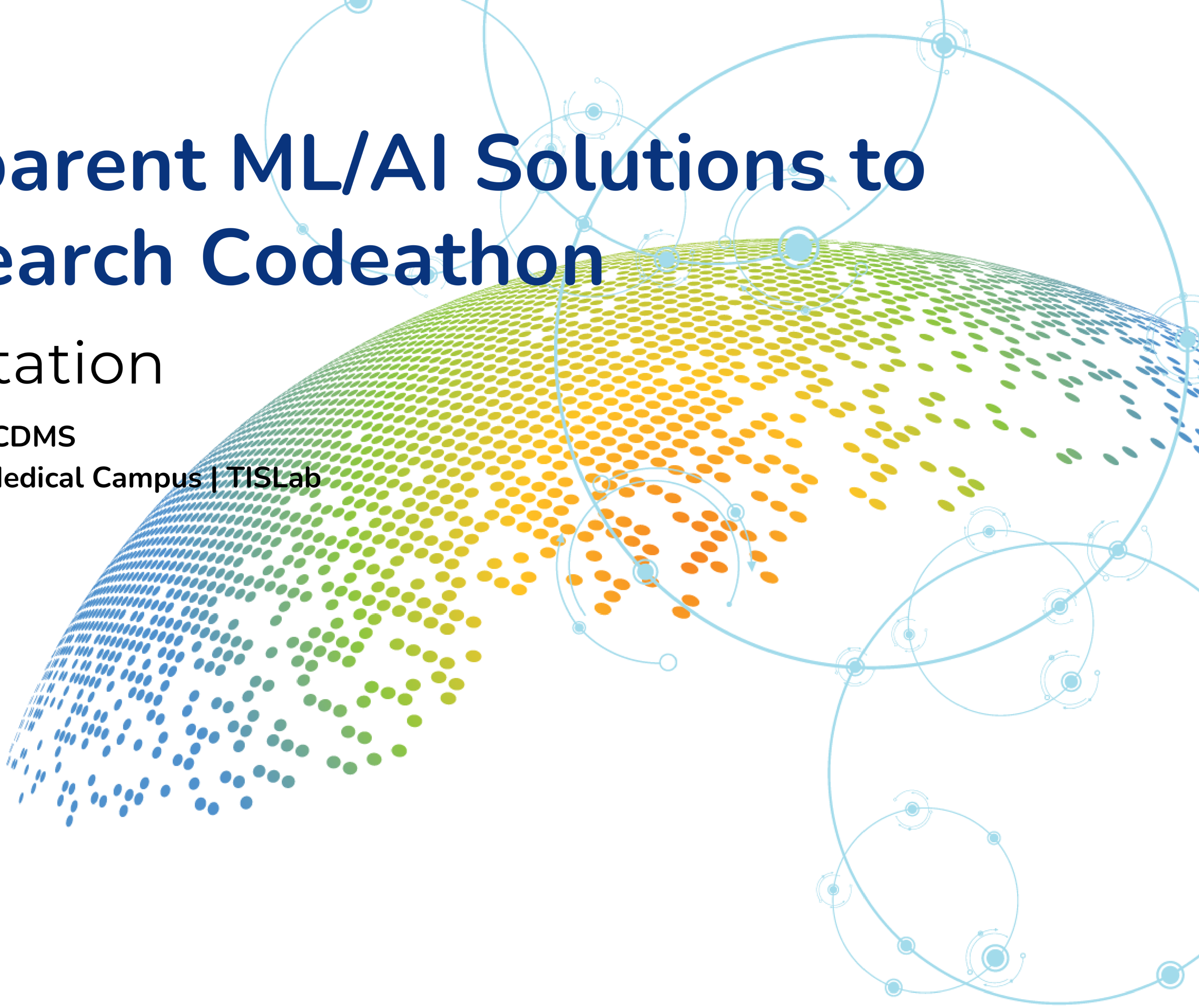
# NCBI - Building Transparent ML/AI Solutions to Advance Bilogical Research Codeathon

## Project RAGVar Final Presentation

- (Team Leader) David Beaumont, RTI International | SSES | CDMS
- (Tech Lead) Corey Cox, University of Colorado | Anschutz Medical Campus | TISLab
- Nathaniel Braswell, RTI International | SSES | CDMS
- Stephen Hwang, RTI International | SSES | CDMS
- Oswaldo Alonso Lozoya, RTI International | SSES | CDMS



Center for Data Modernization Solutions (CDMS)

# Problem Statement

**Abstract**: How do we address the critical challenge of data harmonization in research repositories by aligning new data inputs with existing datasets through retrieval augmented generation and AI reasoning?

**Use Case:**

- Enhanced Data Integration
- Automated Data Harmonization
- Improved Data Quality
- Support for Multidisciplinary Research

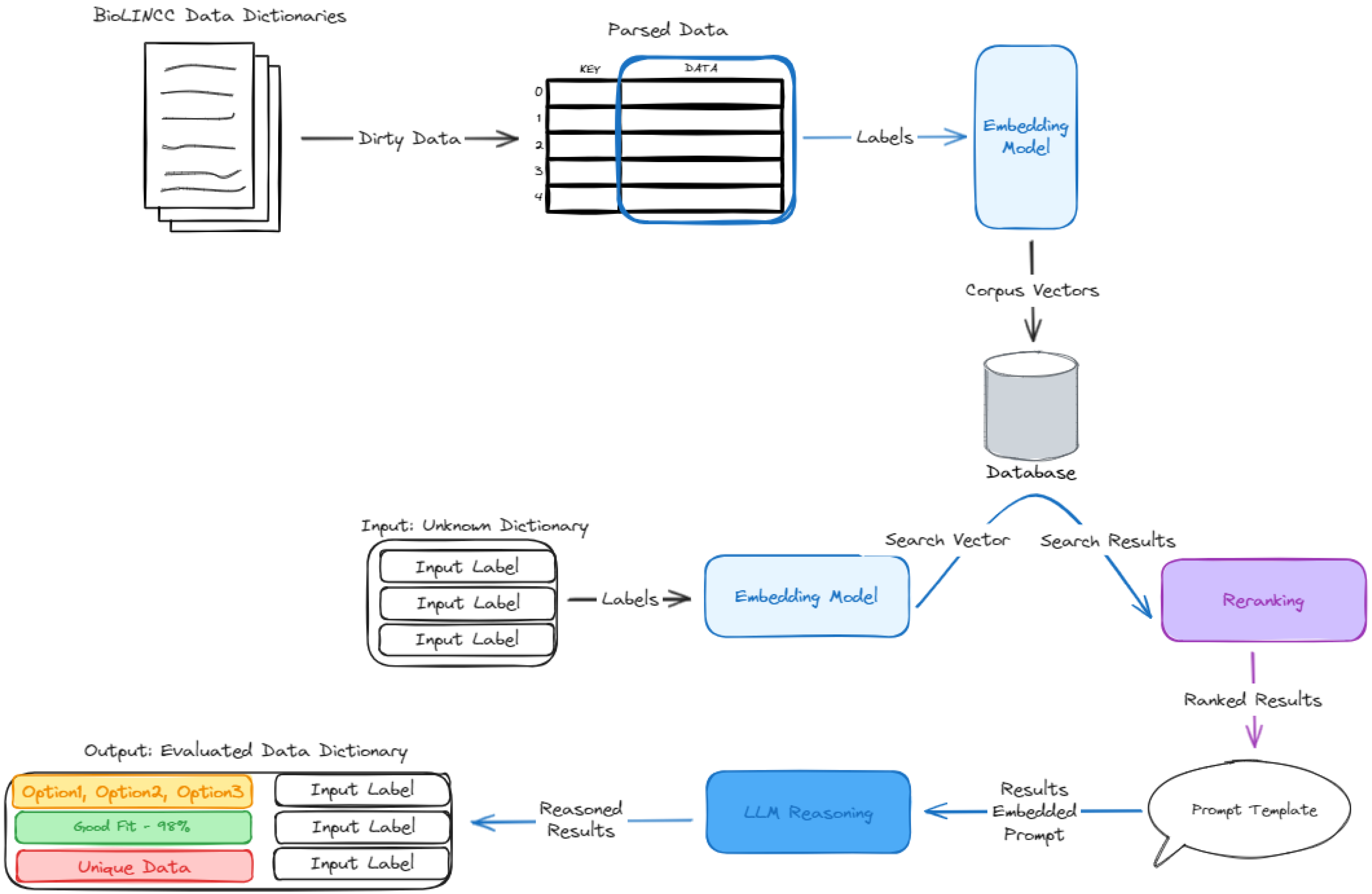
# Technical Deliverables

**Data Extraction**: *Extract study data dictionaries to provide context for search & retrieval process.*

**Build Retrieval System:** *Create embeddings of data dictionary labels, persist in vector store, embed test labels for vector search, use reranking model to score.*

**Evaluation:** *Embed top three search results and test label in prompt template, pass to LLM for evaluation, compare against human evaluation.*

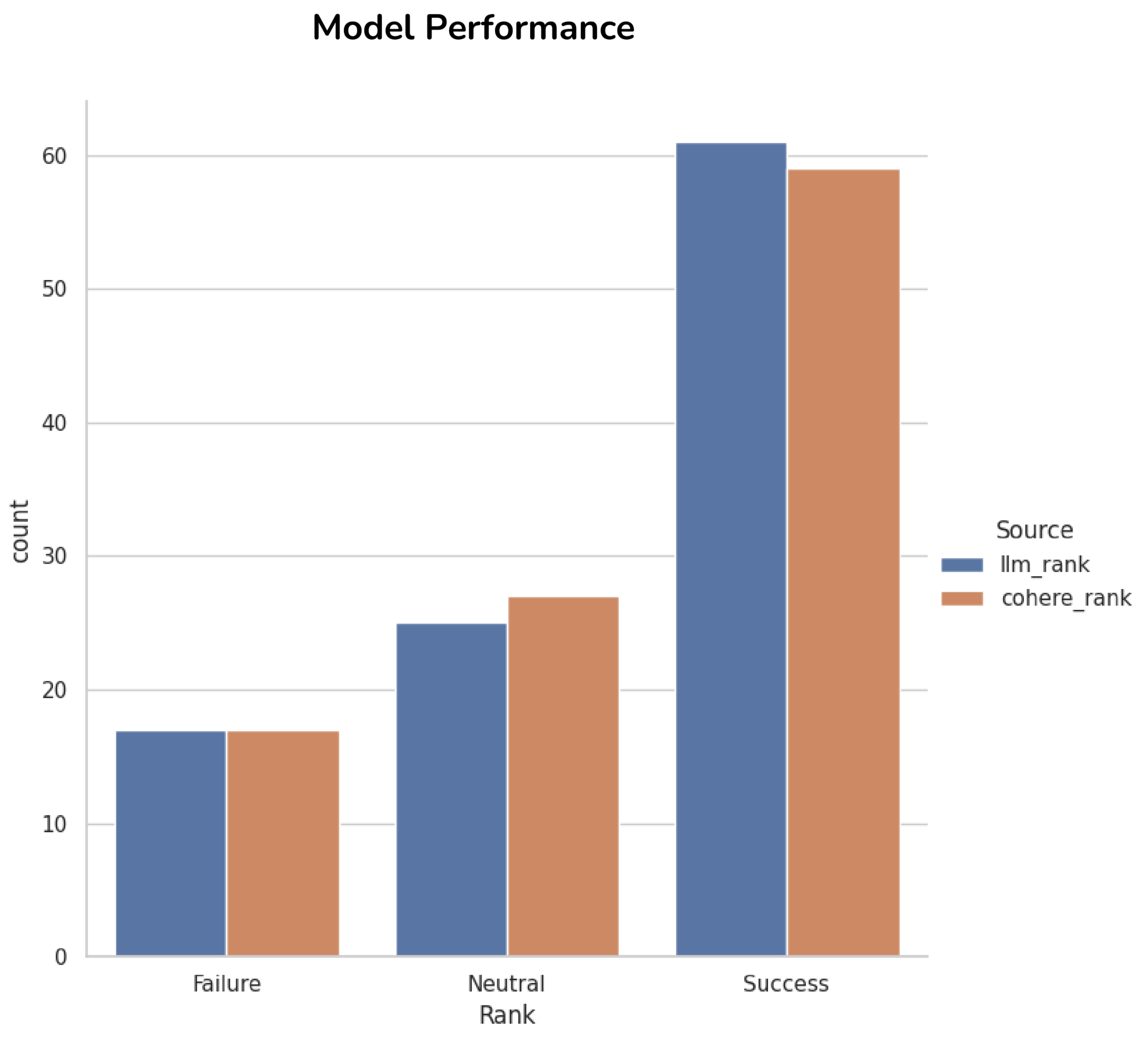
# Architecture

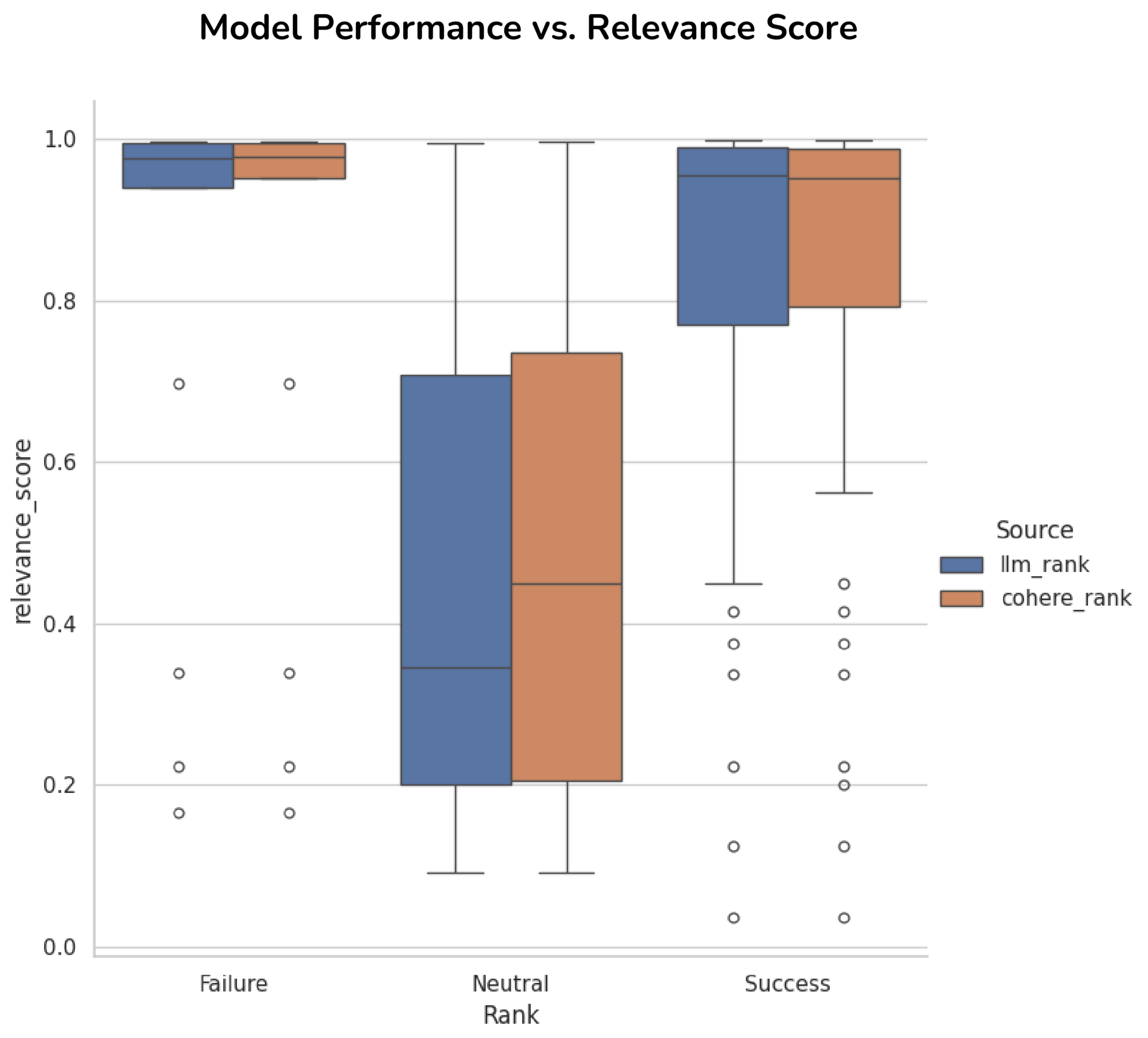


# Retrieval & Evaluation Statistics

Raw Data: ~480,000
Total number of label embeddings: 259,881
Number of search queries evaluated: 104
Number of retrieved labels per search query: 3
Total number of retrievals evaluated: 312

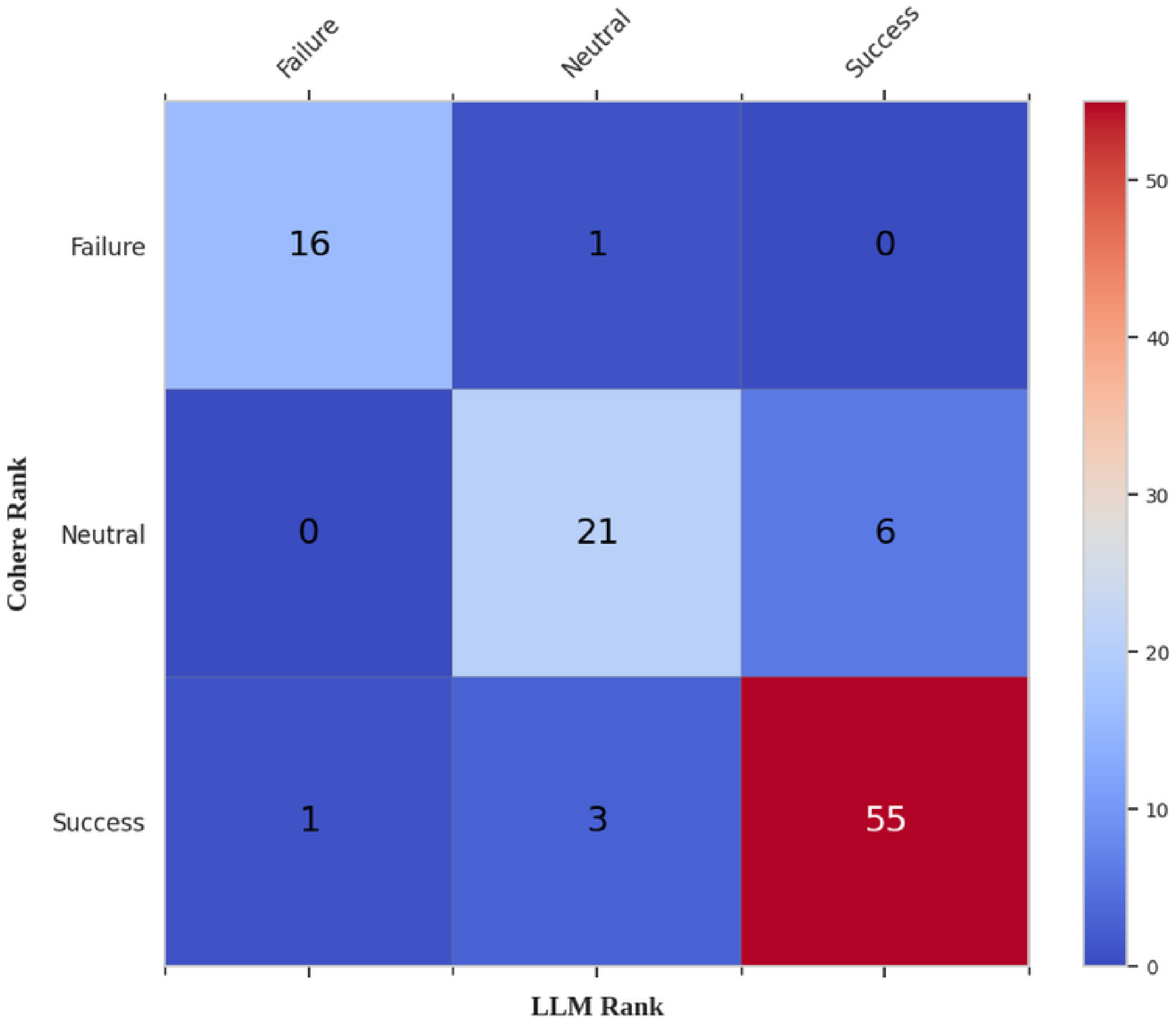
# Results


Model Performance
count
60
50
40
30
20
10
0
Failure
Neutral
Success
Rank
Source
llm_rank
cohere_rank


Model Performance vs. Relevance Score
relevance_score
1.0
0.8
0.6
0.4
0.2
0.0
Failure
Neutral
Success
Rank
Source
llm_rank
cohere_rank

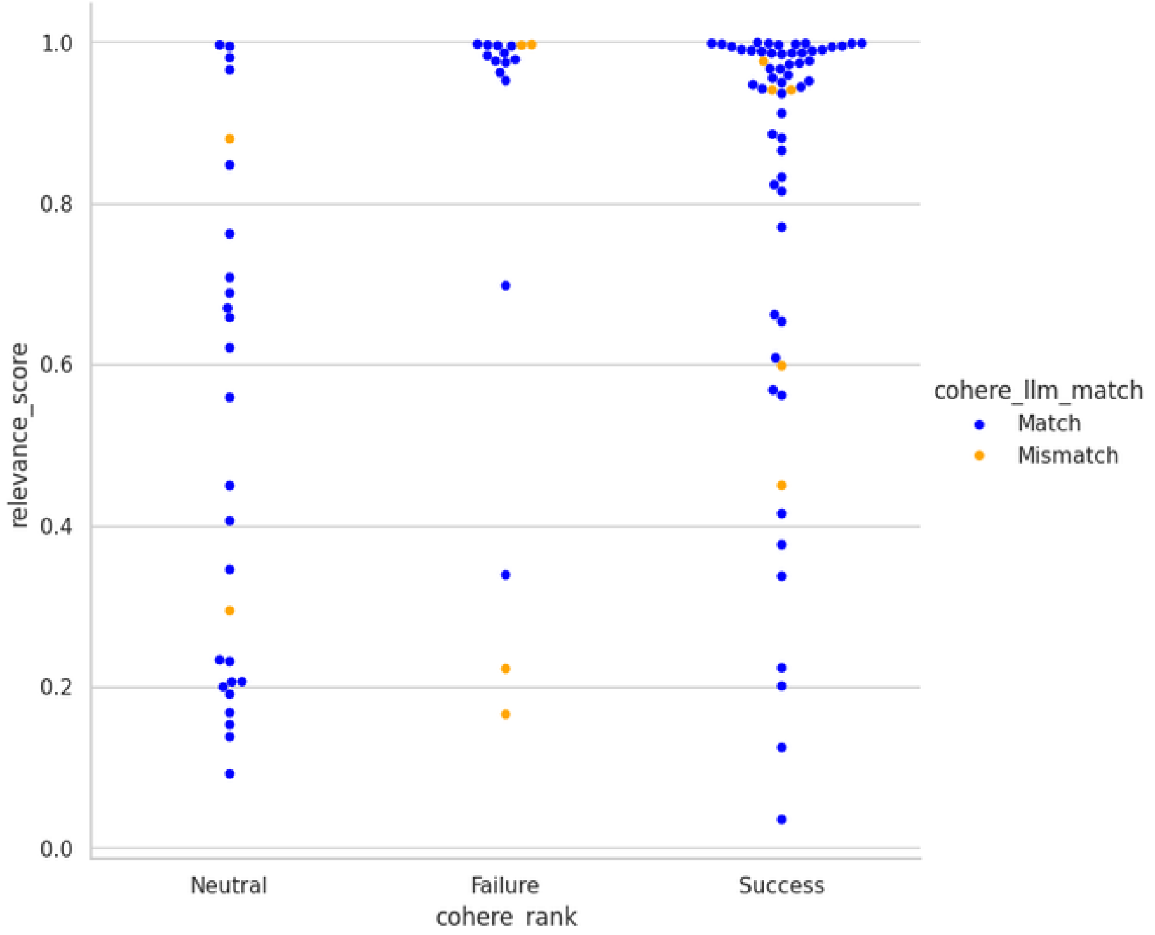
# Results

Cross-tabulation of Cohere Rank vs. LLM Rank

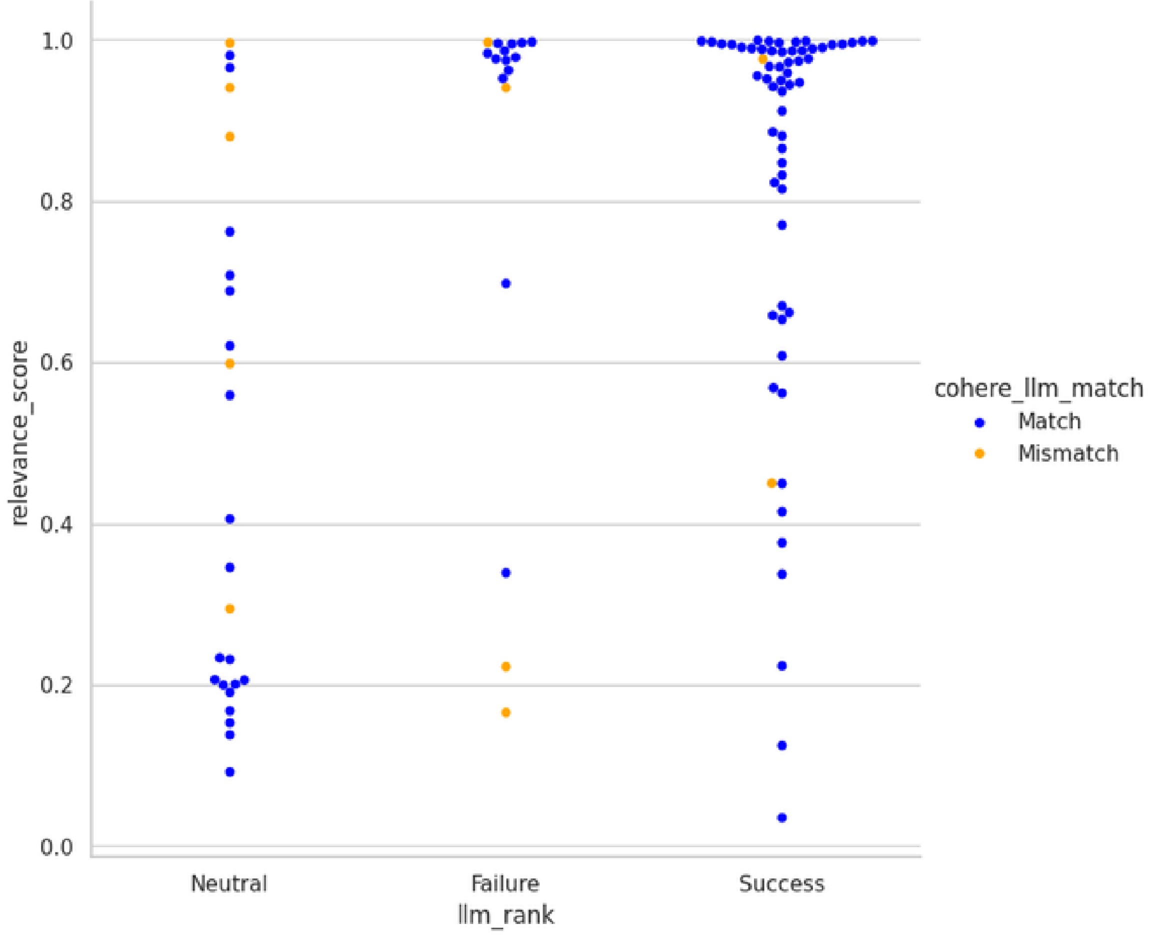


| Cohere Rank \ LLM Rank | Failure | Neutral | Success |
|---|---|---|---|
| Failure | 16 | 1 | 0 |
| Neutral | 0 | 21 | 6 |
| Success | 1 | 3 | 55 |

Distribution of Cohere evaluation vs. relevance score
with cohere_llm_match evaluation
mismatches collapsed(match+2+3 v. mismatch)



Distribution of LLM evaluation vs. relevance score
with cohere_llm_match evaluation
mismatches collapsed (match+2+3 v. mismatch)



# Next Steps / Future Direction

- **Study Descriptions Embeddings**
  - Create separate embeddings based on study descriptions to align studies effectively, aiding the LLM in recognizing relevant variables within study spaces.
- **Prompt Versioning and Engineering**
  - Implement prompt versioning to compare and refine prompts, particularly to enable the LLM to identify and reject non-useful information from Cohere, enhancing user experience.
- **Data Ops and Concept Drift**
  - Address concept drift, where changes in research terminology over time can affect model accuracy. Strategies include evaluating failed high-relevance scores to identify drift, adding data cleaning steps, and leveraging DataOps for continuous model improvement.
- **MLOps for Evolving Data**
  - Utilize user feedback on variable matches to refine embeddings and improve model identification capabilities as new data is added, ensuring the model evolves with the corpus.
- **Enhancing Test Data Set**
  - Improve model evaluation by using test data sets that closely resemble real-world variables, moving away from synthetic variables to better understand model performance.