



ToxPipe

ToxPipe: Chatbots and Retrieval-Augmented Generation on Toxicological Data Streams

Trey Saddler

A screenshot of a code editor interface. At the top, a search bar contains the text "@How can I add a git submodule to a repo?". A context menu is open over the search bar, listing options: Files, Codebase, Code, Docs, Git Diff, Terminal, Problems, Folder, and a highlighted blue button "+ Add more context providers". Below the menu, the text "3. Use the following command to add a submodule:" is visible. A code block contains the command:

```
git submodule add <repository-url> <path>
```

 At the bottom, a note says "Replace <repository-url> with the URL of the repository".

tinue



National Institute of
Environmental Health Sciences
Division of Translational Toxicology

ToxPipe

ToxPipe learns from trusted sources with bespoke AI agents → “expert guidance”



Scott Auerbach



Trey Saddler



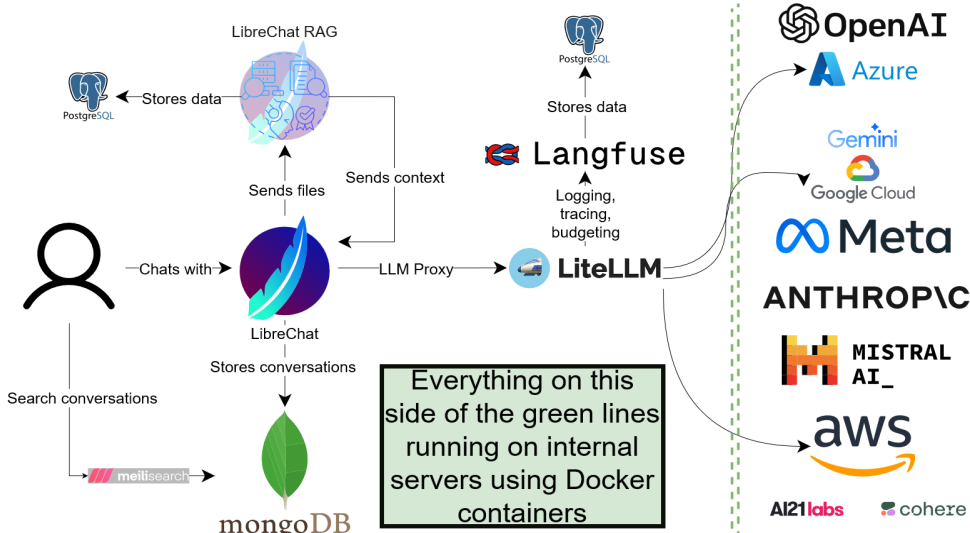
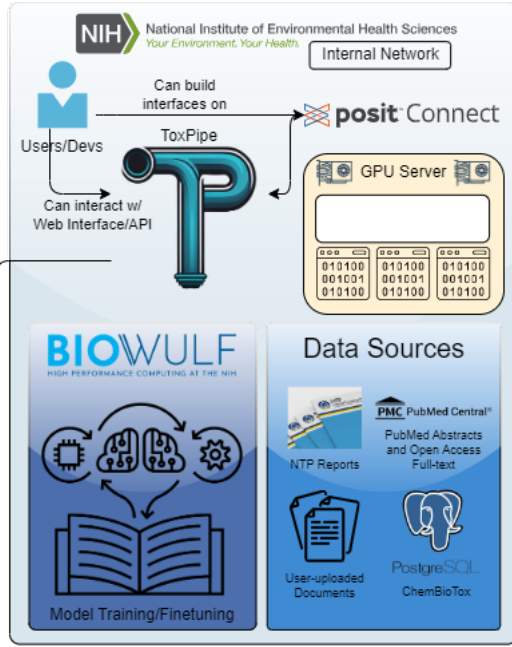
David Reif

Azure Cognitive Search
Power retrieval-augmented generation apps to discover content insights.

Azure Cognitive Search, an AI-powered information retrieval platform, helps developers build rich search experiences and generative AI apps that combine large language models with enterprise data.

gpt-3.5-turbo
gpt-4
text-embedding-ada-002

Internet, Google, PubMed, PubChem, and more...
Comptox Dashboard



ToxPipe LLM Model Testing

niehs.github.io/toxpipe-model-compari

Initial Prompt Testing

AUTHOR: Trey Saddler PUBLISH: March 2

Introduction

These tests were performed using a variety of user i of the markdown syntax was performed to ensure c platforms. However, none of the content of the LLM

Basic Prompt

User Prompt

I see the following genes going up in the liver of rats af acot2, decr1, hmgs2, Cyp4a1, ehhadh, and vnn1. Wha

GPT-4 1106-preview Claude 3 Opus Gemini Mistral Large

The upregulation of these genes in the liver of rat suggests alterations in hepatic metabolism, parti ketogenesis, and detoxification pathways. Here's each mentioned gene:

- Acot1 (Acyl-CoA Thioesterase 1):** Involved i the free fatty acid and coenzyme A (CoA), wh levels of acyl-CoAs and free fatty acids.

LibreChat

chat.toxpipe.niehs.nih.gov/c/new

Bioinformatician/Toxicologist

You

What data streams should be included in a ToxPi model of PFAS health effects? Suggest relative weights for each data stream.

Bioinformatician/Toxicologist

The Toxicological Prioritization Index (ToxPi) model is a versatile framework for integrating diverse types of data to prioritize chemicals based on concerns for human health, ecological health, or other endpoints. Poly- and perfluoroalkyl substances (PFAS) are a diverse group of chemicals with potential health risks, and several data streams can be relevant for assessing their toxicological profile.

Here is a list of potential data streams for a ToxPi model of PFAS health effects, along with suggested relative weights. These weights might be adjusted based on expert consultation, stakeholder input, and the specific objectives of the prioritization exercise:

- In Vivo Animal Toxicity Studies** - This category might include outcomes from rodent assays that investigate various endpoints such as liver toxicity, developmental toxicity, immunotoxicity, and carcinogenicity. Since animal studies can provide direct measures of adverse effects, they could be given a higher weight. [CS: 9]
- Epidemiological Evidence** - Data linking PFAS exposure to health outcomes in humans, such as thyroid dysfunction, elevated cholesterol levels, and certain cancers. Given the direct relevance to human health, this stream may also carry a higher weight. [CS: 9]
- In Vitro Bioactivity** - Includes results from high-throughput screening assays that provide information on bioactivity across a range of biological targets, such as nuclear receptors, stress response pathways, and developmental pathways. These data streams are informative but might be given a lesser weight compared to in vivo data due to uncertainties in extrapolating to whole-organism effects. [CS: 7]
- Physicochemical Properties** - Characteristics that influence the environmental fate and the bioavailability of the substances, including water solubility, vapor pressure, and partition coefficients. These can inform on potential for bioaccumulation and persistence. [CS: 8]
- Exposure Data** - Data on levels of PFAS in environmental media (water, air, soil) and in populations through biomonitoring. Exposure potential is critical to risk, so this category should be weighted significantly. [CS: 9]
- Environmental Persistence and Bioaccumulation** - PFAS are known to be highly persistent in the environment and can bioaccumulate in living organisms, so data on these aspects are essential and should be weighted accordingly. [CS: 10]
- Metabolomics** - Information from metabolomic studies can indicate biochemical

Message Bioinformatician/Toxicologist...

LibreChat v0.6.10 - All AI conversations in one place. Privacy policy Terms of service



Trey Saddler, Parker Combs, Jeremy Erickson, Jonathon Fleming, Charles Schmitt, David Reif, Scott Auerbach

Meet the New LibreChat Resources Hub! 🚀



[Docs](#) [Blog](#) [Demo](#) [Changelog](#)

Search...

CTRL K



Librechat

Every AI in One Place
Built for Everyone

LibreChat is a free, open source AI chat platform. This Web UI offers vast customization, supporting numerous AI providers, services, and integrations. Serves all AI Conversations in one place with a familiar interface, innovative enhancements, for as many users as you need.

[Get Started →](#)

[GitHub ↗](#)

The screenshot shows a LibreChat interface with a dark theme. On the left is a sidebar with a search bar and a list of chat messages. The main chat area shows a conversation between 'You' and 'GPT-4'. 'You' has uploaded a 'README.md' document and asked 'What is this document about?'. 'GPT-4' has responded with a detailed summary of the document's content, including its purpose, system architecture, and ethical considerations. At the bottom, there is a text input field with the text 'Message GPT-4...' and a send button. The footer contains the text 'LibreChat v0.6.10 - All AI conversations in one place. | Privacy policy | Terms of service'.

Chat History:

- Generate Logo for Predictive Tox...
- What Is This Document? [edit] [trash]
- What Is This Document?
- Find Similar Open-Source Proj...
- Complex Liver Response to Tox...
- Testing Assistance Available
- Upregulated Genes in Liver An...
- Upregulation of Genes in Rat Li...
- New Chat
- New Chat
- Neo4j Query for CYP3A4 Induc...
- Is Working Properly

Chat Conversation:

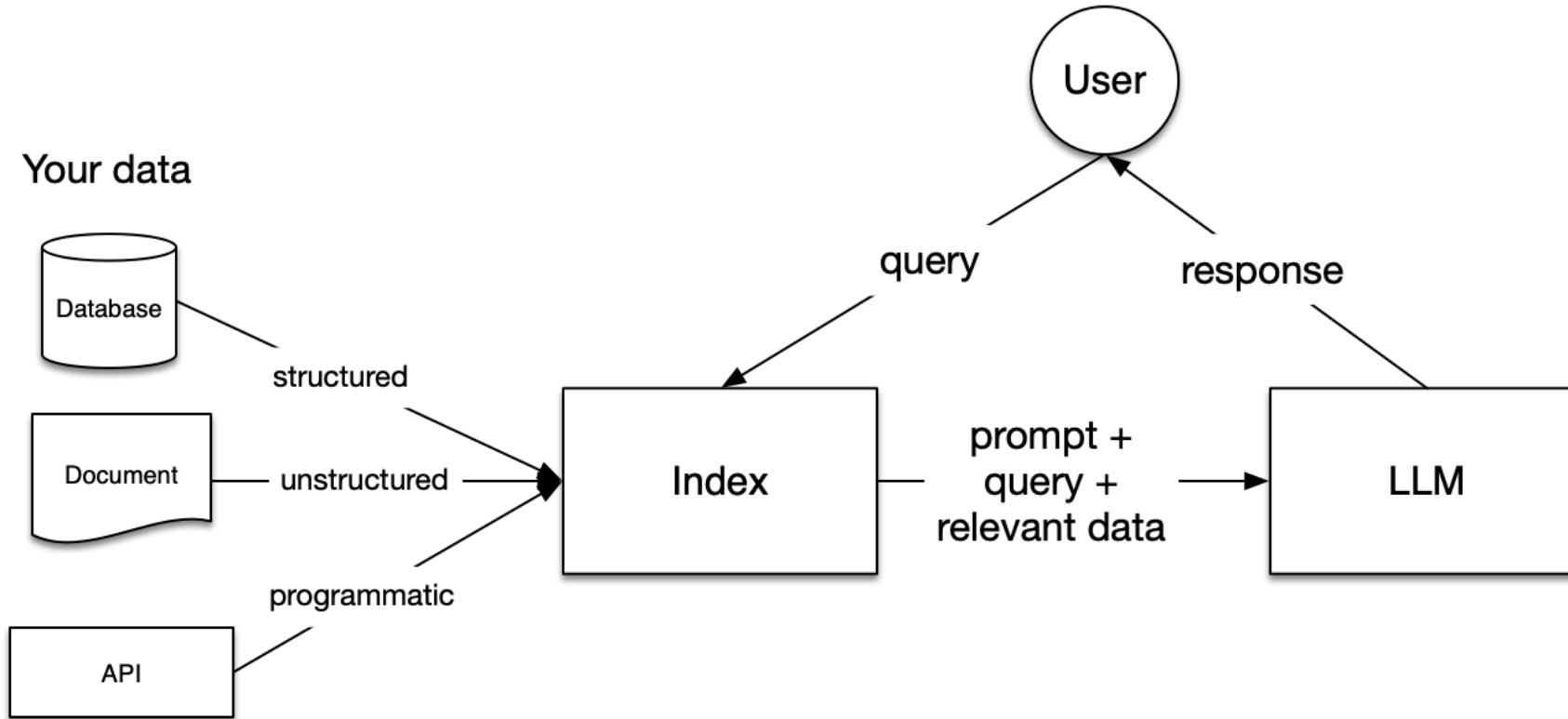
You: [upload] README.md Document
What is this document about?

GPT-4:
The document appears to be a README file that provides an overview of a research project called ToxPipe, associated with the NIEHS-STRIDES initiative. ToxPipe is focused on utilizing expert entrained AI-based systems to analyze and interpret the toxicological properties of various compounds. The document outlines the deployment information of the project, including the use of GitLab for repository management and Azure Cloud for OpenAI API access. It includes useful links related to Azure and GitLab, describes the repository structure, and discusses the use of BioWulf and NIEHS's high-performance computing for training expert models.

Additionally, it mentions the types of tasks ToxPipe aims to accomplish with AI agents in the field of toxicology, such as predicting toxicity, generating narratives, and answering questions about toxicological topics. The document also provides insights into the system architecture of ToxPipe, highlighting that the architectural details are documented in a specific directory and that decision-making rationale is recorded in another.

Finally, the document highlights the importance of investigating ethical considerations when using generative AI for toxicological purposes.

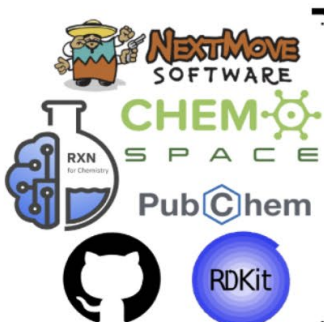
Input: Message GPT-4...



Chem Crow



APIs



Tools



ChemCrow
LangChain
+ LLMs

- Synthesize target
- Explain mechanism
- Find similar molecules

User input / tasks

thought
reason, plan

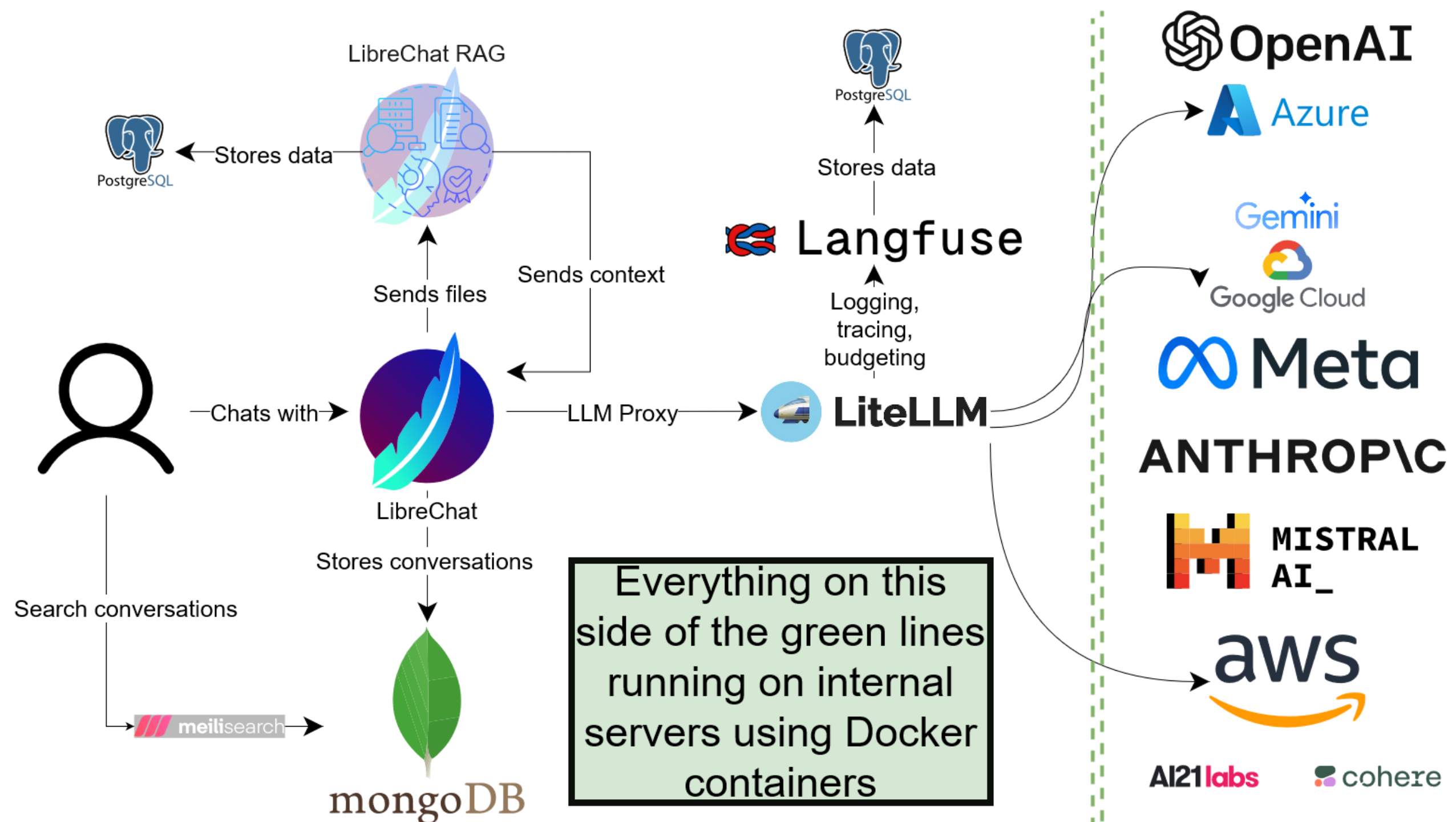
action
select tool

analyse
observation

use tool
action input

**Tool-informed
final answer**

Automatic iterative LLM + tools workflow





Semantic Scholar

PubChem



Open-Source Cheminformatics
and Machine Learning

RDChEMBL @rdch

PubMed



Technical Reports

Long-term studies of the
toxicologic potential of agents
in test animals. [Go »](#)



Chem



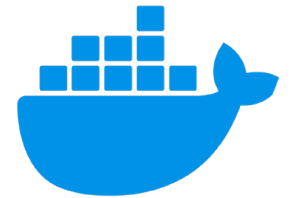
Crow



OpenAI



LangChain

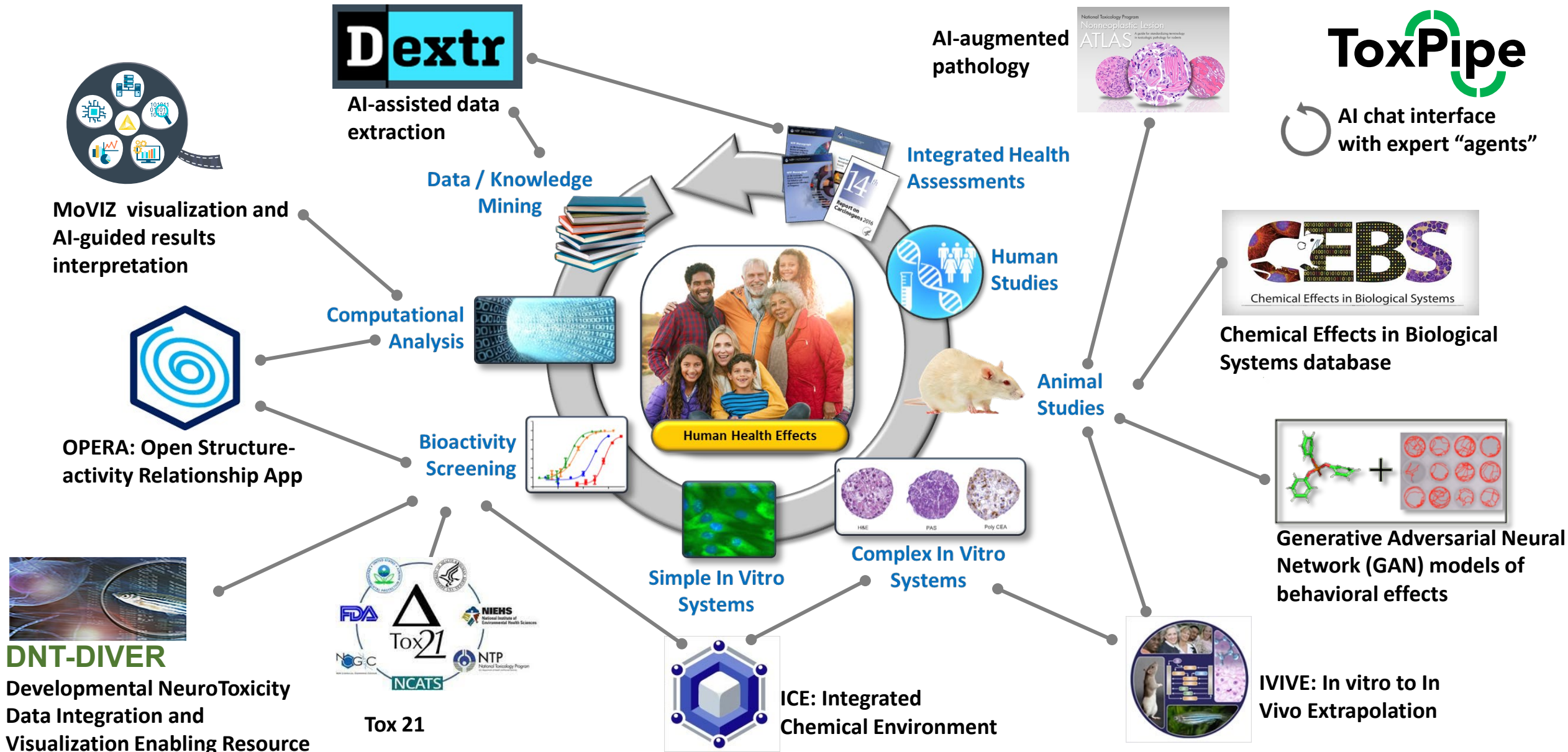


docker



A Microsoft Azure

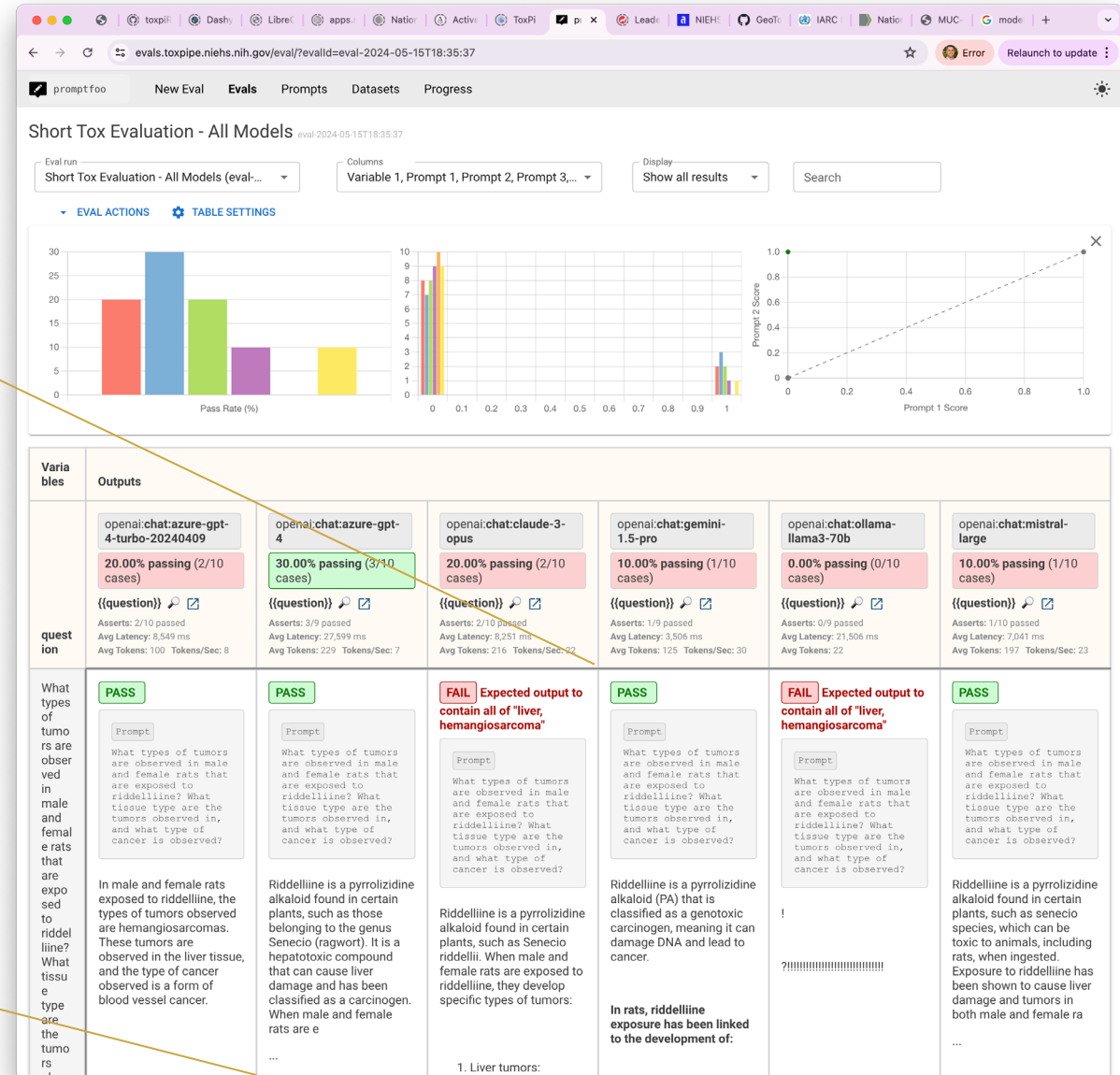
ToxPipe integrates across the Translational Toxicology Pipeline



Formal testing and model evaluation: Does generative AI give meaningful answers?

Generic or *one-size-fits-all* AI chat solutions FAIL at specific translational tests → expert guidance, training and domain-tailored RAG solutions will be key

PASS	FAIL Expected output to contain all of "liver, hemangiosarcoma"
<p>Prompt</p> <p>What types of tumors are observed in male and female rats that are exposed to riddelliine? What tissue type are the tumors observed in, and what type of cancer is observed?</p>	<p>Prompt</p> <p>What types of tumors are observed in male and female rats that are exposed to riddelliine? What tissue type are the tumors observed in, and what type of cancer is observed?</p>
<p>Riddelliine is a pyrrolizidine alkaloid found in certain plants, such as those belonging to the genus Senecio (ragwort). It is a hepatotoxic compound that can cause liver damage and has been classified as a carcinogen. When male and female rats are e</p>	<p>Riddelliine is a pyrrolizidine alkaloid found in certain plants, such as Senecio riddellii. When male and female rats are exposed to riddelliine, they develop specific types of tumors:</p>





AUTHOR
Trey Saddler

PUBLISHED
March 29, 2024

On this page

- [Introduction](#)
- [Basic Prompt](#)
- [Medium Prompt](#)
- [Advanced Prompt](#)

Introduction

These tests were performed using a variety of user interfaces. Therefore, minor editing of the markdown syntax was performed to ensure consistency across the different platforms. However, none of the content of the LLM responses was altered.

Basic Prompt

User Prompt

I see the following genes going up in the liver of rats after treatment with chemical. Acot1, acot2, decr1, hmgcs2, Cyp4a1, ehhadh, and vnn1. What do you surmise is happening?

GPT-4 1106-preview

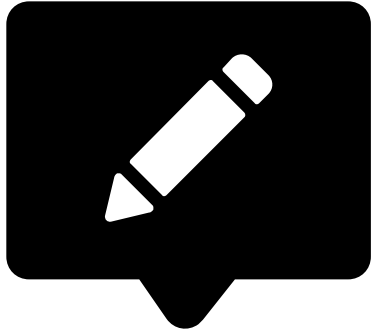
Claude 3 Opus

Gemini Pro 1

Gemini Pro 1.5

Mistral Large

The upregulation of these genes in the liver of rats after exposure to a chemical suggests alterations in hepatic metabolism, particularly in fatty acid oxidation, ketogenesis, and detoxification pathways. Here's a breakdown of the function of each mentioned gene:



promptfoo

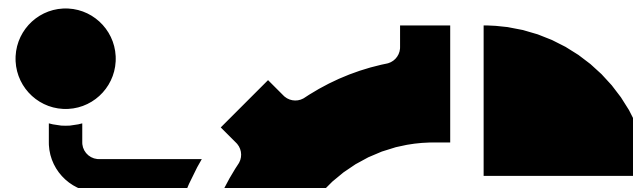
- <https://niehs.github.io/ToxPipe-Model-Comparison/reports/2024-03-29-initial-prompts.html>



- Focused on natural language interactions and conversations with users
- Typically respond to specific queries or commands from users
- Often have a more limited scope of capabilities centered around assisting with tasks like scheduling, answering questions, etc
- May not have as much autonomy or ability to take independent actions




Assistant



- Can operate autonomously without constant human input or supervision
- Designed to take independent actions to achieve goals
- Might learn and improve their performance over time
- May use more advanced reasoning and decision-making capabilities
- Can interact with other systems, tools, and agents to complete complex tasks



Agent

New chat 

Search messages

Bookmarks


Previous 7 days



- Access to toxicology study fil
- Files I Can Access


Previous 30 days

- Accessible Files List
- Document Access and Query
- Summarize Transcript
- Uploaded file - provide conte

AI Mounting Windows Shares St




 **Trey Saddler**

ToxPipe Assistant  




ToxPipe Assistant



How can I help you today?


 | Message AI  

Part of NIEHS ToxPipe Platform | [Privacy policy](#) | [Terms of service](#)

ToxPipe Assistant

 **Agent Builder**

ToxPipe Assistant   **Select**



Name

ToxPipe Assistant
agent_zasCtDMXQHoshRii5p7r9

Description

Optional: Describe your Agent here

Instructions

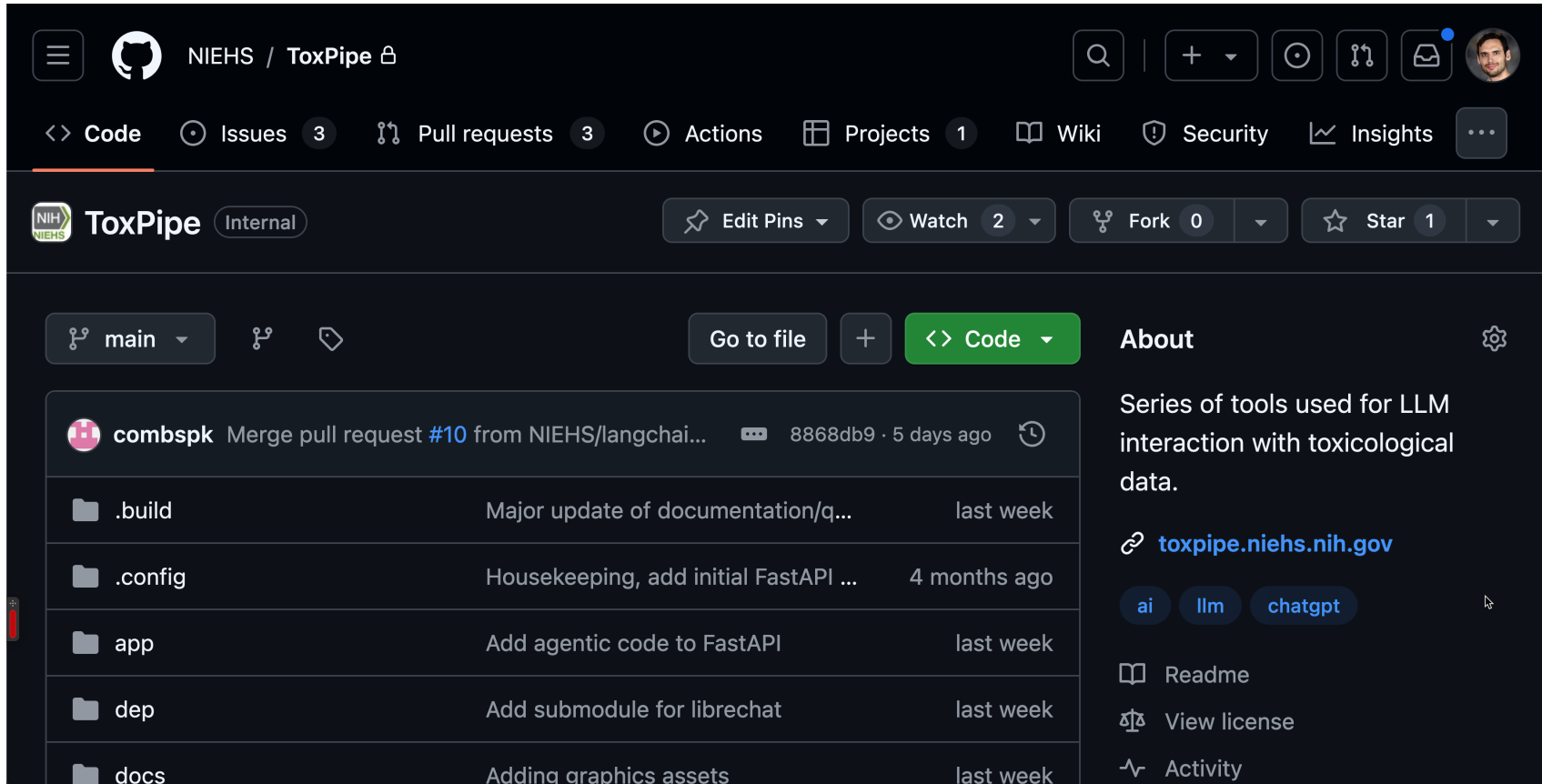
The system instructions that the agent uses

Model *

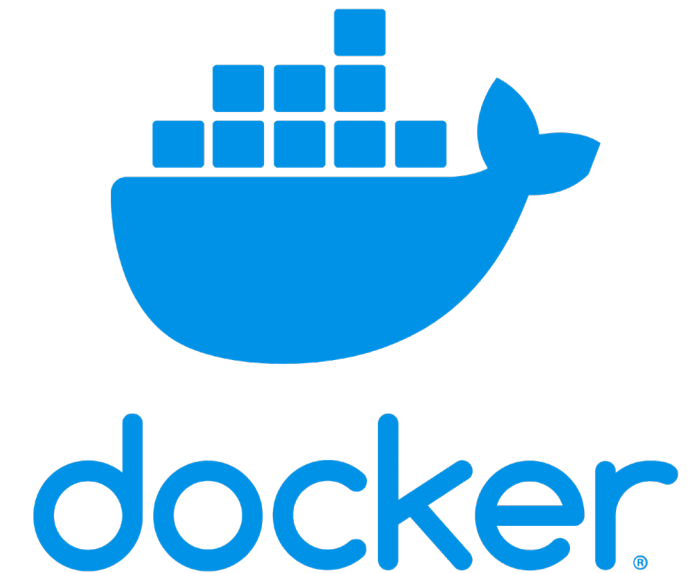
The screenshot shows the configuration page for a ToxPipe Agent. At the top, there is a share icon and the title 'ToxPipe Agent'. Below this, the 'Model' section is set to 'gpt-4o'. The 'Capabilities' section includes a 'File Search' subsection with an unchecked 'Enable File Search' checkbox and an 'Upload files' button. The 'Tools + Actions' section contains 'Add Tools' and 'Add Actions' buttons. At the bottom, there are icons for deleting, sharing, and saving, with a large green 'Save' button. A bottom navigation bar includes 'Prompts', 'Parameters', and 'Attach Files' options.

The screenshot shows the configuration page for an IATA Agent Builder. At the top, there is a share icon and the title 'IATA'. Below this is the 'Agent Builder' section. The main area is titled 'Add actions' with a subtitle: 'Let your Assistant retrieve information or take actions via API's'. The 'Authentication' section is set to 'none'. The 'Schema' section has a dropdown menu set to 'Examples' and a text area containing a Swagger JSON schema for an API endpoint: `"/author/batch"`. The schema details include a query parameter for a comma-separated list of fields to return, with a description explaining that the `authorId` field is always returned and that periods should be used for subfields.

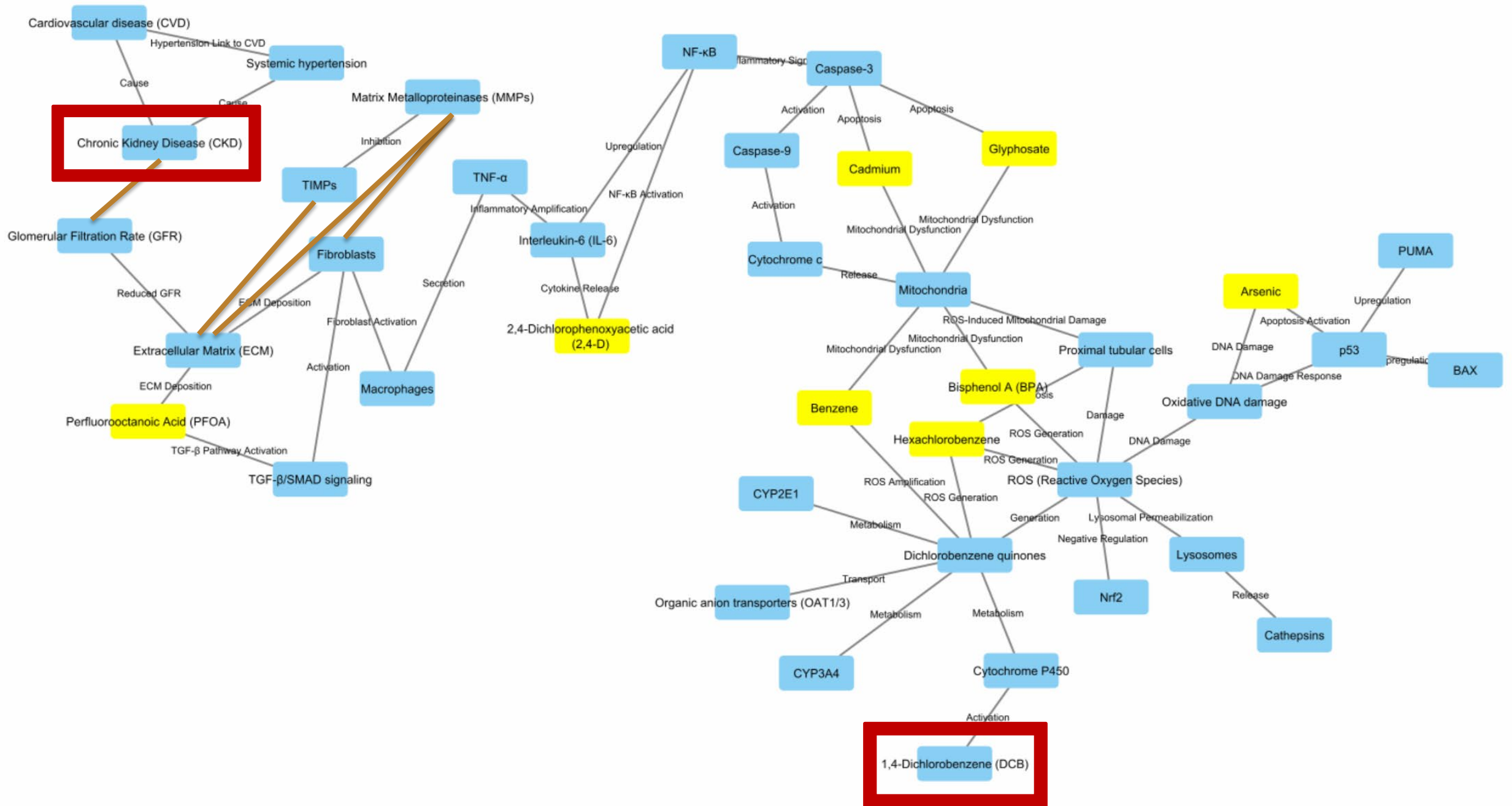
The screenshot displays the ToxPipe Assistant interface. At the top, the title bar reads "ToxPipe Assistant". Below it, a user profile for "Trey Saddler" is shown with the question "What files do you have access to?". A tool titled "Toxicology Study Parameter Extractor" is visible. A modal dialog box titled "Share 'ToxPipe Assistant'" is open in the center, featuring two toggle switches: "Share to all users" (which is turned on) and "Allow other users to edit your agent" (which is turned off). A green "Save" button is located at the bottom right of the dialog. The background interface includes a sidebar on the left with options like "Messages", "Files List", and "Access and Query". On the right, there are settings for "Model" (gpt-4o) and "Capabilities", along with buttons for "Add Tools" and "Add A". At the bottom of the interface, there are links for "Part of NIEHS ToxPipe Platform", "Privacy policy", and "Terms of service".



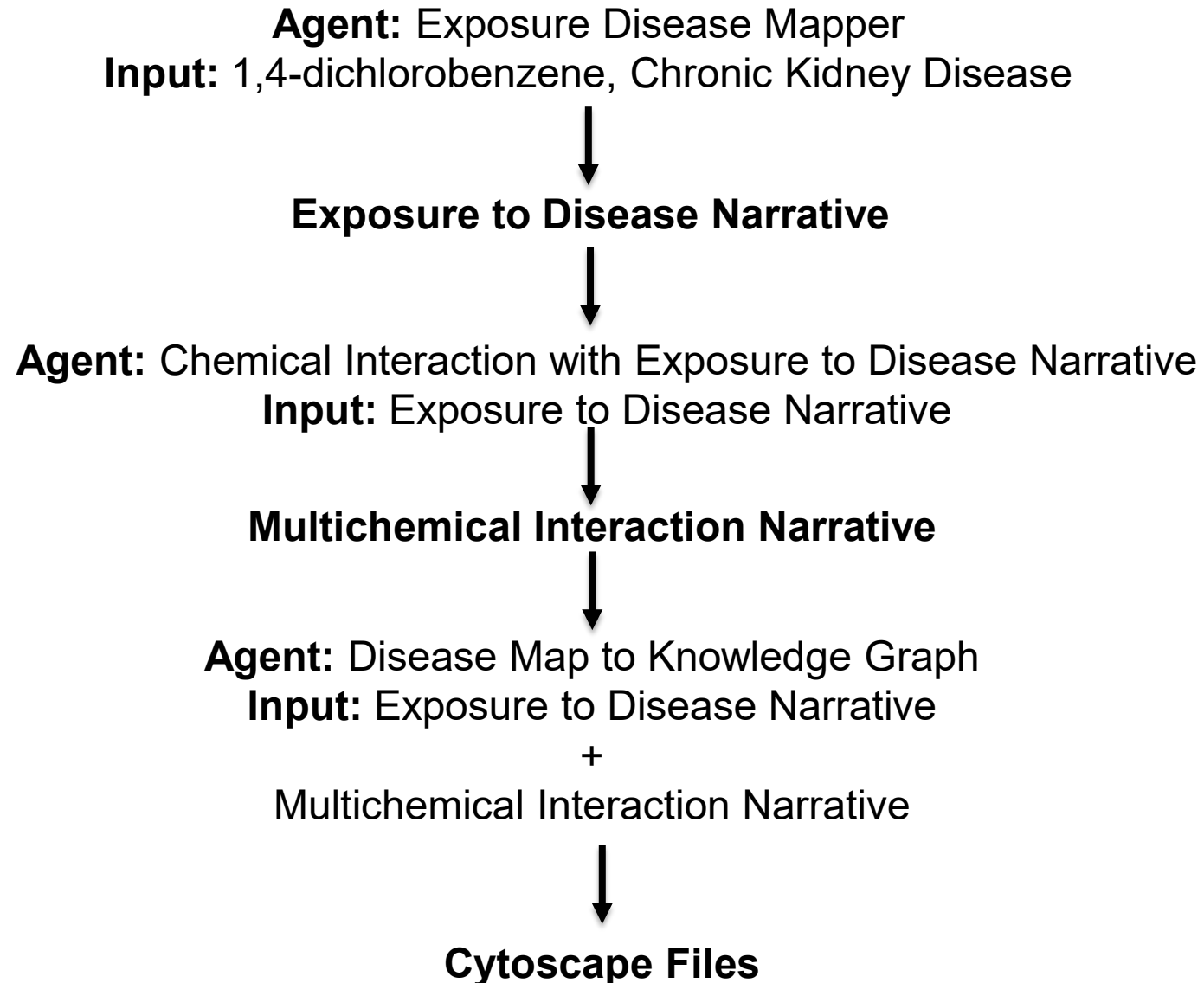
The screenshot shows the GitHub repository page for ToxPipe. At the top, the repository name "ToxPipe" is displayed with a label "Internal". Below this, there are statistics for "Edit Pins", "Watch" (2), "Fork" (0), and "Star" (1). The main content area shows a list of files and folders: ".build" (last week), ".config" (4 months ago), "app" (last week), "dep" (last week), and "docs" (last week). A pull request by "combspk" is also visible. On the right side, the "About" section describes ToxPipe as a "Series of tools used for LLM interaction with toxicological data." and provides a link to "toxpipe.niehs.nih.gov". There are also tags for "ai", "llm", and "chatgpt".



AOP to Knowledge Graph: Dichlorobenzene → Chronic Kidney Disease



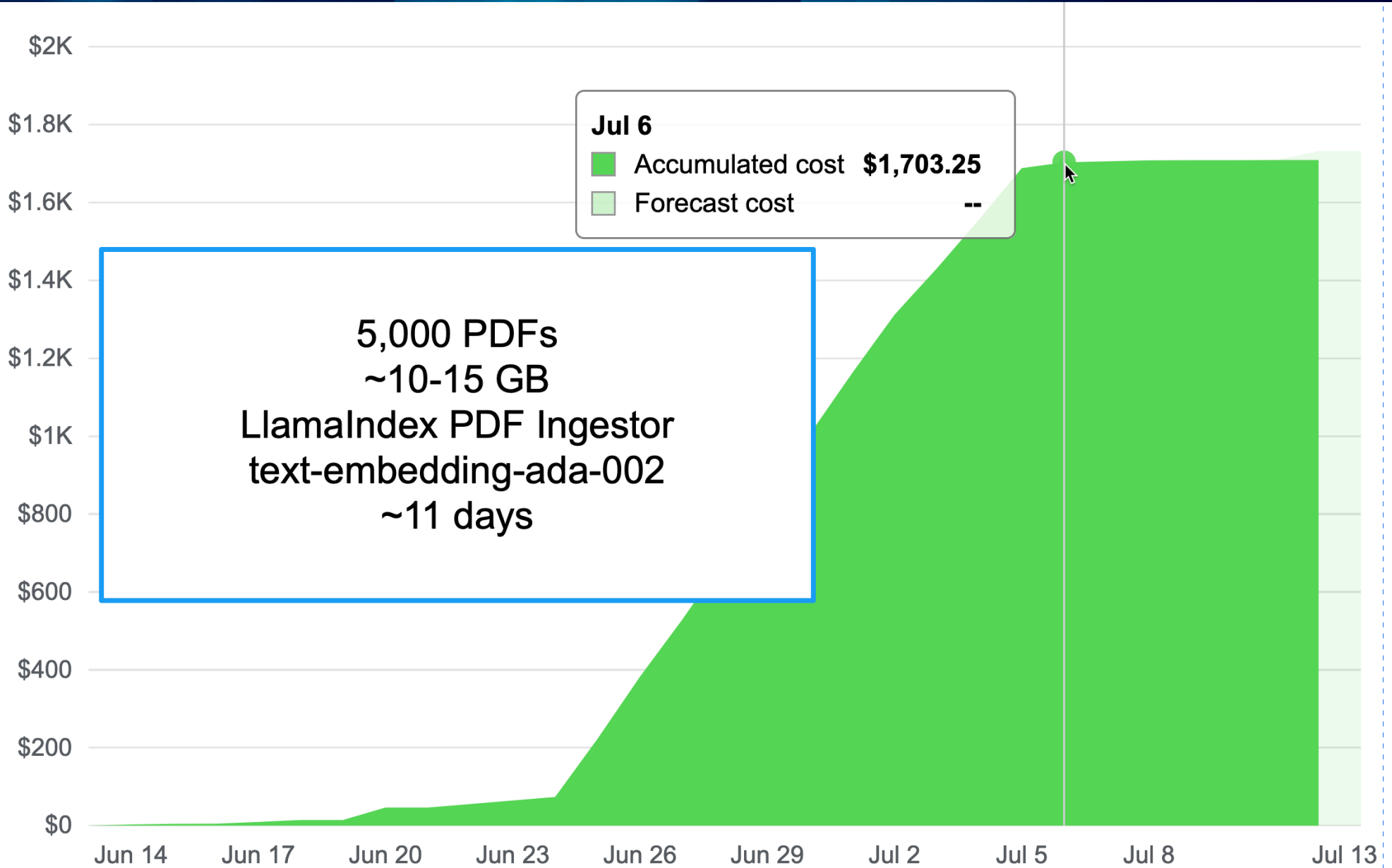
Knowledge Mapping and Mixture Interaction



PDF Information Extraction

- OCRmyPDF
 - Tesseract
- Marker
 - Surya
- PaddleOCR
- Cloud Options (\$\$\$)
 - Azure
 - Amazon Textract
 - Google Cloud Platform

Cost of embedding ~5k PDFs





National Institute of
Environmental Health Sciences
Division of Translational Toxicology

Questions?