

GENAI RAG PROJECT: TOPIC STUDY ASSISTANT [POC]

GENE EXPRESSION ACTIVATION VIA MICRO RNA

USER-FOCUS AND TOPIC CHOICE

▶ Developer's goal:

- ▶ Learn and practice implementing cutting-edge technologies.
- ▶ Create a **pet-project** for portfolio and upscale.
- ▶ Practice user-oriented project design.

▶ User's goal:

- ▶ Get assistance for writing **Master's thesis** in Biology on the topic:

“GENE EXPRESSION ACTIVATION VIA MICRORNA”.

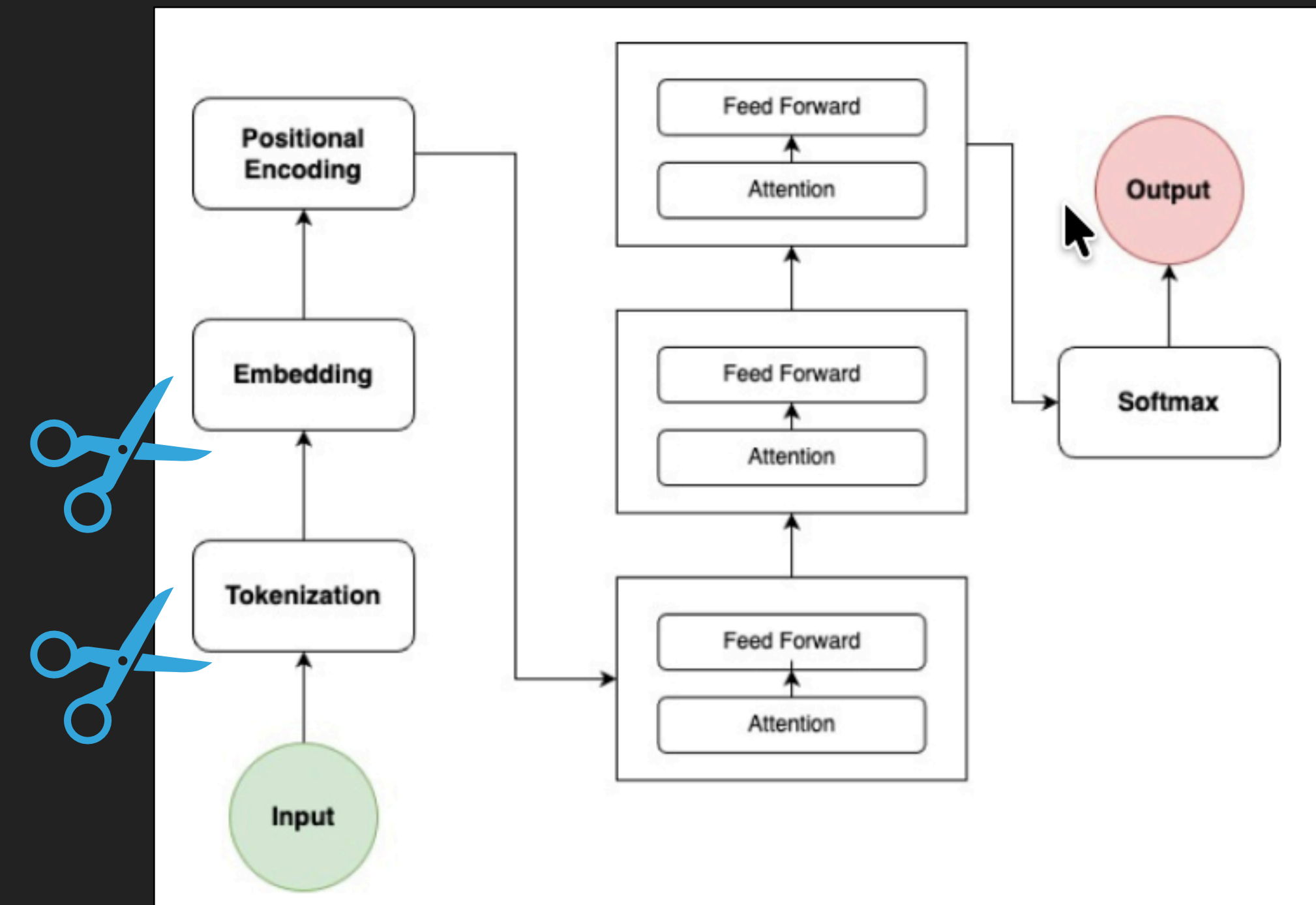
- ▶ Deepen knowledge in the research topic.
- ▶ Start collaborating with data scientists for further steps in Bioinformatics.



TERM DISAMBIGUATION

- ▶ **Generative AI** - AI that generates new data by learning patterns from existing data.
- ▶ **GPT** - Generative Pre-trained Transformer
- ▶ **LLM** - Large Language Model - generates output.
- ▶ **RAG** - Retrieval Augmented Generation
- ▶ **Embedding model** - converts data chunks into a numeric vector.

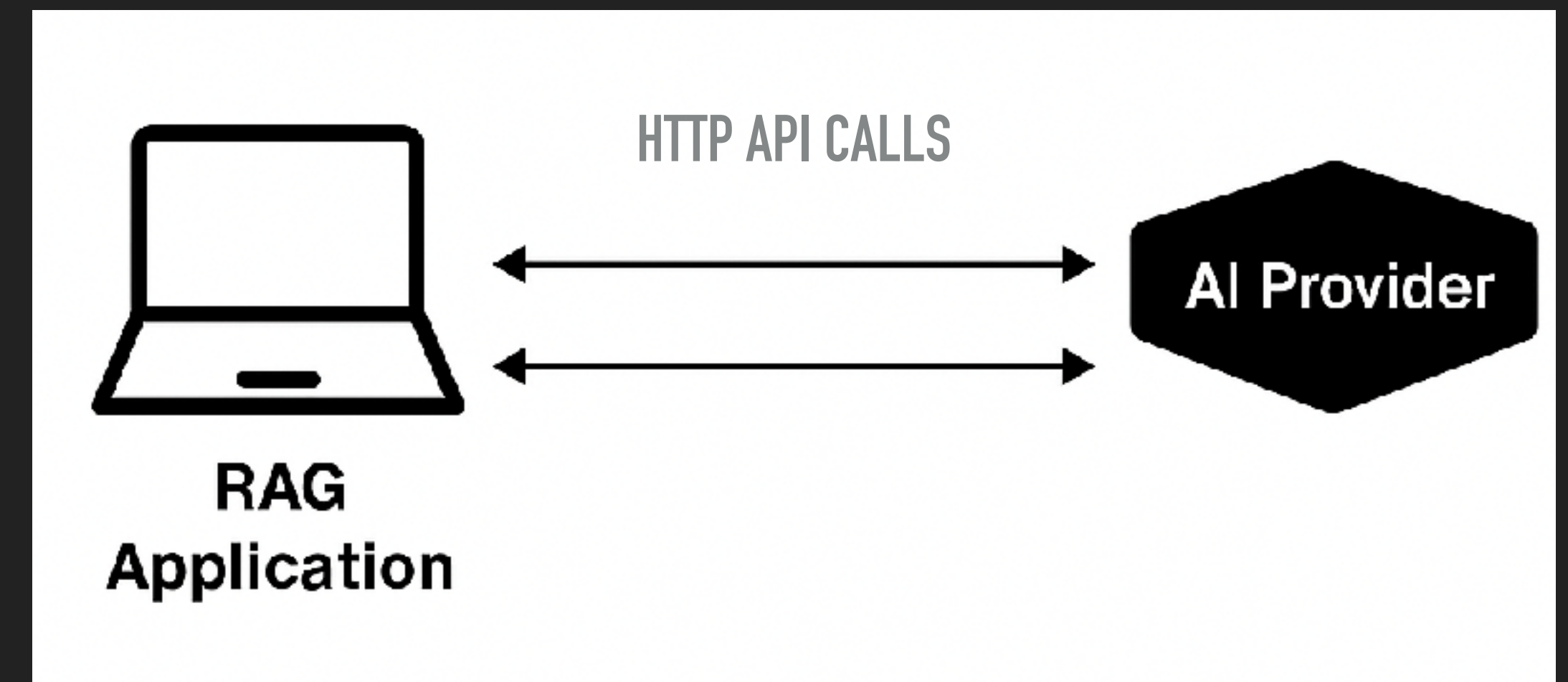
TRANSFORMER / GPT WORKFLOW



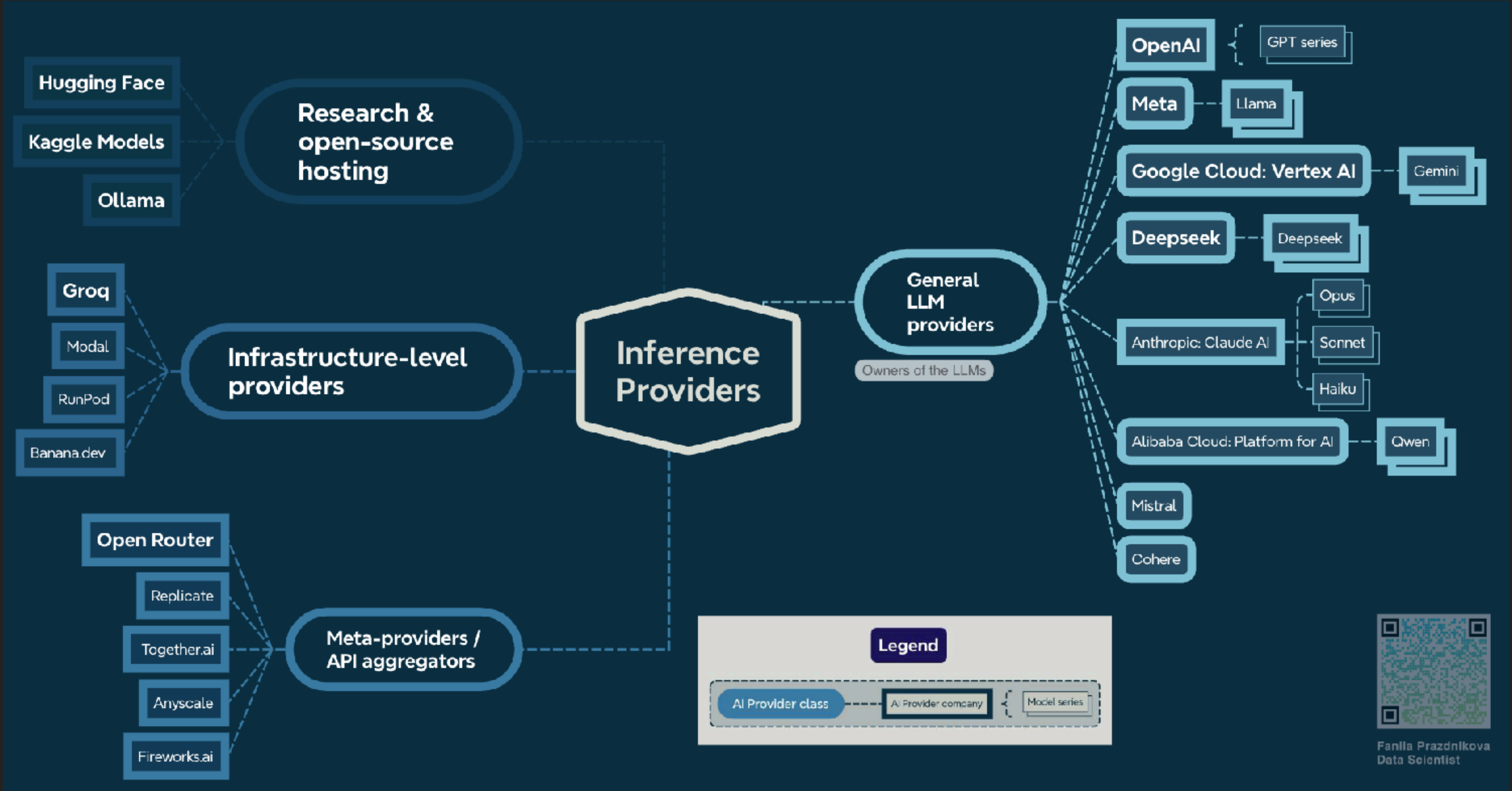
WHAT ARE INFERENCE PROVIDERS

- ▶ Platforms that host, run, and expose AI models via APIs.
- ▶ Handle the **inference** - the actual computation that generates answers, embeddings, images, etc.
- ▶ Provide:
 - ▶ model access (LLMs, embedding models, etc.)
 - ▶ APIs / SDKs for querying these models
 - ▶ Compute infrastructure (GPU clusters for inference)
 - ▶ Model hosting for custom deployments
 - ▶ Monitoring, billing, and authentication

SYSTEM ARCHITECTURE



CHOOSING INFERENCE PROVIDER



RAG PIPELINE

1 Data collection

- ▶ Data types: PDF documents with text, images, formulas
- ▶ POC: data is collected manually by the user
- ▶ MVP: web-scraping or GPT-5 for finding relevant scientific articles

2 Data retrieval and tokenization

- ▶ Extract data from PDF files
- ▶ Preserve text, images, scientific formulas
- ▶ Tokenization: split into chunks

3 Embedding and indexing

- ▶ Embedding model: convert data to vectors
- ▶ Vector DB: store the vectors in DB and build semantic index


4 Data retrieval for answer

- ▶ Using GUI for getting query from a user
- ▶ Query tokenize/encode/embedding using the same model and methods
- ▶ MVP: top-k similarity search, reranking

5 Generating answer

- ▶ Prompt LLM with the retrieved context
- ▶ Return the answer to the user in GUI

CHOOSING AI FRAMEWORK

LangChain	LlamaIndex
tool- and workflow-centric	data-centric
for various LLM projects	specific for RAG projects
more divers functional modules for architecture flexibility and scalability	more efficient in document indexing, data retrieval, semantic search
open-source and free	commercial , price depends on usage
	migration to it can be considered after MVP and efficiency evaluation + donations / monetization prospects

RAG TOOLCHAIN FOR POC PHASE

User Interaction: Type: Web site

- Back-End: none for POC
- Front-End: StreamLit with caching on localhost

AI framework: LangChain

PDF extractor: PyPDFLoader

Vector DB: Chroma DB on localhost

Chunker: RecursiveCharacterTextSplitter

Retriever: Chroma.as_retriever

Embedding model: text-embedding-3-small

LLM: gpt-5-mini (primary), gpt-5-nano (for drafts/tests)

Inference Providers:

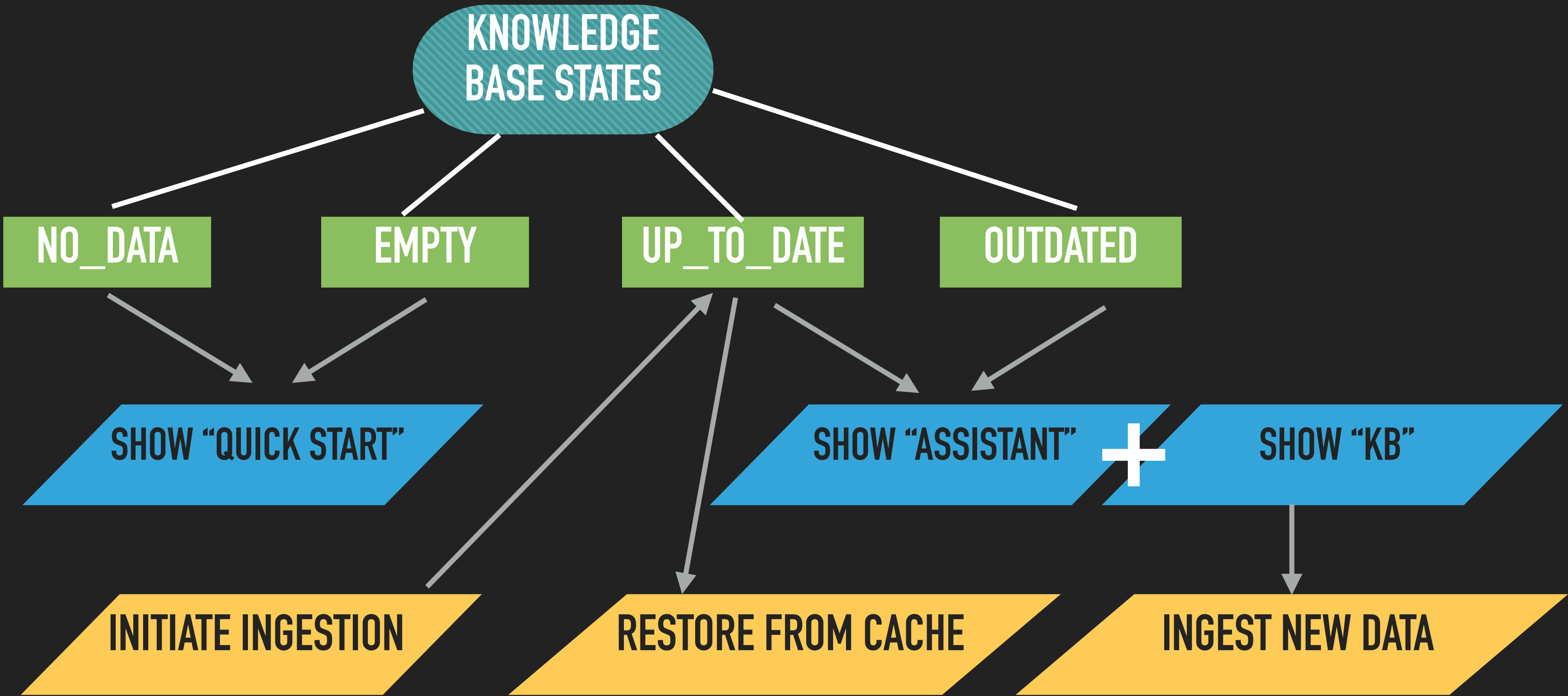
- Embedding model: OpenAI
- LLM: OpenAI

STATE MACHINE


FS Layer

GUI Layer

Logic Layer




KNOWLEDGE-BASE UPDATING


 **Biology topic study assistant**

Gene expression activation via microRNA

Knowledge base status

8% (2/25 documents processed)

 Assistant


 Knowledge Base

Knowledge Base control

Dataset folder: `/Users/`

Total PDFs: 25

Processed: 2

>  New files detected

Update knowledge base

New files detected

- Ago1 Interacts with RNA Polymerase II and Binds to the Promoters of Actively Transcribed Genes in Human Cancer Cells.pdf
- Amplifying gene expression with RNA-targeted therapeutics 2023.pdf
- An Algorithm for Generating Small RNAs Capable of Epigenetically Modulating Transcriptional Gene Silencing and Activation in Human Cells.pdf
- Antisense transcripts are targets for activating small RNAs.pdf
- BOOK Target-Recognition Mechanism and Specificity of RNA Activation.pdf
- Enhancer RNAs participate in androgen receptor-driven looping that selectively enhances gene activation.pdf
- Histone lysine methyltransferase-related neurodevelopmental disorders -current knowledge and saRNA future therapies.pdf
- Nuclear AGO1 Regulates Gene Expression by Affecting Chromatin Architecture in Human Cells 2019.pdf
- Nucleosome repositioning underlies dynamic gene expression.pdf
- Prognostic value and function of KLF4 in prostate cancer -RNAa and vector-mediated overexpression identify KLF4 as an inhibitor of tumor cell growth and migration.pdf
- RNA Activation-A Novel Approach to Therapeutically Upregulate Gene Transcription.pdf
- RNA activation.pdf
- RNA-based therapeutics -an overview and prospectus.pdf
- RNA-mediated gene activation.pdf
- RNAi in cell nuclei -potential for a new layer of biological regulation and a new strategy for therapeutic discovery (Автосохранение)-Копировать.pdf
- RNAi in cell nuclei -potential for a new layer of biological regulation and a new strategy for therapeutic discovery (Автосохранение).pdf
- Reexamining assumptions about miRNA-guided gene silencing.pdf
- Shape of promoter antisense RNAs regulates ligand-induced transcription activation 2021.pdf
- Small activating RNA binds to the genomic target site in a seed-region-dependant manner.pdf
- Small-Activating RNA Can Change Nucleosome Positioning in Human Fibroblasts.pdf
- Small-molecule inducible transcriptional control in mammalian cells.pdf
- Transcriptional Regulation by Promoter Targeted RNAs suzuki2009.pdf
- saRNA-guided Ago2 targets the RITA complex to promoters to stimulate transcription 2016 прочитано.pdf

KNOWLEDGE-BASE UPDATING




Biology topic study assistant

Gene expression activation via microRNA

Knowledge base status

100% (24/24 documents processed)

 Assistant

 Knowledge Base

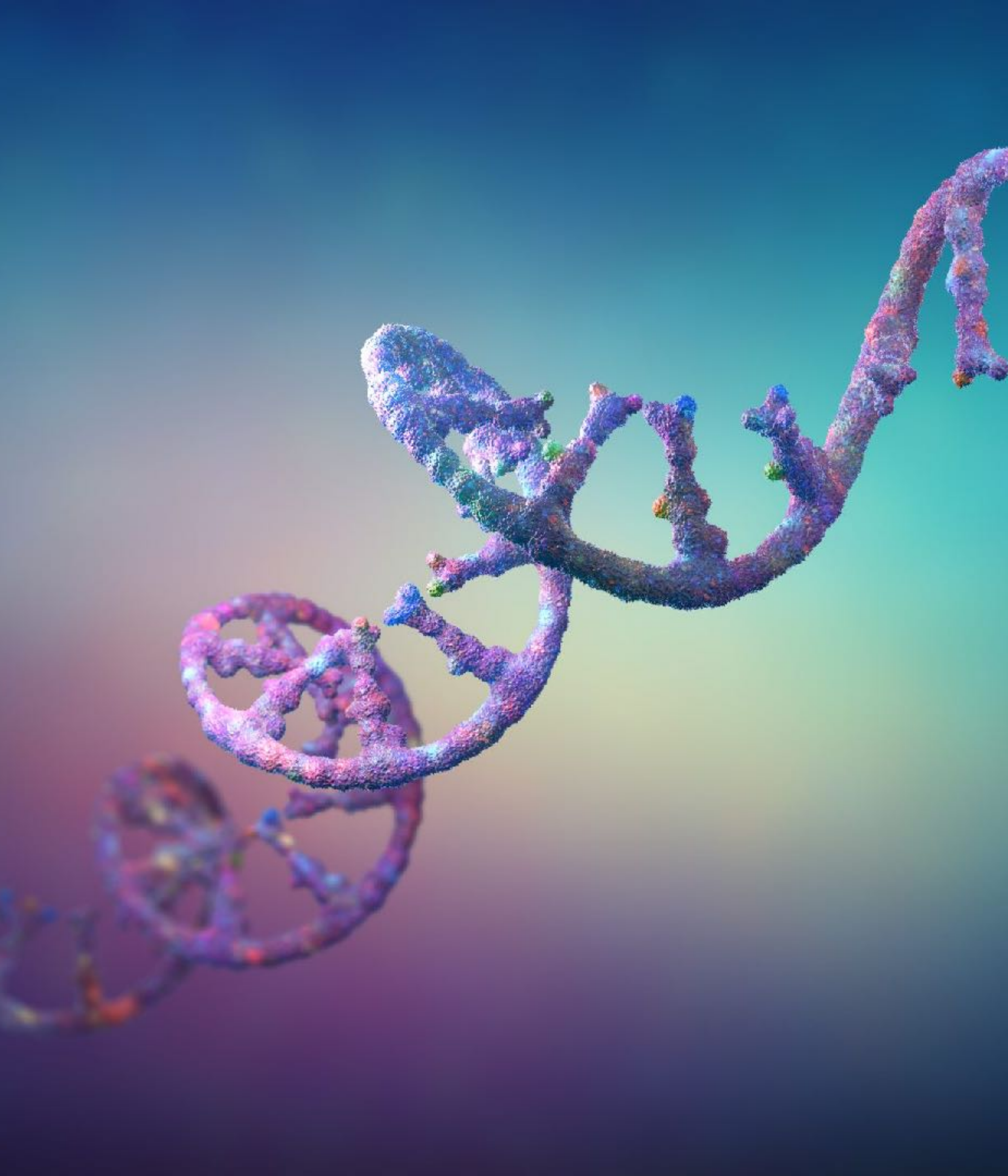
Knowledge Base control

Dataset folder: /Users/bubblegum_doubledrops/Library/Mobile Documents/com~apple~CloudDocs/0prio - Important heavy backups/IronHack/mini_projects/bio-rag-assistant/datasets

Total PDFs: 24

Processed: 24

Update knowledge base



**THANK YOU FOR
ATTENTION**

Faniia Prazdnikova