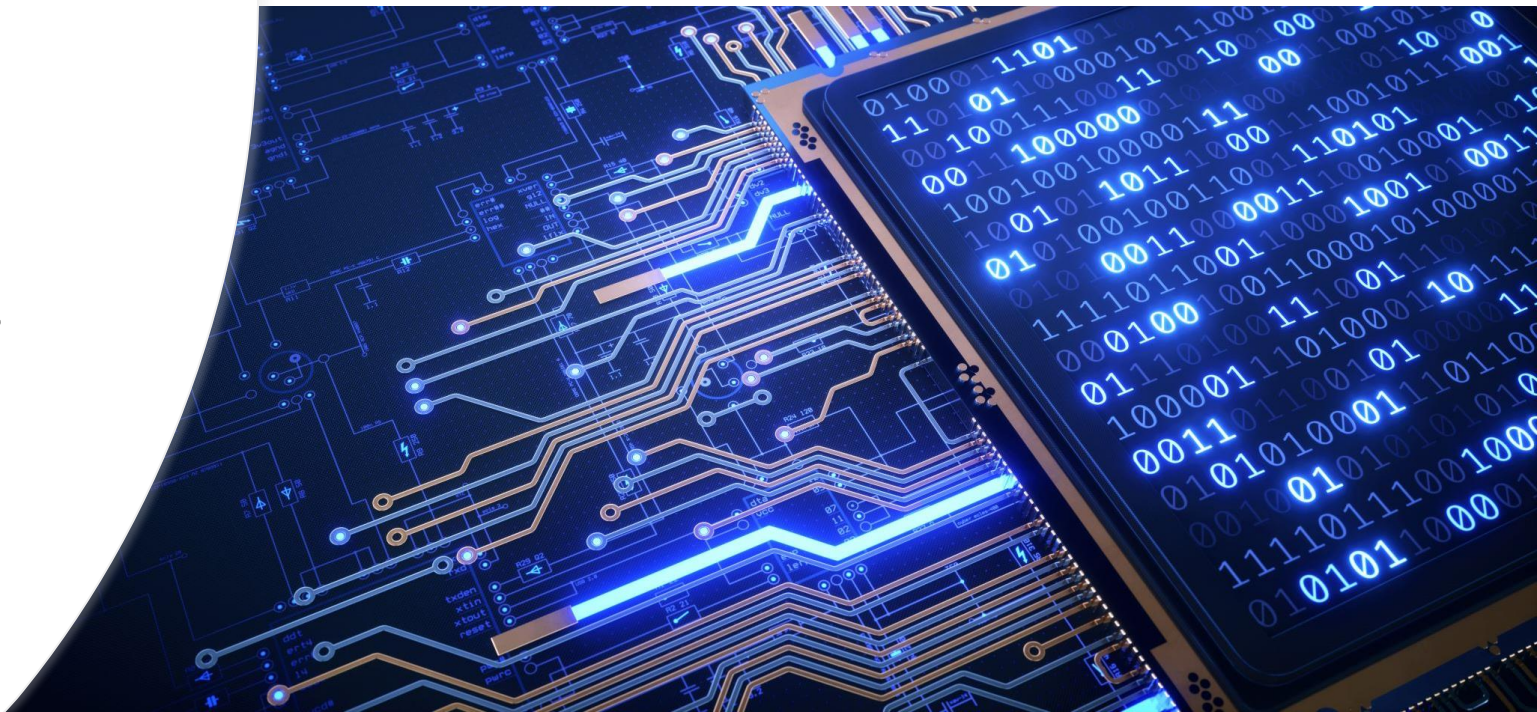


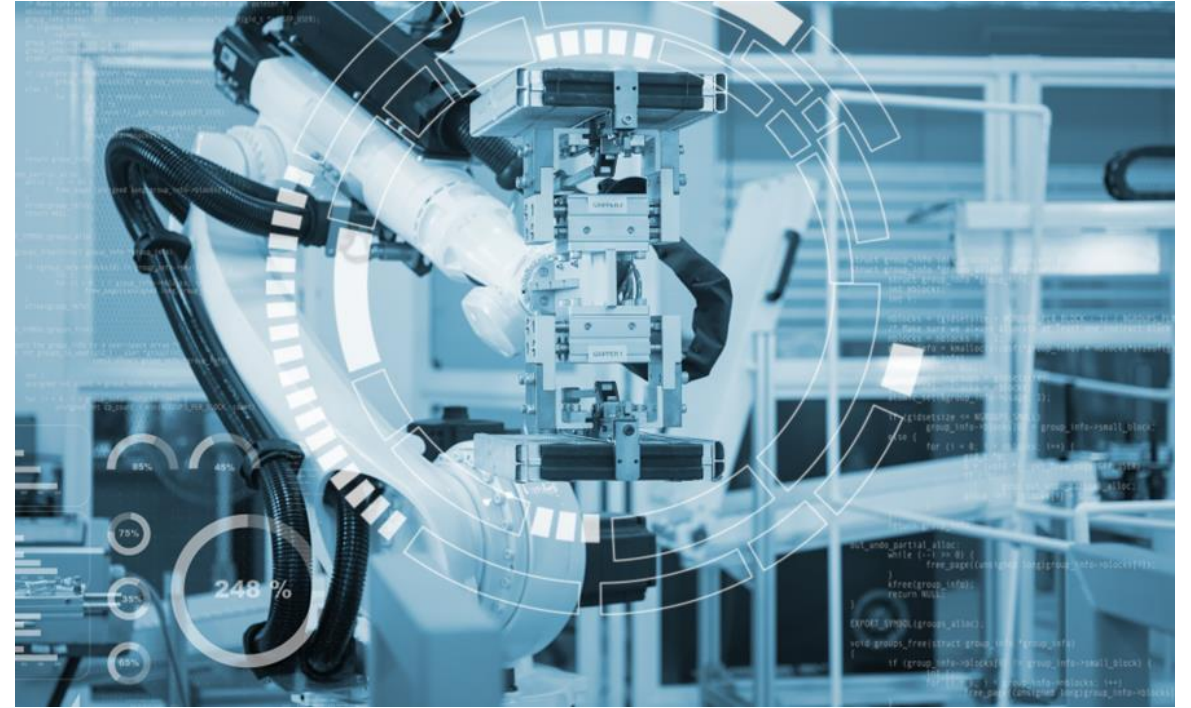
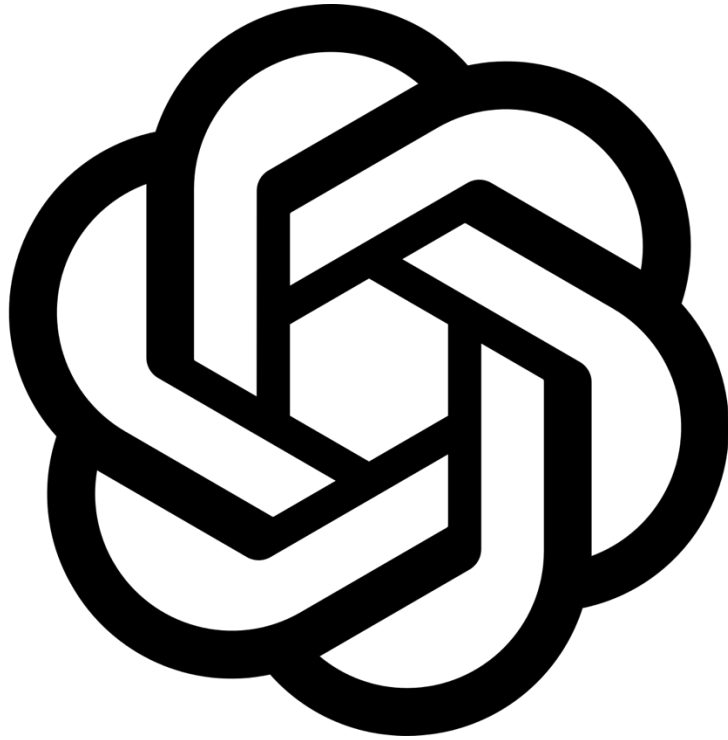
Biomedical Question Answering with Large Language Models

Matthew Moody

Introduction

1. Evaluating the use of Large Language Models (LLMs) for QA tasks
2. Building upon the success achieved by existing LLMs
3. Investigating the limitations of existing LLMs, and how models can be trained to improve the quality of answers





Research Context

- General QA: Llama, ChatGPT
- Biomedical QA: BioBERT, BioLinkBERT, MedPaLM2, PMC-Llama

Aims & Objectives



Implement a biomedical QA system



Explanations for answers, including referring to the information source used



Reduce hallucination of responses by the LLM, and help with identifying hallucinated answers



Retrieval Quality



Response Generation Quality

Data & Resources

- BioASQ 9b datasets
- Llama-2 API



Methods & Experiment Design



Training the Model



Model Testing

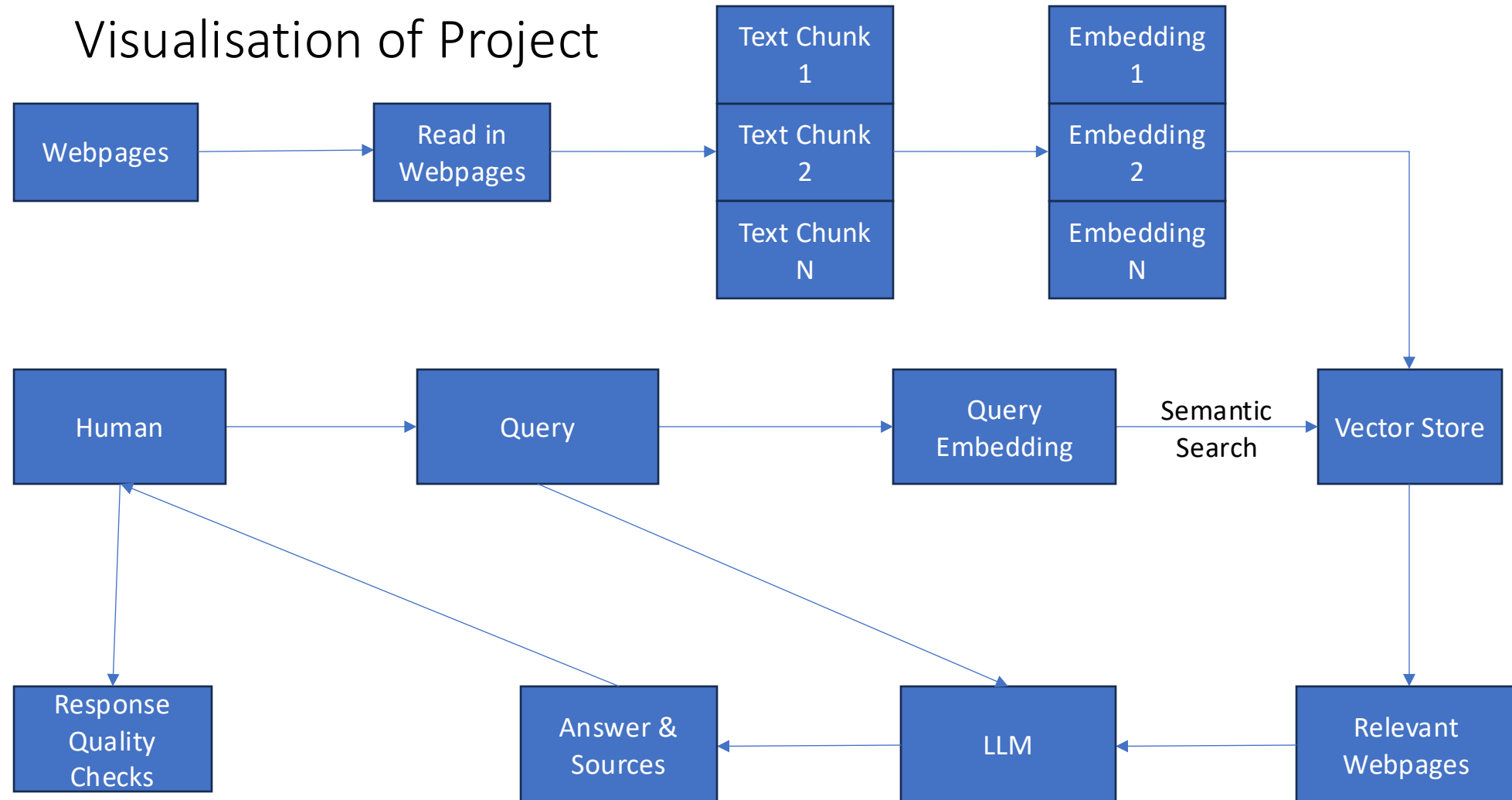


Data Governance &
Ethics



Risk Assessment

Visualisation of Project



Project Execution & Challenges Faced

- Llama 2.0
- Langchain
- ChromaDB
- Challenge with using CPU
- Usage Limits
- Irrelevant text being extracted



Chroma



Example Questions

Is Hirschsprung disease a mendelian or a multifactorial disorder? -
Summary

List signaling molecules (ligands) that interact with the receptor
EGFR? - List

Is the protein Papilin secreted? – Yes/No

Name synonym of Acrokeratosis paraneoplastica. – Factoid

Example

- Running the Model Initialization Code
- Reading in the QA Data
- Extracting the 'documents' part of each question, these will be used to train the retriever part of the model

```
{'body': 'Is Hirschsprung disease a mendelian or a multifactorial disorder?',  
'documents': ['http://www.ncbi.nlm.nih.gov/pubmed/15858239',  
               'http://www.ncbi.nlm.nih.gov/pubmed/15829955',  
               'http://www.ncbi.nlm.nih.gov/pubmed/6650562',  
               'http://www.ncbi.nlm.nih.gov/pubmed/12239580',  
               'http://www.ncbi.nlm.nih.gov/pubmed/21995290',  
               'http://www.ncbi.nlm.nih.gov/pubmed/23001136',  
               'http://www.ncbi.nlm.nih.gov/pubmed/15617541',  
               'http://www.ncbi.nlm.nih.gov/pubmed/8896569',  
               'http://www.ncbi.nlm.nih.gov/pubmed/20598273'],  
'ideal_answer': ["Coding sequence mutations in RET, GDNF, EDNRB, EDN3, and SOX10 are  
involved in the development of Hirschsprung disease. The majority of these genes was shown to  
be related to Mendelian syndromic forms of Hirschsprung's disease, whereas the non-  
Mendelian inheritance of sporadic non-syndromic Hirschsprung disease proved to be complex;  
involvement of multiple loci was demonstrated in a multiplicative model."],  
'type': 'summary'}
```

Example - continued

- Retrieval Augmented Generation
- Ingest data
- Filter data to only include desired sections
- Split data into chunks

```
# 1
web_loader = UnstructuredURLLoader(
    urls=all_docs, mode='elements', strategy='fast',
)
web_doc = web_loader.load()

# 2
updated_web_doc_ = filter_complex_metadata(web_doc)
abstract_index = find(updated_web_doc_, 'Abstract')
updated_web_doc_1 = [updated_web_doc_[a+1] for a in abstract_index]

# 3
text_splitter = RecursiveCharacterTextSplitter(chunk_size=1024, chunk_overlap=128)
qa_docs = text_splitter.create_documents(qa_pairs)
updated_web_doc = updated_web_doc_1 + qa_docs
chunked_web_doc = text_splitter.split_documents(updated_web_doc)
```

Example - continued

- Embedding of text
- Vector database storage
- Creating QA pipeline

```
# 4
model_name = "sentence-transformers/all-mpnet-base-v2"
model_kwargs = {"device": "cuda"}
embeddings = HuggingFaceEmbeddings(model_name=model_name, model_kwargs=model_kwargs)

# 5
vectordb = Chroma.from_documents(documents=chunked_web_doc, embedding=embeddings, persist_directory="chroma_db")

# 6
retriever = vectordb.as_retriever()

qa = RetrievalQA.from_chain_type(
    llm=llm,
    chain_type="stuff",
    retriever=retriever,
    verbose=True
)
```

Example - continued

body	documents	sources
Is Hirschsprung disease a mendelian or a multifactorial disorder?	['http://www.ncbi.nlm.nih.gov/pubmed/15858239', 'http://www.ncbi.nlm.nih.gov/pubmed/15829955', 'http://www.ncbi.nlm.nih.gov/pubmed/6650562', 'http://www.ncbi.nlm.nih.gov/pubmed/12239580', 'http://www.ncbi.nlm.nih.gov/pubmed/21995290', 'http://www.ncbi.nlm.nih.gov/pubmed/23001136', 'http://www.ncbi.nlm.nih.gov/pubmed/15617541', 'http://www.ncbi.nlm.nih.gov/pubmed/8896569', 'http://www.ncbi.nlm.nih.gov/pubmed/20598273']	['http://www.ncbi.nlm.nih.gov/pubmed/15617541', 'http://www.ncbi.nlm.nih.gov/pubmed/15617541', 'http://www.ncbi.nlm.nih.gov/pubmed/23001136', 'http://www.ncbi.nlm.nih.gov/pubmed/23001136', 'http://www.ncbi.nlm.nih.gov/pubmed/21995290']

Example - continued

body	ideal_answer	response	1_fmea	e
Is Hirschsprung disease a mendelian or a multifactorial disorder?	["Coding sequence mutations in RET, GDNF, EDNRB, EDN3, and SOX10 are involved in the development of Hirschsprung disease. The majority of these genes was shown to be related to Mendelian syndromic forms of Hirschsprung's disease, whereas the non-Mendelian inheritance of sporadic non-syndromic Hirschsprung disease proved to be complex; involvement of multiple loci was demonstrated in a multiplicative model."]	Hirschsprung disease is a complex disorder, and the genetics of the disease are highly multifactorial. While some forms of the disease are inherited in a Mendelian manner, others are more complex and involve the interaction of multiple genetic and environmental factors. The disease is known to involve multiple loci, and the genetic architecture of the disease is not fully understood. Therefore, it is not accurate to classify Hirschsprung disease as solely a Mendelian or a multifactorial disorder. Instead, it is a disorder with a complex genetic basis that involves both Mendelian and non-Mendelian inheritance patterns.	0.36943	

Performance Metrics

Hit Rate

$$= \frac{\text{\# of Hits}}{\text{Total \# of Resources Returned}}$$

MRR (Mean Reciprocal Rank)

$$= \frac{1}{N} \sum_{i=1}^N \frac{1}{\text{rank}_i}$$

NDCG (Normalized Discounted Cumulative Gain)

$$= \frac{\sum_{i=1}^K \frac{rel_i}{\log_2(i+1)}}{IDCG}$$

Accuracy

$$= \frac{\text{\# of Correct Responses}}{\text{Total \# of Responses}}$$

F1-Measure

$$= \frac{2 (\text{precision} * \text{recall})}{\text{precision} + \text{recall}}$$

Results



Experiment 1	Number of Questions (Training/Test)	Hit (Training/Test)	Rate (Training/Test)	MRR (Training/Test)	NDCG (Training/Test)	Accuracy (Training/Test)	ROUGE Score – F1- Score (Training/Test)
Yes/No	223/57	91.48%/85.61%		1.000/0.982	0.994/0.980	77.58%/77.19%	0.442/0.261
List	178/45	89.21%/84.00%		1.000/0.917	0.992/0.934	70.98%/39.30%	0.512/0.264
Factoid	235/51	83.64%/72.16%		0.993/0.813	0.985/0.845	61.33%/33.33%	0.422/0.313
Summary	174/47	87.24%/84.68%		1.000/0.922	0.993/0.919	N/A	0.507/0.330
Overall	800/200	87.85%/81.60%		0.998/0.910	0.991/0.921	69.86%/51.43% ¹	0.466/0.291

Results - continued



Experiment 2	Accuracy (Training/Test)	ROUGE Score – F1-Score (Training/Test)
Yes/No	77.58%/77.19%	0.448/0.233
List	N/A	0.544/0.260
Factoid	N/A	0.438/0.313
Summary	N/A	0.507/0.287
Overall	N/A	0.480/0.272

Analysis



STRONG PERFORMANCE ON
RETRIEVAL METRICS



ROOM FOR IMPROVEMENT ON
GENERATION/ACCURACY METRICS

Next Steps/Conclusion

- Increase number of resources from retriever
- Distinct resources requirement
- Additional instructions depending on type of question
- Supervised training of model, manually recognizing correct answers



Questions



Contact Information

- Name: Matthew Moody
- Email Address: matthewmoody@ymail.com
- LinkedIn: [linkedin.com/in/matthew-moody-8992aa168](https://www.linkedin.com/in/matthew-moody-8992aa168)
- For further information, see the supporting paper ML07