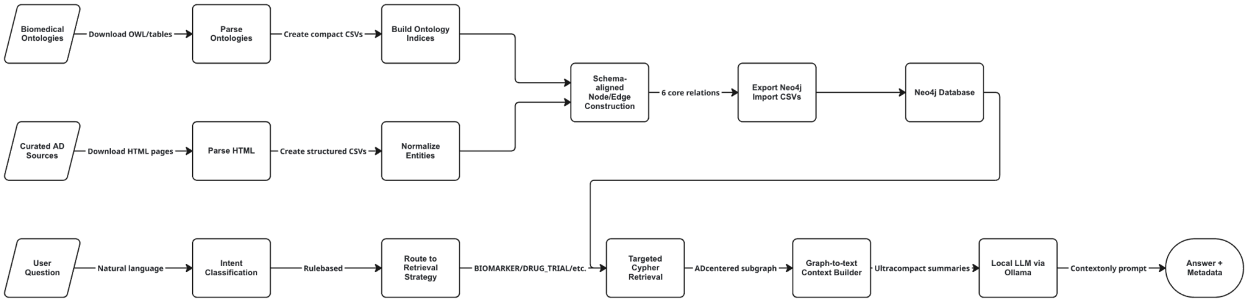

Fig. 1. End-to-end architecture diagram

At the foundation, authoritative biomedical ontologies are used to establish canonical identifiers, synonym resolution, and biological grounding. Curated Alzheimer's disease specific resources provide domain knowledge such as biomarkers, therapeutic trials, and mechanistic targets. These components are normalized into a schema aligned knowledge graph and stored in Neo4j. On top of the graph store, a Graph RAG runtime layer processes user queries through intent classification, structured Cypher retrieval, compact graph to text transformation, and grounded response generation using a local large language model.

A key design principle throughout this system is a strict separation of responsibilities: factual retrieval is handled exclusively by the knowledge graph, while the language model is constrained to summarizing and explaining retrieved evidence rather than inferring new facts.

3.2 Data Sources and Extracted Information

3.2.1 Biomedical Ontologies

Biomedical ontologies form the canonical backbone of the system, providing stable identifiers, controlled vocabularies, and synonym mappings necessary for robust normalization. Each ontology maintains its own identifier namespace, while the knowledge graph connects these namespaces through schema defined relationships.

Ontology	**Primary contribution**	**Terms processed**
MONDO	Disease identifiers and synonyms	1
HPO	Phenotype identifiers and hierarchy	3
GO	Biological processes and pathways	6
PRO	Protein entities and proteoforms	734
ChEBI	Chemical entities and drugs	3
HGNC	Gene symbols and IDs	6

Table 1. Ontologies used and their primary contributions

Ontology subsets were extracted into compact CSV files covering diseases, phenotypes, pathways, proteins, drugs, and genes. Protein entries were further enriched with gene symbol mappings where possible by matching ontology labels and synonyms against HGNC records, strengthening downstream biological connectivity.

3.2.2 Curated Alzheimer’s Disease Resources

To capture domain specific knowledge beyond what ontologies provide, we incorporated curated Alzheimer’s resources from Alzforum. All raw HTML pages were downloaded and stored locally prior to parsing, ensuring reproducibility and allowing re-processing without repeated web access.

Source	**Parsed Entities**	**Volume**
AlzPedia	Gene/protein entities and narrative sections	23
AlzBiomarker	Biomarkers and effect size comparisons	319

Therapeutics	Drug trials, targets, and mechanisms	2171

Table 2. Curated sources and their extracted outputs

AlzBiomarker data were parsed into structured tables separating biomarker definitions from comparison level statistical effects. Therapeutic data were extracted from search-result tables and deduplicated to construct drug level summaries of trial status and biological targets.

3.3 Knowledge Graph Schema Design

The knowledge graph schema is explicitly Alzheimer's disease centric and designed to support explainable, evidence driven retrieval. Core node types include: Disease, Biomarker, Drug, Phenotype, Pathway, Gene, Protein, Trial, Company, Mechanism, and Fluid.

Relationships were selected to capture both clinical associations and mechanistic links.

Relationship	**Direction**	**Purpose**	**Key properties**
HAS_BIOMARKER	Disease → Biomarker	Evidence backed biomarker association	direction, comparison, effect_size, p_value
TREATS	Drug → Disease	Clinical trial association	trial_status, trial_phase_max, trial_count, indication
TARGETS_PROTEIN	Drug → Protein	Mechanistic targeting	action_type, is_primary_target, notes
AFFECTS_PATHWAY	Drug → Pathway	Biological mechanism	action_type, source, target_notes
HAS_PHENOTYPE	Disease → Phenotype	Clinical manifestations	phenotypes
ENCODES	Gene → Protein	Molecular grounding	ontology derived

Table 3. Primary relationship types used in the graph

This schema enables direct and interpretable answers to clinically relevant questions such as biomarker directionality, therapeutic trial status, and drug mechanism of action through explicit graph traversal rather than unstructured inference.

3.4 Entity Normalization and Graph Construction

3.4.1 Canonical Identifier Resolution

All curated entities were normalized using ontology derived lookup indices that map normalized text strings to canonical identifiers (e.g., “alzheimer disease” → MONDO:0004975). Normalization relied on deterministic heuristics including case normalization, punctuation removal, and synonym expansion using ontology provided synonym lists. When no canonical match was available, entities were retained using local identifiers to preserve retrievability.

3.4.2 Node and Edge Construction

Edge construction was driven by structured evidence rather than text co-occurrence. Biomarker relationships were derived directly from meta analysis effect sizes and statistical significance values. Therapeutic relationships were explicitly separated into clinical associations (TREATS) and mechanistic associations (TARGETS_PROTEIN and AFFECTS_PATHWAY). Gene-protein relationships were constructed from ontology mappings to provide biological grounding across molecular layers.

3.4.3 Neo4j Export

The resulting nodes and relationships were exported into Neo4j compatible CSV files.

File	Description	Count
neo4j_nodes.csv	All nodes	2029
neo4j_edges.csv	All relationships	582

Table 4. Resulting graph scale

3.5 Graph RAG Question Answering Pipeline

The Graph RAG runtime converts natural language questions into grounded answers through a structured retrieval pipeline. Incoming queries are first classified using a deterministic keyword based intent classifier into categories such as biomarker, drug trial, phenotype, or general disease overview. Each intent is mapped to a corresponding retrieval strategy.

Cypher queries are executed against the Neo4j graph to retrieve an Alzheimer’s disease centered subgraph relevant to the detected intent. Retrieved nodes and relationships are then converted into ultra compact textual context using schema aware graph to text summarizers. Context is grouped by biologically meaningful dimensions such as biomarker direction and biofluid or drug trial status.

Answer generation is performed using a local llama3.2:3b model accessed via Ollama. The model is constrained by a strict system prompt that enforces exclusive use of the provided context and prohibits the introduction of external knowledge. The final response includes both the generated answer and supporting metadata such as detected intent and retrieval strategy.

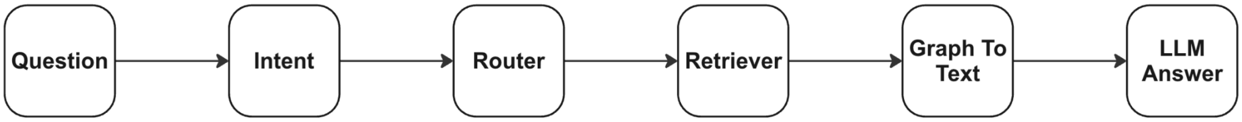

Fig. 2. Graph RAG runtime flow

This design as shown in Fig. 2., deliberately prioritizes structured, schema driven retrieval over vector similarity search, ensuring that all answers remain traceable to explicit knowledge graph facts rather than opaque language model inference.

4. Results and Discussion

4.1. Functional Coverage of the Question Answering System

The implemented Graph RAG system demonstrates reliable performance across a range of Alzheimer's disease focused question types that align with the underlying knowledge graph schema. Because retrieval is driven by explicitly modeled relationships and edge properties, the system is particularly effective for questions that require structured evidence rather than free text synthesis.

In the biomarker domain, the system consistently answers questions about directionality and biofluid specificity, such as identifying biomarkers that decrease in cerebrospinal fluid or increase in plasma or serum. These answers are derived directly from HAS_BIOMARKER relationships, which encode effect direction, statistical significance, and comparison context as edge properties. Similarly, drug and trial related questions are well supported through TREATS relationships, allowing the system to report trial status, maximum phase reached, and clinical indication without relying on unstructured inference.

When mechanistic information is present in the graph, the system can also answer pathway and target oriented questions, including identifying proteins targeted by a given drug or pathways affected by therapeutic interventions. Phenotype related questions are supported through the HAS_PHENOTYPE relationships linking Alzheimer's disease to Human Phenotype Ontology terms. Finally, the system can provide high level disease summaries by aggregating information across biomarkers, drugs, phenotypes, pathways, and representative gene protein mappings.

4.2. Representative Query Outcomes

Query Type	Example Question	Retrieval Strategy	Outcome
Biomarker	Which CSF biomarkers decrease	AD_BIOMARKERS_V1	Grounded list grouped by fluid and

	in AD?		direction
Phenotype	“What phenotypes are associated with AD?”	AD_PHENOTYPES_V1	Phenotype list derived from HAS_PHENOTYPE
Drug / Pathway	“Which Phase 3 drugs affect which pathways?”	AD_DRUGS_PATHWAYS_V1	Drug buckets with associated pathways
General Overview	“Give a high level overview of AD biomarkers, drugs, and pathways.”	Mixed (see discussion)	Partial summary depending on intent routing

Table 5. Summarizes representative query types, example questions, the retrieval strategy used, and the resulting system behavior

These examples illustrate that, when the intent is well defined and aligns with a specific schema slice, the system produces concise, traceable answers grounded in explicit graph evidence.

4.3. Observed Issues and Design Refinements

During evaluation, several limitations and edge cases were identified that inform future refinements. First, Neo4j issued warnings during phenotype retrieval due to references to relationship properties (e.g., onset or frequency) that are not currently populated in the graph schema. This behavior reflects a mismatch between query expectations and stored edge attributes. The issue can be resolved either by simplifying the Cypher queries to reflect the current schema or by extending the graph construction pipeline to populate these additional properties if clinically relevant data become available.

A second issue arises in intent classification for broad, multi domain questions. Queries requesting a general or high level overview may be routed to a biomarker specific strategy when biomarker related keywords dominate the input text. While the resulting answers remain grounded, they may omit relevant drug or pathway information. This limitation suggests the need for lightweight tie breaking logic or explicit detection of multi domain queries, such as routing questions containing terms like “overview” or “high level” to a general Alzheimer’s disease context strategy.

4.4. Strengths of the Approach

Several strengths emerge from the system design and evaluation. First, strict grounding is enforced through structured retrieval and constrained prompting, substantially reducing the risk of hallucinated entities or relationships. Second, all answers are traceable to specific nodes, edges, and properties in the knowledge graph, supporting transparency and interpretability. Third, reproducibility is maintained throughout the pipeline by preserving raw ontology files and

HTML sources and applying deterministic processing steps. Finally, the modular architecture spanning from ingestion, normalization, graph construction, retrieval, and to generation, allows individual components to be extended or replaced without disrupting the overall system.

4.5. Limitations and Scope Constraints

The current system is deliberately designed to prioritize precision, grounding, and reproducibility over broad coverage. As a result, several scope constraints are present. First, the knowledge graph is constructed exclusively from authoritative biomedical ontologies and curated Alzheimer’s specific resources such as Alzforum. While this choice ensures high quality and vetted content, it limits coverage to information that has already been curated by domain experts and does not directly incorporate the rapidly expanding body of primary research literature.

Second, the graph schema currently includes a limited set of core relationship types. Specifically, six primary edge types were implemented to support the most common and practically useful Alzheimer’s disease queries, including biomarker associations, therapeutic trials, mechanistic targets, phenotypes, and gene protein mappings. While this schema is sufficient for answering a wide range of clinically and biologically relevant questions, additional relationship types could be introduced in future work to enable richer cross domain reasoning and more complex graph traversal patterns.

Third, therapeutic information is primarily derived from parsed search result tables rather than comprehensive per drug detail pages or full clinical trial reports. Consequently, certain mechanistic details, trial specific nuances, or negative findings may be absent from the graph. Epidemiological and lifestyle related risk factors are also explicitly excluded from the current scope.

Fourth, retrieval is centered on a single disease anchor: Alzheimer’s disease, rather than supporting generalized multi disease routing. This design choice simplifies retrieval logic and improves answer reliability, but it limits the system’s ability to perform cross disease comparisons. Intent classification is implemented using a deterministic, rule based approach, which improves transparency and interpretability but may misclassify ambiguous or multi domain queries.

Finally, the system uses Neo4j as a labeled property graph rather than an RDF based representation. While this choice facilitates efficient traversal and practical Graph RAG integration, it limits direct interoperability with Semantic Web tooling and formal ontology reasoning frameworks.

4.6. Future Improvements

One important direction for future work is the integration of primary research literature from sources such as PubMed, arXiv, or bioRxiv. Incorporating these sources would allow the knowledge graph to capture emerging findings that may not yet be reflected in curated databases. However, extracting reliable, structured knowledge from long form scientific articles presents substantial challenges. Research papers frequently contain conflicting statements, evolving interpretations, and contextual qualifications, such as refutations of earlier findings or condition specific effects. Addressing this complexity would require document level natural language understanding, evidence tracking, and temporal reasoning, rather than simple sentence level extraction.

In the present work, large-scale triple extraction from full research papers was not implemented due to practical resource constraints, including limited GPU availability and computational capacity. These constraints influenced the decision to focus on curated and structured sources that could be processed deterministically and reproducibly. With access to greater computational resources, future implementations could explore LLM assisted triple extraction pipelines that selectively identify, validate, and prioritize high confidence relations from primary literature.

The graph schema itself could also be expanded beyond the six relationship types currently implemented. Introducing additional biologically and clinically meaningful relations such as disease-stage specific biomarkers, protein-protein interactions, or longitudinal trial outcomes that would enable deeper cross connections and support more advanced graph based analyses.

The graph storage layer could be extended by introducing an RDF or OWL based representation alongside, or in place of, Neo4j. An RDF based implementation would improve interoperability with existing biomedical linked data ecosystems and enable formal reasoning using description logic. A hybrid architecture, in which Neo4j supports efficient traversal and RDF supports semantic inference, may provide a practical balance.

At the retrieval layer, intent classification could be enhanced using a hybrid approach that combines rule based heuristics with an LLM based classifier. Given that the system already operates with a local LLM via Ollama, introducing an auxiliary intent classification step could improve robustness for complex or multi domain queries while retaining deterministic fallbacks for reliability.

Finally, the generation pipeline could be extended to support multi step or multi call LLM reasoning. For example, separate model calls could be used for query decomposition, evidence aggregation, and final answer synthesis. Such an approach could enable more complex analytical responses while maintaining strict grounding in graph derived evidence. Tool based orchestration and explicit reasoning traces would further improve transparency and debuggability.

5. Conclusion

This work presents an end to end Alzheimer's Disease centric Knowledge Graph and Graph RAG question answering system designed to support grounded, interpretable access to complex biomedical knowledge. By integrating authoritative biomedical ontologies with curated Alzheimer's disease resources, the system transforms heterogeneous and fragmented data into a unified, schema driven graph representation stored in Neo4j. On top of this structured foundation, an intent aware Graph RAG pipeline enables natural language questions to be answered through targeted graph retrieval and constrained language model generation.

The results demonstrate that explicitly modeling biomedical entities and relationships at the graph level enables reliable answers to clinically and biologically meaningful questions, including biomarker directionality, therapeutic trial status, and mechanistic drug targets. Compared to document centric RAG approaches, the proposed system improves transparency and reduces hallucination by enforcing strict grounding in retrieved graph evidence.

While the current implementation focuses on a limited but high confidence set of data sources and relationship types, the modular architecture provides a clear path for future extensions. Overall, this work highlights the value of combining ontology grounded knowledge graphs with graph based retrieval and controlled generation as a practical and interpretable approach to question answering in Alzheimer's disease research.

6. References

Köhler, S., Gargano, M., Matentzoglu, N., Carmody, L. C., Lewis-Smith, D., Vasilevsky, N. A., Danis, D., Balagura, G., Baynam, G., Brower, A. M., Callahan, T. J., Chute, C. G., Est, J. L., Galer, P. D., Ganesan, S., Griese, M., Haimel, M., Pazmandi, J., Hanauer, M., … Robinson, P. N. (2021). *The Human Phenotype Ontology in 2021*. Nucleic Acids Research, 49(D1), D1207-D1217. https://doi.org/10.1093/nar/gkaa1043

The Gene Ontology Consortium. (2021). *The Gene Ontology resource: Enriching a GOld mine*. Nucleic Acids Research, 49(D1), D325-D334. https://doi.org/10.1093/nar/gkaa1113

Hastings, J., Owen, G., Dekker, A., Ennis, M., Kale, N., Muthukrishnan, V., Turner, S., Swainston, N., Mendes, P., & Steinbeck, C. (2016). *ChEBI in 2016: Improved services and an expanding collection of metabolites*. Nucleic Acids Research, 44(D1), D1214-D1219. https://doi.org/10.1093/nar/gkv1031

Natale, D. A., Arighi, C. N., Blake, J. A., Bult, C. J., Christie, K. R., Cowart, J., D'Eustachio, P., Diehl, A. D., Drabkin, H. J., Helfer, O., Huang, H., Masci, A. M., Ren, J., Roberts, N. V., Ross, K. E., Ruttenberg, A., Shamovsky, V., Smith, B., … Wu, C. H. (2017). *Protein Ontology (PRO): Enhancing and scaling up the representation of protein entities*. Nucleic Acids Research, 45(D1), D339-D346. https://doi.org/10.1093/nar/gkw1075

Tweedie, S., Braschi, B., Gray, K., Jones, T. E. M., Seal, R. L., Yates, B., & Bruford, E. A. (2021). *Genenames.org: The HGNC and VGNC resources in 2021*. Nucleic Acids Research, 49(D1), D939-D946. https://doi.org/10.1093/nar/gkaa980

Mungall, C. J., McMurry, J. A., Köhler, S., Balhoff, J. P., Borromeo, C., Brush, M., Carbon, S., Conlin, T., Dunn, N., Engelstad, M., Foster, E., Gourdine, J. P., Jacobsen, J. O. B., Keith, D., Laraway, B., Lewis, S. E., NguyenXuan, J., Shefchek, K., … Haendel, M. A. (2017). *The Monarch Initiative: An integrative data and analytic platform connecting phenotypes to genotypes across species*. Nucleic Acids Research, 45(D1), D712-D722. https://doi.org/10.1093/nar/gkw1128

Alzforum. (2024). *AlzPedia: An interactive encyclopedia of Alzheimer's disease research*. https://www.alzforum.org/alzpedia

Alzforum. (2024). *AlzBiomarker: A curated database of Alzheimer's disease biomarkers*. https://www.alzforum.org/alzbiomarker

Alzforum. (2024). *Therapeutics database for Alzheimer's disease*. https://www.alzforum.org/therapeutics

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-T., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). *Retrieval-augmented generation for knowledge-intensive NLP tasks*. Advances in Neural Information Processing Systems, 33, 9459-9474. https://doi.org/10.48550/arXiv.2005.11401

Robinson, I., Webber, J., & Eifrem, E. (2015). *Graph Databases: New Opportunities for Connected Data* (2nd ed.). O'Reilly Media.

Lamy, J.-B. (2017). *Owlready2: Ontology-oriented programming in Python with automatic classification and high-level constructs*. Artificial Intelligence in Medicine, 80, 11-28. https://doi.org/10.1016/j.artmed.2017.07.002

Neo4j, Inc. (2024). *Neo4j Graph Database Platform*. https://neo4j.com

Ollama. (2024). *Ollama: Run large language models locally*. https://ollama.com

Jin, Q., Dhingra, B., Liu, Z., Cohen, W. W., & Lu, X. (2019). *PubMedQA: A dataset for biomedical research question answering*. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP) (pp. 2567-2577). https://doi.org/10.18653/v1/D19-1259

Romano, J. D., Truong, V., Kumar, R., et al. (2024). *The Alzheimer's Knowledge Base: A knowledge graph for Alzheimer disease research*. Journal of Medical Internet Research, 26, e46777. https://doi.org/10.2196/46777

Spasov, I., Lazarova, S., & Petrova-Antonova, D. (2024). *Ontology-based knowledge graph for Alzheimer's disease using Neo4j*. In Proceedings of the International Conference on Data Analytics and Management (pp. 65-77). Springer.

Li, D., Yang, S., Tan, Z., & Zhang, Y. (2024). Dynamic co-augmentation of large language models and knowledge graphs for answering Alzheimer's disease questions with scientific literature. arXiv preprint arXiv:2405.04819. https://arxiv.org/abs/2405.04819

Liang, Y., Ouyang, C., & Wang, Y. (2025). Large language models and structured knowledge graphs for biomedical question answering: A review of methodologies and challenges. Journal of Biomedical Informatics, 150, 104523. https://pubmed.ncbi.nlm.nih.gov/40520893/

Xu, T., Feng, J., Melendez, J., & Chen, Y. (2025). *Addressing accuracy and hallucination of LLMs in Alzheimer's disease research through knowledge graphs*. arXiv preprint arXiv:2508.21238. https://arxiv.org/abs/2508.21238

Dobreva, J., Simjanoska Misheva, M., Mishev, K., Trajanov, D., & Mishkovski, I. (2025). *A unified framework for Alzheimer's disease knowledge graphs: Architectures, principles, and clinical translation*. Brain Sciences, 15(5), 523. https://doi.org/10.3390/brainsci15050523

Appendix: Implementation and Reproducibility

A. Code and Data Availability

All code used in this project is publicly available: https://github.com/Athish49/alzheimers_kg

The repository contains scripts for ontology ingestion, curated Alzheimer's data processing, knowledge graph construction, Neo4j export, and Graph RAG question answering.

B. System Implementation Overview

The implementation follows a modular, stage based pipeline:

1. Ontology ingestion and processing:
   Authoritative biomedical ontologies are downloaded and subsetted into AD relevant entities, providing canonical identifiers and synonym mappings.
2. Curated Alzheimer's data parsing:
   Curated resources from Alzforum (AlzPedia, AlzBiomarker, Therapeutics) are downloaded as raw HTML and converted into structured CSV representations.
3. Knowledge graph construction:
   Ontology backed entities are normalized and connected using a schema aligned graph model. Nodes and relationships are exported as Neo4j compatible CSV files.
4. Graph RAG question answering:
   Natural language queries are processed through intent classification, targeted Neo4j retrieval, compact graph to text transformation, and grounded answer generation using a local llama3.2:3b model.

This separation ensures that factual retrieval is handled exclusively by the knowledge graph, while language generation is limited to summarizing retrieved evidence.

C. Reproducibility Instructions

The complete pipeline can be reproduced using the following steps:

1. Run ontology ingestion and processing scripts
2. Download and parse curated Alzheimer's disease resources
3. Normalize entities and generate graph edges
4. Import generated CSVs into Neo4j
5. Execute the Graph RAG pipeline using the provided test script

D. Environment and Dependencies

All Python dependencies are specified in requirements.txt.

The system uses a locally hosted Neo4j instance and a local llama3.2:3b model accessed via Ollama. Configuration parameters are controlled through environment variables.

Statutory Declaration

I hereby declare that the paper presented is my own work and that I have not called upon the help of a third party. In addition, I affirm that neither I nor anybody else has submitted this paper or parts of it to obtain credits elsewhere before. I have clearly marked and acknowledged all quotations or references that have been taken from the works of others. All secondary literature and other sources are marked and listed in the bibliography. The same applies to all charts, diagrams and illustrations as well as to all Internet resources. Moreover, I consent to my paper being electronically stored and sent anonymously in order to be checked for plagiarism. I am aware that the paper cannot be evaluated and may be graded “failed” if the declaration is not made.

Athish Gopal Rajesh
December 17, 2025