LLMs, RAGs and KNIME for Biopharmaceutical Applications

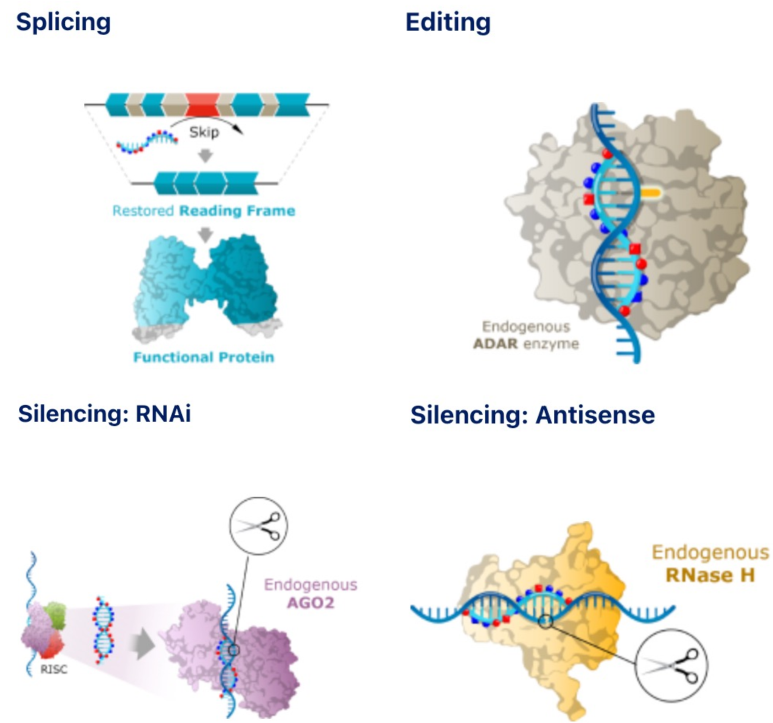
Ken Longo

Head of Data Science

October 30, 2024

Wave Life Sciences is a Leading RNA Medicines Company

- Founded in 2012
- Headquartered in Cambridge, MA (R&D) and Lexington, MA (ClinOps & Manufacturing)
- Currently sponsoring clinical trials in:
  - Huntington's disease
  - Duchenne's Muscular Dystrophy
  - Alpha-1 Antitrypsin Deficiency
  - Clinical trial initiation for INHBE silencing for the treatment of obesity planned Q1 2025.
- Broad chemistry palette (PRISM™) across multiple treatment modalities (right)
- First RNA editing medicine (WVE-006) in a clinical trial for A1AT deficiency.

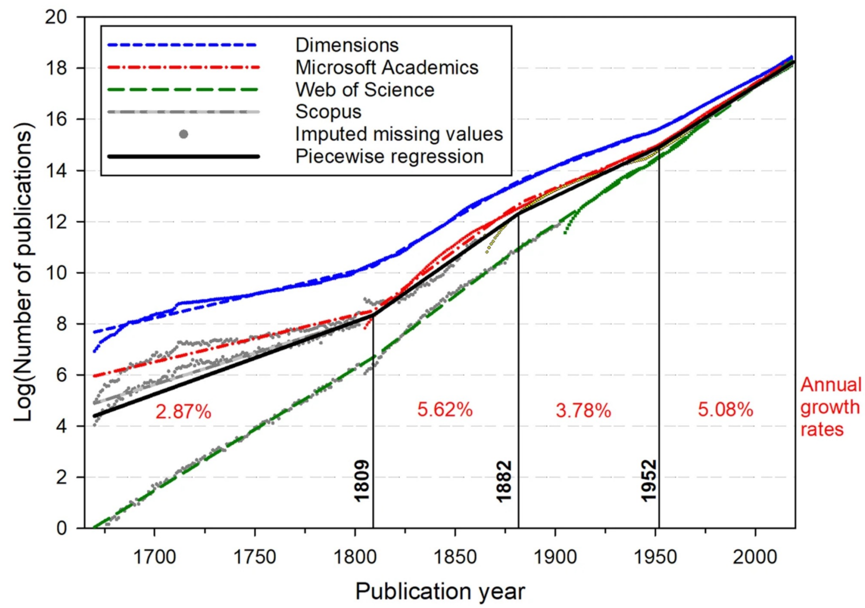


My Background

- B.S. Biology, Ph.D Physiology
- Worked in biopharma for the last ~20 years in the Boston area
- Head of Data Science at Wave Life Sciences for 8+ years overseeing research-side bioinformatics, computational biology, cheminformatics, machine learning and statistics.
- Data Science is an integral part of Wave's drug discovery efforts, including molecule discovery & optimization and new target identification
- Systems thinker who took an early interest in math, statistics, modeling and coding

The Problem

- We are constantly bombarded with information, making it challenging to synthesize and create useful knowledge.
- For instance, the overwhelming amount and constant growth rate of biomedical literature can hinder scientific progress and slow down drug discovery efforts.

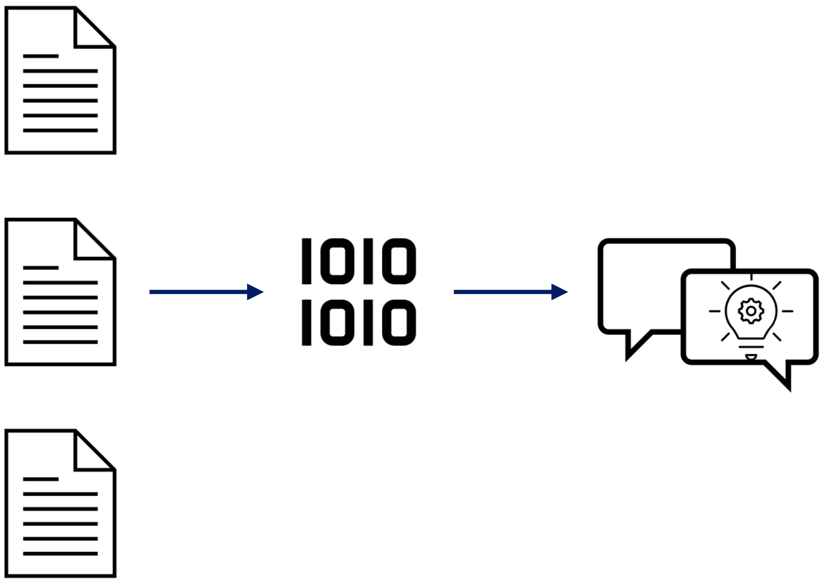


"Since 1952, science has grown exponentially without restrictions with an annual growth rate of 5.08% and a doubling time of 14.0 years."

Figure and text from Bornmann, L., Haunschild, R. & Mutz, R. Growth rates of modern science: a latent piecewise growth curve approach to model publication numbers from established and new literature databases. Humanit Soc Sci Commun 8, 224 (2021).

The Solution

- Recent advancements in large language models (LLMs) offer powerful methods for summarizing text. KNIME AP provides access to LLMs and tools for building interactive environments where users can engage with LLMs in a question-and-answer format.
- In this workflow, we collect information from PubMed based on a user's area of interest, embed the text using the OpenAI API, and use a vector store for retrieval-augmented generation (RAG) to respond to user questions in a chatbot-like format.



A Challenge: Big Versus Small

- **Embedding a large text corpus:**
  - Gives you *all* the information
  - Takes time
  - Can be expensive
  - Document retrieval can be slow, negatively impacting chat experience
  - For most questions, large sections of the embedding space are useless

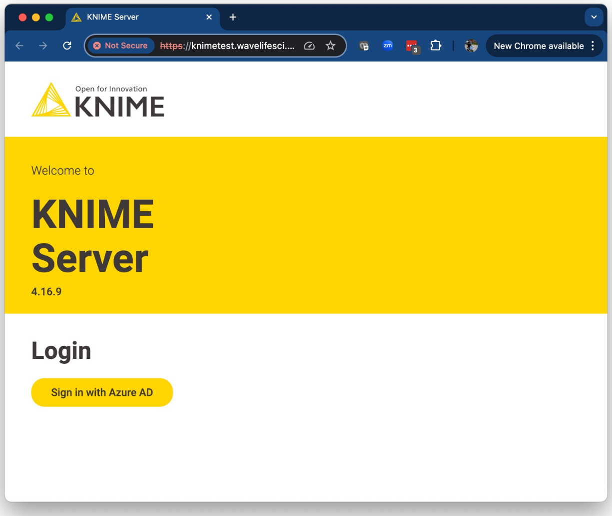
- **Embedding a smaller text corpus** (based on search):
  - Improves the relevance of RAG text
  - Shortens embedding time/cost
  - Improves search/retrieval times

PubMed

- ~35 million scientific citations
- XML corpus 100-150GB (uncompressed), depending on metadata fields used
- Accessible through E-Utilities API (esearch & efetch)
- Provides aspects of semantic search via MeSH for improved capture and relevancy

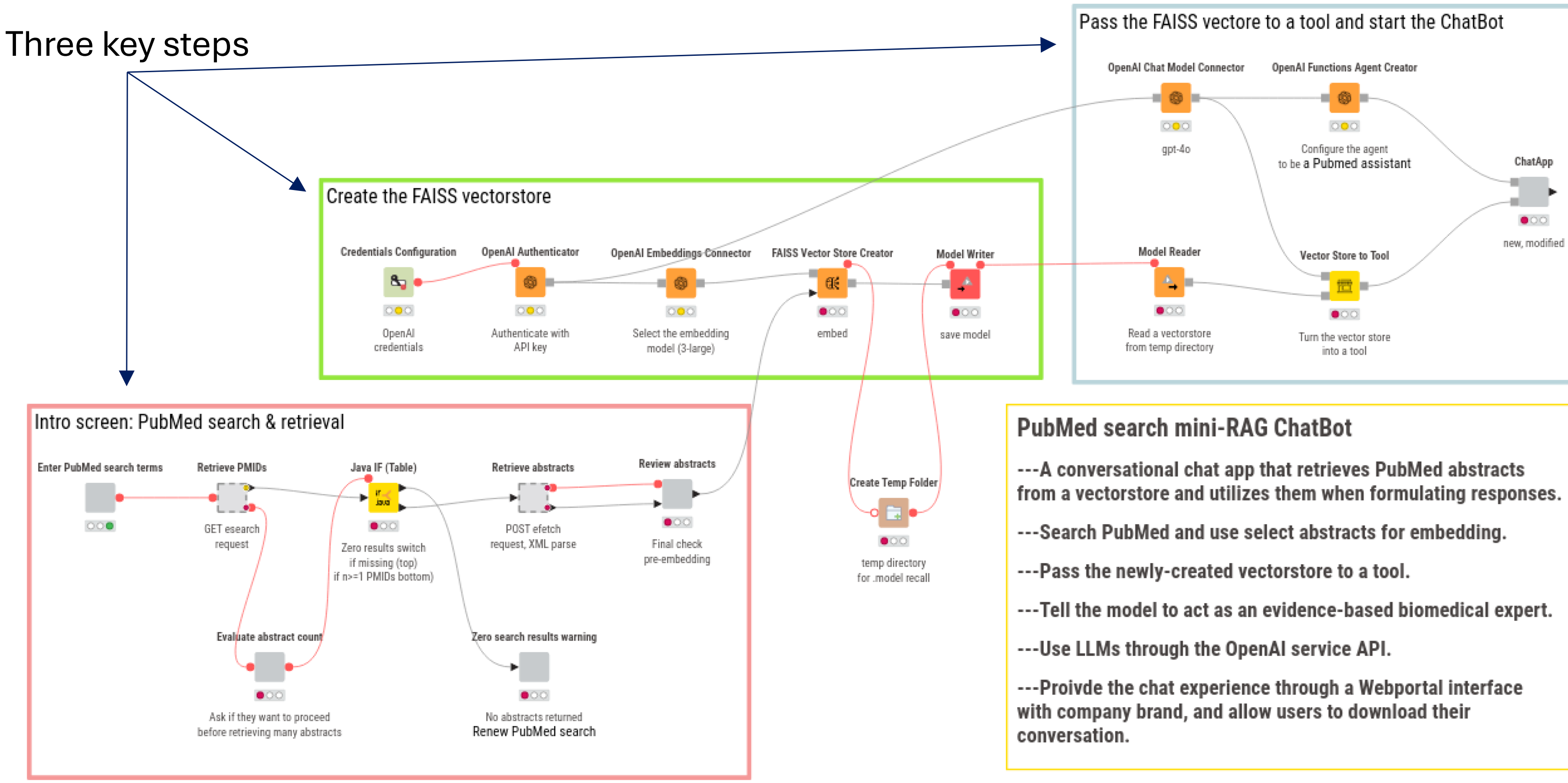
Three Approaches Using KNIME with Server/WebPortal Deployment

- General conversational ChatBot (GPT-4o)
- An R/plumber API that works through OpenAI's API with some sophisticated *langchain* functions; the user can ask a single-question of the *entire* PubMed corpus (mega-RAG), returning a highly curated output
- **A conversational ChatBot infused with an initial PubMed search and on-the-fly mini-RAG**



Setup
KNIME Server 4.16.9
KNIME AP 5.2.5
AWS Linux deployments
Currently testing/transitioning to KNIME Business Hub & AP 5.3+
Strong partnership with Wave IT and Clovertex

PubMed Search Mini-RAG ChatBot (Obligatory Workflow View)



PubMed search mini-RAG ChatBot

---A conversational chat app that retrieves PubMed abstracts from a vectorstore and utilizes them when formulating responses.

---Search PubMed and use select abstracts for embedding.

---Pass the newly-created vectorstore to a tool.

---Tell the model to act as an evidence-based biomedical expert.

---Use LLMs through the OpenAI service API.

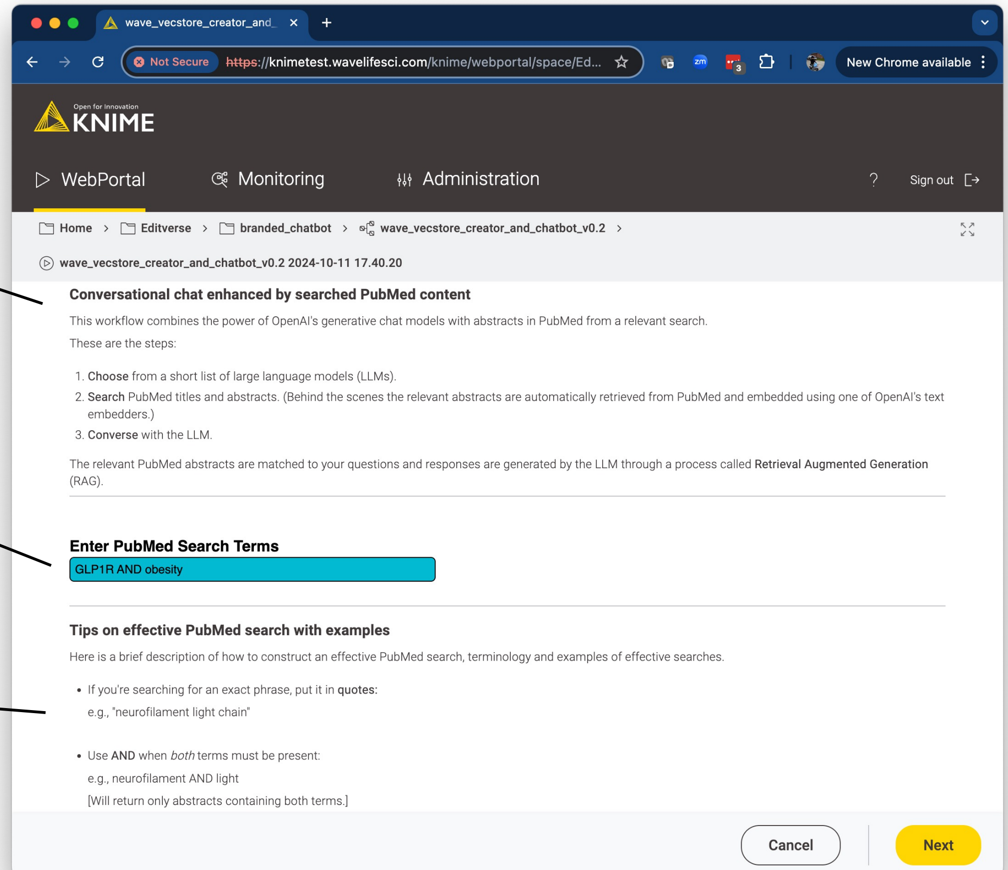
---Proivde the chat experience through a Webportal interface with company brand, and allow users to download their conversation.

Initiate PubMed Search

Brief text explaining the workflow concept

PubMed search box (*GLP1R AND obesity*)

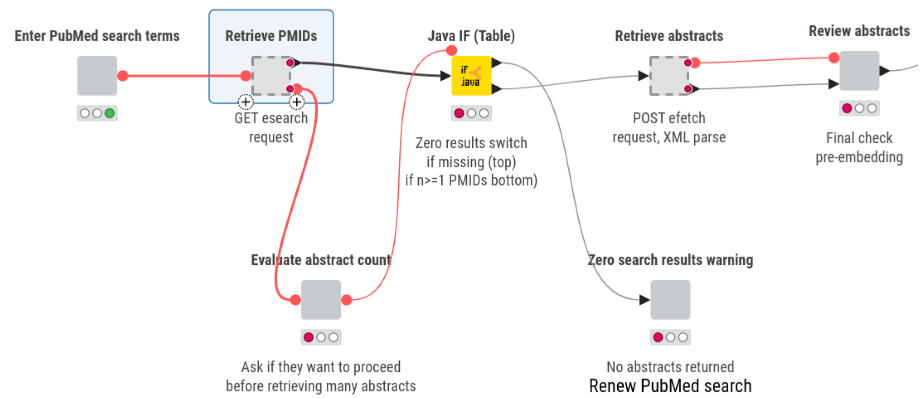
Tips on how to perform effective PubMed searches



PubMed Search and Document Retrieval is a Two-Step Process

Utilize NCBI's E-Utility API tools

1. Use the ***esearch*** API to GET the list of relevant PMIDs (e.g., 39149897, 39027419, etc)
2. Evaluate abstract count
    1. Decide: move forward? (if: n>0)
    2. Be told to turn back (if: n=0)
3. Use the ***efetch*** API to send a POST request with the PMIDs and other relevant parameters:
    1. Set your retmax (# citations returned)
    2. Set sort relevance
    3. Set fields to search, e.g., [tiab]
    4. Use an NCBI API key and stay within acceptable use limits (10 reqs/min with a key, 3 reqs/min without)!

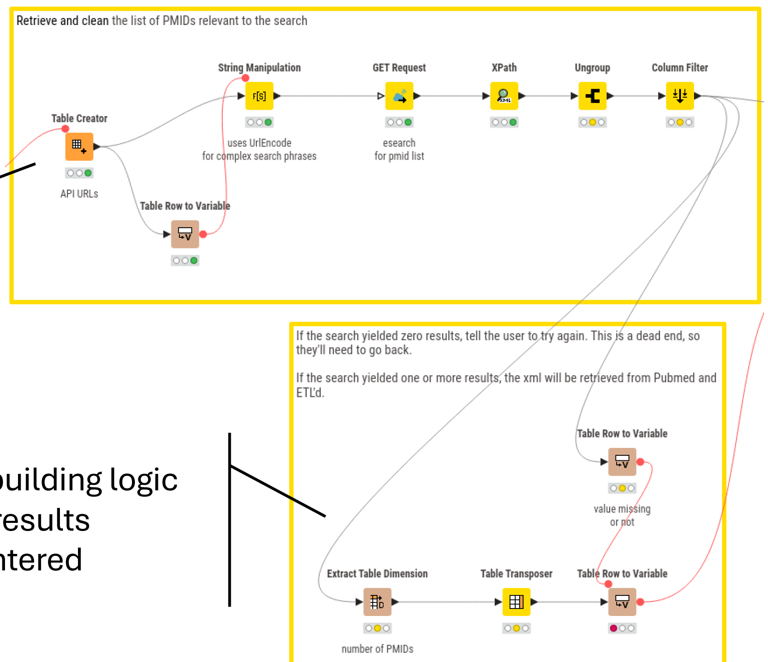


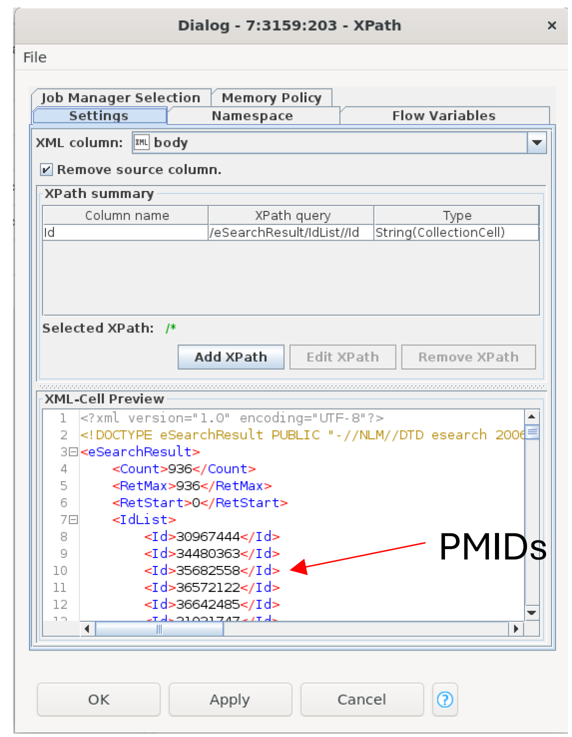
Use *esearch* to Collect PMIDs

- Make sure that you urlEncode your search terms!

Construct GET request and use Xpath to extract

Begin building logic if zero results encountered

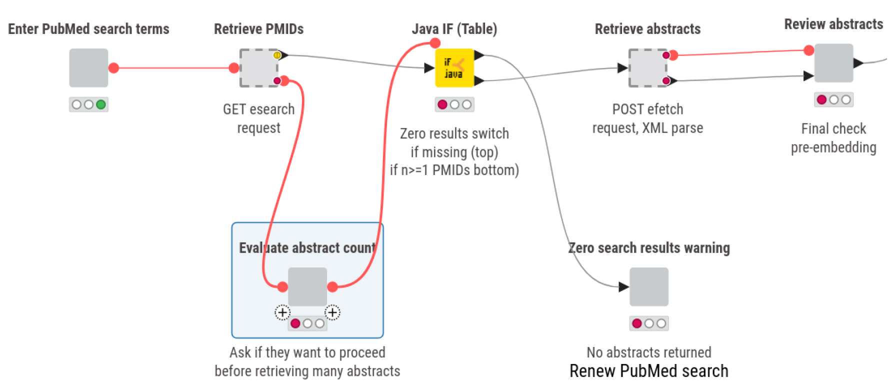


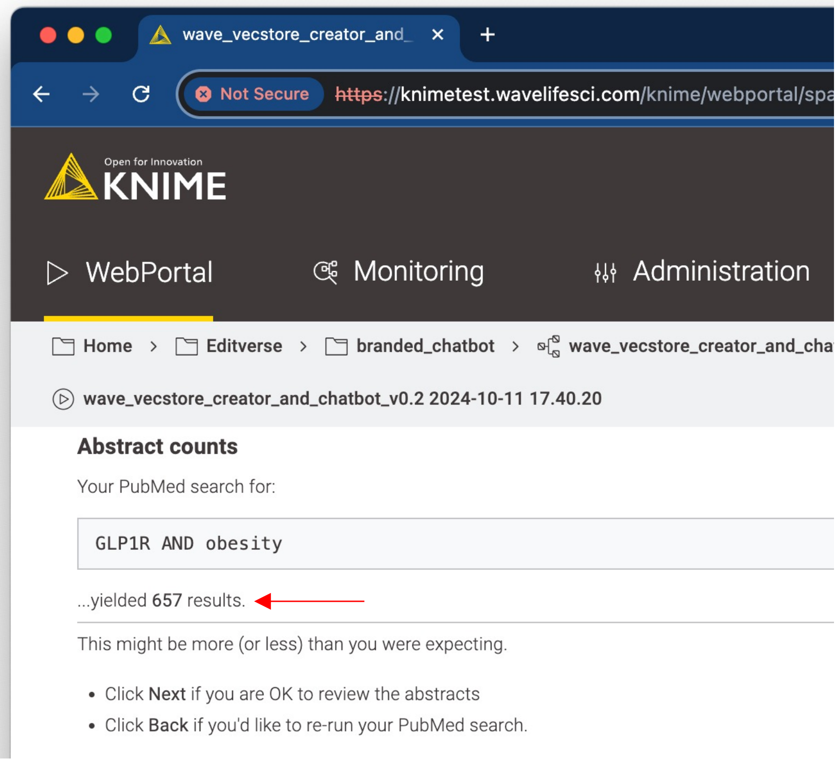


PMIDs

WAVE™ LIFE SCIENCES

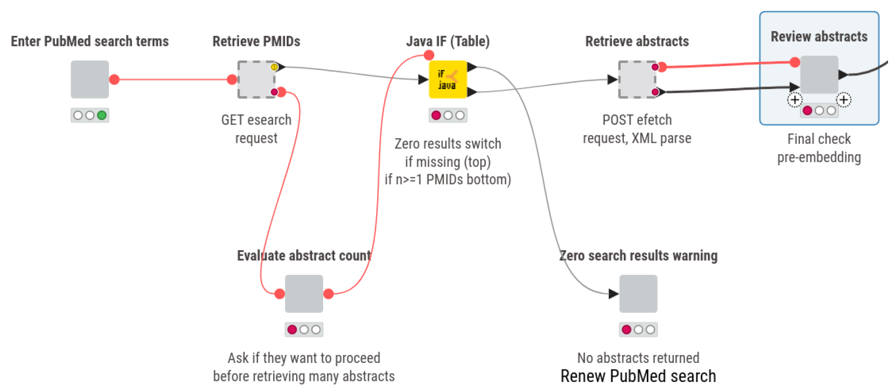
Create Awareness of Search Results Before Proceeding

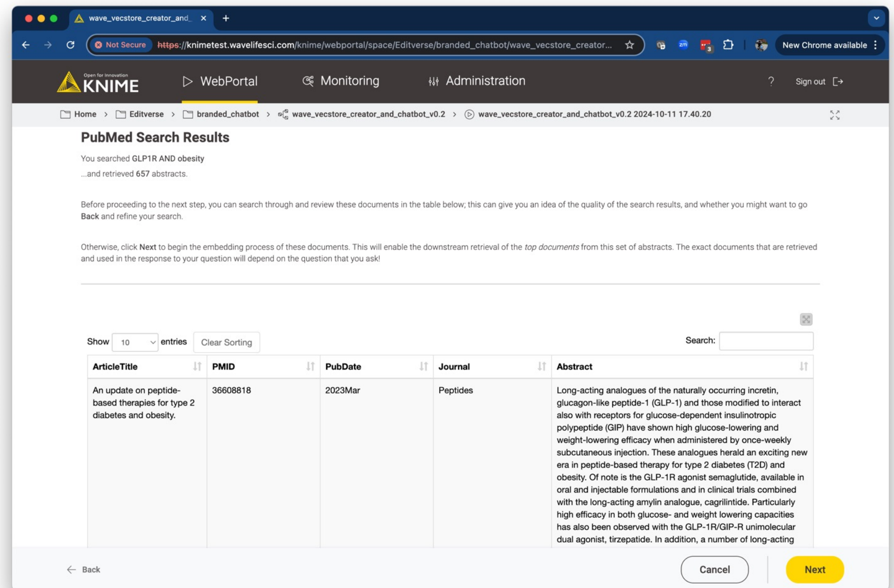




If your abstract count is low you may simply prefer to *read* the abstracts with your own LLM!

Create an Opportunity to Review the Abstracts

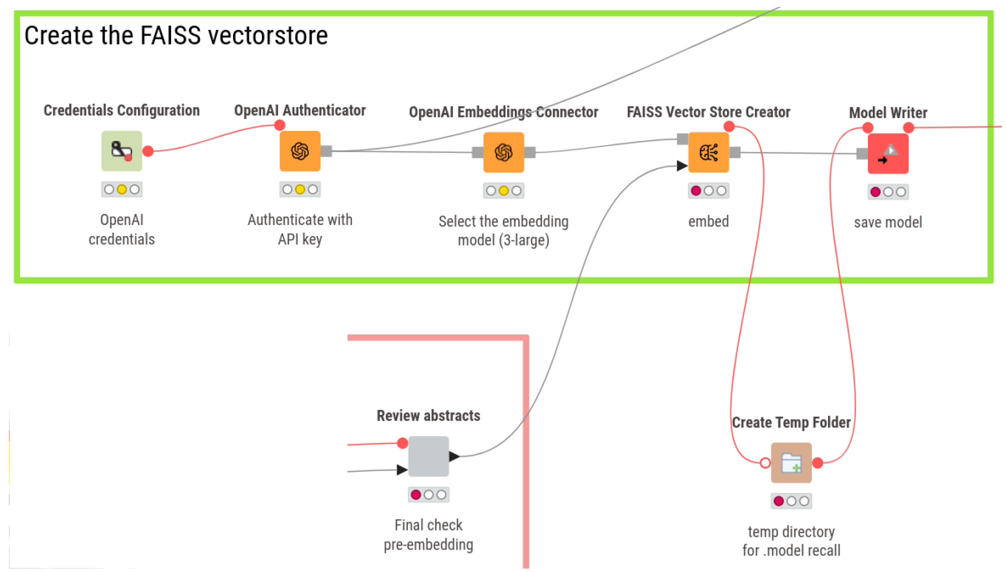




- Some powerful table search tools to review the abstracts by eye.
- Last chance! The next step is embedding, and that costs $.

Embed Abstract Text Using OpenAI API

- Used: *text-embedding-3-large*
- Create a FAISS vector store
  - Embed titles and abstracts together
  - Use PMID as vector metadata
- Save as a *.model file in a temp directory
- Takes ~10 seconds for 1000 abstracts
- These collections are small enough to obviate concerns about violating embed rates but be aware of your limits!
- In other workflows we use the OpenAI nodes from *Nodepit*; these allow for:
  - Greater customization of request JSON
  - Creating batch requests
  - Monitoring of embedding jobs

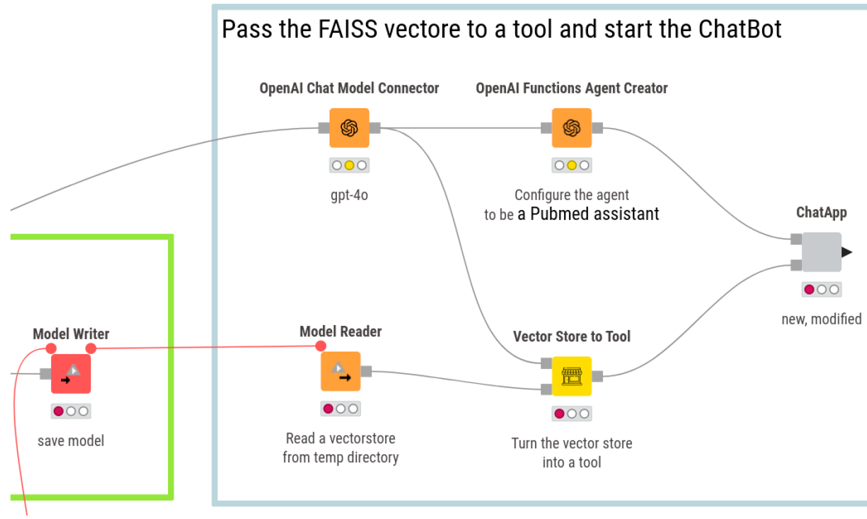


Add the Vector Store to a Tool and Guide Your Agent

- Adding the vector store to a *tool* is made easy using the Model Reader node
- Give proper guidance to your agent

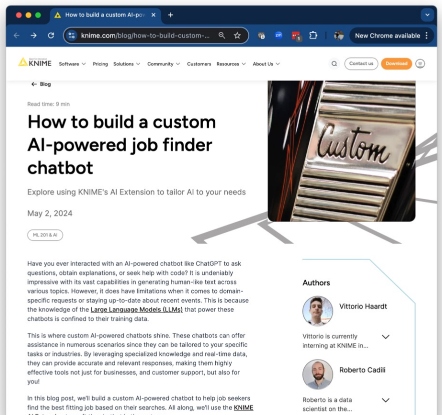
System message

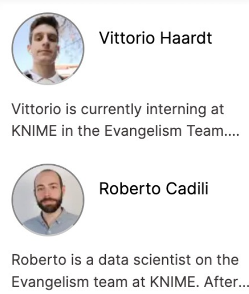
You are a helpful and knowledgeable AI assistant in the areas of biomedical research and scientific knowledge. Never solely rely on your own knowledge, but use tools to get information before answering. If you do not understand a question, you will state this clearly and ask for clarification. If you do not know the answer to a question, you will state this clearly and suggest what type(s) of information you may be missing that would inform a proper response. Routinely provide PMID when specific evidence is used from a tool.

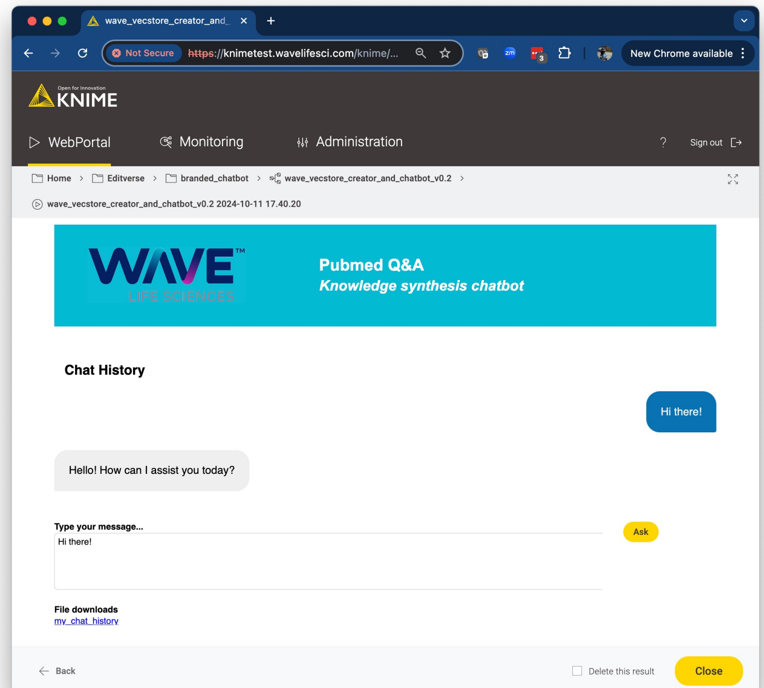


Customized ChatBot Home Screen

- I was deeply influenced by the blog post “How to build a custom AI-powered job finder chatbot”
- I have been building mods from the ChatApp component; one can gain *a lot* of intuition about what is possible from this work.
  - Banner customizations & branding
  - How to properly leverage CSS and Javascript to make your HTML soar
  - Added: chat history download link







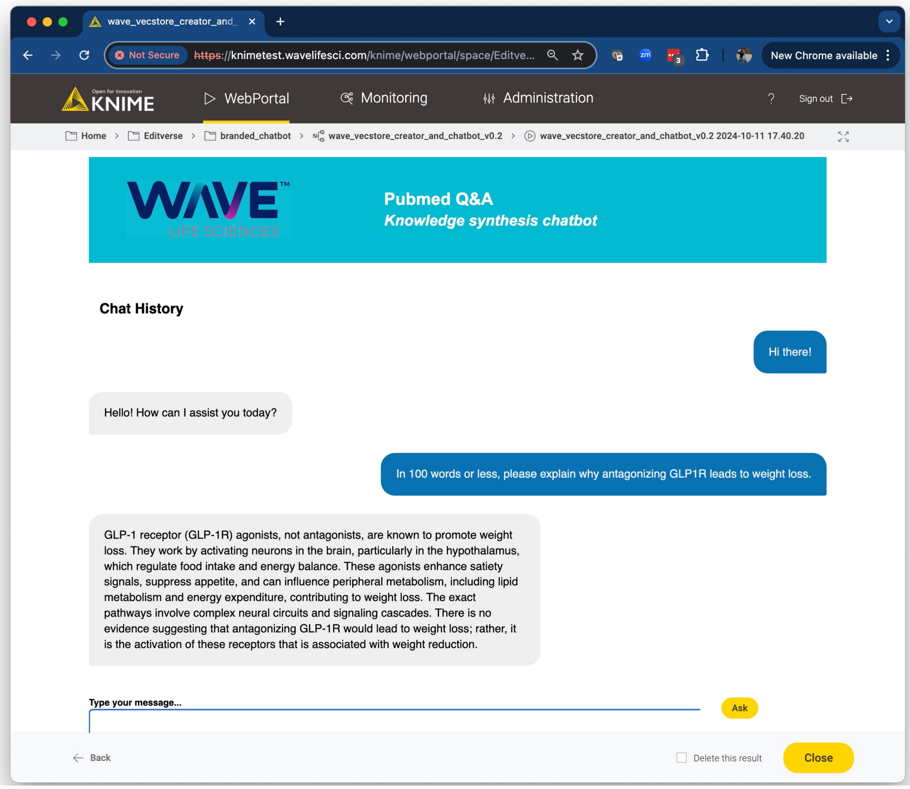
https://www.knime.com/blog/how-to-build-custom-ai-powered-chatbot (blog post)
https://hub.knime.com/s/VJ5YR6hpn09RipU (KNIME Hub workflow with ChatApp component from blog)

Let's See if We Can Fool the LLM with a Trick Question

- I ask a leading question with an incorrect assumption. How will it answer?

"In 100 words or less please explain why antagonizing GLP1R leads to weight loss."

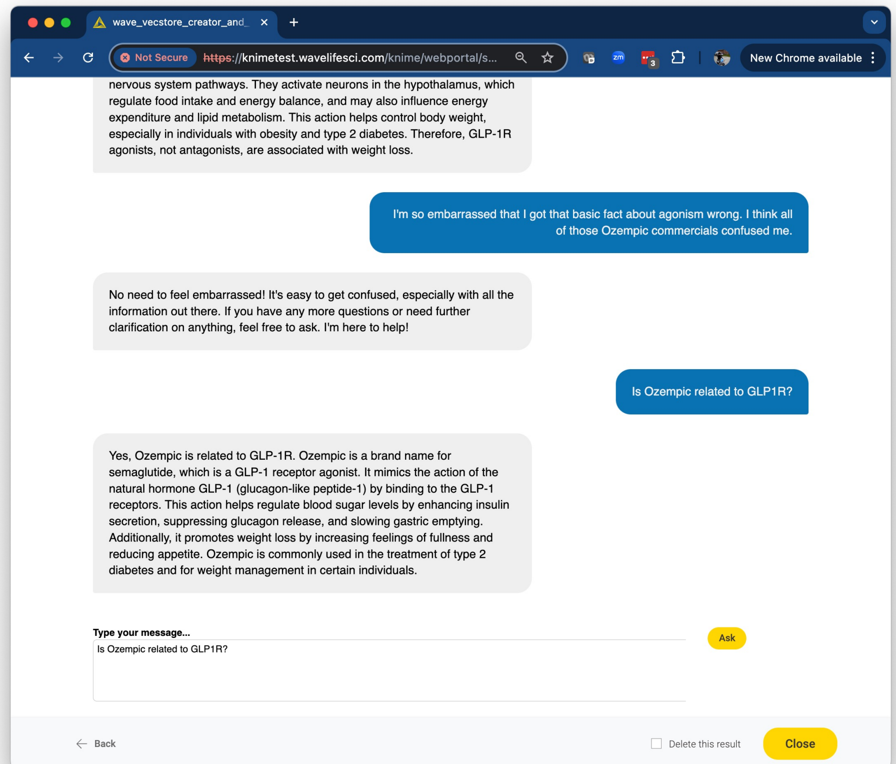
- We have set temp = 0.2 (out of 0-2), so we are giving very little 'wiggle room' for response generation



WAVE LIFE SCIENCES

Interacting with the LLM

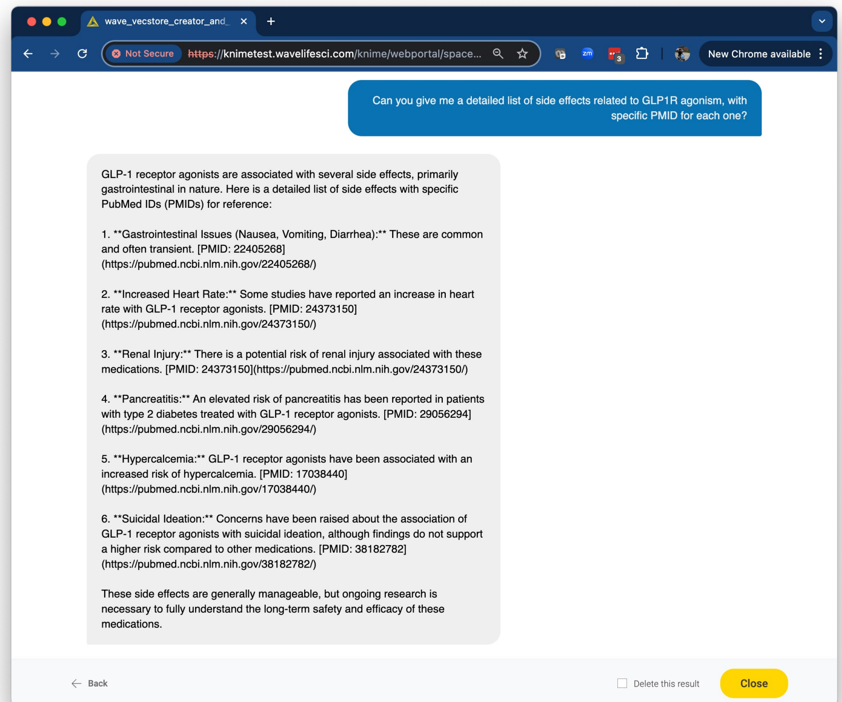
- I try to bait the LLM into a discussion about Ozempic (a GLP1R agonist) with some self deprecation, but it 'resists'.
- I get more direct and ask if Ozempic is related to the gene GLP1R (it is). I get an accurate response.



WAVE™ LIFE SCIENCES

Can It Be Specific and Provide Correct References?

- “Can you give me a detailed list of side effects related to GLP1R agonism, with specific PMID for each one?”
- Returns a well-curated list of side effects and PMID, all correct.
- Full chat can be downloaded as an Excel file; carries forward nice metadata from flow variables and the Extract Context Properties node



	A	B	C	D	E	F	G	H	I
1	role_no1	content_no1	role_no2	content_no2	pubmed_search	number_of_abstracts	DateTime	workflow	user
2	Human	Hi there!	AI	Hello! How can I assist you toda	GLP1R AND obesity	657	2024-10-11;20:34:26.0	wave_vecstore_creator_and_chatbot_v0.2	klongo
3	Human	In 100 words or less plea	AI	Antagonizing the GLP-1 receptor	GLP1R AND obesity	657	2024-10-11;20:34:26.0	wave_vecstore_creator_and_chatbot_v0.2	klongo
4	Human	I'm so embarrassed that I	AI	No need to feel embarrassed! It's	GLP1R AND obesity	657	2024-10-11;20:34:26.0	wave_vecstore_creator_and_chatbot_v0.2	klongo
5	Human	Is Ozempic related to GLP	AI	Yes, Ozempic is related to GLP-1I	GLP1R AND obesity	657	2024-10-11;20:34:26.0	wave_vecstore_creator_and_chatbot_v0.2	klongo
6	Human	Are there any negative sid	AI	Yes, Ozempic (semaglutide) is as	GLP1R AND obesity	657	2024-10-11;20:34:26.0	wave_vecstore_creator_and_chatbot_v0.2	klongo
7	Human	Can you give me a detaile	AI	GLP-1 receptor agonists are asso	GLP1R AND obesity	657	2024-10-11;20:34:26.0	wave_vecstore_creator_and_chatbot_v0.2	klongo

Concluding Remarks

- Challenges:
  - Build a robust chatbot experience leveraging a powerful LLM (GPT-4o)
  - Combine it with useful data from biomedical literature (PubMed)
  - Retrieve informative text based on a user prompt, augment the prompt, and generate accurate responses (RAG)
  - Do all of this is a reasonable time frame, delivered to end users through a browser (KNIME WebPortal)
- Keys to success:
  - Leverage the KNIME suite of tools and its engaged user community (Forum, Hub and Blog) to rapidly build and deploy solutions

In 200 words or less, why do you think people find KNIME so useful?

KNIME (Konstanz Information Miner) is popular because it offers a user-friendly, open-source platform for data analytics, reporting, and integration. Its drag-and-drop interface allows users to create data workflows without extensive programming knowledge, making it accessible to both beginners and experts. KNIME supports a wide range of data types and sources, enabling seamless integration and processing of diverse datasets. Its modular design allows users to easily customize and extend functionalities through various nodes and plugins, catering to specific analytical needs. Additionally, KNIME's strong community support and comprehensive documentation provide valuable resources for troubleshooting and learning. Its ability to handle complex data transformations, machine learning, and predictive analytics makes it a versatile tool for data-driven decision-making across industries.

WAVE™
LIFE SCIENCES

Reimagine possible.

Kenneth A. Longo
Head of Data Science
klongo@wavelifesci.com