NOVARTIS

NEST

NURTURING EXCELLENCE,
STRENGTHENING TALENT

TEAM NEWBIE NEURONS

Problem Statement 1

Ritesh
- iit2023060@iiita.ac.in

Ranjeet
- iit2023064@iiita.ac.in

Priyam
- iit2023147@iiita.ac.in

Aayush
- iib2023013@iiita.ac.in

ABSTRACT

PROBLEM : Semantic grouping of clinical studies for retrieval & strategic insights

Obtaining a matched subset of clinical trials for only a combination of 4 unstructured features (textual data: Study Title, Primary outcome measure, Secondary outcome measure and Criteria) and using the patterns found in them after applying human logical thought process. Expectation is to design the approach, model, method and/or algorithm to enable this.

Key Assumptions

1. ClinicalTrials.gov data provides meaningful and reliable context.
2. Embeddings from Stella 1.5B improve semantic understanding of relationships.
3. FAISS efficiently maps relationships in large datasets.
4. Neo4j enables scalable graph-based analysis for recommendations.

Approach

The Knowledge Graph pipeline enables clinical trial recommendations through:

1. Data Preprocessing: Relationships extracted using Gemma 2 in structured format.
2. Node Normalization: Stella 1.5B embeddings and FAISS identify and merge similar nodes.
3. Knowledge Graph Creation: Neo4j constructs a scalable graph database from relationships.
4. Enhanced Recommendations: Jaccard similarity in Neo4j's GDS identifies top similar trials.

Results

- Performance: The pipeline ensured efficient and accurate semantic grouping, enabling meaningful insights and recommendations.
- Scalability: Neo4j effectively managed large-scale graph.

Limitations

- Computational Constraints: The pipeline required significant computational resources. Although Gemma 2 is an open-source model that can be run locally at no cost with sufficient computational power, limited local resources necessitated running it on the Groq cloud.
- Lack of Ground Truth Samples: The absence of verified ground truth samples posed challenges in assessing the accuracy of the results. This limited the ability to fully validate the recommendations and insights generated.

PROJECT INTRODUCTION

Background

Designing clinical trial protocols is complex and critical for success in the pharmaceutical industry. Researchers often reference multiple studies to enhance predictability and quality. However, retrieving relevant historical trials is challenging, leading to inefficiencies, delays, and increased costs.

Problem Statement

Business Context

Researchers struggle to identify relevant trials categorized by drugs, diseases, and criteria. Manual searches through unstructured data like study titles and outcomes are time-consuming and inconsistent.

Technical Challenge

The task is to develop an AI solution to retrieve semantically similar trials using a dataset of 450,000 clinical trials from ClinicalTrials.gov. This involves:

1. Cleaning and preprocessing data.
2. Semantic feature engineering for grouping.
3. Building AI models to retrieve 10 unique, relevant trials per query based on criteria.
4. Ensuring explainability for domain alignment.

Proposed Solution

A knowledge graph powered by Gemma 2 and Neo4j, semantically groups and retrieves clinical trials efficiently. This scalable and explainable solution enhances trial design by providing fast, accurate, and contextually relevant recommendations, benefiting the pharmaceutical industry significantly.

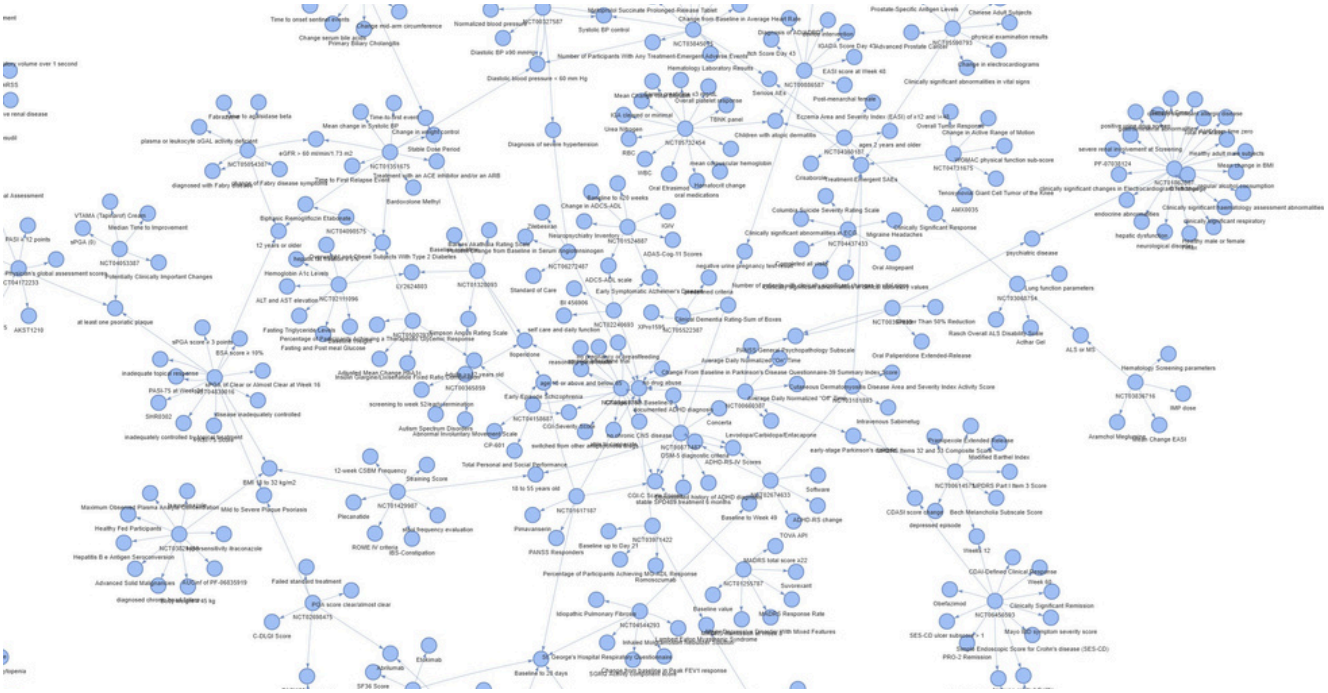
METHODOLOGY

This project aims to build a scalable AI-driven solution for semantic grouping and retrieval of clinical trials *using free and open-source models* for accessibility and cost efficiency. The methodology involves four key steps: data preprocessing, feature engineering, knowledge graph construction, and recommendation generation. Below is a concise explanation of the approach:

1. **Data Preprocessing and Relationship Extraction**
   - Goal: Extract structured relationships from unstructured clinical trial data.
   - Steps:
     - Merge textual data from columns like Study Title, Primary Outcome Measures, and Criteria.
     - Use the Gemma 2 open-source model to extract relationships into a structured format.
     - Clean and organize the relationships into CSV files for downstream use.
   - Output: relationships.csv with structured relationships and Object_Value_Counts2.csv with object frequency counts.
2. **Node Deduplication and Normalization**
   - Goal: Reduce redundancy by merging similar nodes.
   - Steps:
     - Generate embeddings for objects using the Stella 1.5B model.
     - Use cosine similarity to merge nodes with a similarity threshold (e.g., 0.8) via FAISS.
     - Serialize normalized results for scalability.
   - Output: filtered_results_with_similars.csv containing deduplicated nodes.
3. **Knowledge Graph Construction**
   - Goal: Build a queryable graph to represent clinical trial relationships.
   - Steps:
     - Load preprocessed data into Neo4j, creating nodes and edges.
     - Optimize graph construction with concurrent transactions.
   - Output: A Neo4j knowledge graph enabling structured queries.
4. **Recommendation Generation**
   - Goal: Identify similar clinical trials using graph algorithms.
   - Steps:
     - Project the graph in Neo4j's Graph Data Science (GDS) library.
     - Compute Jaccard similarity between trials using 'gds.nodeSimilarity.stream'
     - Allow user inputs to retrieve top recommendations for a given trial ID.
   - Output: Ranked recommendations for similar trials.
5. **Evaluation and Results**

- Scalability: Neo4j effectively handled large-scale graphs with seamless integration.
- Performance: The pipeline demonstrated high accuracy in semantic grouping with an average cosine similarity of 0.857, aiding decision-making in research and development.

EXPLAINABILITY & VISUALIZATION



NCT ID: NCT00385736

Recommendations:

Rank	Trial ID	Cosine Similarity
1.	NCT02289417	0.9686
2.	NCT00408629	0.9588
3.	NCT03029143	0.9241
4.	NCT02065557	0.9136
5.	NCT01620255	0.9134
6.	NCT05731128	0.908
7.	NCT00488631	0.9069
8.	NCT01482884	0.8995
9.	NCT00659802	0.8847
10.	NCT03221036	0.8812
Average		0.9159

NCT ID: NCT00386607

tecommendations:

Rank	Trial ID	Cosine Similarity
1.	NCT01456169	0.8961
2.	NCT00402103	0.8956
3.	NCT00923091	0.8902
4.	NCT00281580	0.869
5.	NCT00698646	0.8487
6.	NCT00151775	0.8479
7.	NCT01204398	0.8422
8.	NCT00435162	0.8347
9.	NCT00841672	0.8049
10.	NCT06174766	0.7229
Average		0.8452

NCT ID: NCT03518073

Recommendations:

Rank	Trial ID	Cosine Similarity
1.	NCT00762411	0.9351
2.	NCT02754830	0.8588
3.	NCT00477659	0.8577
4.	NCT00843518	0.8554
5.	NCT05310071	0.8435
6.	NCT02670083	0.8416
7.	NCT04994483	0.8223
8.	NCT00428090	0.8164
9.	NCT01849055	0.6608
10.	NCT02091362	0.6143
Average		0.8106

- We validated the reliability of our results by comparing Jaccard similarity outcomes with cosine similarity derived from vector embeddings, confirming consistency through the average cosine similarity score.
- Each recommendation is backed by clear node connections and similarity scores, enabling researchers to trace the reasoning behind a trial match.
- Neo4j's native visualization tools provide an intuitive way to explore the Knowledge Graph, showing connections between nodes (e.g., trials, diseases, and interventions).