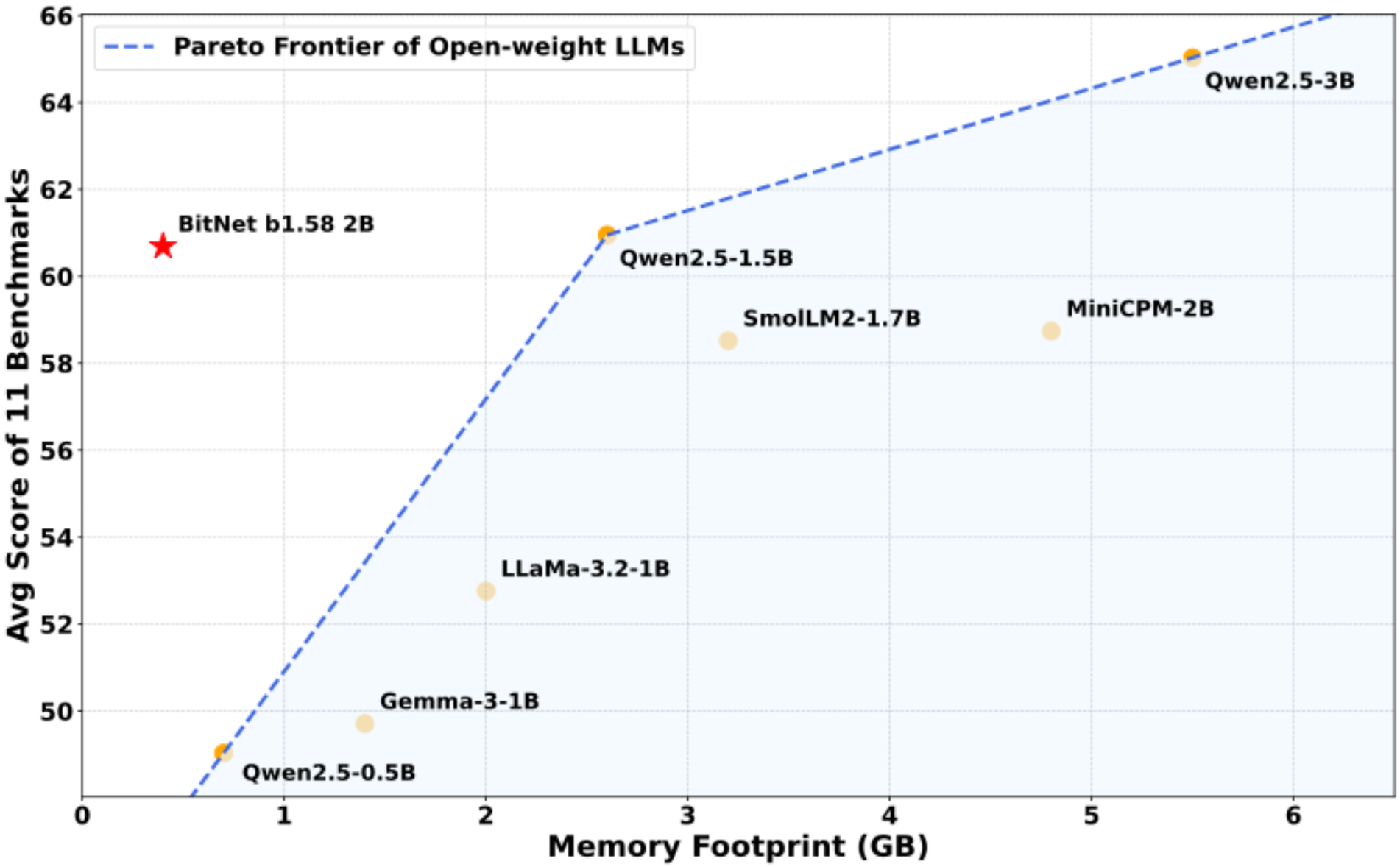
CS 696 Applied Large Language Models
Spring Semester, 2025
Doc 22 RAG
Apr 17, 2025

BitNet - 1-bit LLMs

https://github.com/microsoft/BitNet

https://arxiv.org/abs/2410.16144



BitNet - 1-bit LLMs

https://github.com/microsoft/BitNet

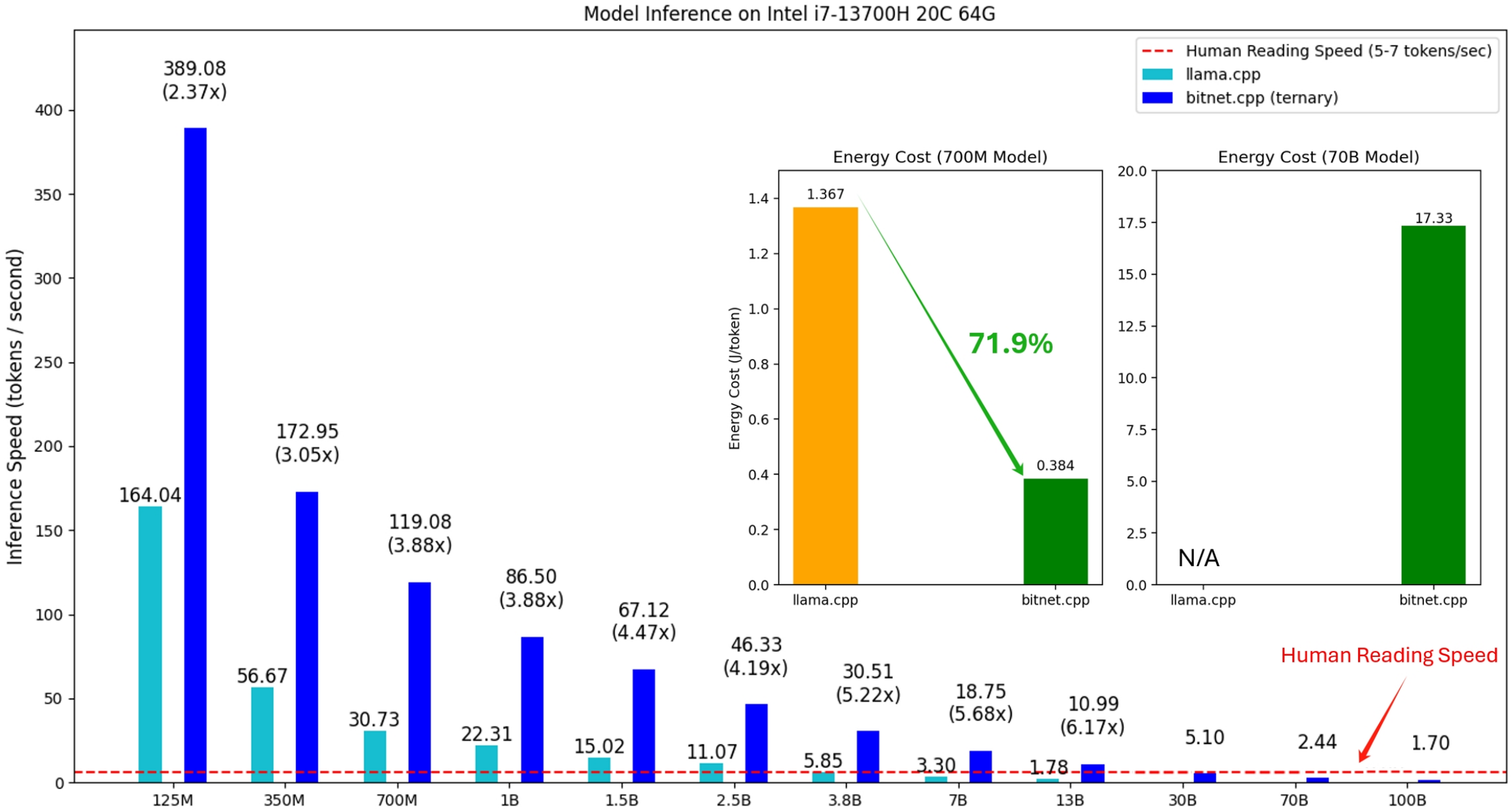
https://arxiv.org/abs/2410.16144

Benchmark (Metric)	LLaMA 3.2 1B	Gemma-3 1B	Qwen2.5 1.5B	SmolLM2 1.7B	MiniCPM 2B	BitNet b1.58 2B
Memory (Non-emb)	2GB	1.4GB	2.6GB	3.2GB	4.8GB	**0.4GB**
Latency (CPU; TPOT)	48ms	41ms	65ms	67ms	124ms	**29ms**
Energy (Estimated)	0.258J	0.186J	0.347J	0.425J	0.649J	**0.028J**
Training Tokens (Pre-training)	9T (pruning & distillation)	2T (distillation)	18T	11T	1.1T	4T

BitNet - 1-bit LLMs

https://github.com/microsoft/BitNet

https://arxiv.org/abs/2410.16144



Differentiable Programming from Scratch

https://thenumb.at/Autodiff/

Automatic Differentiation - autodiff

Two ways to compute the derivative of a Python function

Forward Mode
Backward Mode

Testing 18 RAG Techniques to Find the Best

https://levelup.gitconnected.com/testing-18-rag-techniques-to-find-the-best-094d166af27f

Simple RAG
Semantic Chunking
Context Enriched Retrieval
Contextual Chunk Headers
Document Augmentation
Query Transformation
Re-Ranker
RSE
Contextual Compression
Feedback Loop
Adaptive RAG
Self RAG
Knowledge Graph
Hierarchical Indices
HyDE
Fusion
Multi Model
Crag

Did not use LangChain or FAISS

Paper & Github code

Test Query and LLMs

Test query and its true answer.

PDF document on which RAG will be applied.

Embedding generation model.

Response and validation LLM.

Reference Document

16+ page PDF generated by Claude 3.5 Thinking model,

Understanding Artificial Intelligence

Chapter 1: Introduction to Artificial Intelligence

Artificial intelligence (AI) refers to the ability of a digital computer or computer-controlled robot to perform tasks commonly associated with intelligent beings. The term is frequently applied to the project of developing systems endowed with the intellectual processes characteristic of humans, such as the ability to reason, discover meaning, generalize, or learn from past experience. Over the past few decades, advancements in computing power and data availability have significantly accelerated the development and deployment of AI.

Historical Context

The idea of artificial intelligence has existed for centuries, often depicted in myths and fiction. However, the formal field of AI research began in the mid-20th century. The Dartmouth Workshop in 1956 is widely considered the birthplace of AI. Early AI research focused on problem-solving and symbolic methods. The 1980s saw a rise in expert systems, while the 1990s and 2000s brought advancements in machine learning and neural networks. Recent breakthroughs in deep learning have revolutionized the field.

Modern Observations

Response generation and validation

Paper

Embeddings
TaylorAI/gte-tiny

Response & Validation
LLaMA-3.2–3B Instruct

GitHub code

Embeddings
BAAI/bge-en-icl

Response & Validation
LLaMA-3.2–3B Instruct

My Runs

Embeddings	Response	Validation
text-embedding-3-large	gpt-4.1-nano	gpt-4

```
from openai import OpenAI
client = OpenAI(
    base_url="https://api.studio.nebius.com/v1/",
    api_key=os.getenv("OPENAI_API_KEY")
)
```

Test Query - Paper

test query:

How does AI’s reliance on massive data sets act as a double-edged sword?

True Answer:

It drives rapid learning and innovation while also risking the amplification of inherent biases, making it crucial to balance data volume with fairness and quality.

Gihup Test Queries

question:

What is 'Explainable AI' and why is it considered important?

ideal_answer:

Explainable AI (XAI) aims to make AI systems more transparent and understandable, providing in
into how they make decisions. It's considered important for building trust, accountability, and ens
fairness in AI systems

reference:

Chapter 5: The Future of Artificial Intelligence - Explainable AI (XAI); Chapter 19: AI and Ethics,

has_answer:

true

reasoning:

The document directly defines and explains the importance of XAI.

Gihup Test Queries

What is 'Explainable AI' and why is it considered important?

Can AI be used to predict earthquakes?

What are some of the ethical concerns related to AI-powered facial recognition?

How does AI contribute to personalized medicine?

Does the document mention any specific companies developing AI technology?

What is the role of AI in smart grids?

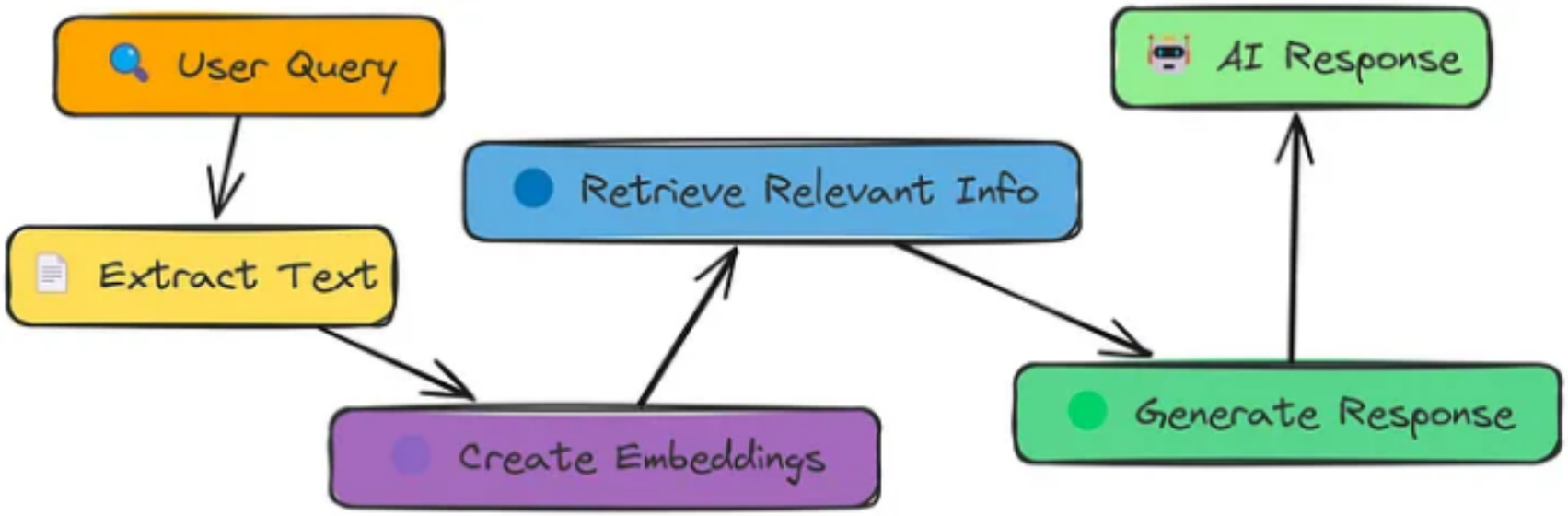
Can AI write a complete, original novel?

What is a 'cobot'?

What is Direct Air Capture (DAC) used for?

Is AI currently being used to control nuclear weapons systems?

Simple RAG



Simple RAG workflow (Created by Fareed Khan)

Extract text from a PDF.

Split the text into smaller chunks.

- Chunks based on length only
- 1000 chars, 200 char overlap

Convert the chunks into numerical embeddings.

Search for the most relevant chunks based on a query.

- Use top two chunks

Generate a response using the retrieved chunks.

Compare the response with the correct answer to evaluate accuracy.

semantic_search(query, text_chunks, embeddings, k=5)

query (str): The query for the semantic search.
text_chunks (List[str]): A list of text chunks to search through.
embeddings (List[dict]): A list of embeddings for the text chunks.
k (int): The number of top relevant text chunks to return. Default is 5.

```
    query_embedding = create_embeddings(query).data[0].embedding
    similarity_scores = []

    # Calculate similarity scores between the query embedding and each text chunk embedding
    for i, chunk_embedding in enumerate(embeddings):
        similarity_score = cosine_similarity(np.array(query_embedding),
                                             np.array(chunk_embedding.embedding))
        similarity_scores.append((i, similarity_score))

    similarity_scores.sort(key=lambda x: x[1], reverse=True)
    top_indices = [index for index, _ in similarity_scores[:k]]
    return [text_chunks[index] for index in top_indices]
```

End of first chunk

> However, the formal field of AI research b**egan** in the mid-20th century. The Dartmouth Workshop in 1956 is widely considered the birthplace of AI. Early AI research focused on problem-solving and symbolic methods. The 1980s saw a rise in **exp**

Start of second chunk

egan in the mid-20th century. The Dartmouth Workshop
in 1956 is widely considered the birthplace of AI. Early AI research focused on problem-solving and symbolic methods. The 1980s saw a rise in **exp**ert systems, while the 1990s and 2000s

System Prompt

"You are an AI assistant that strictly answers based on the given context. If the answer cannot be derived directly from the provided context, respond with: 'I do not have enough information to answer that.'"

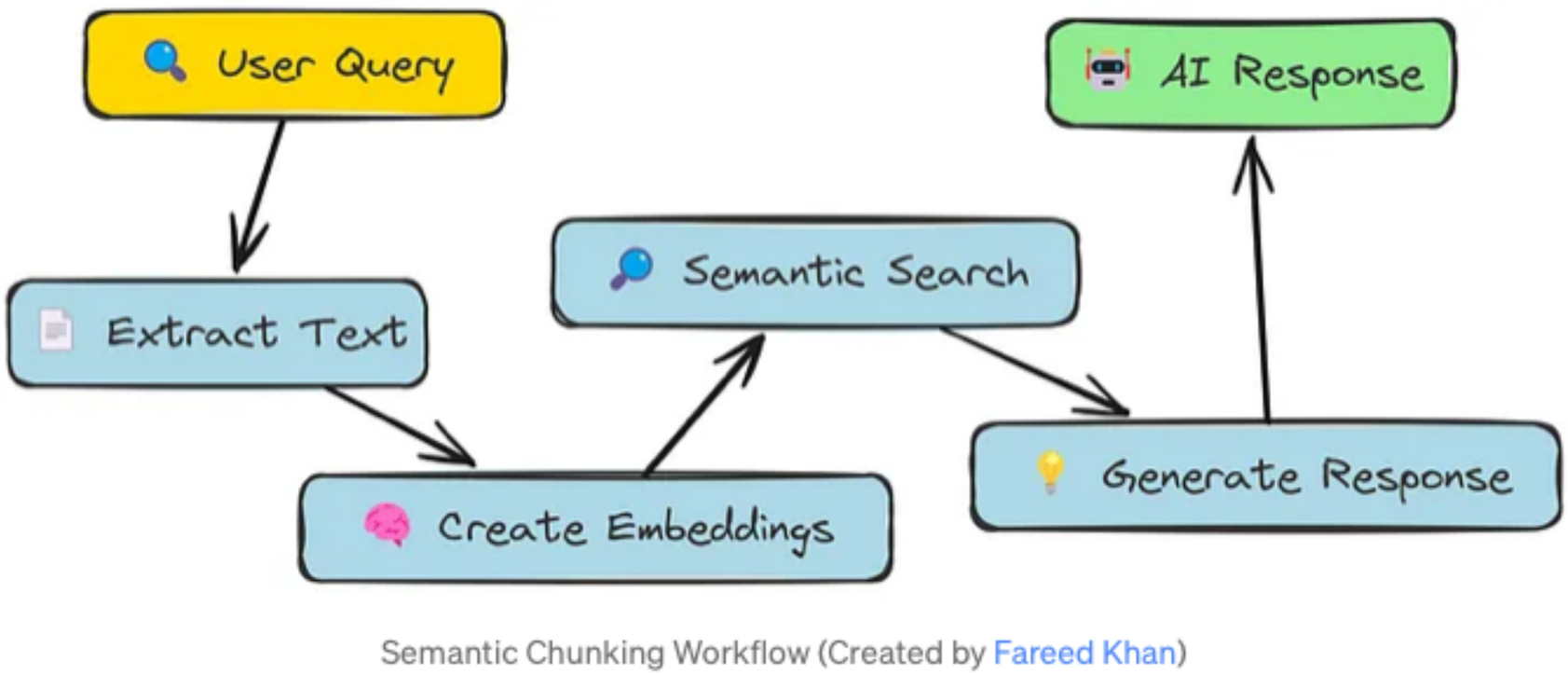
User Prompt

"Context 1:\nsystems. Explainable AI (XAI) \ntechniques aim to make AI decisions more understandable, enabling users to assess their \nfairness and accuracy. \nPrivacy and Data Protection \nAI systems often rely on large amounts of data, raising concerns about privacy and data \nprotection. Ensuring responsible data handling, implementing privacy-preserving techniques, \nand complying with data protection regulations are crucial. \nAccountability and Responsibility \nEstablishing accountability and responsibility for AI systems is essential for addressing potential \nharms and ensuring ethical behavior. This includes defining roles and responsibilities for \ndevelopers, deployers, and users of AI systems. \nChapter 20: Building Trust in AI \nTransparency and Explainability \nTransparency and explainability are key to building trust in AI. Making AI systems understandable \nand providing insights into their decision-making processes helps users assess their reliability \nand fairness. \nRobustness and Reliability \n\n==========\n\n
Context 2:\nnt aligns with societal values. Education and awareness campaigns inform the public \nabout AI, its impacts, and its potential. \nChapter 19: AI and Ethics \nPrinciples of Ethical AI \nEthical AI principles guide the development and deployment of AI systems to ensure they are fair, \ntransparent, accountable, and beneficial to society. Key principles include respect for human \nrights, privacy, non-discrimination, and beneficence. \n \n \nAddressing Bias in AI \nAI systems can inherit and amplify biases present in the data they are trained on, leading to unfair \nor discriminatory outcomes. Addressing bias requires careful data collection, algorithm design, \nand ongoing monitoring and evaluation. \nTransparency and Explainability \nTransparency and explainability are essential for building trust in AI systems. Explainable AI (XAI) \ntechniques aim to make AI decisions more understandable, enabling users to assess their \nfairness and accuracy. \nPrivacy and Data Protection \nAI systems often rely on la\n==========\n\n
Question: What is 'Explainable AI' and why is it considered important?"

Rating From 0 to 1

	Paper	Code	Me
Simple	0.3	0.8	1.0

Semantic Chunking



Semantic Chunking Workflow (Created by Fareed Khan)

Divide into sentences - split on periods

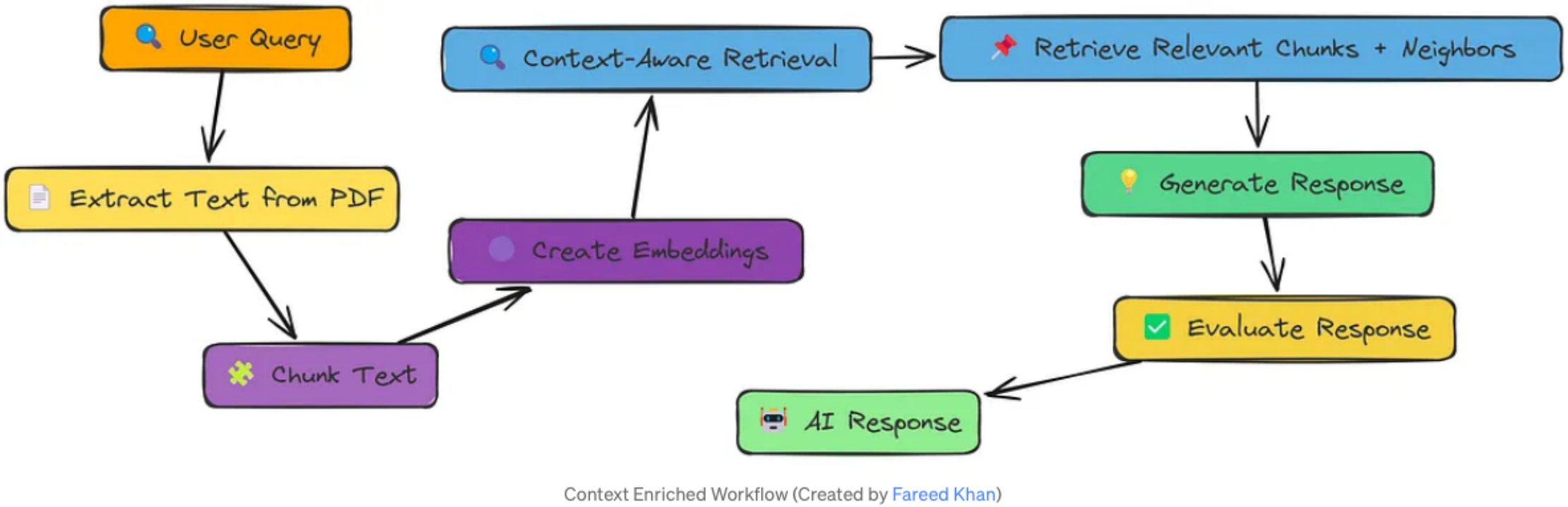
If adjacent sentences are similar enough, combine them

Average chunk size 144 characters

Rating From 0 to 1

	Paper	Github Code	Me
Simple	0.3	0.8	1.0
Semantic Chunking	0.2	0.5	1.0

Context-Enriched Retrieval



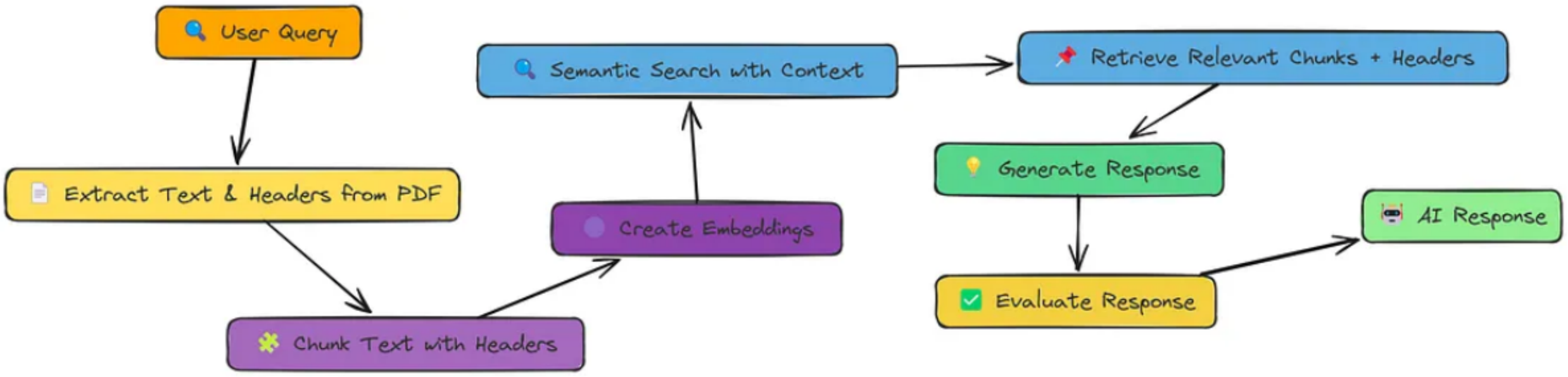
Context Enriched Workflow (Created by Fareed Khan)

Chunk Size = 1,000, Overlap = 200

Select two most relevant chunks and their surrounding chunks

Type	Paper	Github Code	Me
Simple	0.3	0.8	1.0
Semantic Chunking	0.2	0.5	1.0
Context-Enriched Retrieval	0.6	0.8	

Contextual Chunk Headers



Contextual Chunk Headers (Created by Fareed Khan)

Paper indicates:

Generate descriptive header for each chuck in Context-Enriched Retrieval

Code does:

Chunk: 1000 chars, 200 char overlap
Generate a descriptive header for each chunk using LLM

Sample Chunk

Header:

"Introduction to Artificial Intelligence: Understanding the Foundations and Evolution"

Content:

Understanding Artificial Intelligence

Chapter 1: Introduction to Artificial Intelligence

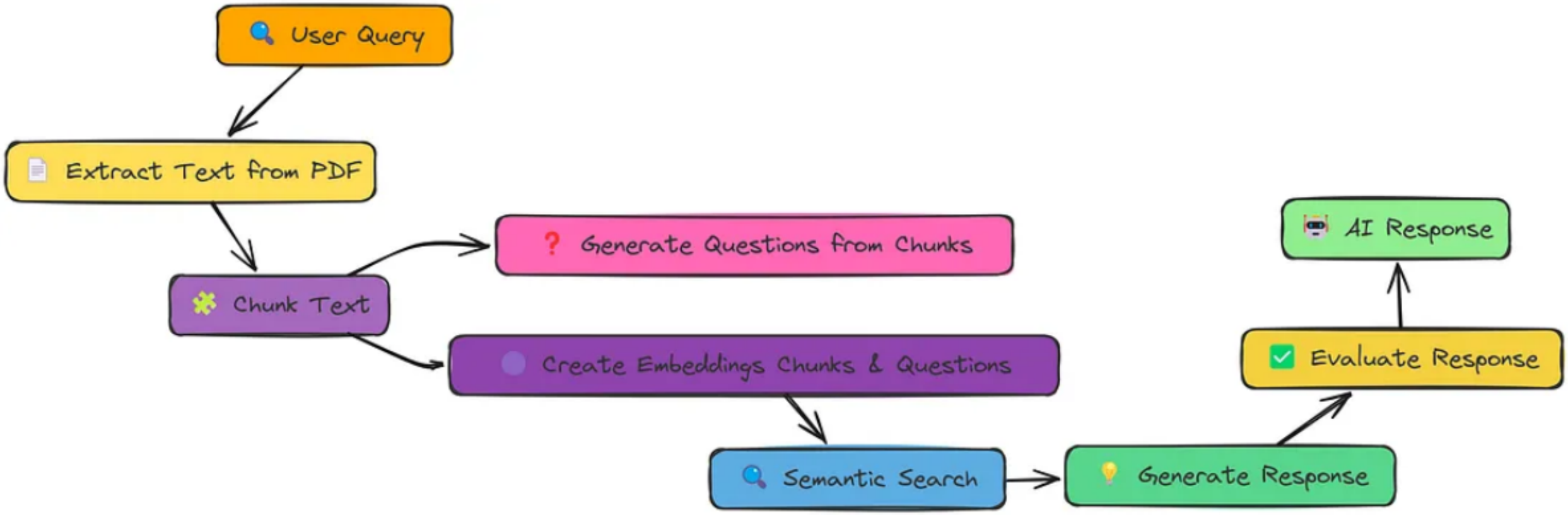
Artificial intelligence (AI) refers to the ability of a digital computer or computer-controlled robot to perform tasks commonly associated with intelligent beings. The term is frequently applied to the project of developing systems endowed with the intellectual processes characteristic of humans, such as the ability to reason, discover meaning, generalize, or learn from past experience. Over the past few decades, advancements in computing power and data availability have significantly accelerated the development and deployment of AI.

Historical Context

The idea of artificial intelligence has existed for centuries, often depicted in myths and fiction. However, the formal field of AI research began in the mid-20th century. The Dartmouth Workshop in 1956 is widely considered the birthplace of AI. Early AI research focused on problem-solving and symbolic methods. The 1980s saw a rise in exp

Type	Paper	Github Code	Me
Simple	0.3	0.8	1.0
Semantic Chunking	0.2	0.5	1.0
Context-Enriched Retrieval	0.6	0.8	
Contextual Chunk Headers	0.5	0.5	

Document Augmentation



Document Augmentation Workflow (Created by Fareed Khan)

Have an LLM generate three questions for each chunk

Questions are added to vector database with chunk id

Prompts for Questions

system_prompt = "You are an expert at generating relevant questions from text. Create concise questions that can be answered using only the provided text. Focus on key information and concepts."

Define the user prompt with the text chunk and the number of questions to generate

user_prompt = f"""
Based on the following text, generate {num_questions} different questions that can be answered using only this text:

{text_chunk}

Format your response as a numbered list of questions only, with no additional text.
"""

Query: What is 'Explainable AI' and why is it considered important?

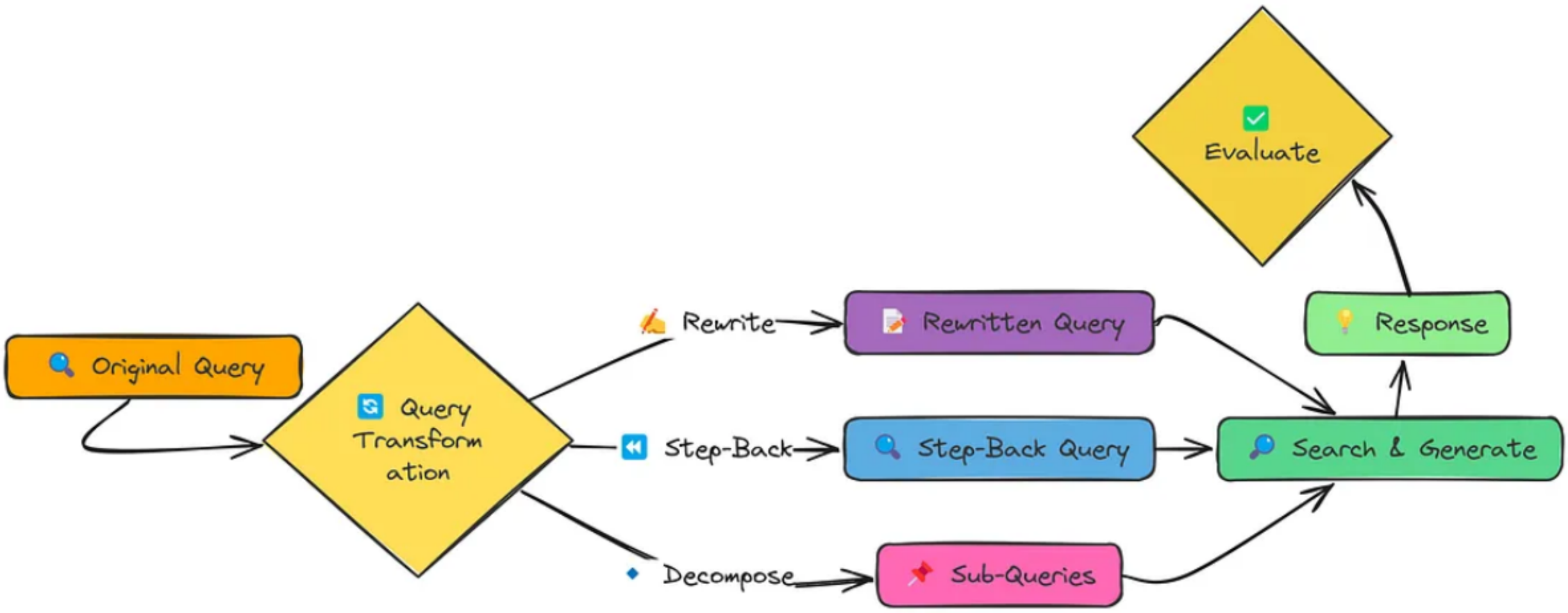
Search Results:

Relevant Document Chunks:

Matched Questions:
Question 1 (similarity: 0.8629):
What is the main goal of Explainable AI (XAI)?
From chunk 10
==========
Question 2 (similarity: 0.8488):
What is the primary goal of Explainable AI (XAI) techniques?
From chunk 37
==========
Question 3 (similarity: 0.8414):
What is the focus of research on Explainable AI (XAI)?
From chunk 29
==========
Question 4 (similarity: 0.7995):
Why are transparency and explainability essential for building trust in AI systems?
From chunk 36
==========
Question 5 (similarity: 0.7841):
Why is transparency and explainability essential in building trust and accountability with AI systems?
From chunk 9
==========

Type	Paper	Github Code	Me
Simple	0.3	0.8	1.0
Semantic Chunking	0.2	0.5	1.0
Context-Enriched Retrieval	0.6	0.8	
Contextual Chunk Headers	0.5	0.5	
Document Augmentation	0.8	0.8	

Query Transformation



Query Transformation Workflow (Created by Fareed Khan)

Generate three different types of queries

Search the Vector database using each query separately

Send each result separately to LLM

Rewrite Query - Prompt to LLM

system_prompt = "You are an AI assistant specialized in improving search queries. Your task is to rewrite user queries to be more specific, detailed, and likely to retrieve relevant information."

Define the user prompt with the original query to be rewritten

user_prompt = f"""

Rewrite the following query to make it more specific and detailed. Include relevant terms and concepts that might help in retrieving accurate information.

Original query: {original_query}

Rewritten query:

"""

```
response = client.chat.completions.create(
        model=model,
        temperature=0.0,  # Low temperature for deterministic output
        messages=[
            {"role": "system", "content": system_prompt},
            {"role": "user", "content": user_prompt}
        ]
    )
```

Original Query: What are the impacts of AI on job automation and employment?

Rewritten Query:

"What are the current and projected impacts of artificial intelligence (AI) on job automation and employment, including the types of jobs most susceptible to automation, the skills required to remain employable in an AI-driven economy, and the potential effects on unemployment rates, social welfare systems, and the gig economy?"

Step-back Prompting - Prompt to LLM

system_prompt = "You are an AI assistant specialized in search strategies. Your task is to generate broader, more general versions of specific queries to retrieve relevant background information."

```
    # Define the user prompt with the original query to be generalized
    user_prompt = f"""
    Generate a broader, more general version of the following query that could help retrieve useful background information.

    Original query: {original_query}

    Step-back query:
    """

response = client.chat.completions.create(
        model=model,
        temperature=0.1,  # Slightly higher temperature for some variation
        messages=[
            {"role": "system", "content": system_prompt},
            {"role": "user", "content": user_prompt}
        ]
    )
```

Original Query: What are the impacts of AI on job automation and employment?

Step-back Query:

"Effects of automation and artificial intelligence on the modern workforce and labor market, including trends, challenges, and potential implications for employment and economic growth."

Sub-query Decomposition - Prompt to LLM

system_prompt = "You are an AI assistant specialized in breaking down complex questions. Your task is to decompose complex queries into simpler sub-questions that, when answered together, address the original query."

Define the user prompt with the original query to be decomposed

user_prompt = f"""

Break down the following complex query into {num_subqueries} simpler sub-queries. Each sub-query should focus on a different aspect of the original question.

Original query: {original_query}

Generate {num_subqueries} sub-queries, one per line, in this format:

1. [First sub-query]

2. [Second sub-query]

And so on...

"""

temperature=0.2

Original Query: What are the impacts of AI on job automation and employment?

Sub-queries:

1. What are the primary job roles that are most susceptible to automation by AI?

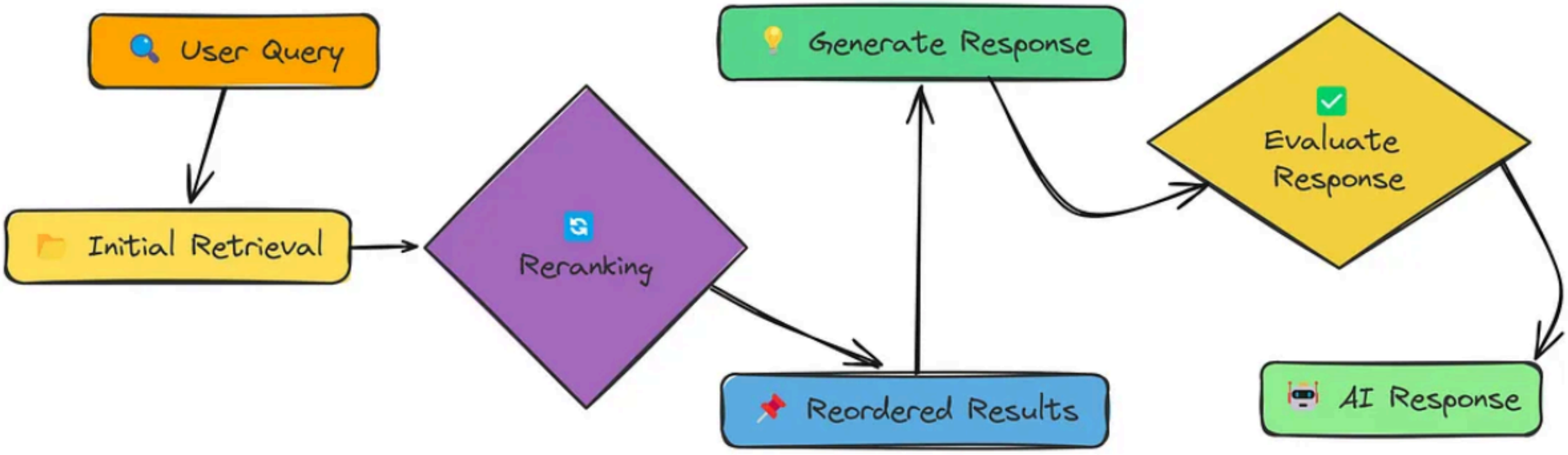
2. How does AI-driven automation affect the overall job market, including the creation of new job opportunities?

3. What are the potential consequences of widespread AI-driven job automation on employment rates and workforce demographics?

4. How do governments, industries, and individuals respond to and mitigate the impacts of AI on job automation and employment?

Type	Paper	Github Code	Me
Simple	0.3	0.8	1.0
Semantic Chunking	0.2	0.5	1.0
Context-Enriched Retrieval	0.6	0.8	
Contextual Chunk Headers	0.5	0.5	
Document Augmentation	0.8	0.8	
Query Transformation	0.5		
Query Transformation Original Query		0.8	
Query Transformation: Rewrite Query		0.9	
Query Transformation: Step_back		0.85	
Query Transformation: Decompose Query		0.95	

Reranker



Reranker (Created by Fareed Khan)

Have an LLM rerank the chunks from the Vector Store

Type	Paper	Github Code	Me
Simple	0.3	0.8	1.0
Semantic Chunking	0.2	0.5	1.0
Context-Enriched Retrieval	0.6	0.8	
Contextual Chunk Headers	0.5	0.5	
Document Augmentation	0.8	0.8	
Query Transformation	0.5		
Query Transformation Original Query		0.8	
Query Transformation: Rewrite Query		0.9	
Query Transformation: Step_back		0.85	
Query Transformation: Decompose Query		0.95	
Reranker	0.7	0.6	

Relevant Segment Extraction (RSE)

Chunk the document

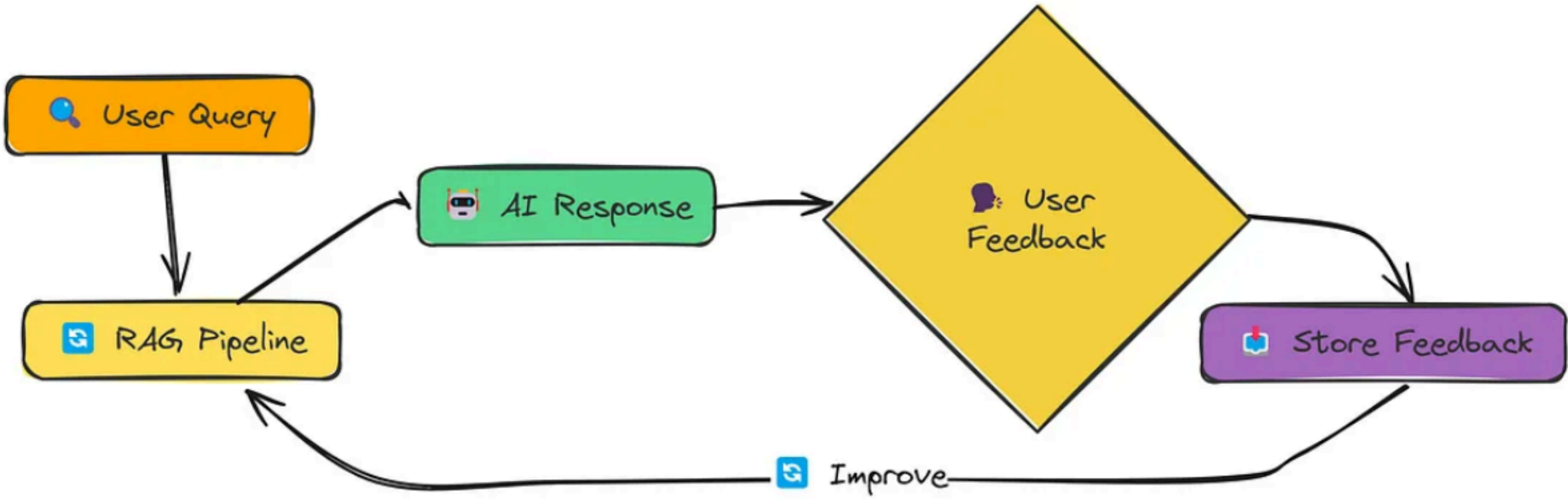
Find the best segments in chunks relevant to the query

Uses a clever algorithm to find the best contiguous segments of chunks

Combines those segments into a context

Type	Paper	Github Code	Me
Simple	0.3	0.8	1.0
Semantic Chunking	0.2	0.5	1.0
Context-Enriched Retrieval	0.6	0.8	
Contextual Chunk Headers	0.5	0.5	
Document Augmentation	0.8	0.8	
Query Transformation	0.5		
Query Transformation Original Query		0.8	
Query Transformation: Rewrite Query		0.9	
Query Transformation: Step_back		0.85	
Query Transformation: Decompose Query		0.95	
Reranker	0.7	0.6	
RSE	0.8		

Feedback Loop



Feedback Loop (Created by Fareed Khan)

The user provides feedback on the RAG system's response (e.g., good/bad, relevant/irrelevant).
The system stores this feedback.
Future retrievals use this feedback to improve.

Result

He ran the same query twice

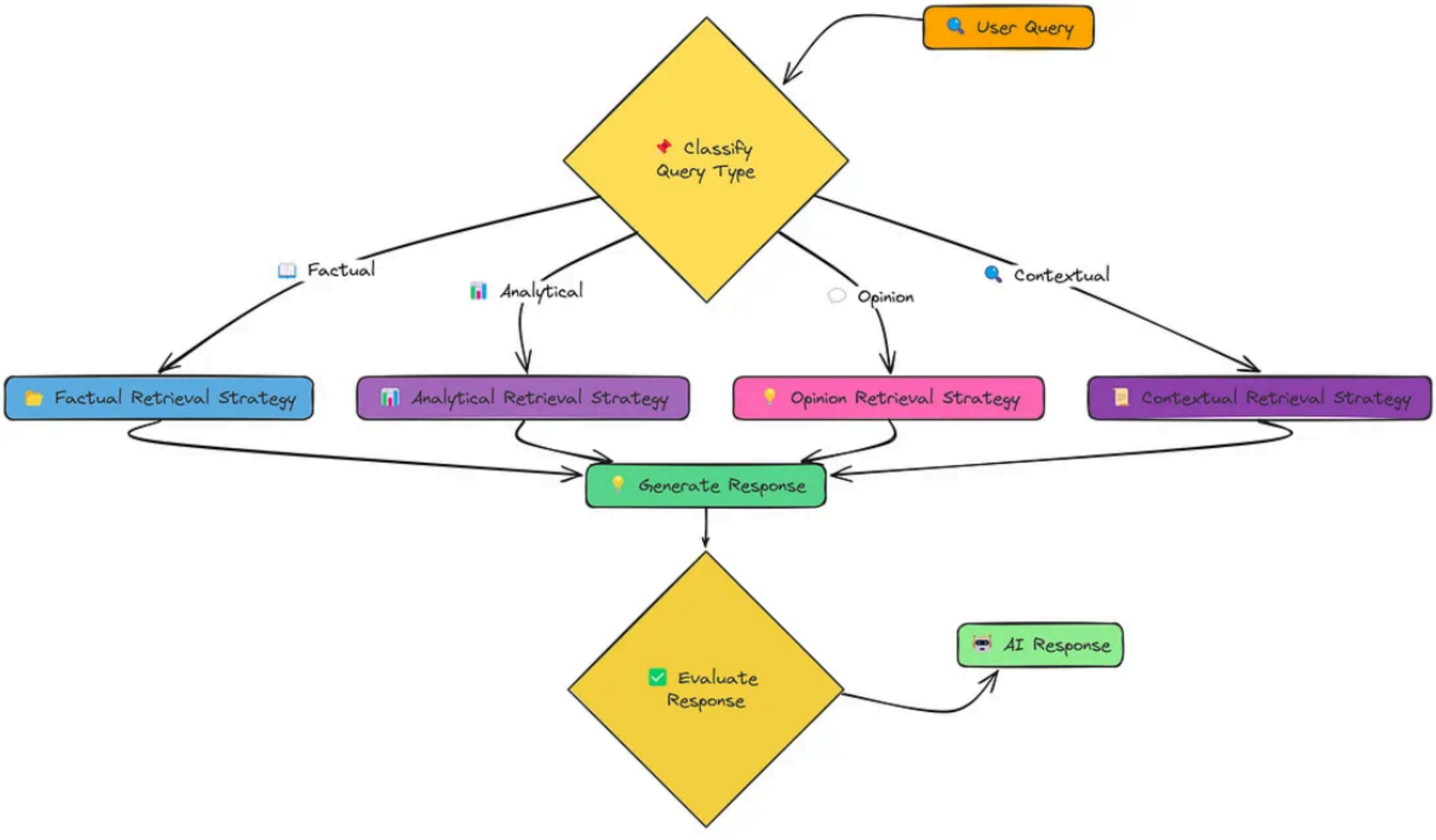
First with no feedback

Second time with feedback from user

The feedback loop has significantly improved the response quality by:

* Providing a clear and concise explanation of how neural networks function
* Providing more specific examples of applications
* Addressing the limitations of the standard RAG response (e.g., lack of clarity and concision)
* Providing a more accurate explanation of the components of a neural network

Adaptive RAG



Adaptive Rag Workflow (Created by Fareed Khan)

Type	Paper	Github Code	Me
Simple	0.3	0.8	1.0
Semantic Chunking	0.2	0.5	1.0
Context-Enriched Retrieval	0.6	0.8	
Contextual Chunk Headers	0.5	0.5	
Document Augmentation	0.8	0.8	
Query Transformation	0.5		
Query Transformation Original Query		0.8	
Query Transformation: Rewrite Query		0.9	
Query Transformation: Step_back		0.85	
Query Transformation: Decompose Query		0.95	
Reranker	0.7	0.6	
RSE	0.8		
Adaptive RAG	~0.856		