



# Leveraging AI for Video Script Creation

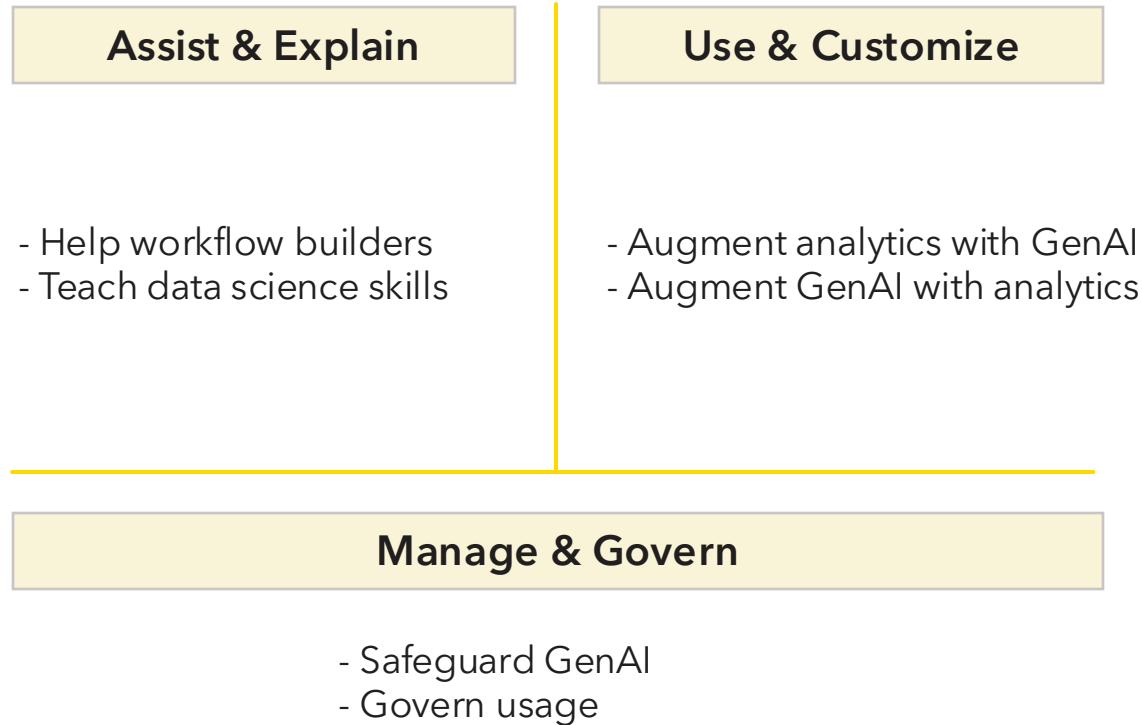
Corey Weisinger, Data Scientist @ KNIME  
September 26, 2024

# Agenda

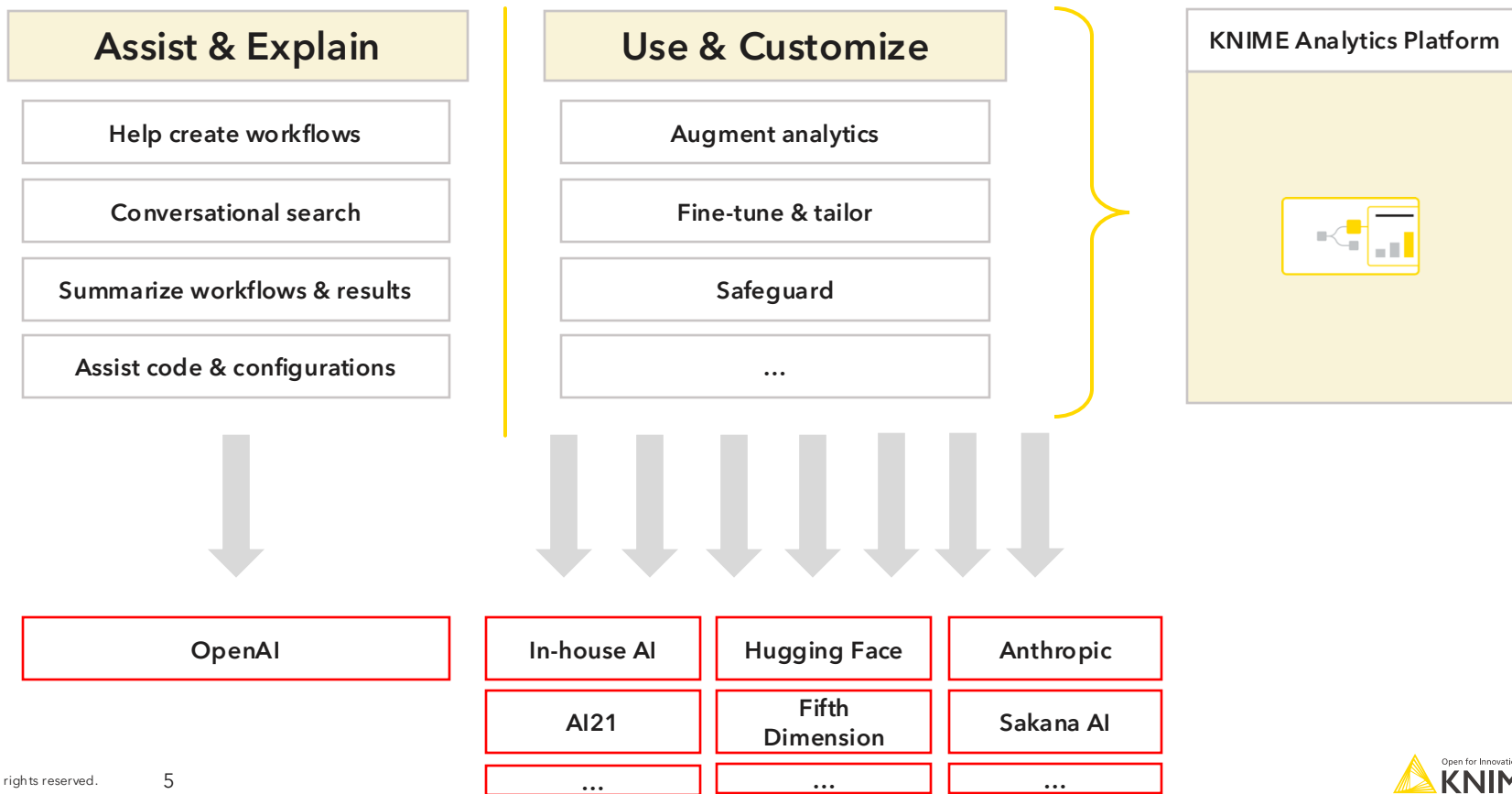
1. AI and KNIME
2. AI for video script generation?
3. What is a RAG Model?
4. Embedding Models and Vector Stores
5. Conversational Retrieval Agents
6. AI-powered script generation
7. Wrap Up

# AI and KNIME

# GenAI and Data Science



# GenAI for everybody: Flexibility



# Assisting and Explaining

## Assist & Explain

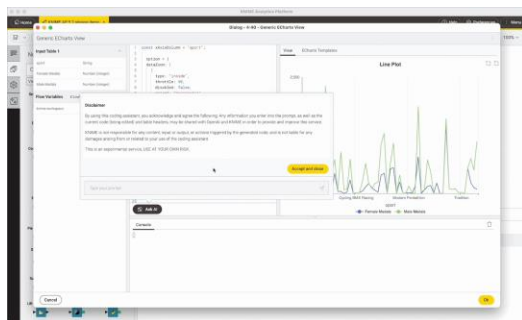
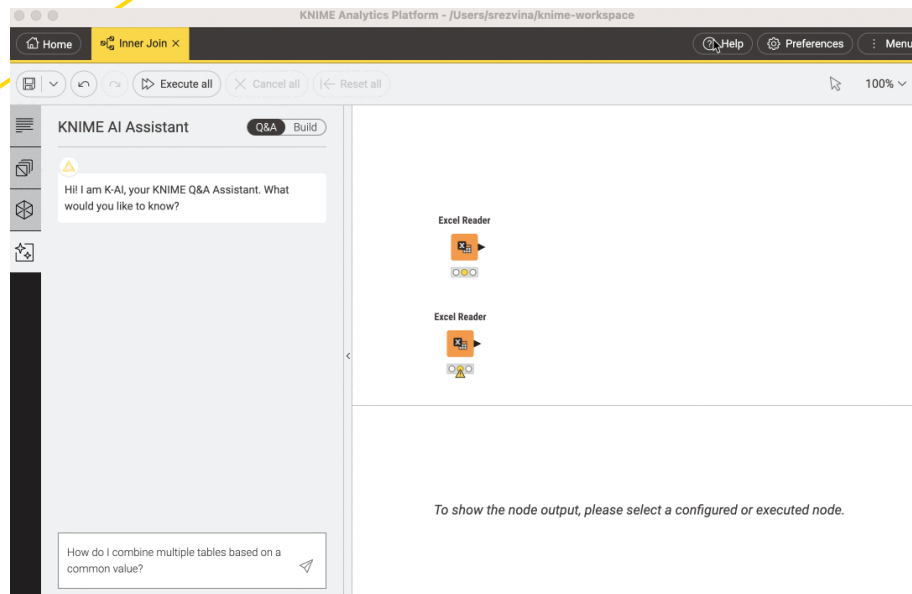
Help create workflows

Conversational search

Summarize workflows & results

Assist code & configurations

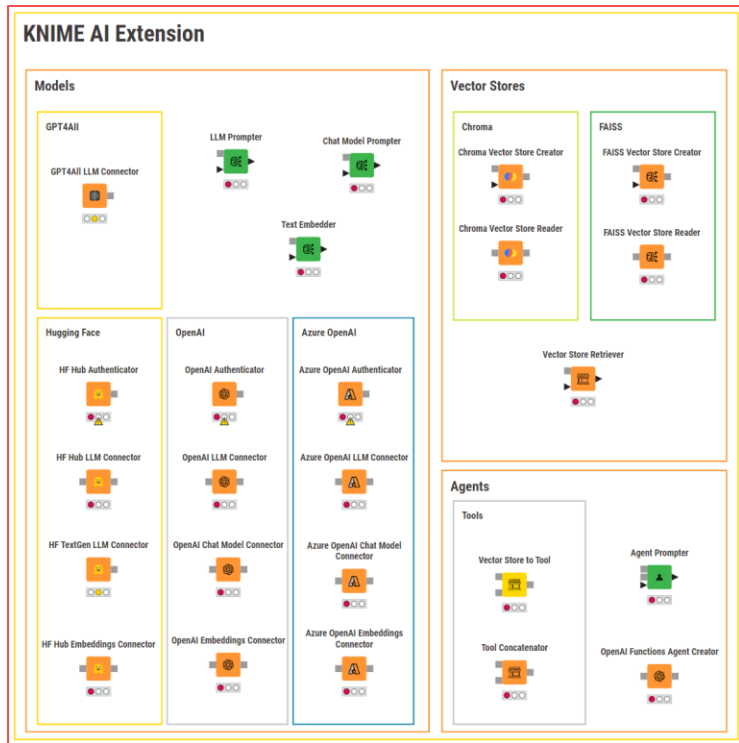
K-AI: Q&A and Build mode



Copilots for Python, eCharts  
[ tbdev: R, SQL, ... ]

# Using and Customizing

## GenAI Extensions



## Use & Customize

Augment Analytics

Fine-tune & Tailor

Safeguard

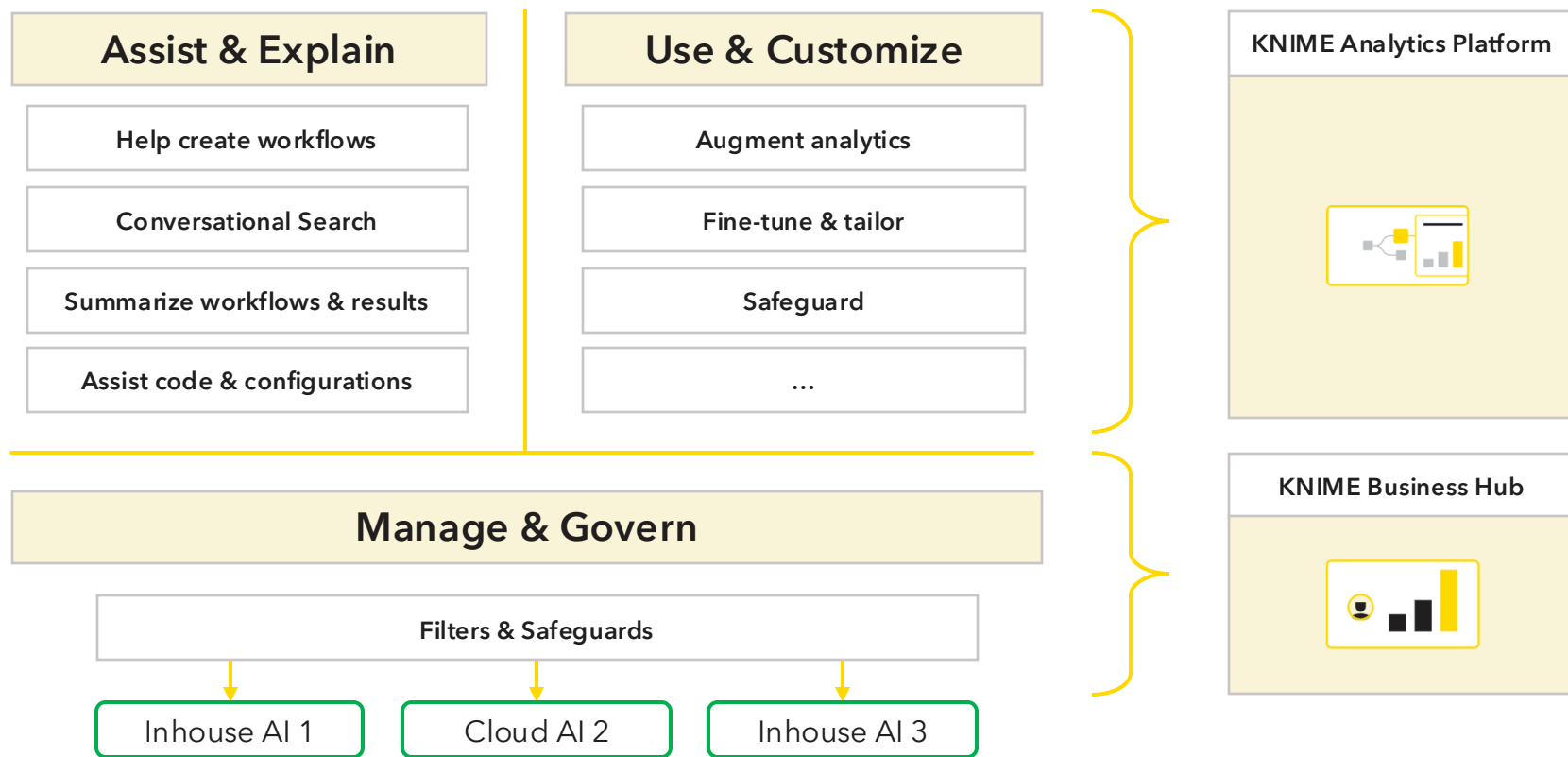
...

## AI Extension Example Workflows

Home

- 1) Large Language Models
- 2) Chat Model
- 3) Vector Stores
- 4) Agents

# GenAI in the Enterprise: Flexibility & Governance

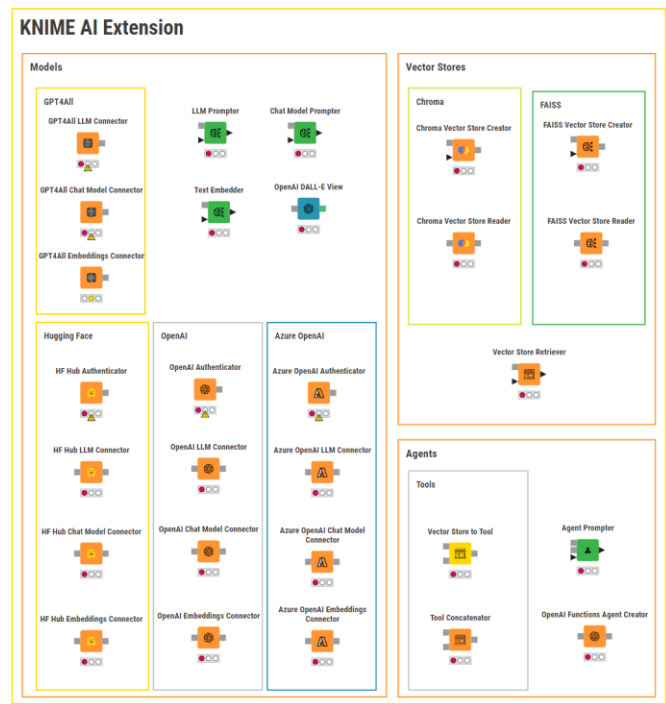




# KNIME AI Extension

LLMs can be leveraged with the [KNIME AI Extension](#) for connecting to open-source and closed-source models via API, or to open-source local models

- Authenticators
- LLM Connectors
- Chat Model Connectors
- Embeddings Connectors
- LLM and Embeddings Prompters
- Local Connectors and Embedders
- Vector Stores
- Model Fine-Tuner
- Agents



Abstract geometric lines in the top right corner of the slide, consisting of several overlapping triangles and polygons in a light yellow color.

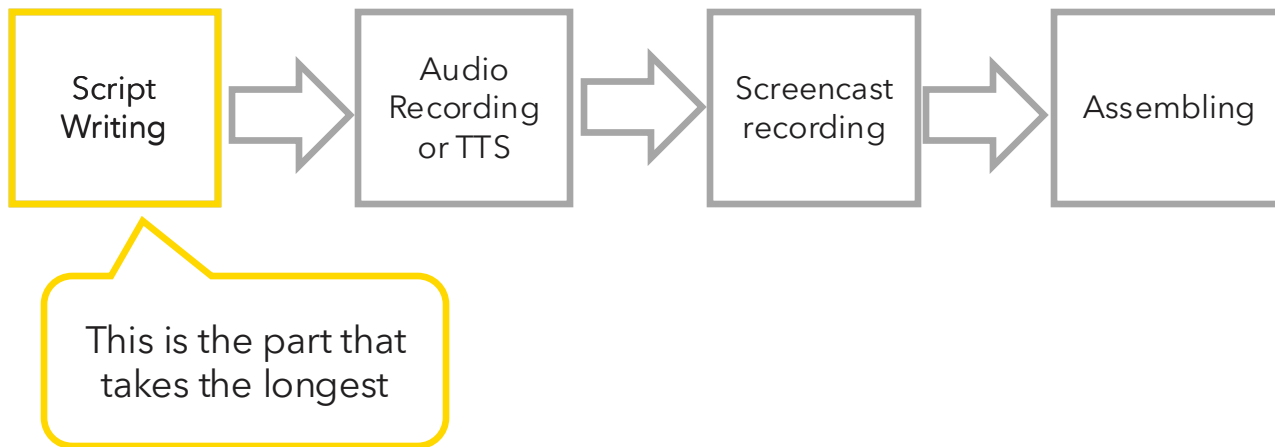
## **AI for video script generation?**

# The KNIME TV Channel on YouTube

The screenshot shows the KNIME TV YouTube channel page. At the top, there's a search bar with 'knime tv' entered. Below the search bar is the channel banner with the text 'Make sense of data, together' and 'Analyze. Upskill. Scale. No coding required.' The channel name 'KNIMETV' is displayed, along with subscriber and video counts. Below this, there's a 'For You' section featuring three video thumbnails: 'Data Science Pronto! Why is KNIME Open Source?', 'Real Time Fraud Detection: Luigi's Journey with KNIME', and 'What's New in KNIME Analytics Platform 5.3'. A 'Testimonials' section follows, showing two video thumbnails. Below that is the 'Intro to KNIME Analytics Platform Version 5' playlist, which includes videos like 'What is KNIME Analytics Platform?', 'What is a Node? What is a Workflow?', 'How to Install KNIME Analytics Platform', 'Tour of the User Interface on KNIME Analytics Platform', 'Build Your First Workflow with KNIME Analytics...', and 'Import and Export KNIME Workflows'. The 'KNIME Summits' section displays a row of video thumbnails for various summits, including 'KNIME Spring Summit 2023', 'KNIME Fall Summit 2022', 'KNIME Spring Data Talks 2022', 'KNIME Fall Data Talks 2021', 'KNIME Lab Data Talks 2021', and 'KNIME Spring Data Talks 2021'.

This screenshot shows the continuation of the KNIME TV YouTube channel page. It features the 'KNIME Summits' section, which includes a 'View all' link and a row of video thumbnails for summits from 2021 to 2023. Below this is the 'Data Science Pronto!' section, which has a 'Play all' link and a row of video thumbnails for various data science topics. The 'My Data Guest' section follows, featuring a 'Play all' link and a row of video thumbnails for guest appearances. Finally, the 'KNIME Webinars' section is shown at the bottom, with a 'Play all' link and a row of video thumbnails for various webinar topics.

# The video making process



# What is a RAG Model?

# Retrieval Augmented Generation

- **Retrieval augmented generation** (RAG) is an AI framework that enhances the generation of human-like responses, reducing the likelihood of generating inaccurate or misleading responses.
- **The Goal:**
  - Make LLMs more knowledgeable.
  - Make their response more relevant to the user/application.
  - Mitigating risks of hallucinations, biases or non-factual information.
- **How?**
  - Giving generalist LLMs access to **user-curated** and **domain-relevant knowledge bases** (e.g., data sources with specific knowledge, terminology, context or up-to-date information) to customize responses for specific applications.

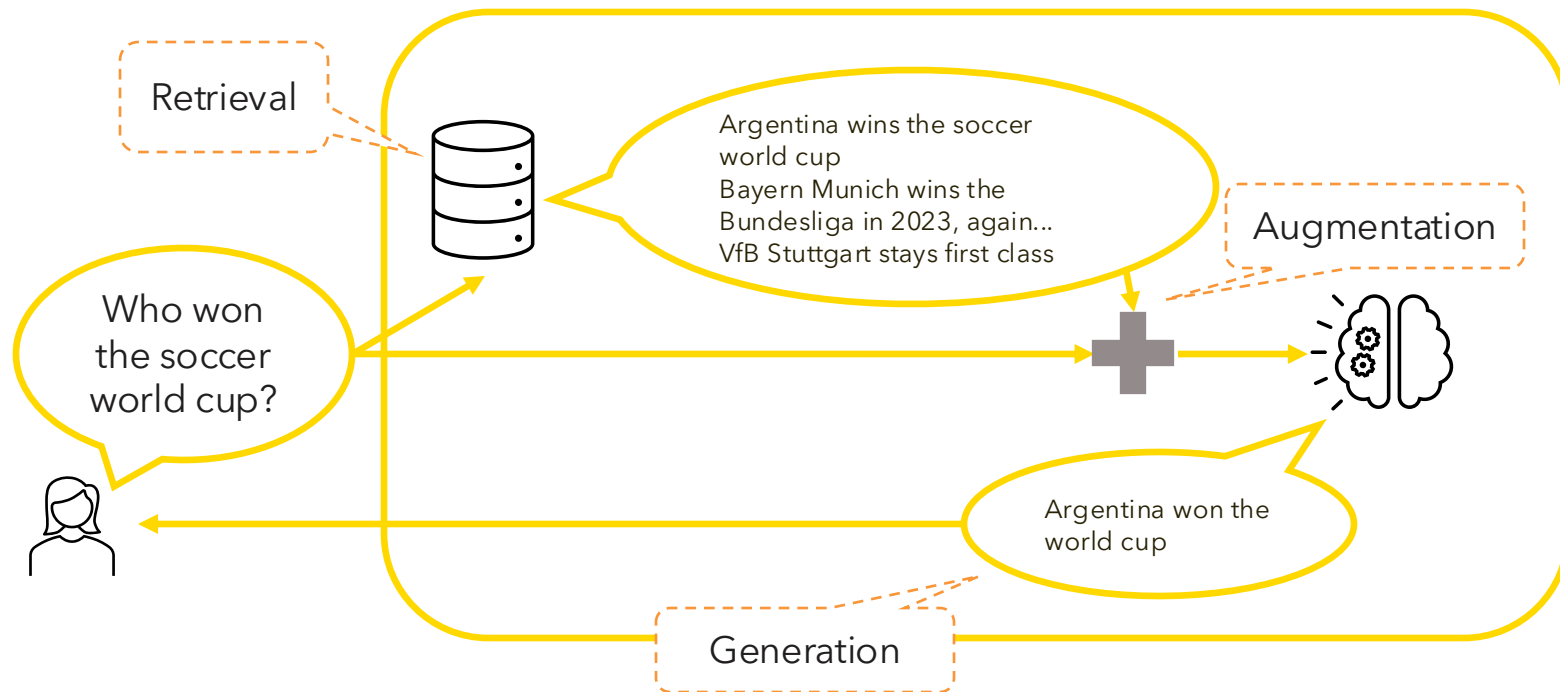
RAG is LLMs speaking  
after thinking

# Retrieval Augmented Generation

The RAG process involves three steps:

- **Retrieval.** Retrieve relevant information from a knowledge base.
- **Augmentation.** Augment the user prompt with the retrieved information. This enhances the model's understanding by providing additional context from the retrieved sources.
- **Generation.** Generate a more informed and contextually rich response based on the augmented input, leveraging the generative power of the model.

# The RAG process visualized





Abstract geometric lines in the top right corner of the slide, consisting of several overlapping triangles and polygons in a light yellow color.

# Embedding Models and Vector Stores

# What is an Embeddings model?

The Oscar for best  
actress goes to ...



Embeddings  
Model



5	7	-2	3
---	---	----	---

# What is an Embeddings model?

The Oscar for best actress goes to ...



Embeddings Model



5	7	-2	3
---	---	----	---

The Oscar for best actress goes to ...



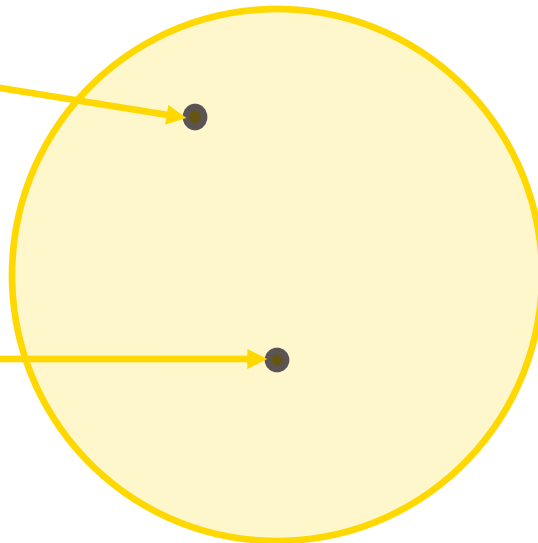
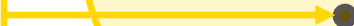
Embeddings Model



Argentina wins the Soccer world cup

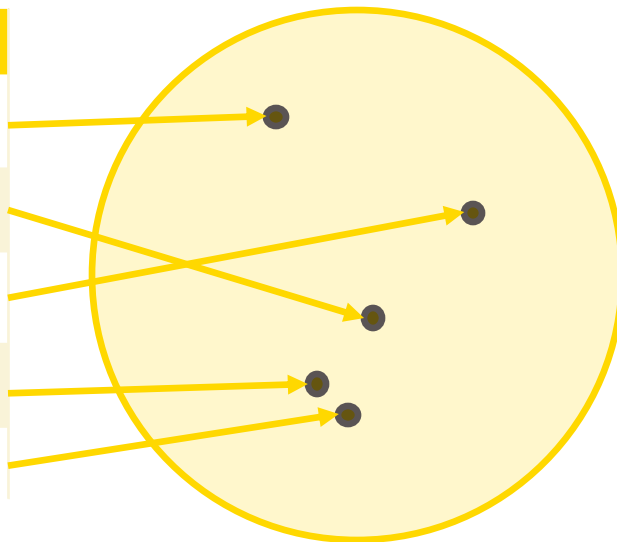


Embeddings Model



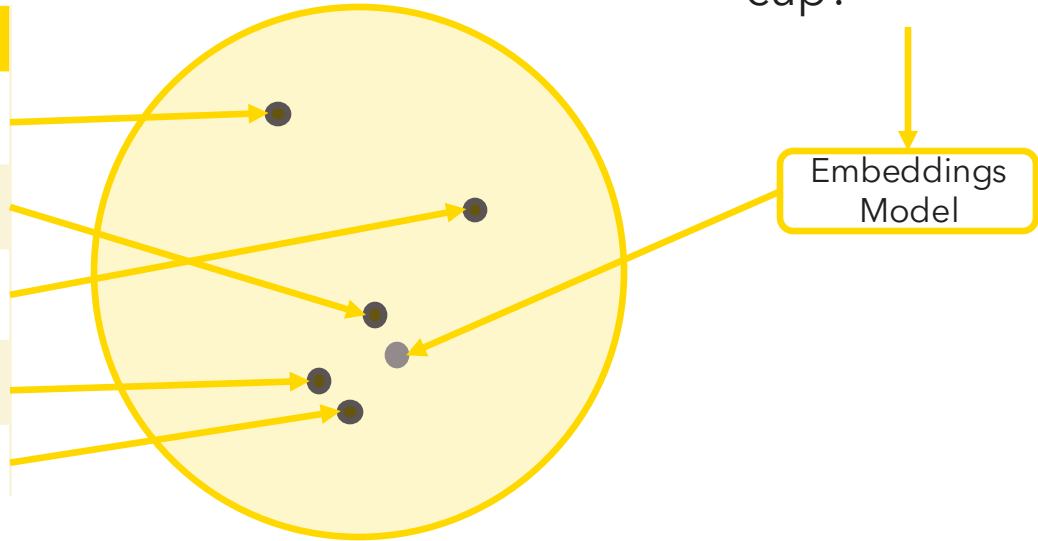
# What is a Vector Store?

Document	Embedding
The 2023 Oscar for best actress goes to ...	5, 7, -2, 3
Argentina wins the soccer world cup	4, 6, -2, 4
The Kansas City Chiefs win the superbowl in 2023	4, 6, 5, -1
Bayern Munich wins the Bundesliga in 2023, again...	0, 4, 8, 9
VfB Stuttgart stays first class	0, 1, 9, 8



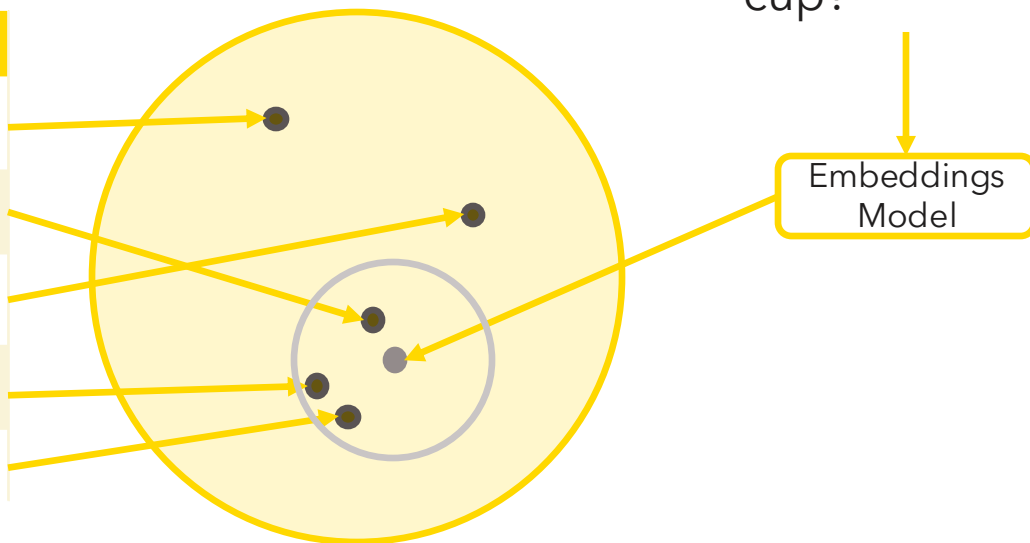
# How can we use it for Semantic Search?

Document	Embedding
The 2023 Oscar for best actress goes to ...	5, 7, -2, 3
Argentina wins the soccer world cup	4, 6, -2, 4
The Kansas City Chiefs win the superbowl in 2023	4, 6, 5, -1
Bayern Munich wins the Bundesliga in 2023, again...	0, 4, 8, 9
VfB Stuttgart stays first class	0, 1, 9, 8

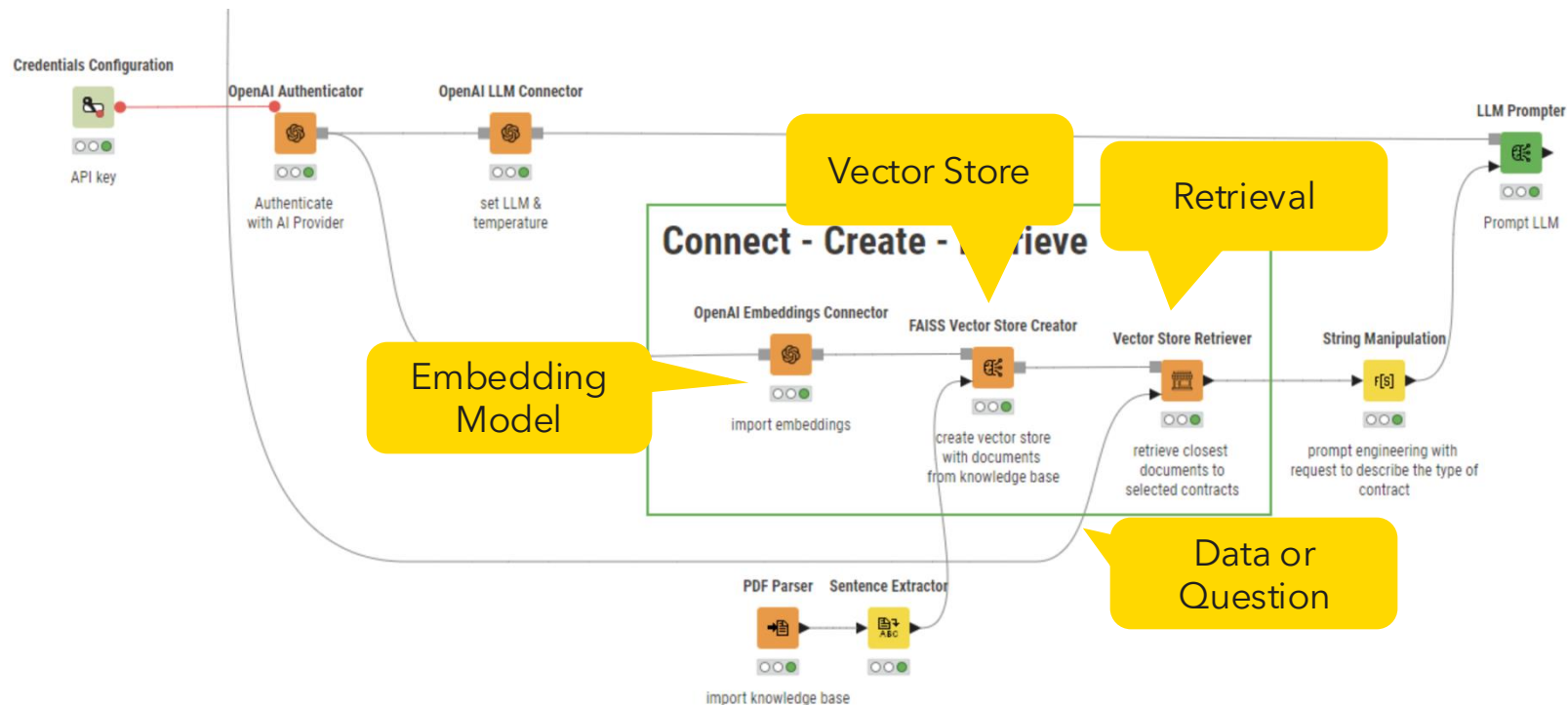


# How can we use it for Semantic Search?

Document	Embedding
The 2023 Oscar for best actress goes to ...	5, 7, -2, 3
Argentina wins the soccer world cup	4, 6, -2, 4
The Kansas City Chiefs win the superbowl in 2023	4, 6, 5, -1
Bayern Munich wins the Bundesliga in 2023, again...	0, 4, 8, 9
VfB Stuttgart stays first class	0, 1, 9, 8



# RAG: Simple Prompting + Connect - Create - Retrieve



Abstract geometric lines in the top right corner of the slide, consisting of several overlapping, thin, light-yellow lines forming a complex, star-like pattern.

# Conversational Retrieval Agents



# RAG is cool but is not *intelligent*

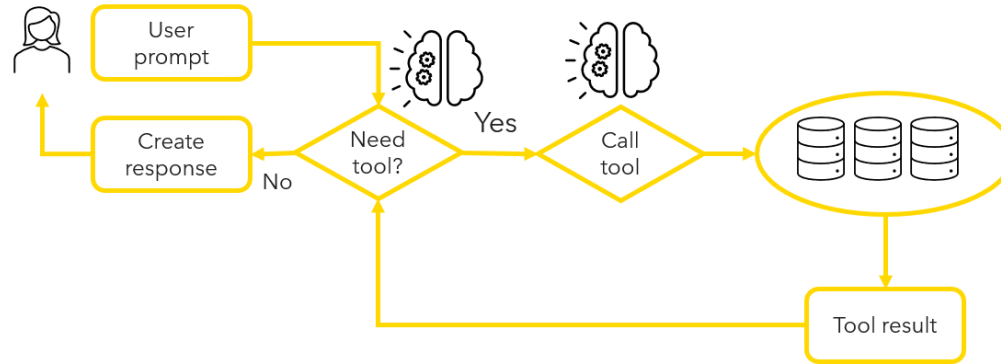
- RAG allows us to obtain domain-specific answers but it's not smart about which tools or vector stores, if any, to use.
- It requires a hard-coded behavior to obtain the answer we expect
- Can we do better? Can we have a truly *intelligent* system that can interact with us and choose smartly where to get answers to our questions?

# Conversational Retrieval Agents

- Conversational Retrieval Agents enhance the LLM with the ability to chat and use tools to answer specific questions.
- By "agents", we mean a system where the **sequence of steps or reasoning behavior is not hard-coded**, fixed or known ahead of time, but is rather determined by a language model.
- Agents rely on the conversational capabilities of generalist LLMs but are also endowed with a suite of specialized tools (usually one or more vector stores).

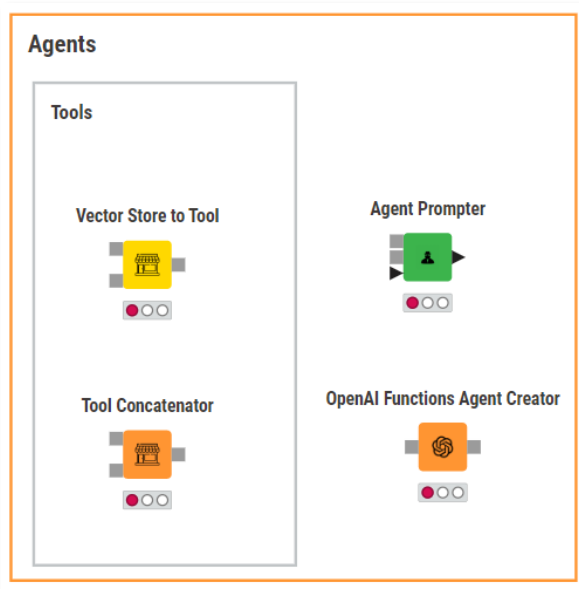
# Conversational Retrieval Agents

- Depending on the user's prompt and hyperparameters, **the agent understands which, if any, of the tools to employ** to best provide an improved response.
- Agents can be instructed to perform specific functions or roles in a certain way.
  - For example, an agent can be prompted to write a political text as if it was a poet of the Renaissance or a soccer commentator.



# Conversational Retrieval Agents in KNIME

- The KNIME AI Extension allows the creation of agents!
- Currently, only possible with OpenAI or Azure Open AI



# Conversational Retrieval Agents in KNIME

- Decide how the agent should behave
  - Act as a...
  - How verbose answers should be.
  - Force it to always use the tool first vs. let it decide.
- Give it the tools (=vector stores)
  - Provide a tool name and description.
  - Decide how many doc should be retrieved.
- Prompt the agent!
  - Provide function and tools.
  - Provide conversation history.
  - Provide user prompt.

OpenAI Functions Agent Creator



Vector Store to Tool



Agent Prompter



OpenAI Functions Agent Creator



single document

Vector Store to Tool



Turn the vector store into a tool



Conversation history

Agent Prompter



Dialog - 4:908 - Agent Prompter

**Conversation Settings**

Message role  
No value selected

Messages  
No value selected

**Prompt Settings**

Message

[Show advanced settings](#)

Cancel Ok

Abstract geometric lines in the top right corner of the slide, consisting of several overlapping triangles and polygons in a light yellow color.

# **AI-powered video script generation**

# Building the RAG powered chatbot

1. Collect data representing the knowledge base
2. Chunk data to a reasonable size. Is the entire blog or document relevant to each question or would sections or paragraphs be more appropriate?
3. Store the data using a vector embedding
4. Use prompt engineering best practices to establish desired outputs

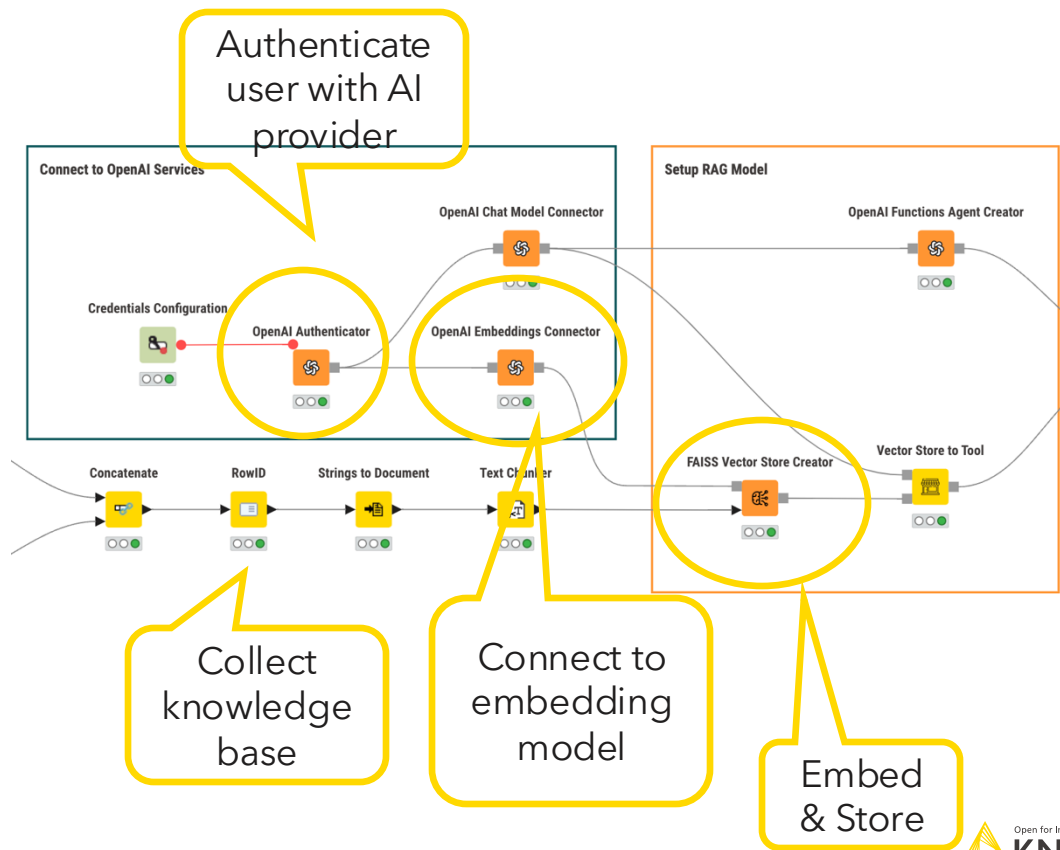
# Data Collection and RAG

## ■ Data Collection

- Blogs
- Documentation pages
- Node descriptions
- Old video scripts
- Course content

## ■ Embedding

- OpenAI text embedding model
- Create vector store





# Chatbot Creation

## ■ Connect

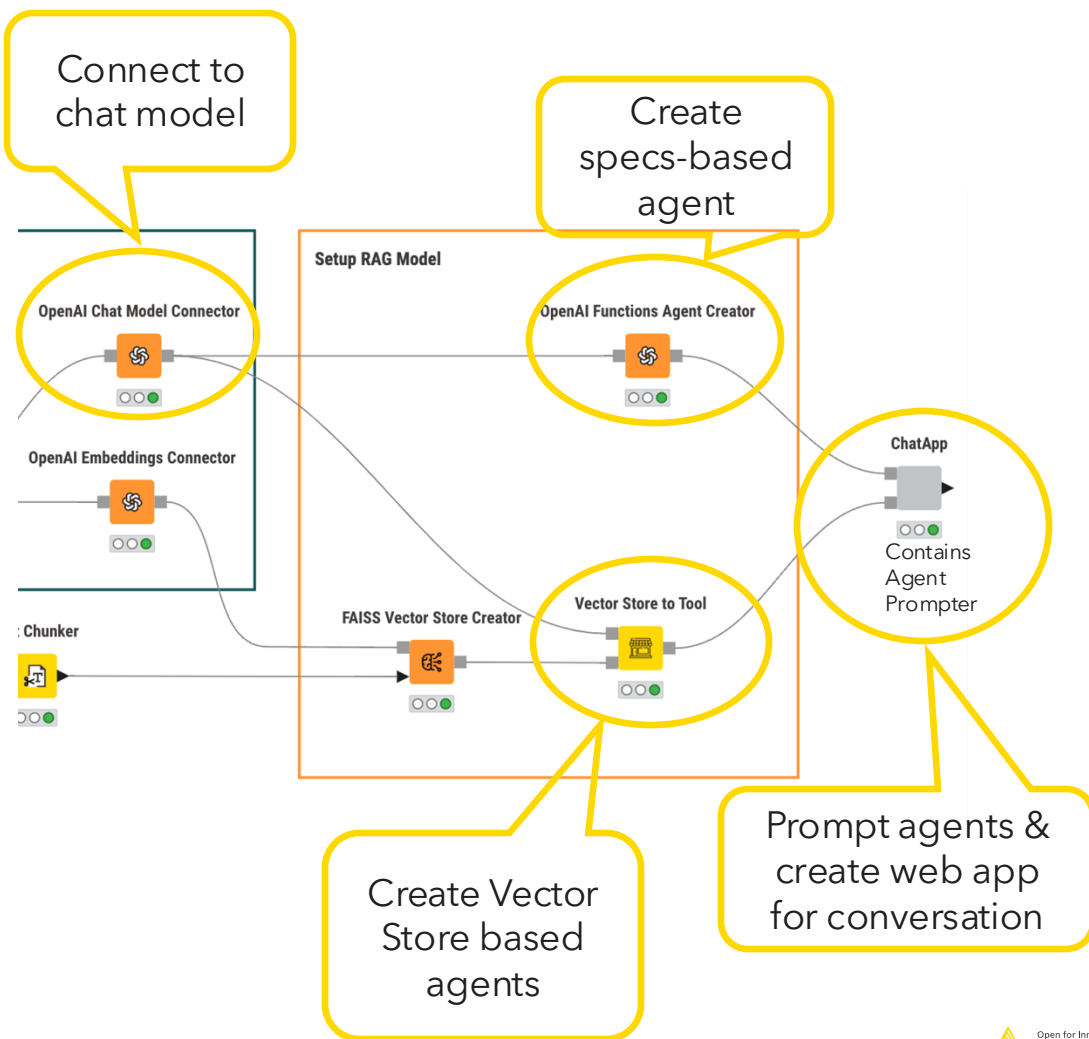
- To chat model
- To Vector Store

## ■ Create

- Specs based agents
- Vector store based agents

## ■ Prompt

- Agents



# Prompt Engineering: The Art of Communication

## ■ Clarity

- You are an experienced Data Scientists and writer...
- You are writing video scripts for social media...
- The video scripts should be about 1 minute long...

## ■ How?

- Either directly in the model prompt or in a system message that is always referenced by the LLM

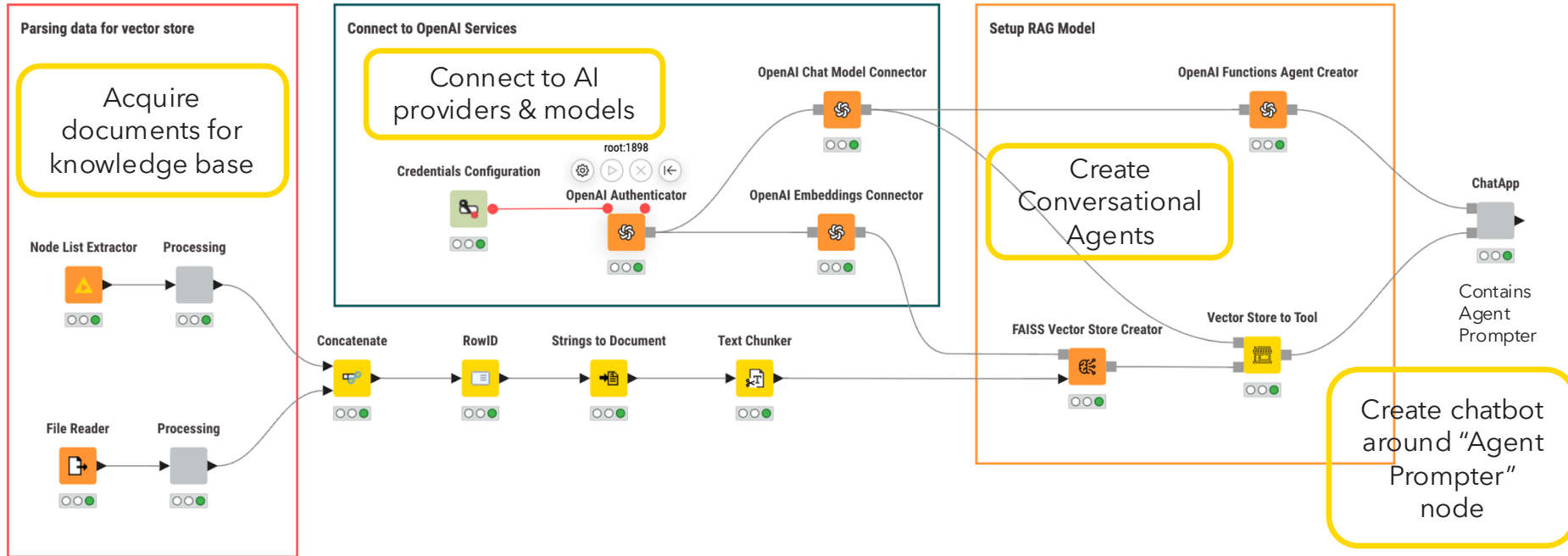
### System message

[Instructions] (You are a computer model who is an expert in video direction for data science. You will write a script based on the topic or question provided by the user.);

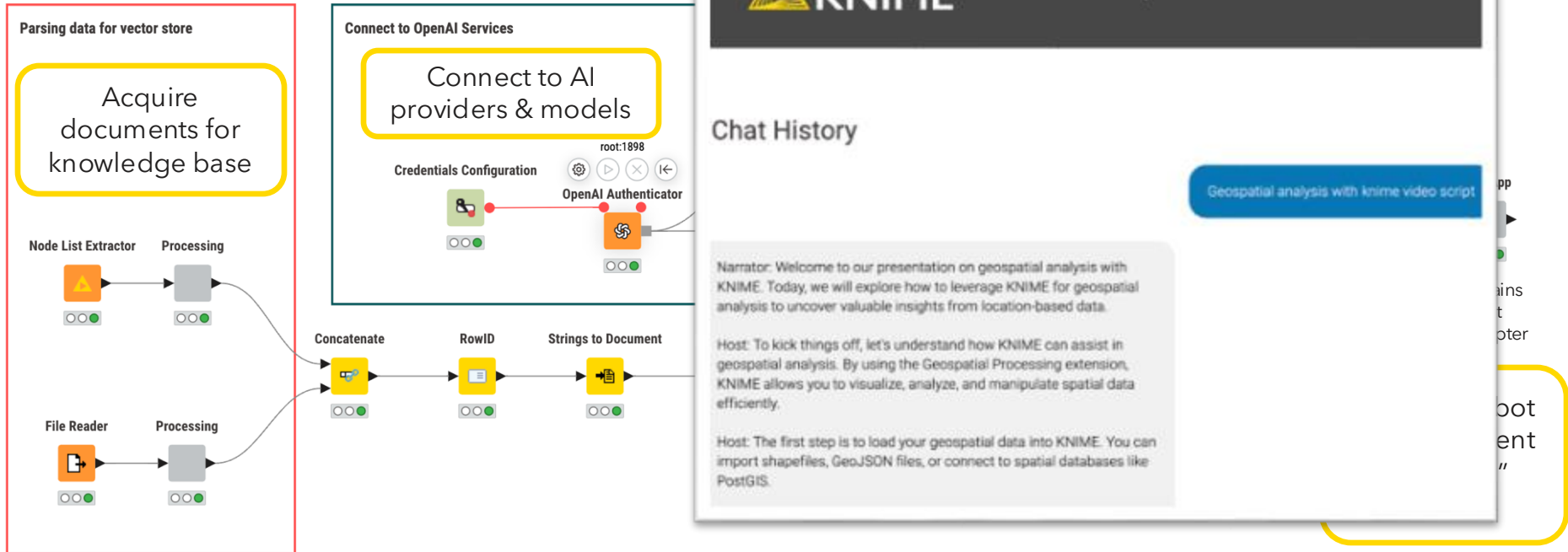
[Script Instructions](The script should be engaging, concise, and fit within a 1 minute and 30 seconds time frame. the content of the script topic should be derived from KNIME\_expert, explain the steps as to how to use that concept, provide a short example, KEEP THE TONE INFORMATIVE and avoid lengthy introductions.

Format it so that every time a person speaks, it is preceded by their name. For example: Narrator: Welcome to our presentation.)

# From Theory to Practice: Let's see it in action



# From Theory to Practice: Let's see it in action



# The Result: AI-Generated Script with TTS

Hello and welcome! Today, we're diving into the fascinating world of geospatial analytics, a powerful tool for organizations worldwide. Just imagine being able to analyze and visualize location-based data to uncover valuable insights and make informed decisions. At the recent KNIME Fall Summit 2022, Prof. Wendy Guan from Harvard's CGA team showcased the importance of geospatial data and the challenges of accessing and analyzing it. To address these challenges, the CGA team and KNIME...



**Wrap up**



# What we talked about

- AI Integration in KNIME Analytics Platform:
- Vector Stores:
- RAG (Retrieval Augmented Generation):
- AI Agents:
- Building a Chatbot:
  
- Readings on KNIME Blog:
  - A beginner's guide to LLM-based solutions → <https://www.knime.com/blog/guide-to-build-your-own-LLM-solutions>
  - What are AI hallucinations & how to prevent them → <https://www.knime.com/blog/ai-hallucinations>
  - Mitigate hallucinations with RAG in KNIME → <https://www.knime.com/blog/mitigate-hallucinations-in-LLMs-with-RAG>
  - "How to build a custom AI powered job finder chatbot" → <https://www.knime.com/blog/how-to-build-custom-ai-powered-chatbot>

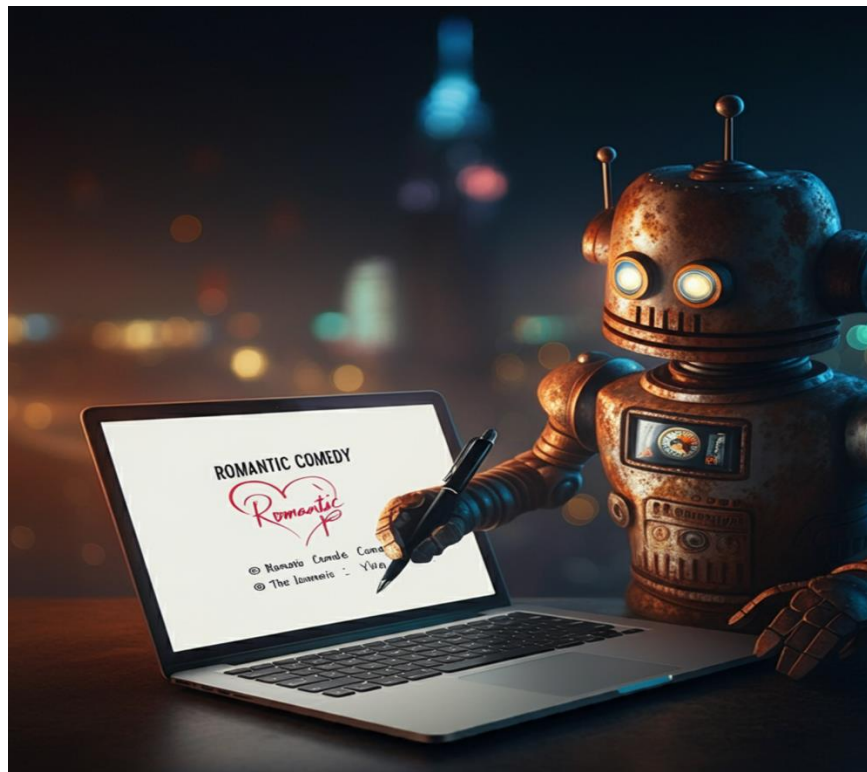
# What's next?

## ■ Existing Processes

- Use it to check for tone in new writings such as blogs or technical documentations
- Automatically create potential outlines for blog topics

## ■ New Processes

- Expand support for more languages by using translation models with software and industry "knowledge"



*Created by Imagen3*



# Stay Connected with KNIME



## **Blog:**

[knime.com/blog](https://knime.com/blog)



## **KNIME Self-Paced Courses:**

[knime.com/knime-self-paced-courses](https://knime.com/knime-self-paced-courses)



## **Forum:**

[forum.knime.com](https://forum.knime.com)



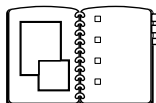
## **Email:**

[education@knime.com](mailto:education@knime.com)



## **KNIME Hub:**

[hub.knime.com](https://hub.knime.com)



## **Medium Journal:**

[medium.com/low-code-for-advanced-data-science](https://medium.com/low-code-for-advanced-data-science)

Follow us on  
social media:





## Questions?

KNIME Hub

[hub.knime.com/corey](https://hub.knime.com/corey)



[Corey.Weisinger@knime.com](mailto:Corey.Weisinger@knime.com)



[Linkedin.com/in/corey-weisinger](https://www.linkedin.com/in/corey-weisinger)