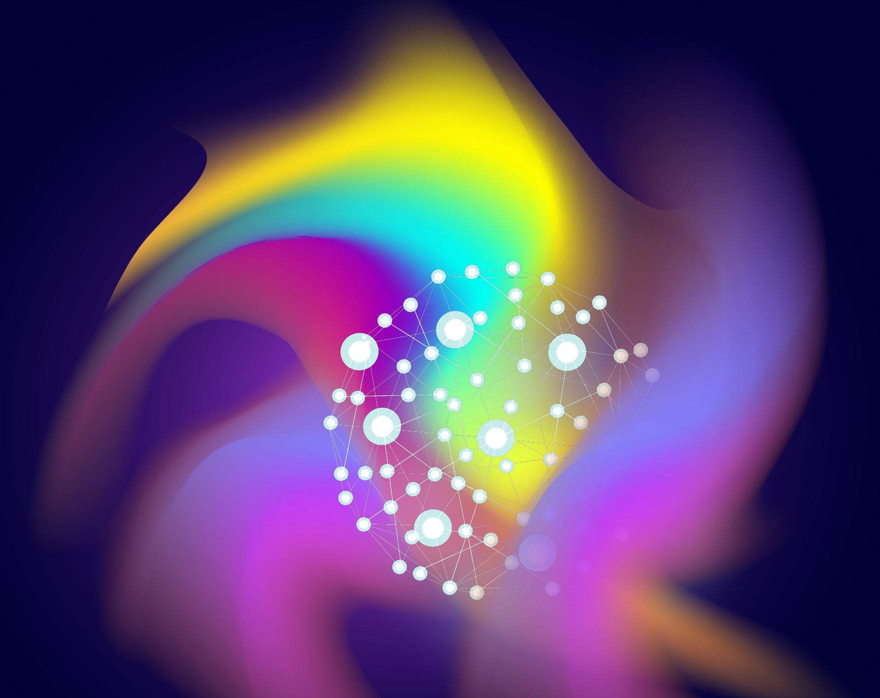
arXiv Chatbot

A RAG-based model to query arXiv papers

Erdos Deep Learning Bootcamp
June - Aug 2024



Team Members

Tantrik Mukerji
Ketan Sand
Xiaoyu Wang
Tajudeen Mamadou Yacoubou
Guoqing Zhang

Outline

- Introduction
- RAG + LLM pipeline
- App design and web deployment
- Summary, Demo and future Work

RAG + Large Language Model (LLM) Intro

Ideal for:

- Research and academic fields needing precise, trustworthy information.
- Niche questions that require specific, detailed answers.

What is RAG?

- **R**etrieval **A**ugmented **G**eneration
- Combines real-time data retrieval with user query to generate answers.

How It Works

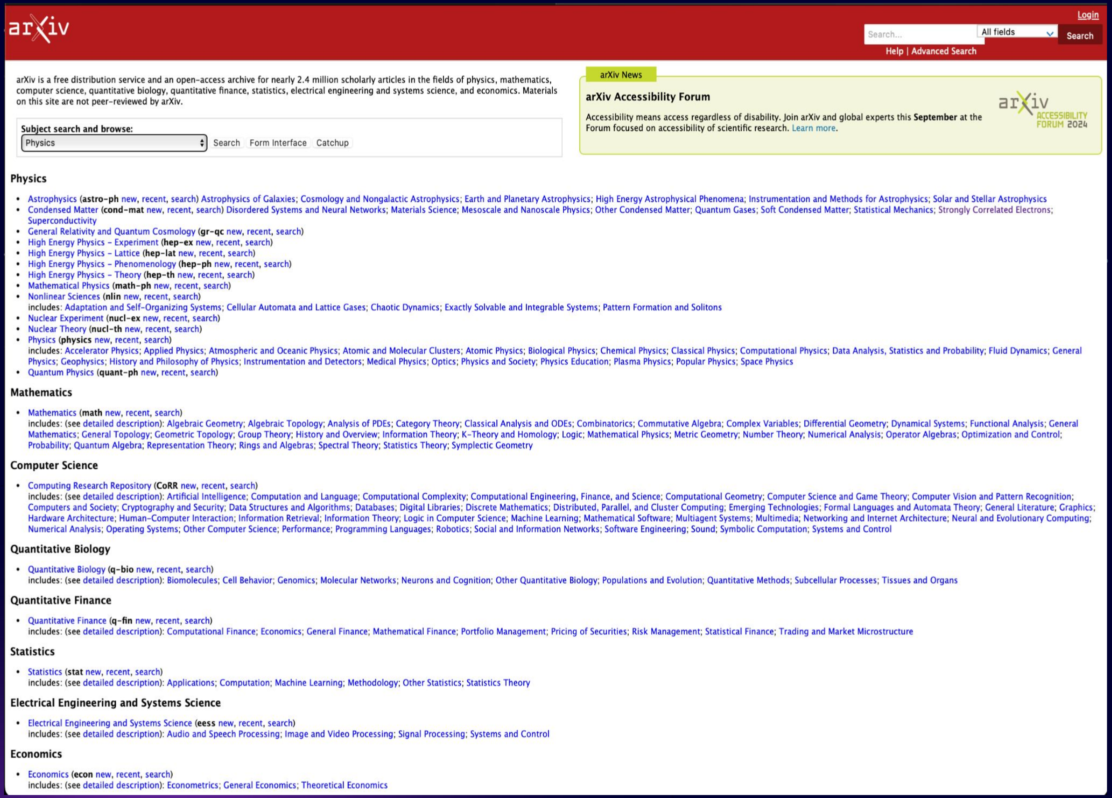
- **Data Retrieval:** Fetches relevant, real-time information from external sources based on semantic search.
- **LLM Processing:** Generates accurate and contextualized responses using the retrieved data.

Why Use RAG + LLM?

- **Reduces Errors:** Minimizes hallucinations common in standard AI models.
- **Up-to-Date:** Provides the latest information, unlike static pre-trained models.
- **Citable Sources:** Delivers responses with verifiable references.
- **Flexible:** Can be adapted to any database

What is arXiv

- Research papers spanning 8 major disciplines
- 2.4 million articles
- Data goldmine for recent and factual information in science
- Multiple use cases combining it with a LLM


arXiv

Login

Search... All fields Search

Help | Advanced Search

arXiv is a free distribution service and an open-access archive for nearly 2.4 million scholarly articles in the fields of physics, mathematics, computer science, quantitative biology, quantitative finance, statistics, electrical engineering and systems science, and economics. Materials on this site are not peer-reviewed by arXiv.

Subject search and browse: Physics | Search | Form Interface | Catchup

arXiv News

arXiv Accessibility Forum

Accessibility means access regardless of disability. Join arXiv and global experts this **September** at the Forum focused on accessibility of scientific research. Learn more.

arXiv ACCESSIBILITY FORUM 2024

Physics

- Astrophysics (**astro-ph** new, recent, search) Astrophysics of Galaxies; Cosmology and Nongalactic Astrophysics; Earth and Planetary Astrophysics; High Energy Astrophysical Phenomena; Instrumentation and Methods for Astrophysics; Solar and Stellar Astrophysics
- Condensed Matter (**cond-mat** new, recent, search) Disordered Systems and Neural Networks; Materials Science; Mesoscale and Nanoscale Physics; Other Condensed Matter; Quantum Gases; Soft Condensed Matter; Statistical Mechanics; Strongly Correlated Electrons; Superconductivity
- General Relativity and Quantum Cosmology (**gr-qc** new, recent, search)
- High Energy Physics – Experiment (**hep-ex** new, recent, search)
- High Energy Physics – Lattice (**hep-lat** new, recent, search)
- High Energy Physics – Phenomenology (**hep-ph** new, recent, search)
- High Energy Physics – Theory (**hep-th** new, recent, search)
- Mathematical Physics (**math-ph** new, recent, search)
- Nonlinear Sciences (**nlin** new, recent, search)
  includes: Adaptation and Self-Organizing Systems; Cellular Automata and Lattice Gases; Chaotic Dynamics; Exactly Solvable and Integrable Systems; Pattern Formation and Solitons
- Nuclear Experiment (**nucl-ex** new, recent, search)
- Nuclear Theory (**nucl-th** new, recent, search)
- Physics (**physics** new, recent, search)
  includes: Accelerator Physics; Applied Physics; Atmospheric and Oceanic Physics; Atomic and Molecular Clusters; Atomic Physics; Biological Physics; Chemical Physics; Classical Physics; Computational Physics; Data Analysis, Statistics and Probability; Fluid Dynamics; General Physics; Geophysics; History and Philosophy of Physics; Instrumentation and Detectors; Medical Physics; Optics; Physics and Society; Physics Education; Plasma Physics; Popular Physics; Space Physics
- Quantum Physics (**quant-ph** new, recent, search)

Mathematics

- Mathematics (**math** new, recent, search)
  includes: (see detailed description): Algebraic Geometry; Algebraic Topology; Analysis of PDEs; Category Theory; Classical Analysis and ODEs; Combinatorics; Commutative Algebra; Complex Variables; Differential Geometry; Dynamical Systems; Functional Analysis; General Mathematics; General Topology; Geometric Topology; Group Theory; History and Overview; Information Theory; K-Theory and Homology; Logic; Mathematical Physics; Metric Geometry; Number Theory; Numerical Analysis; Operator Algebras; Optimization and Control; Probability; Quantum Algebra; Representation Theory; Rings and Algebras; Spectral Theory; Statistics Theory; Symplectic Geometry

Computer Science

- Computing Research Repository (**CoRR** new, recent, search)
  includes: (see detailed description): Artificial Intelligence; Computation and Language; Computational Complexity; Computational Engineering, Finance, and Science; Computational Geometry; Computer Science and Game Theory; Computer Vision and Pattern Recognition; Computers and Society; Cryptography and Security; Data Structures and Algorithms; Databases; Digital Libraries; Discrete Mathematics; Distributed, Parallel, and Cluster Computing; Emerging Technologies; Formal Languages and Automata Theory; General Literature; Graphics; Hardware Architecture; Human-Computer Interaction; Information Retrieval; Information Theory; Logic in Computer Science; Machine Learning; Mathematical Software; Multiagent Systems; Multimedia; Networking and Internet Architecture; Neural and Evolutionary Computing; Numerical Analysis; Operating Systems; Other Computer Science; Performance; Programming Languages; Robotics; Social and Information Networks; Software Engineering; Sound; Symbolic Computation; Systems and Control

Quantitative Biology

- Quantitative Biology (**q-bio** new, recent, search)
  includes: (see detailed description): Biomolecules; Cell Behavior; Genomics; Molecular Networks; Neurons and Cognition; Other Quantitative Biology; Populations and Evolution; Quantitative Methods; Subcellular Processes; Tissues and Organs

Quantitative Finance

- Quantitative Finance (**q-fin** new, recent, search)
  includes: (see detailed description): Computational Finance; Economics; General Finance; Mathematical Finance; Portfolio Management; Pricing of Securities; Risk Management; Statistical Finance; Trading and Market Microstructure

Statistics

- Statistics (**stat** new, recent, search)
  includes: (see detailed description): Applications; Computation; Machine Learning; Methodology; Other Statistics; Statistics Theory

Electrical Engineering and Systems Science

- Electrical Engineering and Systems Science (**eess** new, recent, search)
  includes: (see detailed description): Audio and Speech Processing; Image and Video Processing; Signal Processing; Systems and Control

Economics

- Economics (**econ** new, recent, search)
  includes: (see detailed description): Econometrics; General Economics; Theoretical Economics

Why arXiv with RAG + LLM?

Quick Retrieval Without Manual Searching:

- **Automated Paper Retrieval:** With RAG + LLM, users can bypass the need to manually search for papers by entering specific keywords or topics. The model automatically pulls relevant papers from arXiv.
- **Semantic Search:** Instead of relying on traditional keyword searches, RAG uses a semantic search approach, ensuring that the papers retrieved are not just keyword matches, but contextually relevant to the query.

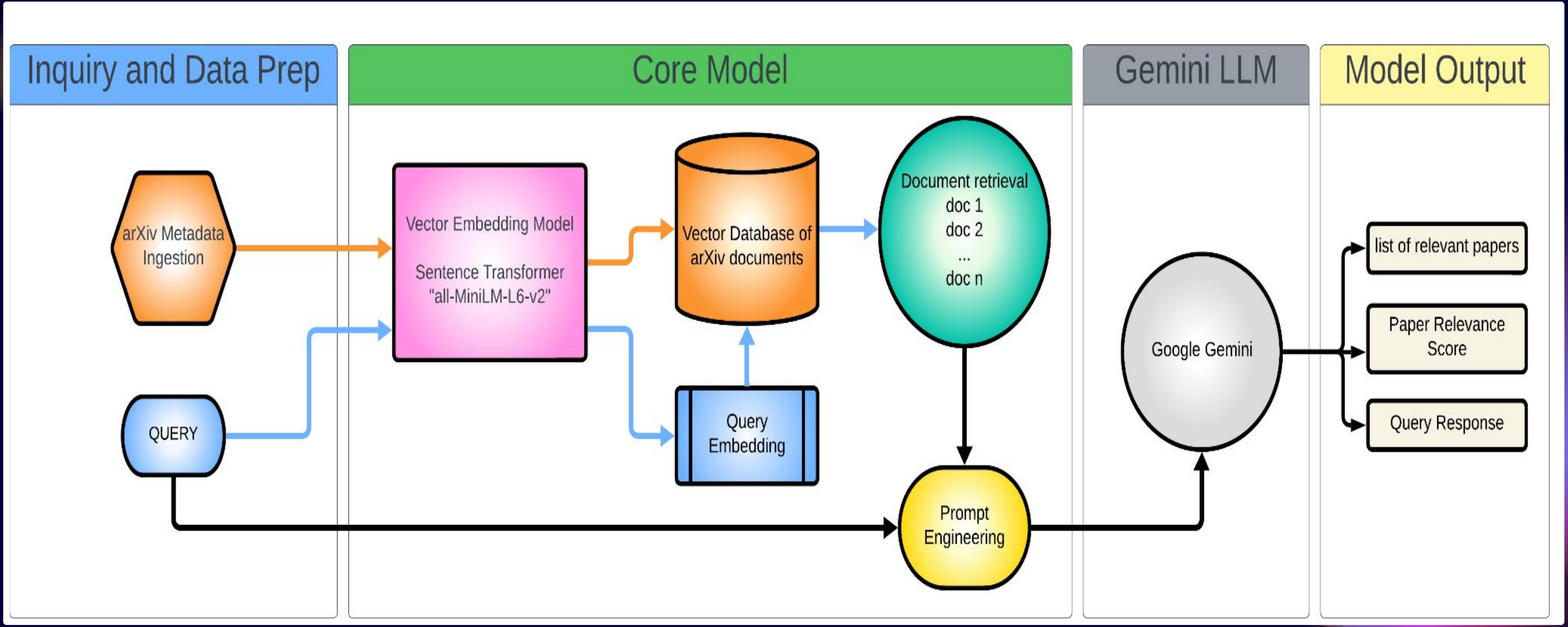
Instant Access to the Latest Research:

- **Real-Time Data Fetching:** The RAG pipeline is connected to arXiv's live database, allowing for immediate access to all query-relevant publications.
- **Dynamic Updating:** As new papers are added to arXiv, RAG + LLM can seamlessly incorporate them, ensuring that users always receive the latest research without needing to manually check for updates.

Simplifying Complex Queries:

- **Contextual Understanding:** RAG’s ability to understand the context of a query means that it can retrieve documents that are highly relevant, even for complex or niche topics.
- **Efficient Document Ranking:** The pipeline ranks the documents based on their relevance to the query, presenting the most pertinent papers first, i.e. those with the highest relevance score.

Pipeline


Inquiry and Data Prep
Core Model
Gemini LLM
Model Output
arXiv Metadata Ingestion
QUERY
Vector Embedding Model
Sentence Transformer "all-MiniLM-L6-v2"
Vector Database of arXiv documents
Query Embedding
Document retrieval
doc 1
doc 2
...
doc n
Prompt Engineering
Google Gemini
list of relevant papers
Paper Relevance Score
Query Response

Document preparation

- User input a topic such as “quantum physics”, “dolphins”, author name, …
- 100 document summaries (title, author, abstract, identifier etc) are loaded using the `arXivLoader` module from `langchain-community`, based on traditional keyword matching approach.

Sentence Transformer

- Creates a high dimensional vector for each document summary. Document summaries longer than N tokens are truncated.
- We use a pre-trained sentence transformer model `all-MiniLM-L6-v2`, which encodes a maximum of 256 tokens, with vector embedding dimension d=384.
- Typical arXiv document abstract length is ~200 words, compatible with `all-MiniLM-L6-v2`.

Document retrieval accuracy (KPI)

- We use relevance score to measure document retrieval accuracy.
- It is obtained from the cosine similarity between vector embeddings of the user query and document summaries.
- A score close to 1 means high relevance.
- Including entirety of document summaries improves the relevance score to 0.5~0.8
  - Upper end achieved by more precise user queries

```python
query = ['research articles about graphene']
documents = [
    'Phases and phase transitions in a dimerized spin-12 XXZ chain ',
    'Strongly interacting Hofstadter states in magic-angle twisted bilayer graphene',
    'Constraints imposed by symmetry on pairing operators for the iron pnictides',
    'Interplay between tetragonal magnetic order, stripe magnetism, and superconductivity in iron-based materials',
    'Visualizing the nonlinear coupling between strain and electronic nematicity in the iron pnictides by elasto-scanning tunneling spectroscopy',
    'Strong-coupling expansion of multi-band interacting models: Mapping onto the transverse-field J1-J2 Ising model'
]
model = SentenceTransformer("all-MiniLM-L6-v2")
doc_embedding = model.encode(documents)
query_embedding = model.encode(query)
scores = model.similarity(query_embedding,doc_embedding)
for i in range(len(documents)):
    print(documents[i]+': '+ f'{scores[0][i].item():.2f}')
```

[14] ✓ 0.5s Python

```
Phases and phase transitions in a dimerized spin-12 XXZ chain : 0.13
Strongly interacting Hofstadter states in magic-angle twisted bilayer graphene: 0.47
Constraints imposed by symmetry on pairing operators for the iron pnictides: 0.12
Interplay between tetragonal magnetic order, stripe magnetism, and superconductivity in iron-based materials: 0.13
Visualizing the nonlinear coupling between strain and electronic nematicity in the iron pnictides by elasto-scanning tunneling spectroscopy: 0.21
Strong-coupling expansion of multi-band interacting models: Mapping onto the transverse-field J1-J2 Ising model: 0.10
```

Prompt Engineering

```
prompt = """
    You are a question-answer bot that provides answers in the scientific domain.
    Given the provided context: {rag_context}
    Answer user's question "{query}" on the topic of {user_input}.
    When answering the question, try to make use of arxiv_specifiers provided in the context.\n
"""
```

Choice of Commercial LLM

Gemini:

- We use Gemini, a family of large language models developed by Google.
- This is the language model that is used to generate answers to the users queries.
- Gemini advertises as being capable of complex reasoning and offers an easy to obtain API, and has a large free query allowance of up to 1500.

Deployment & Demo (streamlit)

Project summary

- We stress tested our RAG+LLM pipeline on a broad spectrum of topics covered by arXiv.org. The results are generally **satisfactory in the eyes of** team members who are **domain experts**
- The retrieved documents have **relevance scores in the 0.5 to 0.8 range**. Higher scores are typically obtained if the user query contains keywords that are also in the document summaries.
- The final generated response is contextualized with the retrieved documents, **providing accurate answers while also citing relevant sources**.
- The pipeline is deployed as a **web app** at https://erdos-arxiv-chatbot.streamlit.app/, which has a clean user interface, and instructions on how to use it.

Future Work

- The future work will revolve around two aspects: improvements to the core RAG module, and adding new features.
- These improvements and new features include
  - Pre-built vector database
  - Whole document ingestion
  - Customized sentence LLM
  - Bibliography construction
  - Abstract Generation

Acknowledgement

Erdos Institute Deep Learning Bootcamp

Instructor: Lindsay Warrenburg
TA: Marcos Ortiz

Deployment: streamlit.io