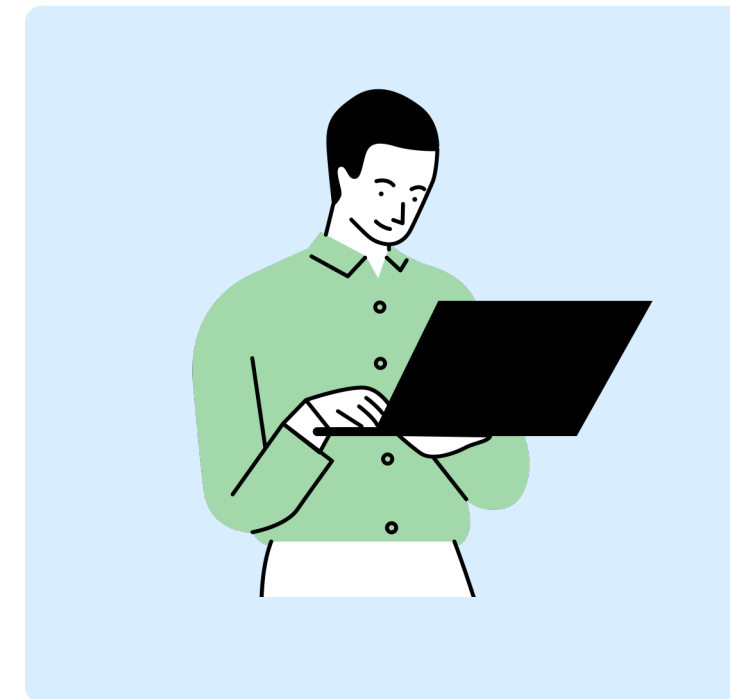
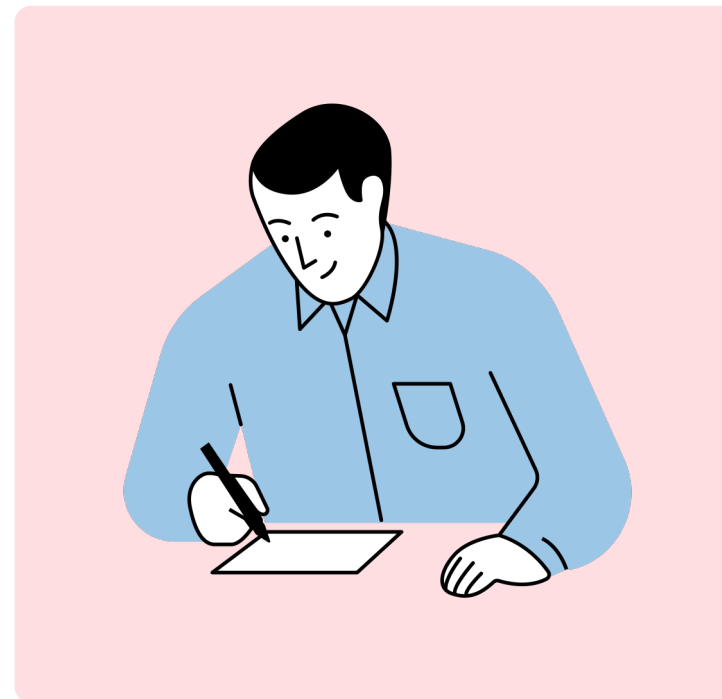
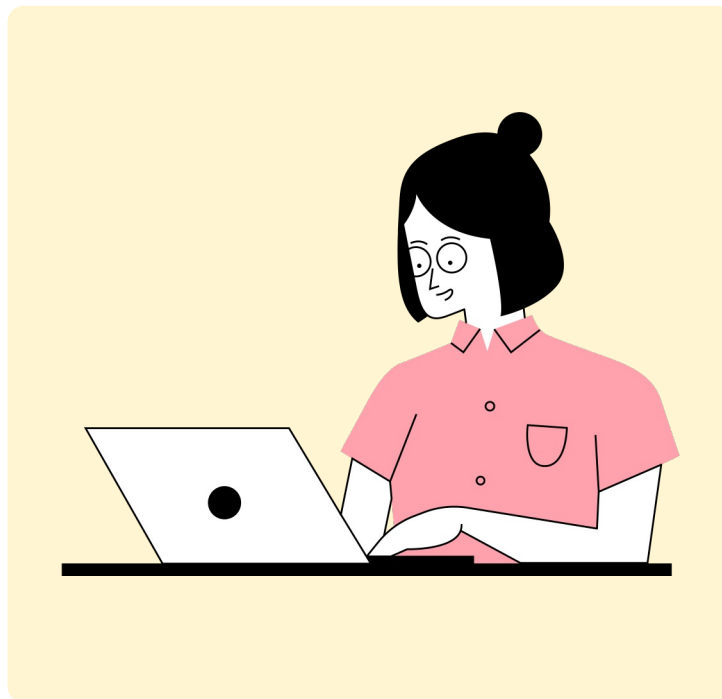


AI-assisted Communication



Instructors: Lingzi Hong & Xiaoying Song

Presenter Profile

Dr. Lingzi Hong is an Assistant Professor of Data Science in the College of Information and an affiliated Assistant Professor of Computer Science at the University of North Texas.

Research focus: AI-driven methods and systems responsive to information needs

- Computational linguistics
- Human-centered artificial intelligence
- LLM applications
- Crisis informatics



Presenter Profile

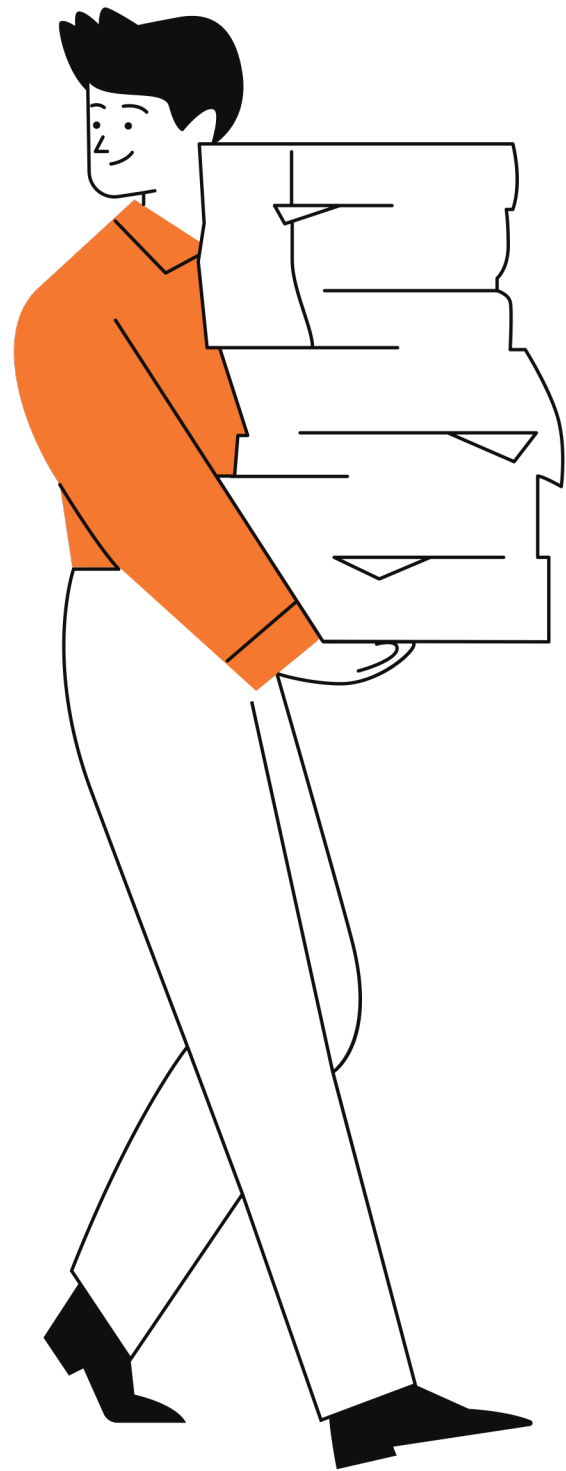


Xiaoying Song is a Research Assistant in Data Science at the College of Information at the University of North Texas.

Research focus:

- Human-centered artificial intelligence
- Crisis computing
- Natural language processing

Outline



1

Introduction

2

Technical Foundations

3

Library Service Applications

4

Case Demonstrations

5

Closing and Q&A

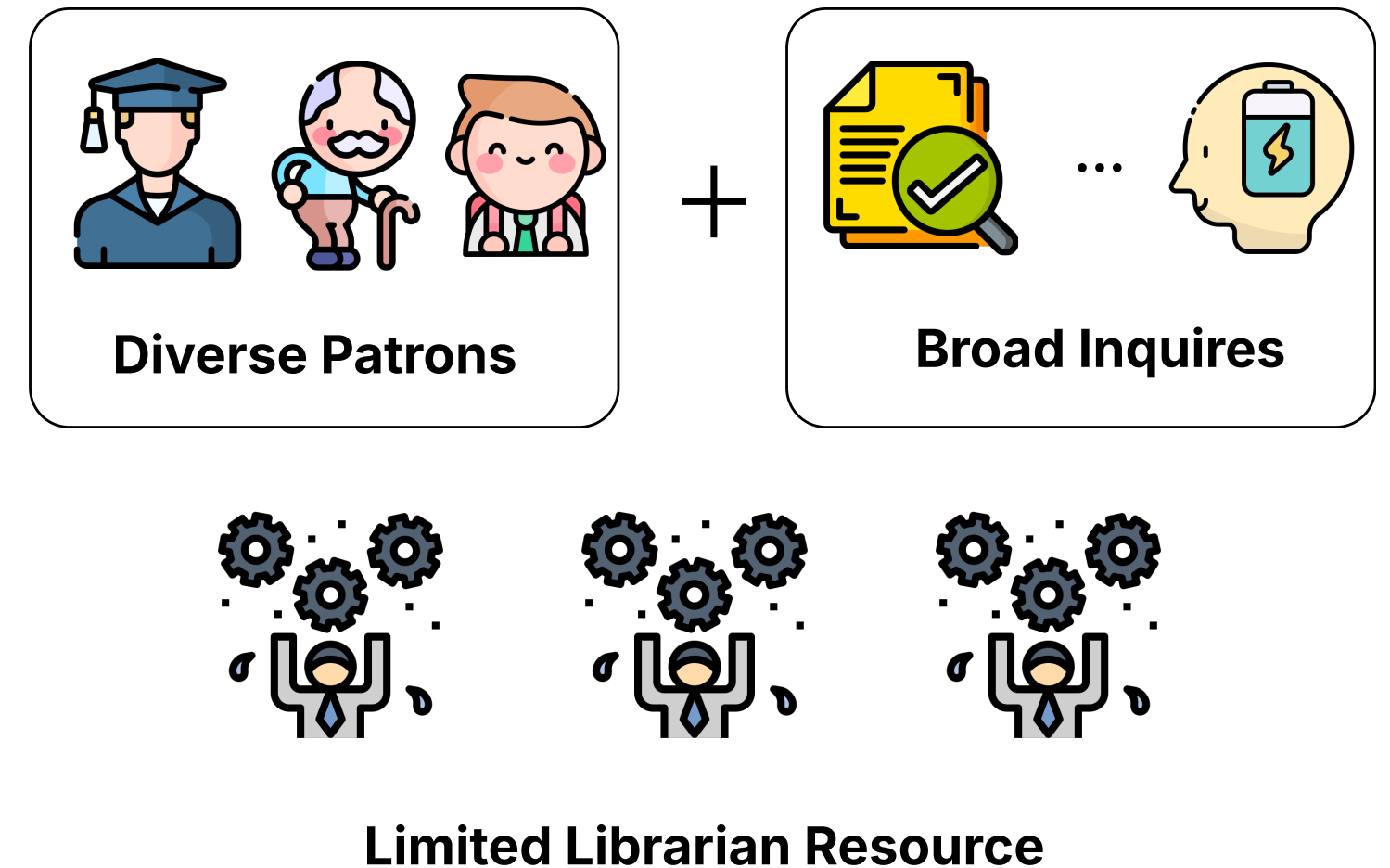
Library Service and Challenges

Library Service

- a. Reference and information services
- b. User education
- c. Community engagement

Communication challenges

- a. Diverse patron needs
- b. Broad range of inquiries
- c. Resource constraints



Patron-Centered Communications

Tailoring responses for different patrons matters



Improve information comprehension and accessibility (Ayre et al., 2024)

- A high school student asks how to cite sources for a history project. Instead of only pointing them to a long academic guide, the librarian provides a simple step-by-step infographic and an online citation generator.



Increase engagement and trust (Çakmak & Uğurluoğlu, 2024).)

- A new immigrant who speaks limited English needs help finding job resources. The librarian listens carefully, uses clear language, and shares bilingual materials. Because of this personalized support, the patron feels welcomed and returns regularly for other services.



Address equity and inclusion (Truong et al., 2014)

- A patron with visual impairment wants to access e-books. The librarian ensures screen-reader-friendly formats are available and teaches the patron how to use the library's accessibility features. This shows the library values inclusivity and makes resources equitable for all.

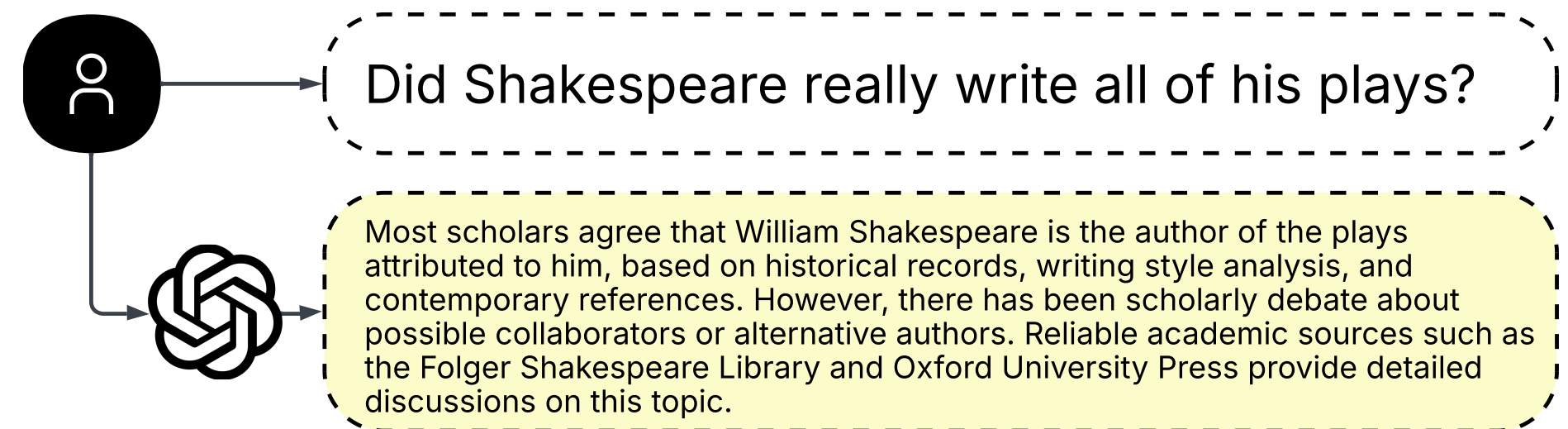
Large Language Models (LLMs)

LLMs refer to transformer-based neural language models that contain tens to hundreds of billions of parameters, which are pretrained on massive text data (Minaee et al., 2024).

Examples: ChatGPT, LLaMA, Qwen. etc

LLM Capabilities:

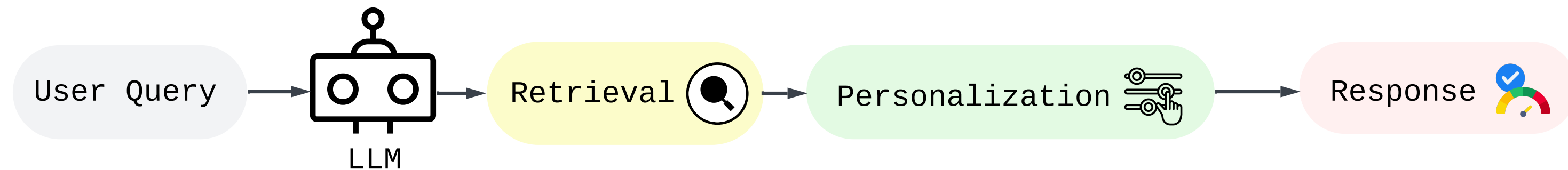
- Natural language understanding
- Natural language generation
- Multilingual support ...



Interaction with AI chatbot (response to user query)

How LLMs Support Library Communications

1. Information retrieval from trusted sources
2. Personalization and tone adjustment
3. Drafting and recommending user responses



LLMs respond without/ with retrieved documents

LLM Without Retrieved Documents

Example Question: “What are the latest CDC guidelines on COVID-19 vaccines?”

LLM Answer:

“Vaccines are recommended to reduce severe illness. Booster shots may be needed, especially for high-risk groups.”

Limitations:

- May be outdated (knowledge cutoff).
- Can produce vague or incorrect answers.
- Lacks citations.

LLM With Retrieval (RAG)

Example Question: “What are the latest CDC guidelines on COVID-19 vaccines?”

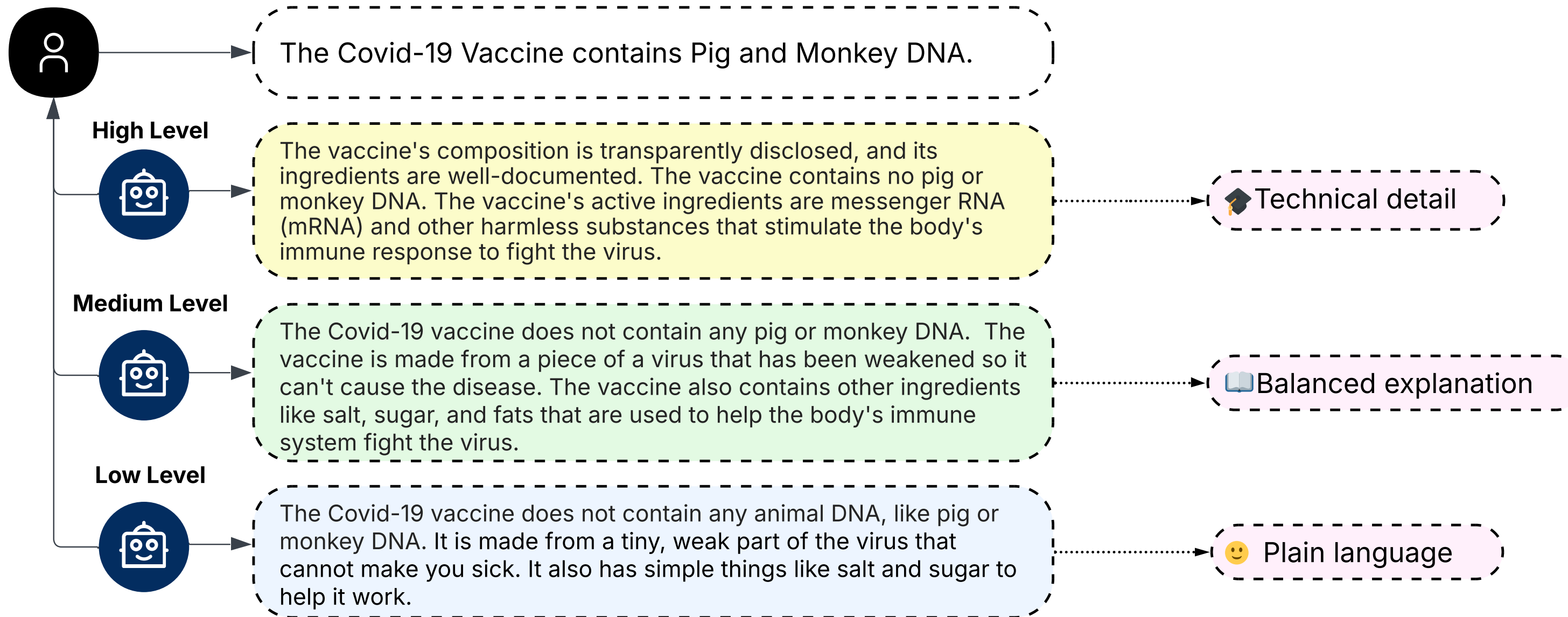
LLM Answer (with sources):

“As of August 2025, CDC recommends updated COVID-19 vaccines for everyone aged 6 months and older. People 65+ should receive an additional booster.” [CDC.gov, Aug 2025]

Strengths:

- Provides up-to-date info.
- Can cite sources.
- Builds user trust.

LLMs tailoring responses for different Literacy Levels





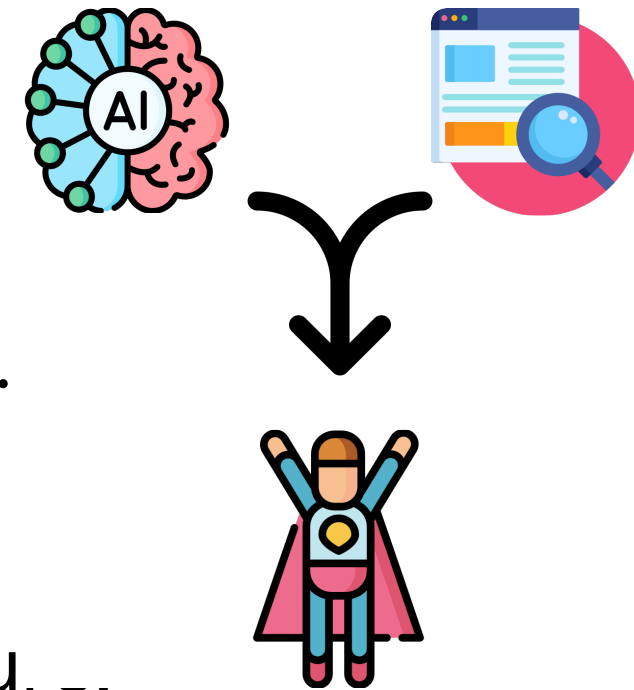
Technical Foundations

Retrieval-Augmented Generation (RAG)

RAG is the process of optimizing the output of a large language model. It retrieves an external knowledge base outside of its training data sources before generating a response.

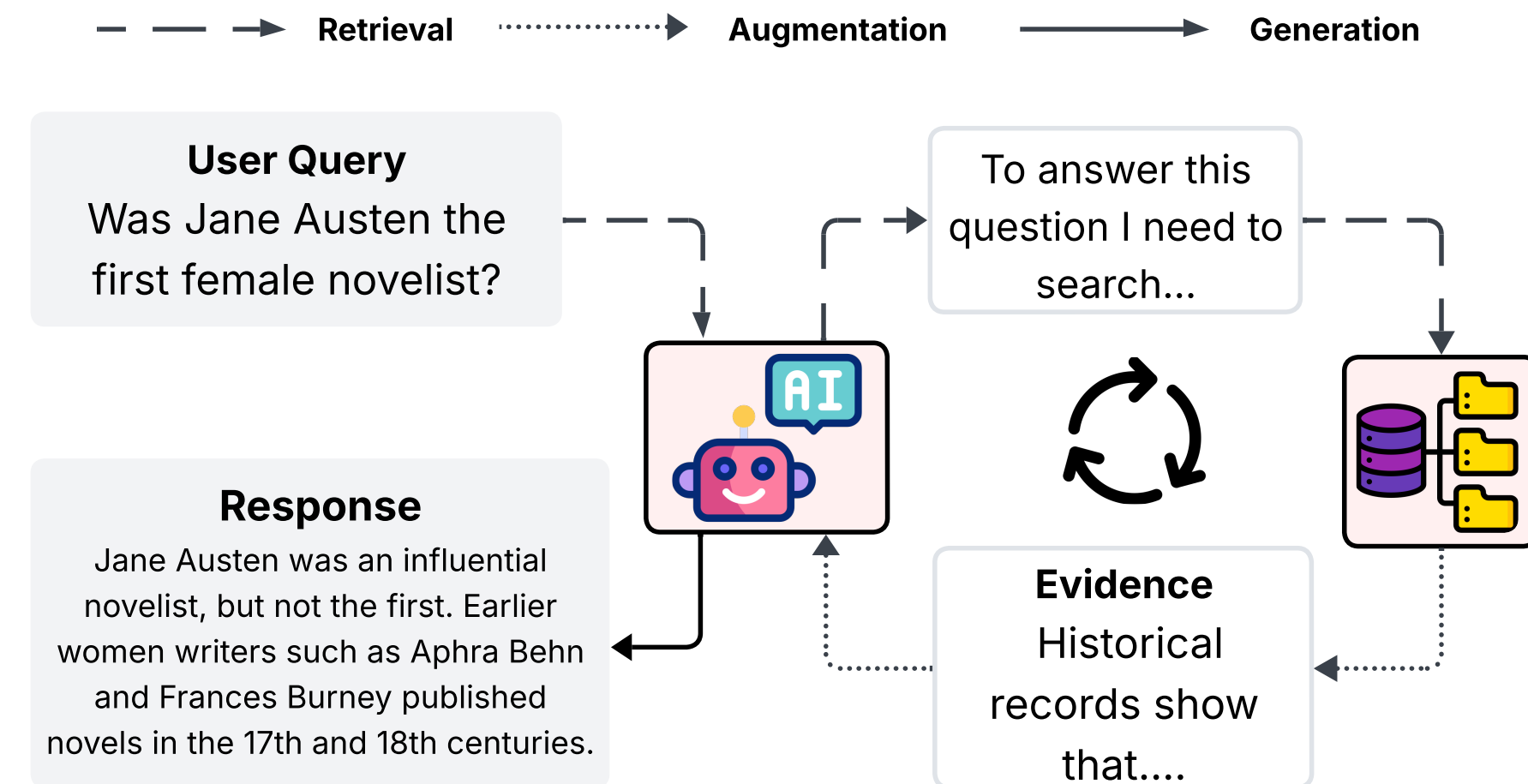
Why RAG Exists

- LLMs are static:
 - Once trained, they cannot automatically learn new facts without retraining.
- LLMs may “hallucinate”:
 - They can produce plausible-sounding but incorrect information when unsu. . .
- Knowledge can be incomplete or outdated:
 - Real-world information changes, and training data may not reflect the latest updates.



How does RAG work?

- ❖ **Retrieval:** Searching for relevant information from an external knowledge source (e.g., databases, websites, PDFs, library archives) that the LLM did not see during training.
- ❖ **Augmentation:** Supplying the retrieved documents as context to the LLM before it answers.
- ❖ **Generation:** The LLM then uses both its own knowledge and the retrieved evidence to produce a response that is more factually grounded and up-to-date.



Key Points

Retrieval

- ❑ LLMs can forget or miss specific facts. Retrieval ensures the model has fresh, domain-specific, or verified information before answering.

Augmentation

- ❑ The LLM doesn't work alone. It's boosted by adding the retrieved documents to its prompt so it can reference them when generating an answer.

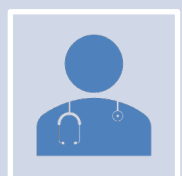
Generation

- ❑ The LLM still writes the answer in natural language, but now it is guided by real evidence, not just memory.

RAG System Design–Knowledge Base Construction

- ❖ This is about building a trusted memory for the system.
- ❖ It represents **what the model can “look up”** when it needs information outside its training data.
- ❖ The knowledge base is usually static but updateable, and it reflects a specific domain (e.g., library policies, historical archives, research literature).
- ❖ Think of it as **creating a specialized bookshelf** that only contains high-quality, relevant resources.

Knowledge Base Construction



Collect relevant documents from reliable source



Clean and preprocess the content

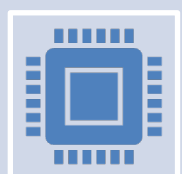
Remove irrelevant metadata, ads, or formatting artifacts.

Convert documents into plain text or structured formats



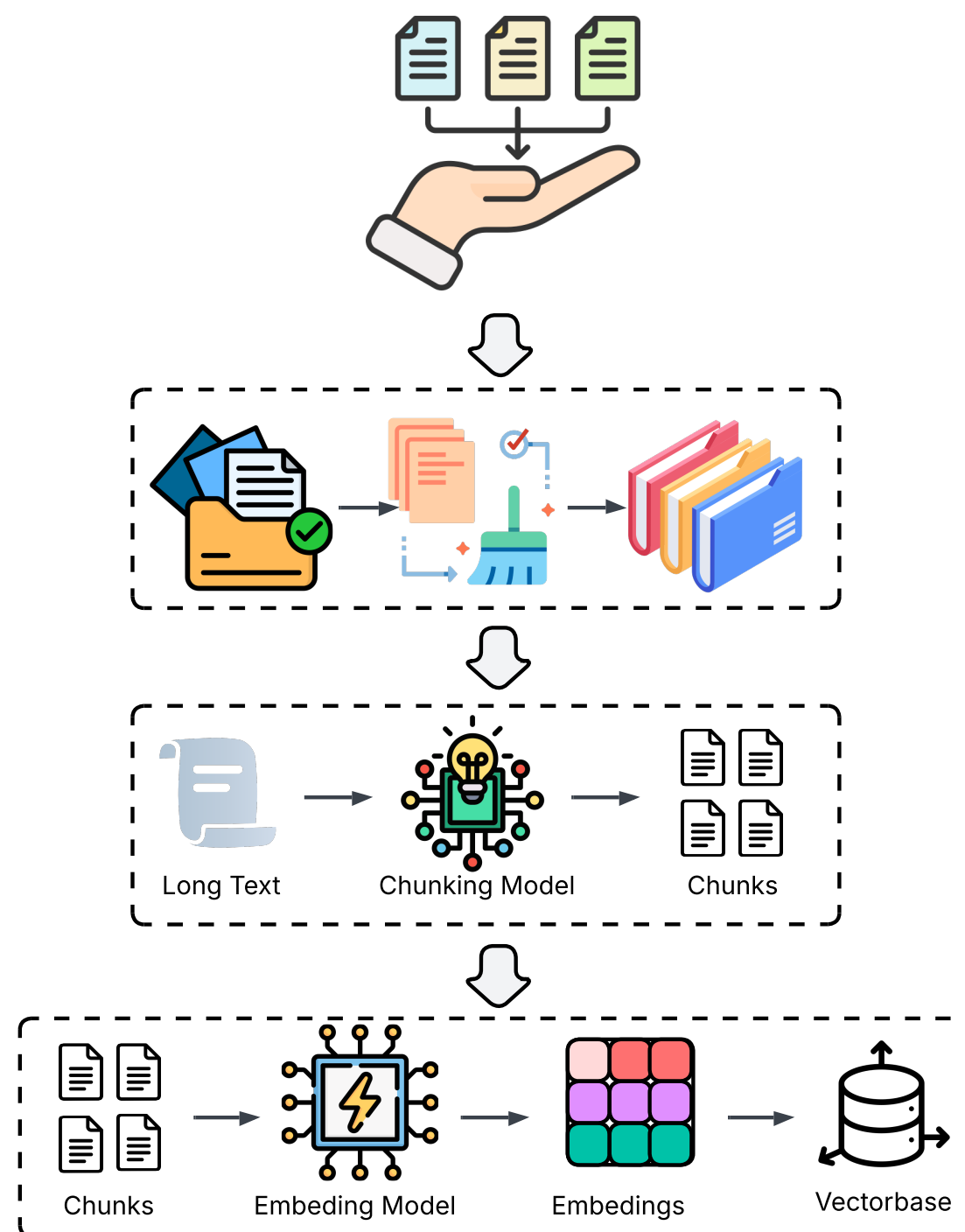
Chunk and index the text

Divide content into manageable “chunks”.
Assign each chunk a unique identifier.
Index using semantic embedding models to support fast similarity search.



Embed the text with the retriever model

Use a dense embedding model to convert chunks into vector representations.
Store these vectors in a vector database.

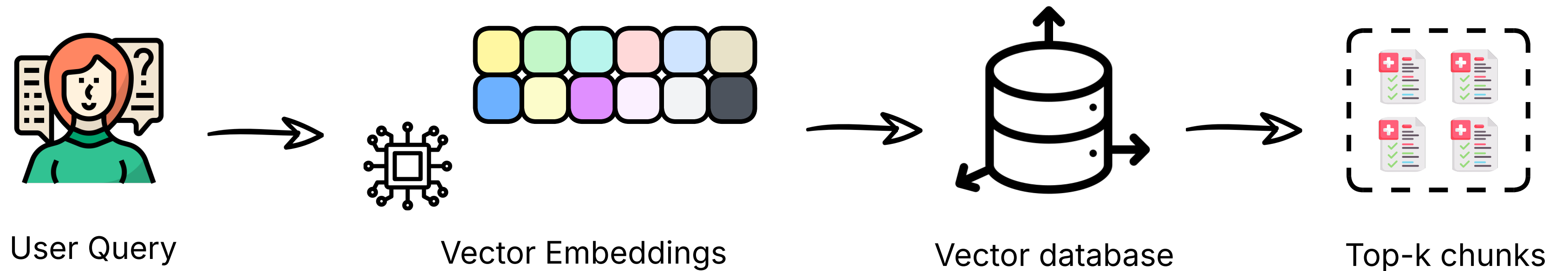


RAG System Design–Evidence Retrieval

- ❖ This is the search phase, the model is not yet answering, **it's finding the most relevant information.**
- ❖ It bridges the gap between a vague human question and precise stored knowledge by **using semantic similarity or other criteria.**
- ❖ Retrieval focuses on **precision and relevance**, ensuring the model starts from accurate and contextually appropriate evidence.
- ❖ This step is like a librarian **locating the right books and opening them to the most useful pages.**

Evidence Retrieval

- User query as the input
- Converts the query into a **dense vector representation**
- Search in the vector database
- Retrieves the best-matching chunks

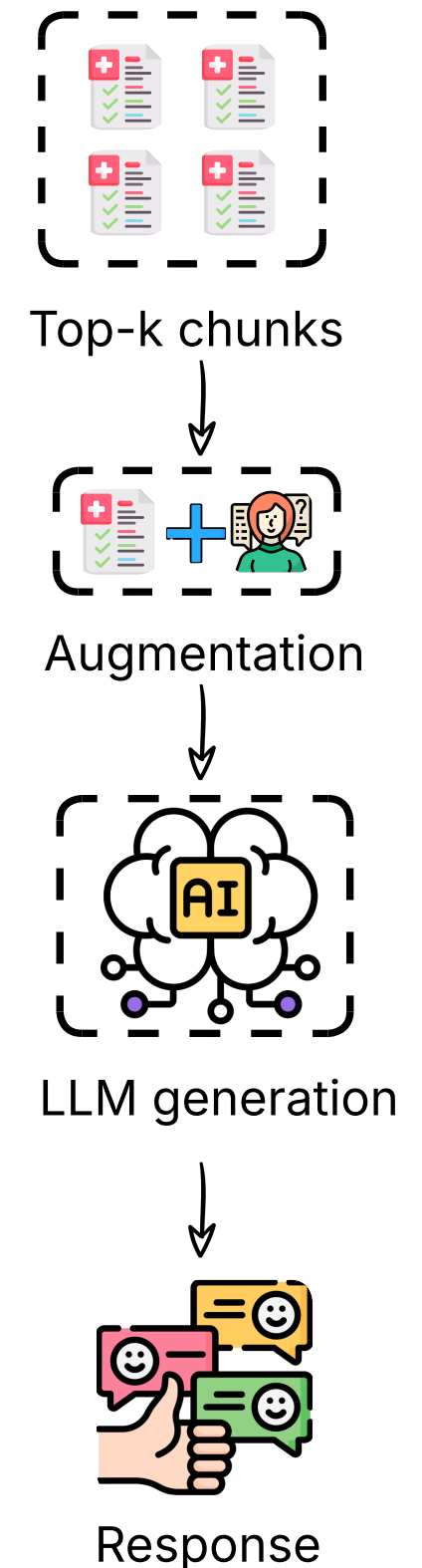


RAG System Design–Response Generation

- ❖ This is the composition phase, the model now **interprets and integrates the retrieved evidence into a fluent, human-readable answer.**
- ❖ It's where reasoning, summarizing, and tailoring the message to the audience happen.
- ❖ While retrieval is factual, generation is narrative and adaptive, shaping the information for clarity, completeness, and engagement.
- ❖ It's like a librarian **explaining the content of the books in a way the patron can understand and use.**

Response Generation

- **Augmentation** (query + retrieved chunks)
 - The system augments the user's query by combining it with the retrieved content.
 - This augmented input becomes the prompt fed into the language model.
 - The goal is to “ground” the generation in factual, retrieved evidence to avoid hall
- **Generation**
 - A LLM takes the augmented input and generates a coherent, fluent, and grounded response.
 - The model doesn't guess from scratch. It's guided by the real evidence retrieved earlier.



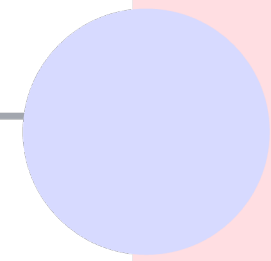
Brief Example

- **Patron asks:** "When was the first photograph of the Moon taken?"
- **Knowledge Base Construction:** astronomy books, NASA reports, and historical records.
- **Evidence Retrieval:** AI searches the knowledge base and finds an 1840 daguerreotype by John William Draper.
- **Response Generation:** AI answers: "The first known photograph of the Moon was taken in 1840 by John William Draper using a daguerreotype process."

How to tailor response for diverse users

- **Knowledge base selection**
 - Choose the right source for the right audience.
 - Ensures foundational content matches the user's comprehension capacity before generation.
- **Evidence filtering**
 - Filter retrieved documents to match the user's profile and context.
 - Only relevant and understandable chunks are passed into the LLM.
- **Prompt adaption**
 - Customize the tone, structure, and style of the generation process.





Library Service Applications

LLM Applications



1. Information Retrieval & Question Answering

- Use case: Answering factual or domain-specific queries from large document collections.
- Example: Librarian AI retrieves academic papers on climate change impact on agriculture and provides a concise summary.

2. Summarization

- Use case: Condensing large volumes of text into easy-to-read summaries.
- Example: Summarizing 200 pages of municipal regulations into a 2-page briefing for policy makers.

3. Content Generation

- Use case: Producing original text for specific purposes.
- Example: Drafting library blog posts, exhibition descriptions, or instructional guides.

5. Translation & Multilingual Support

- Use case: Breaking language barriers in accessing resources.
- Example: Translating historical documents from French to English for a research project

Benefits of Integrating LLMs in Library Services

Enhanced Information Services

- ❖ Provides evidence-grounded answers to user queries.
- ❖ Supports health, legal, and academic reference services.

Personalized Support

- ❖ Tailor responses to different users using adaptive prompting.
- ❖ Helps bridge the digital divide.

Scalable Knowledge Access

- ❖ Uses large language models with live document repositories.

Challenges of Integrating LLMs in Library Workflows

Privacy & Compliance

- ❖ Handling sensitive queries (e.g., health) raises ethical and legal concerns.
- ❖ Must comply with data protection laws (e.g., HIPAA, GDPR).

Maintenance Overhead

- ❖ Needs continuous update of document repositories and models.
- ❖ May require librarian retraining in digital tool usage.

Evaluation & Monitoring

- ❖ Must establish criteria for response readability, factual accuracy, and user satisfaction.
- ❖ Challenging to benchmark quality across domains and audiences.



Case Demonstration

- System Design
- Implementation

Background & Questions

Background

- ❖ Health misinformation is common, especially during the Covid-19 pandemic.
- ❖ Different people have different levels of health literacy.
- ❖ A “one-size-fits-all” response may fail to inform or build trust.

Question

How can LLMs provide evidence-based responses that are tailored to different health literacy levels?

Data Preparation

- Collect data from Reddit
 - Utilize PRAW API to collect health misinformation from Reddit
 - Focus on topics related to the coronavirus disease (COVID-19), influenza (Flu), and human immunodeficiency virus (HIV), as these topics have been the center of conspiracy theories and misinformation
- Annotate health misinformation
 - Employ annotators to label health misinformation



COVID shots are 'essentially murder'.

Health Misinformation

Knowledge Base Construction

- Search for relevant documents from the CDC, WHO, and so on.
- Employ LLMs and humans to label the health literacy level of documents

Literacy Level	Definition	Documents & Source	Content Features
Low Health Literacy	Basic ability to read, write, and understand simple health messages.	COVID-19-Vaccines-A-Plain-Language-Guide (HealthMatters Program)	- Short sentences - No jargon - Clear, direct instructions
Medium Health Literacy	Ability to process, communicate, and apply health information in new or evolving situations.	covid19-teen-info-sheet (CDC)	- Some technical terms (explained) - Balanced tone
High Health Literacy	Can critically analyze health risks and interpret clinical/scientific data.	CORD-19 (COVID-19 Open Research Dataset , Allen Institute for AI)	- Data tables, statistics, citations - Technical precision

Load & Split Documents

```
from llama_index.core import SimpleDirectoryReader
from llama_index.core.node_parser import TokenTextSplitter
```

Load and split documents

```
documents =
SimpleDirectoryReader("./knowledge_base/low_level/").load_data()
splitter = TokenTextSplitter(chunk_size=1024, chunk_overlap=20)
nodes = splitter.get_nodes_from_documents(documents)
```

❖ **chunk_size=1024**
Meaning: Each document will be split into chunks of 1024 tokens.

❖ **chunk_overlap=20**
Meaning: Each chunk will overlap by 20 tokens with the next chunk.

Evidence Retrieval

a. Input: User's health misinformation statement

- Example: "COVID-19 vaccine causes infertility."

b. Tokenization:

- The sentence transformer splits the input into tokens the model understands.
- Example: ["COVID", "-", "19", "vaccine", "causes", "infertility", "."]

c. Embedding Generation:

- Model encodes the input into a vector representing meaning.
- Example vector start: [0.123, -0.456, 0.789, ...]

d. Document Embeddings:

- Each document chunk in the knowledge base already has a stored embedding (computed when we built the index).

e. Similarity Computation:

- Compare the query embedding with all document embeddings using cosine similarity.
- Higher score → more semantically relevant document.

Semantic Retrieval Setup

```
from llama_index.core import VectorStoreIndex
from llama_index.core.retrievers import VectorIndexRetriever
from llama_index.core import StorageContext

# Build and store document nodes

storage = StorageContext.from_defaults()
storage.docstore.add_documents(nodes)

# Build vector index for semantic search

vector_index = VectorStoreIndex(nodes, storage_context=storage)

# Create retriever

vector_retriever = VectorIndexRetriever(index=vector_index)
```

❖ **StorageContext**

Holds the document store and manages in-memory storage

❖ **VectorStoreIndex**

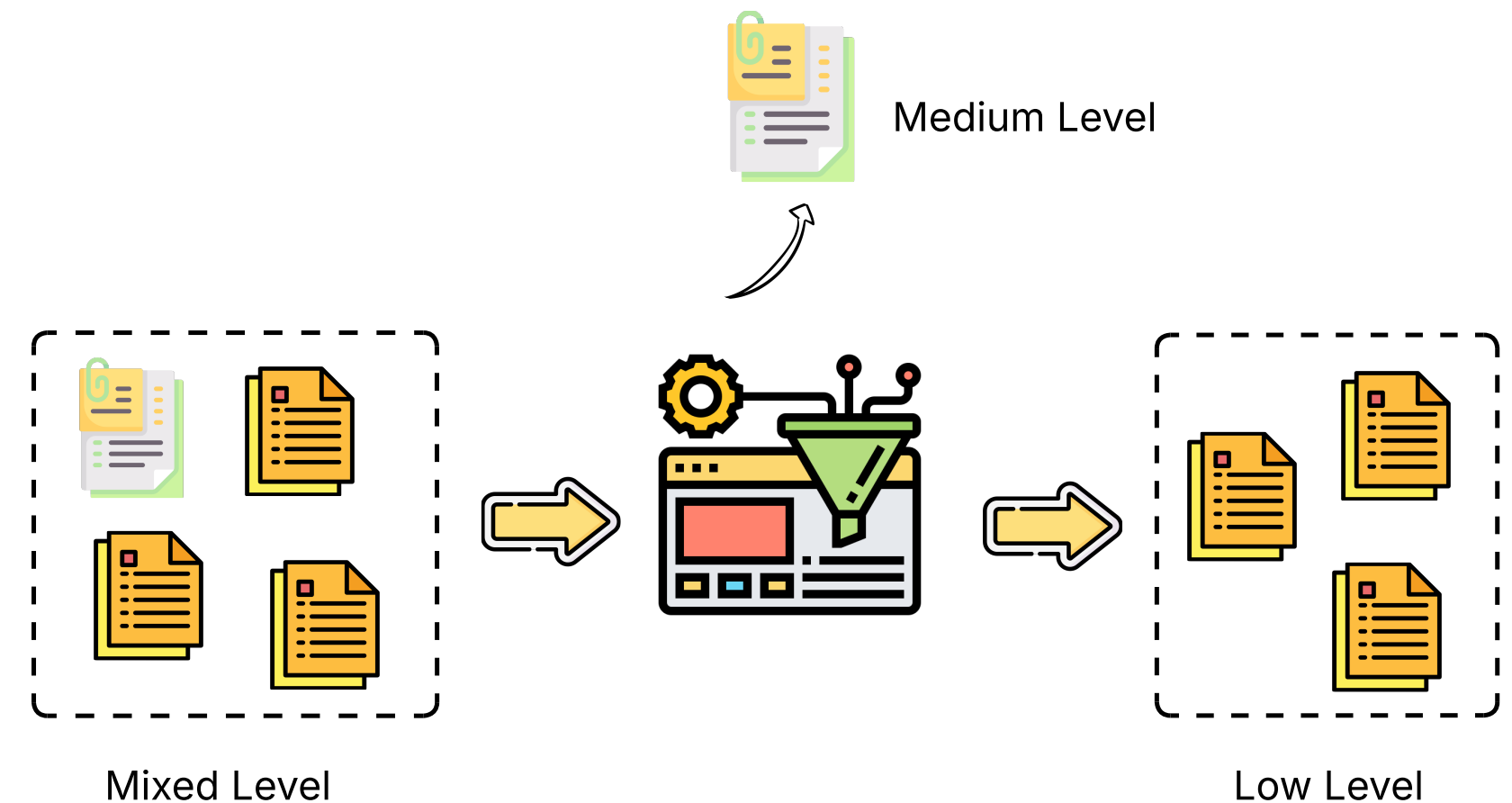
Builds a searchable vector index for semantic matching

❖ **VectorIndexRetriever**

Retrieves top-k relevant chunks based on query similarity

Evidence Filtering

- Flesch–Kincaid Reading Ease (FKRE) score as a practical, scalable proxy for estimating the functional dimension of health literacy. FKRE quantifies text difficulty based on syntactic features, which directly influence comprehension.
- Collapse the FKRE score into three categories (Roein et al., 2023), **easy (80–100)**, **medium (60–79)**, and **hard (0–59)**






Prompt Drafting

Design prompts that instruct the LLM to generate fact-based, respectful counterspeech tailored to the user's health literacy level, tone preferences, and communication context.

- **Include role: frame the LLM as a supportive expert**
 - "You are a public health librarian helping a patron understand the facts."
- **Explicitly state the task**
 - Use clear task descriptors like: "Provide a factual, respectful correction to the following health claim..."
- **Specify target literacy level**
 - Low Literacy: Use short, simple sentences and no medical jargon. Explain like you're talking to someone with a 6th-grade reading level...
- **Set the tone and style**
 - Add descriptors like: "Use an empathetic tone."

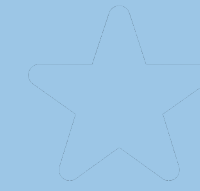
Prompt Example

 You are an expert in health communication and plain language, specializing in engaging audiences with low health literacy who struggles to comprehend basic health information , such as medication labels, appointment slips, or preventive care guidelines. They may have difficulty understanding medical jargon or following treatment plans. Your task is to generate a counterspeech response to a piece of health misinformation .

Your response should meet the following criteria:

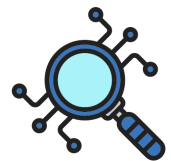
- ✓ Simple and clear language: Use straightforward, everyday language that avoids medical jargon and complex terms, ensuring it is easy for anyone to understand.
- ✓ Evidence-Based correction: Correct the misinformation using clear, research-backed health information drawn from the provided documents.
- ✓ Clarity and accessibility: Organize your response in a simple, well-structured manner with clear examples or analogies that make the information relatable and easy to grasp.
- ✓ Respectful tone: Maintain a respectful and professional tone that encourages critical evaluation and further inquiry.

Response Generation



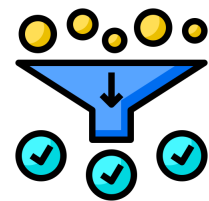
User Query
(Misinformation)

“COVID-19 kills people
in seconds.”



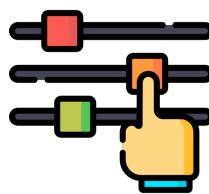
Retriever fetches
evidence

Excerpt from The
Lancet, CDC guidance



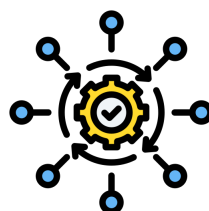
Evidence Filter

Filter evidence for target
user



Prompt adapts by
literacy level + tone

Prompt crafted for
low/med/high literacy

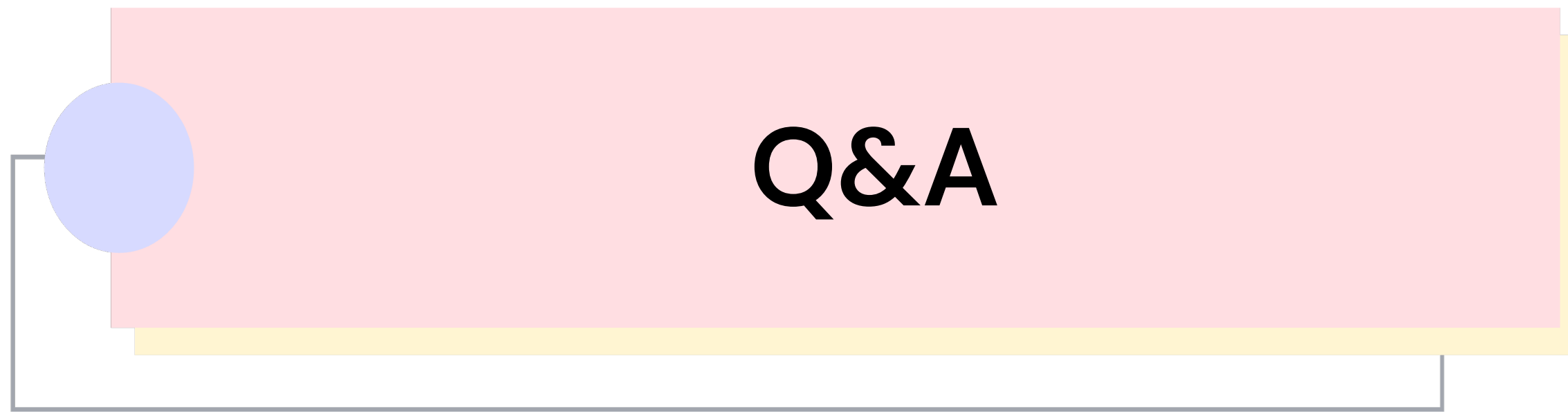


LLM generates
counterspeech

Response tailored to
user's level and style

Response Evaluation

Dimension	What It Measures	Why It Matters	Measurement
Readability	Is the response easy to understand for the target user?	Ensures users of different literacy levels can comprehend it	<ul style="list-style-type: none">• Flesch–Kincaid Grade Level• FKRE Score (Flesch Reading Ease)• SMOG Index
Factual Accuracy	Is the content factually correct and based on reliable sources?	Avoids spreading false or misleading information	Expert review or citation check
Relevance	Does it directly address the misinformation or user's query?	Keeps the response focused and useful	<ul style="list-style-type: none">• Human rating or semantic similarity to query• LLM-based relevance score
User Preference	Is the tone appropriate and aligned with how the user prefers to receive information?	Builds trust and increases engagement	<ul style="list-style-type: none">• Feedback surveys• A/B testing different response tones (e.g., empathetic vs. neutral)• LLM-based simulations



References:

- Ayre, J., Bonner, C., Muscat, D. M., Cvejic, E., Mac, O., Mouwad, D., ... & McCaffery, K. J. (2024). Online plain language tool and health information quality: a randomized clinical trial. *JAMA Network Open*, 7(10), e2437955–e2437955.
- Çakmak, C., & Uğurluoğlu, Ö. (2024). The effects of patient-centered communication on patient engagement, health-related quality of life, service quality perception and patient satisfaction in patients with cancer: a cross-sectional study in Türkiye. *Cancer Control*, 31, 10732748241236327.
- Truong, M., Paradies, Y., & Priest, N. (2014). Interventions to improve cultural competency in healthcare: a systematic review of reviews. *BMC health services research*, 14(1), 99.
- Minaee, S., Mikolov, T., Nikzad, N., Chenaghlu, M., Socher, R., Amatriain, X., & Gao, J. (2024). Large language models: A survey. *arXiv preprint arXiv:2402.06196*.

Thank you!

