



**watsonx.ai™**



# Implementing your own ChatGPT : developing a RAG solution with IBM Watsonx.ai

**Jarlath Quinn & John McConnell**

[www.sv-europe.com](http://www.sv-europe.com)

A SELECT INTERNATIONAL COMPANY



**watsonx.ai™**



Just waiting for all attendees to join...

# Implementing your own ChatGPT : developing a RAG solution with IBM Watsonx.ai

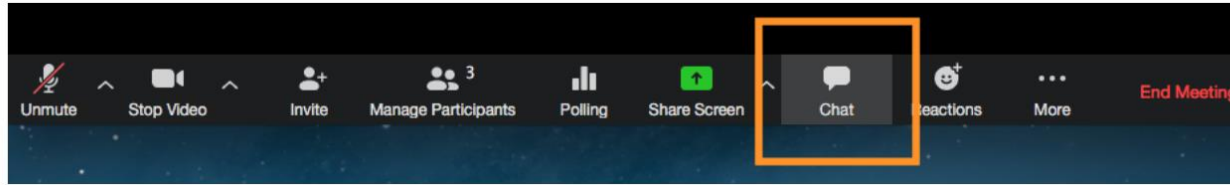
**Jarlath Quinn & John McConnell**

[www.sv-europe.com](http://www.sv-europe.com)

A SELECT INTERNATIONAL COMPANY

# FAQ's

- Is this session being recorded? Yes
- Can I get a copy of the slides? Yes, we'll email links to download materials after the session has ended.
- Can we arrange a re-run for colleagues? Yes, just ask us.
- How can I ask questions? All lines are muted so please use the chat panel – if we run out of time we will follow up with you.





- Gold accredited partner to IBM, Predictive Solutions and DataRobot specialising in statistics, advanced analytics & AI technologies
- Work with open-source technologies (R, Python, Spark etc.)
- Team each has 15 to 30 years of experience working in the advanced analytics industry

- Deep experience of applied data science applications across sectors
  - Retail
  - Healthcare/Pharma
  - Finance/Insurance
  - Media/Telecoms
  - Utilities
  - FMCG
  - Charity/Housing/Government



# Agenda

- What is a RAG solution?
- The background to ChatSPSS
- Key elements of a RAG solution
- Working with Wastonx.ai – Prompt Lab
- Working with Wastonx.ai – Auto AI RAG
- Deploying RAG models



# What is a RAG solution?

# What is a RAG Solution?



**Retrieval-Augmented Generation (RAG):** Combines generative AI with a retrieval system, allowing a model to pull relevant facts and context from large databases or document collections before generating an answer



It helps to reduce hallucinations by grounding AI responses in up-to-date, verifiable data, even across vast unstructured text sources.



It makes it easy to query huge collections of data (like reports, logs, or research papers) with a single prompt, improving efficiency, accuracy, and transparency in a secure setting

# Why Build Your Own RAG?

- Data/user control & compliance
- Quality control that can be customised and measured
- Swap models and change parameters easily
- Ensure relevance, versioning and use of the most up-to-date data
- Integration & automation with your own systems

Accurate    Compliant    Current    Controlled



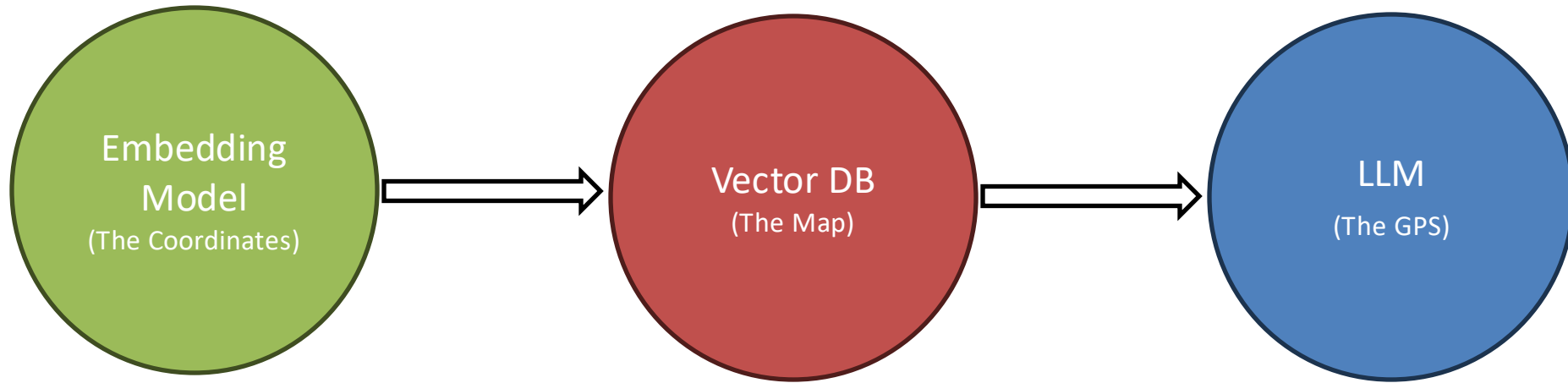
# Real world applications of RAG solutions

- **Customer Support:** Powers chatbots that provide accurate, context-rich responses by retrieving information from FAQs, product guides, and support tickets.
- **Healthcare & Life Sciences:** Enables clinicians and researchers to query medical literature, patient records, data without manually sifting through unstructured text.
- **Legal & Compliance:** Helps lawyers or compliance officers find relevant case law, regulations, or contractual clauses efficiently.
- **Financial Services:** Retrieves and summarizes relevant regulations, policies, or market reports to assist analysts and advisors.
- **E-commerce & Retail:** Enhances product recommendation systems or answer customer questions using information drawn from product catalogues and reviews.
- **Government & Policy:** Allows policymakers and analysts to extract relevant insights from large archives of policy documents, laws, or public records.

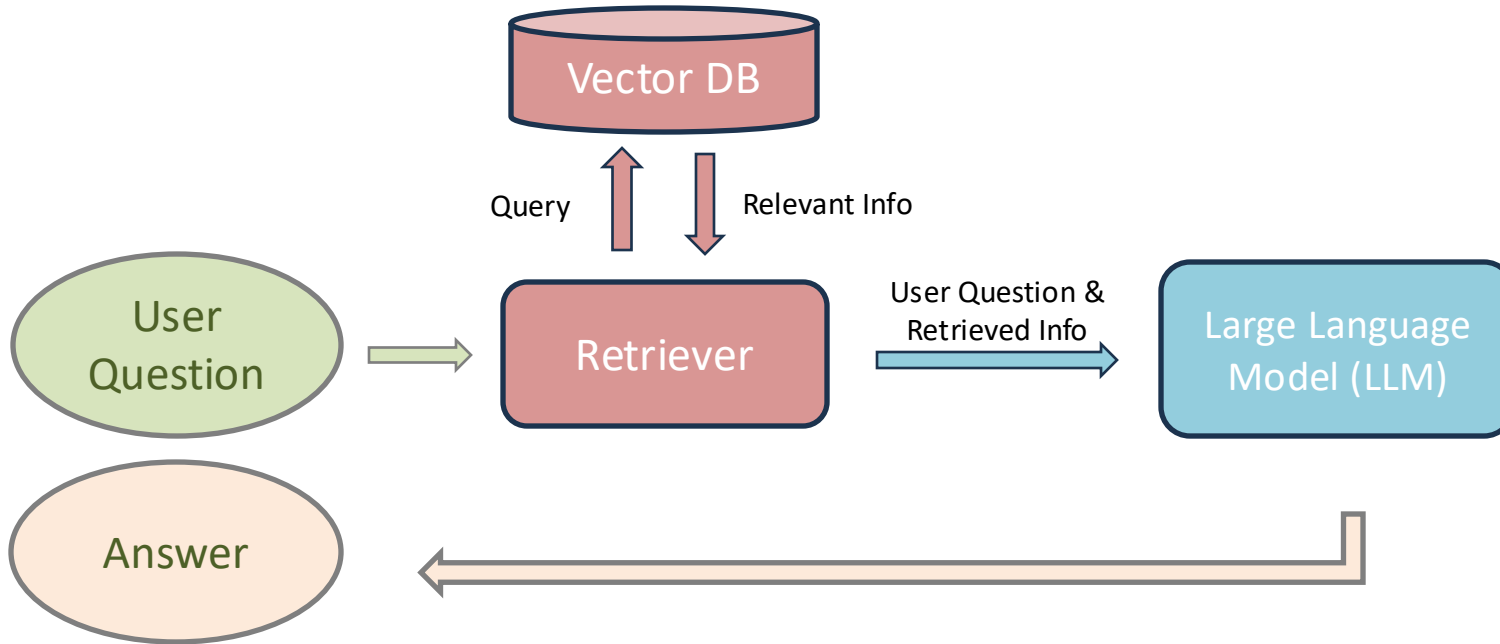
# Key elements of a RAG solution

- **Embedding Model** - Instead of storing raw text, the embedding model converts documents, paragraphs, or sentences into vectors (numerical representations).
- **Vector Database** - After an embedding model converts documents into vectors, these are stored in the vector database instead of plain text. You can think of a vector as a GPS coordinate in meaning-space — instead of longitude/latitude, it's hundreds of numbers that describe the semantic position of a text fragment
- **LLM (Large Language Model):**
  - The LLM takes the query and the retrieved documents as input
  - The retriever passes a set of relevant text snippets (i.e. the context) along with the user's question to the LLM
  - The LLM then processes both the query and the retrieved passages, integrating them into its reasoning and generates a natural language response.

# Key elements of a RAG solution (and Generative AI)

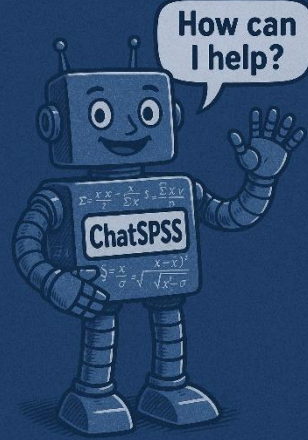


# Querying a Simple RAG solution





watson $x$ .ai™























## The background to ChatSPSS

# The background to ChatSPSS

- Opportunity to learn more about Watsonx.ai and its RAG capabilities
- Picked a test area where we already had a lot of subject matter expertise and available data, to act as an example and demonstrator
- Decided that if it we were successful, we would deploy it to the wider community of SPSS users

# Project data: IBM SPSS Statistics User Guides

- Based on a collection of 20 pdf documents covering SPSS Statistics Base plus assorted add-on modules and admin docs.

 IBM_SPSS_Statistics_Brief_Guide.pdf	26/03/2025 10:47	Chrome HTML Do...	5,609 KB
 IBM_SPSS_Custom_Tables.pdf	26/03/2025 10:42	Chrome HTML Do...	2,402 KB
 IBM_SPSS_Statistics_Core_System_User_Guide.pdf	26/03/2025 10:49	Chrome HTML Do...	2,237 KB
 IBM_SPSS_Exact_Tests.pdf	26/03/2025 11:09	Chrome HTML Do...	1,898 KB
 IBM_SPSS_Statistics_Base.pdf	26/03/2025 10:40	Chrome HTML Do...	1,649 KB
 IBM_SPSS_Advanced_Statistics.pdf	26/03/2025 10:41	Chrome HTML Do...	1,521 KB
 IBM_SPSS_Regression.pdf	26/03/2025 10:44	Chrome HTML Do...	1,093 KB
 IBM_SPSS_Statistics_Batch_Facility_User_Guide.pdf	26/03/2025 11:12	Chrome HTML Do...	586 KB
 IBM_SPSS_Complex_Samples.pdf	26/03/2025 10:41	Chrome HTML Do...	571 KB
 IBM_SPSS_Forecasting.pdf	26/03/2025 10:44	Chrome HTML Do...	571 KB
 IBM_SPSS_Categories.pdf	26/03/2025 10:41	Chrome HTML Do...	490 KB
 IBM_SPSS_Neural_Network.pdf	26/03/2025 10:44	Chrome HTML Do...	419 KB
 IBM_SPSS_Data_Preparation.pdf	26/03/2025 10:42	Chrome HTML Do...	408 KB
 IBM_SPSS_Decision_Trees.pdf	26/03/2025 10:42	Chrome HTML Do...	400 KB
 IBM_SPSS_Direct_Marketing.pdf	26/03/2025 10:42	Chrome HTML Do...	395 KB
 IBM_SPSS_Missing_Values.pdf	26/03/2025 10:44	Chrome HTML Do...	393 KB
 IBM_SPSS_Conjoint.pdf	26/03/2025 10:41	Chrome HTML Do...	339 KB
 Authorized_User_License_Administrator_Guide.pdf	26/03/2025 11:15	Chrome HTML Do...	331 KB
 IBM_SPSS_Bootstrapping.pdf	26/03/2025 10:41	Chrome HTML Do...	318 KB
 Getting_Started_with_Installation_and_Licensing.pdf	26/03/2025 11:16	Chrome HTML Do...	288 KB

## CONVERSATIONS

- ▼ How can I remove unwanted LOG output?
  - How can I remove unwanted LOG output
  - Whats the difference between ANOVA and ANCOVA

To remove unwanted LOG output in IBM SPSS Statistics, you can hide the log messages in the Viewer. Here's how to do it:

1. From the menus, choose:  
Edit > Options...
2. Click the **Viewer** tab.
3. Double-click the **Log** icon to hide log messages in the output.
4. Click **OK** to save your changes and close the dialog box.

This will hide the log output from the Viewer, helping to declutter your output display.

Whats the difference between ANOVA and ANCOVA?

## Difference Between ANOVA and ANCOVA

ANOVA (Analysis of Variance) and ANCOVA (Analysis of Covariance) are both statistical methods used to compare the means of different groups. However, they differ in their approach and application:

- **ANOVA:** This method is used to compare the means of three or more groups to see if there are any statistically significant differences between them. It does not take into account any other variables that might influence the dependent variable.
- **ANCOVA:** This method is an extension of ANOVA that includes one or more continuous independent variables, known as covariates. ANCOVA adjusts the means of the groups based on the covariates, which can help reduce the error variance and increase the power of the test.

In summary, ANCOVA is used when you want to control for the effects of one or more continuous variables (covariates) while comparing the means of different groups, whereas ANOVA is used when you only want to compare the means without considering any other variables.

For more detailed information, please consult external sources or the official SPSS documentation.





# Meet ChatSPSS – Your AI-Powered SPSS Tutor

Get instant, accurate answers to your SPSS and statistics questions – anytime you need them.

Smart Vision Europe is proud to introduce **ChatSPSS**, the first AI-enabled tutor designed specifically for SPSS users. Whether you're new to SPSS or an experienced analyst, ChatSPSS helps you get the very most out of your software, saving you time and boosting your confidence with data analysis.

## What can ChatSPSS do for you?

- ✓ Guide you through loading and preparing data in SPSS
- ✓ Show you how to summarise results and run statistical tests
- ✓ Explain outputs from regression models and other advanced techniques
- ✓ Answer almost any SPSS or statistics-related question, on demand

## Why use ChatSPSS?

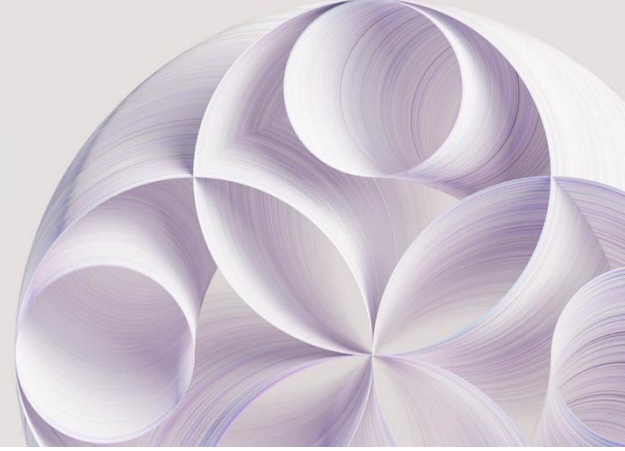
Unlike generic AI tools, ChatSPSS has been optimised with expert-verified SPSS resources. That means you get **reliable, relevant, and accurate answers** tailored to your analysis.

### Ready to try it?

👉 [Register now at ChatSPSS.com](https://www.chat-spss.com) and start getting better answers from your data today. ChatSPSS is completely free to use and, like any large language model, is all about improvement so we welcome any feedback or comment you may have once you've used it.



IBM  
watsonx.ai



# Working with Wastonx.ai – Prompt Lab

## What do you want to do?

Select a task based on your goal. You'll use a tool to create an asset for that goal.

All

Prepare data

Work with models

Search for a task or tool

### Recents



Build machine learning or RAG solutions automatically

with AutoAI



Ground gen AI with vectorized documents

[Learn more](#)

with Vector indexes

A good place to start...

### Prepare data ⓘ



Connect to a data source



Document and organize your insights,



Ground gen AI with vectorized

# watsonx.ai™ Prompt Lab

- Here we can design, test, and refine prompts against a library of IBM-hosted foundation models or even tuned models you've already deployed.
- We can use it to experiment across models and different applications like summarization, Q&A, extraction and classification tasks with ready-made examples and adjustable parameters.
- You can try out different prompts, compare outputs, and lock in effective settings without writing code.

 Webinar Test ×

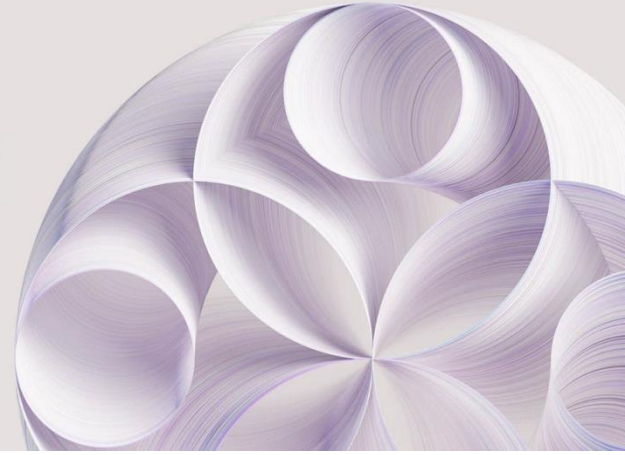


Type something...





IBM  
watsonx.ai



# Working with Wastonx.ai – Auto AI RAG

# watsonx.ai™ Auto AI RAG

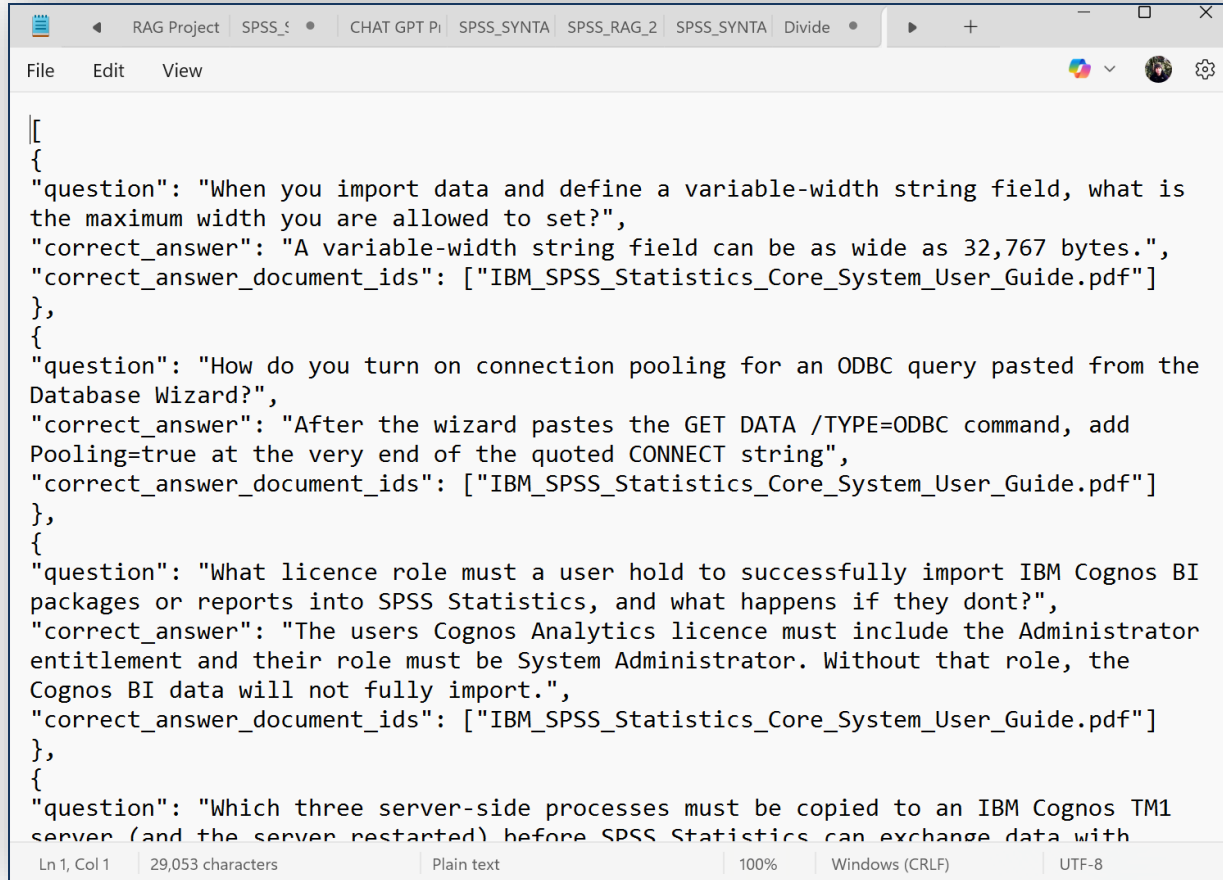
- Here we can set up experiments that try out lots of parameter settings (embedding model, chunking, LLM etc) and ranks them according to how they performed.
- One way to do this, is to compare how each model scored against a 'perfect' set of *example questions and answers* as part of an **evaluation file**.
- Any of the resultant models can then be directly deployed to the outside world for use.

RAG Pattern leaderboard ▾

	Rank	↑	Name	Model name	Answer faithfulness (Optimized)	Retrieval method	Hybrid ranker strategy	Number of chunks	Chunk size	Chunk overlap
★	1		Pattern 16	granite-3-8b-instruct slate-125m-english-rtrvr-v2	<div><div></div></div> 0.932	Window Size: 4	N/A	5	2048	0
	2		Pattern 17	granite-3-8b-instruct multilingual-e5-large	<div><div></div></div> 0.916	Window Size: 4	N/A	5	2048	0
	3		Pattern 13	granite-3-8b-instruct multilingual-e5-large	<div><div></div></div> 0.906	Window Size: 4	N/A	3	2048	0
	4		Pattern 11	granite-3-8b-instruct slate-125m-english-rtrvr-v2	<div><div></div></div> 0.895	Window Size: 4	N/A	3	2048	0
	5		Pattern 18	granite-3-8b-instruct slate-125m-english-rtrvr-v2	<div><div></div></div> 0.894	Window Size: 4	N/A	5	2048	512



# AutoAI RAG – Evaluation File



```
[
{
  "question": "When you import data and define a variable-width string field, what is the maximum width you are allowed to set?",
  "correct_answer": "A variable-width string field can be as wide as 32,767 bytes.",
  "correct_answer_document_ids": ["IBM_SPSS_Statistics_Core_System_User_Guide.pdf"]
},
{
  "question": "How do you turn on connection pooling for an ODBC query pasted from the Database Wizard?",
  "correct_answer": "After the wizard pastes the GET DATA /TYPE=ODBC command, add Pooling=true at the very end of the quoted CONNECT string",
  "correct_answer_document_ids": ["IBM_SPSS_Statistics_Core_System_User_Guide.pdf"]
},
{
  "question": "What licence role must a user hold to successfully import IBM Cognos BI packages or reports into SPSS Statistics, and what happens if they dont?",
  "correct_answer": "The users Cognos Analytics licence must include the Administrator entitlement and their role must be System Administrator. Without that role, the Cognos BI data will not fully import.",
  "correct_answer_document_ids": ["IBM_SPSS_Statistics_Core_System_User_Guide.pdf"]
},
{
  "question": "Which three server-side processes must be copied to an IBM Cognos TM1 server (and the server restarted) before SPSS Statistics can exchange data with
```

Ln 1, Col 1 | 29,053 characters | Plain text | 100% | Windows (CRLF) | UTF-8



# Evaluating a RAG model

- **Answer faithfulness:** This measures whether the generated answer is consistent with the provided retrieved context. It minimises the likelihood of the response including information that wasn't in the source documents (hallucinations) .
- **Answer correctness:** This evaluates the factual correctness of the information presented in the final answer. While faithfulness links the answer to the context, correctness verifies the answer against known facts in the documents and the evaluation file. Here the focus is on enhanced accuracy and reliability.
- **Context correctness:** This is a *retrieval-quality* metric. In **AutoAI RAG**, Context correctness shows how well the retriever pulled the right chunk(s) of context compared to the benchmark. As such, it evaluates the retriever, not the generated answer.



# Evaluating a RAG model: practical advice

- If you want to minimise hallucinations, then **Answer faithfulness** is a good optimum but watch correctness as a secondary metric.
- If you need to match *exact* answers (e.g. as in safety and compliance applications), try to optimise **Answer correctness**.
- You can use **Context correctness** to debug the retriever itself. Low scores here indicate that parameters such as chunking, embedding, or retrieval may need tuning.


# Auto AI Rag Experiment progress map

[Projects](#) / [SV SPSS RAG Project - London](#) / [SPSS\\_SYNTAX\\_Auto\\_AI\\_RAG\\_Build\\_31\\_July\\_28\\_v1](#)

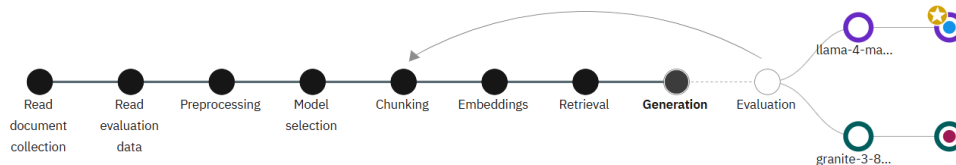
## Experiment summary

### Setting importance


## Pattern comparison

★ Rank by: Answer correctness | Mean 

Progress map ⓘ



## Experiment status

Experiment in progress 

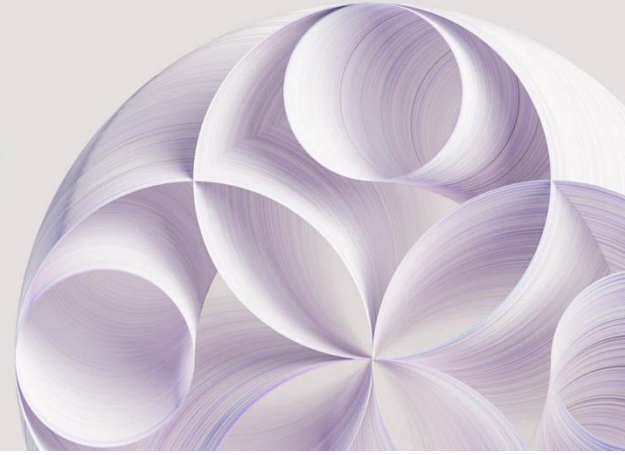
Text generation completed

RAG Pattern leaderboard 

Rank	Name	Model name	Answer correctness (Optimized)	Retrieval method	Hybrid ranker strategy	Number of chunks	Chunk size
1	Pattern 1	llama-4-maverick-17b-128e-instruct-fp8-slate-125m-english-rtvr-v2	0.918	Window Size: 3	N/A	5	2048



IBM  
watsonx.ai



# Deploying RAG Models

# Deployment options - Notebook

## Save notebook

Save a pattern as a notebook, so you can review, edit, or run the code for the RAG pattern. Or, save as an AI service to create a deployable asset based on the current experiment settings.

## Choose your objective

### Index building, retrieval, and generation



Build an index knowledge base using ChromaDB, retrieve relevant passages from the knowledge base, evaluate the retrieval quality, and build and deploy a function as an endpoint to inference with the RAG Pattern.

## Define details

Asset type

- ☒ Notebook  
☐ AI Service

Name

Pattern 3: SPSS\_SYNTAX\_Auto\_AI\_RAG\_Build\_31\_Jul...

Deployment space

SV\_RAG\_DEPLOYMENT\_LONDON



Why don't I see all of my spaces? ⓘ

Runtime environment

Runtime 24.1 on Python 3.11 AutoAI-M (4 vCPU and 16 GB RAM)



# Deployment options – AI Service

## Save AI Service

Save a pattern as a notebook, so you can review, edit, or run the code for the RAG pattern. Or, save as an AI service to create a deployable asset based on the current experiment settings.

### Choose your objective

#### Index building, retrieval, and generation

Build an index knowledge base using ChromaDB, retrieve relevant passages from the knowledge base, evaluate the retrieval quality, and build and deploy a function as an endpoint to inference with the RAG Pattern.

### Define details

Asset type

- ☐ Notebook
- ☒ AI Service

Name

Pattern 3: SPSS\_SYNTAX\_Auto\_AI\_RAG\_Build\_31\_Jul...

Description (optional)

AI Service description

Tags

Add tags to make assets easier to find.

Add a tag

☒ Promote and deploy AI Service to deployment space

Cancel

Create and deploy

# Deployment options – AI Service deploying

Deployment spaces / SV\_RAG\_DEPLOYMENT\_LONDON / Pattern 3: SPSS\_SYNTAX\_Auto\_AI\_RAG\_Build\_31\_Jul...

Deployments

Code

Schema

Search

Name	Type	Status	Tags
<div><div></div>Pattern 3: SPSS_SYNTAX_Auto_AI_RAG_Build_31_Jul...</div>	Online	<div></div> Initializing	<div>wx-autoai-rag</div> <div></div>

# API endpoint of deployed AI Service

Pattern 3: SPSS\_SYNTAX\_Auto\_AI\_RAG\_Build\_31\_Jul... ✓ Deployed Online

API reference Test Preview

## Endpoints for inferencing ⓘ

### Private endpoint

[https://eu-gb.ml.cloud.ibm.com/ml/v4/deployments/0f2eb07b-5931-442b-8a53-469c62329ce8/ai\\_service?version=2021-05-01](https://eu-gb.ml.cloud.ibm.com/ml/v4/deployments/0f2eb07b-5931-442b-8a53-469c62329ce8/ai_service?version=2021-05-01)

[https://eu-gb.ml.cloud.ibm.com/ml/v4/deployments/0f2eb07b-5931-442b-8a53-469c62329ce8/ai\\_service\\_stream?version=2021-05-01](https://eu-gb.ml.cloud.ibm.com/ml/v4/deployments/0f2eb07b-5931-442b-8a53-469c62329ce8/ai_service_stream?version=2021-05-01)

### Public endpoint

[https://eu-gb.ml.cloud.ibm.com/ml/v4/deployments/0f2eb07b-5931-442b-8a53-469c62329ce8/ai\\_service?version=2021-05-01](https://eu-gb.ml.cloud.ibm.com/ml/v4/deployments/0f2eb07b-5931-442b-8a53-469c62329ce8/ai_service?version=2021-05-01)

[https://eu-gb.ml.cloud.ibm.com/ml/v4/deployments/0f2eb07b-5931-442b-8a53-469c62329ce8/ai\\_service\\_stream?version=2021-05-01](https://eu-gb.ml.cloud.ibm.com/ml/v4/deployments/0f2eb07b-5931-442b-8a53-469c62329ce8/ai_service_stream?version=2021-05-01)

[Learn more](#) about the 2021-05-01 version query parameter

## Code snippets

cURL

Java

JavaScript

Python

Scala

```
import requests

# NOTE: you must manually set API_KEY below using information retrieved from your IBM Cloud account (https://eu-gb.dataplatform.cloud.ibm.com/docs/content/wsj/analyze-data/ml-authentication.
API_KEY = "<your API key>"
token_response = requests.post('https://iam.cloud.ibm.com/identity/token', data={"apikey": API_KEY, "grant_type": 'urn:ibm:params:oauth:grant-type:apikey'})
mltoken = token_response.json()["access_token"]

header = {'Content-Type': 'application/json', 'Authorization': 'Bearer ' + mltoken}
```

## About this deployment

### Name

Pattern 3:  
SPSS\_SYNTAX\_Auto\_AI\_RAG\_Build\_31\_Ju  
...

### Description

No description provided.

### Deployment Details

Deployment ID: 0f2eb07b-5931-44...

Serving name:

No serving name

Software specification:

[runtime-24.1-py3.11](#)

Hardware specification:

Extra extra small: 1 CPU and 2 GB RAM

Copies:

1

### Tags

wx-autoai-rag

### Associated asset

[Pattern 3: SPSS\\_SYNTAX\\_Auto\\_AI\\_RA...](#)

5b51ff4b-84c6-43d3-86cc-09b359cab643

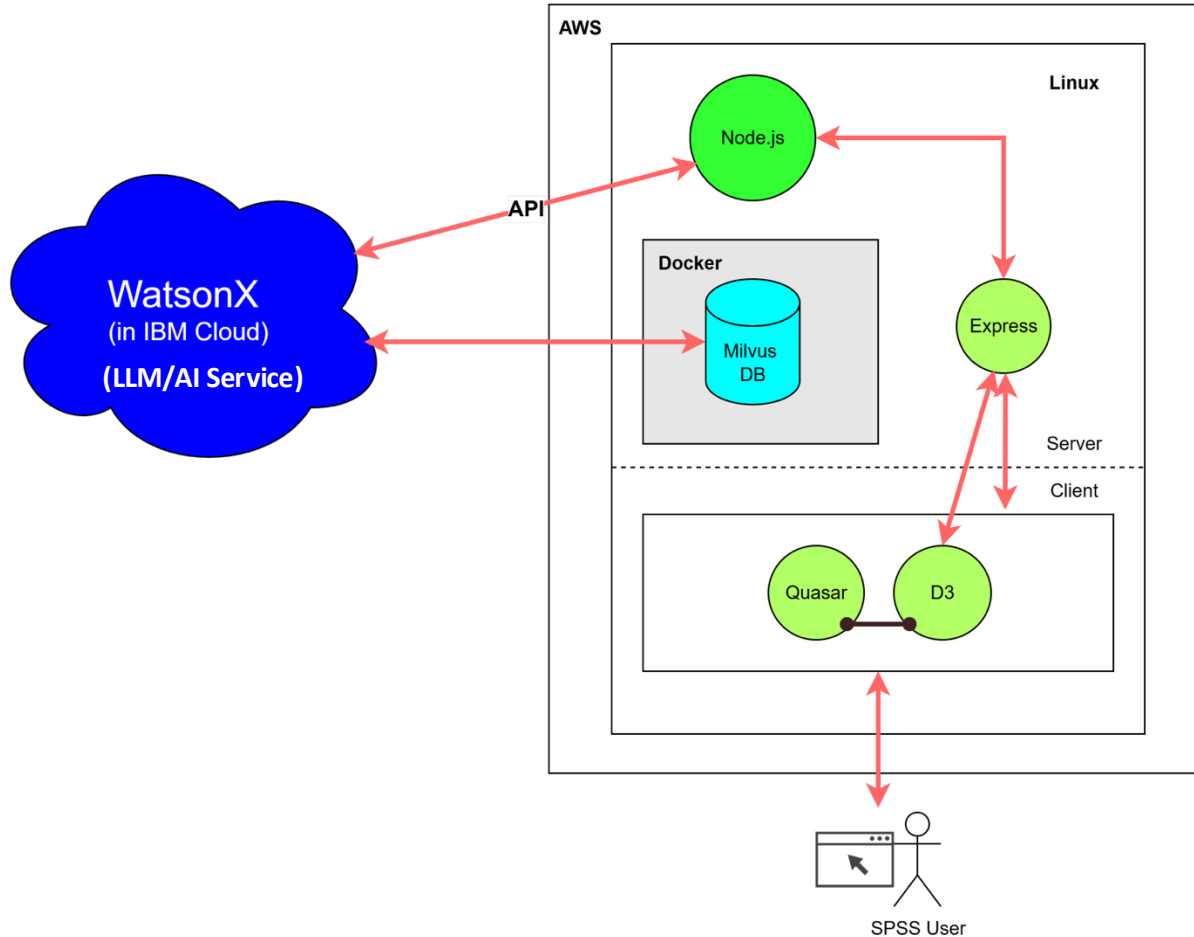
Last modified

13 minutes ago

Created on

Sep 22, 2025

# chatSPSS deployment on Amazon (AWS)



Open-source components on AWS:

[D3](#)

[Quasar](#)

[Express](#)

[Node.js](#)

[Milvus](#)



An aerial photograph of the London skyline at sunset. The city is densely packed with buildings, and the River Thames is visible on the right. The sky is a mix of orange, yellow, and blue, with some clouds. The Shard is prominent on the right side of the image.

# Smart Vision Europe

Your one stop shop for advanced analytics software,  
training and consulting

[Find out more](#)

## Working with Smart Vision Europe

# Working with Smart Vision Europe

- *Ready to start your Retrieval Augmented Generation AI journey?*
- We can assist with every aspect of your RAG project
  - Help with solution design and technical architecture
  - Selection of appropriate tools and technology components
  - Ensuring appropriate governance and security
- As experts in Data Science and AI we will
  - Offer “skills transfer” consulting
  - Collaborate with your internal resources to deliver a complete solution that you will own
  - Offer advice and guidance on engineering (separating code from content), content selection and security relating to RAG projects

# Smart Vision Europe: Other services

- **Sourcing Software**
  - You can buy your analytical software from us often with discounts
  - Assist with selection, pilot, implementation & support of analytical tools
  - <http://www.sv-europe.com/buy-spss-online/>
- **Training**
  - Formal classroom/virtual training
  - Online self-paced training resources
- **Advice and Support**
  - ‘No strings attached’ technical and business advice relating to analytics
  - Tracked technical support services around the IBM SPSS product line



Contact us:

+44 (0)207 786 3568

[info@sv-europe.com](mailto:info@sv-europe.com)

Twitter: @sveurope

[Follow us on Linked In](#)

[Sign up for our Newsletter](#)



# Thank you