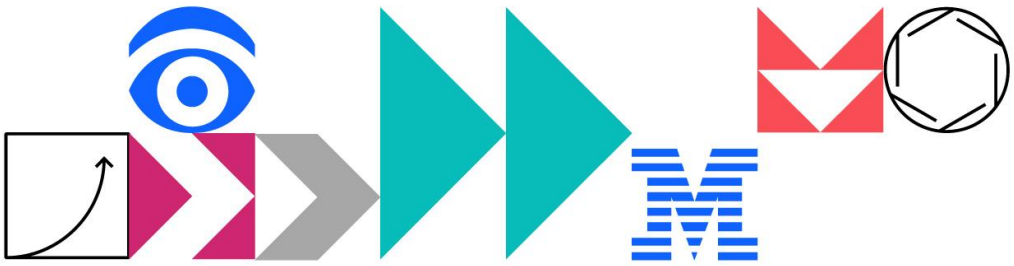
IBM TechXchange



Yet Another RAG - LLM assembled but also verified answers, Hands-on Lab

Session 1531

Lab Exercise Guide

Martin Ryšánek

Technical Pre-Sales Professional, Data&AI, IBM Czech Republic

martin_rysanek@cz.ibm.com

Notices and disclaimers

U.S. Government Users Restricted Rights — use, duplication or disclosure restricted by GSA ADP Schedule Contract with IBM.

This document is current as of the initial date of publication and may be changed by IBM at any time. Not all offerings are available in every country in which IBM operates.
Information in these presentations (including information relating to products that have not yet been announced by IBM) has been reviewed for accuracy as of the date of initial publication and could include unintentional technical or typographical errors. IBM shall have no responsibility to update this information.

This document is distributed "as is" without any warranty, either express or implied. In no event, shall IBM be liable for any damage arising from the use of this information, including but not limited to, loss of data, business interruption, loss of profit or loss of opportunity. IBM products and services are warranted per the terms and conditions of the agreements under which they are provided. The performance data and client examples cited are presented for illustrative purposes only. Actual performance results may vary depending on specific configurations and operating conditions.

IBM products are manufactured from new parts or new and used parts.
In some cases, a product may not be new and may have been previously installed. Regardless, our warranty terms apply."

Any statements regarding IBM's future direction, intent or product plans are subject to change or withdrawal without notice.

Performance data contained herein was generally obtained in a controlled,
isolated environments. Customer examples are presented as illustrations of how those customers have used IBM products and the results they may have achieved. Actual performance, cost, savings or other results in other operating environments may vary.

References in this document to IBM products, programs, or services does not imply that IBM intends to make such products, programs or services available in all countries in which IBM operates or does business.

Workshops, sessions and associated materials may have been prepared by independent session speakers, and do not necessarily reflect the views of IBM. All materials and discussions are provided for informational purposes only, and are neither intended to, nor shall constitute legal or other guidance or advice to any individual participant or their specific situation.

It is the customer's responsibility to ensure its own compliance with legal requirements and to obtain advice of competent legal counsel as to the identification and interpretation of any relevant laws and regulatory requirements that may affect the customer's business and any actions the customer may need to take to comply with such laws. IBM does not provide legal advice or represent or warrant that its services or products will ensure that the customer follows any law.

Notices and disclaimers (Continued)

Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products. IBM does not warrant the quality of any third-party products, or the ability of any such third-party products to interoperate with IBM's products. **IBM expressly disclaims all warranties, expressed or implied, including but not limited to, the implied warranties of merchantability and fitness for a purpose.** The provision of the information contained herein is not intended to, and does not, grant any right or license under any IBM patents, copyrights, trademarks or other intellectual property right.

IBM, the IBM logo, and ibm.com are trademarks of International Business Machines Corporation, registered in many jurisdictions worldwide. Other product and service names might be trademarks of IBM or other companies. A current list of IBM trademarks is available on the Web at "Copyright and trademark information" at: www.ibm.com/legal/copytrade.shtml.

Table of Contents

1 Overview

NeuralSeek is an AI-powered answer-generation engine that empowers businesses. NeuralSeek provides a clickable path to fact AI generative RAG Pattern solution, which can online AI generate answers for given questions by taking a user question and assembling a corpus of backing information from a KnowledgeBase like Watson Discovery or ElasticSearch. It is the business solution to use AI in a professional workplace.

Generating answers ONLINE is a common way of implementation of RAG Pattern solution. However, we do encounter some **inaccuracies in the answers due to several reasons**. Inaccuracies comes from weak users questions, not well prepared KnowledgeBase documents, some incorrectly generated text by Large Language Model (e.g. due to LLM hallucinations), not well searched and selected document passages for generation of answers.

Therefore, while some of our clients want the answers to questions to be automatically generated by RAG online, they also require that the answer be human-checked and possibly manually edited before being sent to the end user. Neuralseek is the ultimate orchestration tool for implementation of RAG, capable of not only to answer user queries online, but also aggregate answers to similar queries into intents. It then allows a subject matter expert to edit intent answers and propagate them as verified and validated answers for given queries.

It leads to system of Frequently Answered Questions generated by Neuralseek and verified by human. Instead of an automated answer from the RAG / LLM, the user is provided with a set of prepared answers corresponding to the question asked, but which have been reviewed and modified by a subject matter expert.

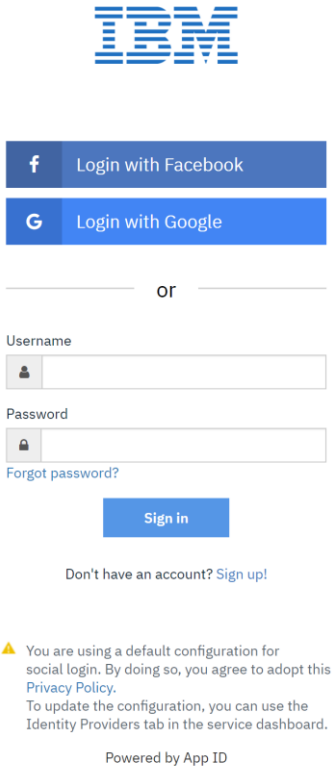
This lab includes an introduction to the capabilities of NeuralSeek Lite in the areas of answer generation, answer verification and editing, grouping of several questions into intent, and Neuralseek's ability to export the question-and-answer intent system to IBM Watsonx Assistant. We modify the IBM Watson Assistant into a system that searches for the most likely validated answers to a question (FAQs) and that records questions that are not sufficiently answered in this way and will be batch processed by Neuralseek and verified offline by subject matter expert.

2 IBM Cloud Login

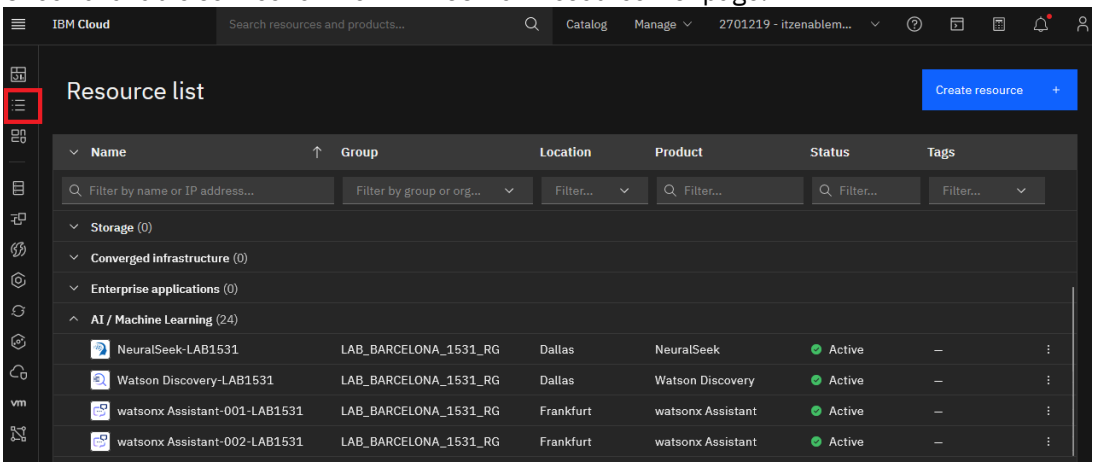
All services for LAB 1531 students are available from IBM Cloud. So the first step is to log in to the IBM Cloud and find the LAB 1531 services.

1. You will be given your **student ID**, which follows the pattern student0xx (e.g. student013).
2. Log in to LAB 1531 portal

   https://cloud.ibm.com/authorize/b08c84ff99dd4f0f9cbc032750158984



3. Fill given labuser0xx as Username and Password (0xx is YOUR_ID)
4. After successful login, you will get to IBM Cloud for LAB 1531 with several provisioned services.
5. Check available service for the LAB 1531 on Resource List page:



3 Concept of RAG with Verified Answers (FAQ)

The entire LAB will be done using IBM Cloud services, some of which are already provisioned and configured and some of which we will have to configure and connect with other services.

You can test and experiment with the entire running solution using the services at the following URL:

https://pages.github.ibm.com/martin-rysanek/BARCELONA-WEB-LAB1531

Additional configuration files and keys, you can find on the following URL:

https://github.ibm.com/martin-rysanek/BARCELONA-LAB1531

There are steps, which we need to follow in next chapters:

1. Neuralseek as RAG solution is used as shared service among the LAB students and it is already provisioned and setup just for your review. You can open the service from LAB resource list under the name NeuralSeek-LAB1531 and review its configuration and search results stored there. Please do not change its configuration.

2. We have also already provisioned and setup Watson Discovery as KnowledgeBase document search engine under the name Watson Discovery-LAB1531. The service is available for your review from administrator perspective, so please keep configuration untouched.

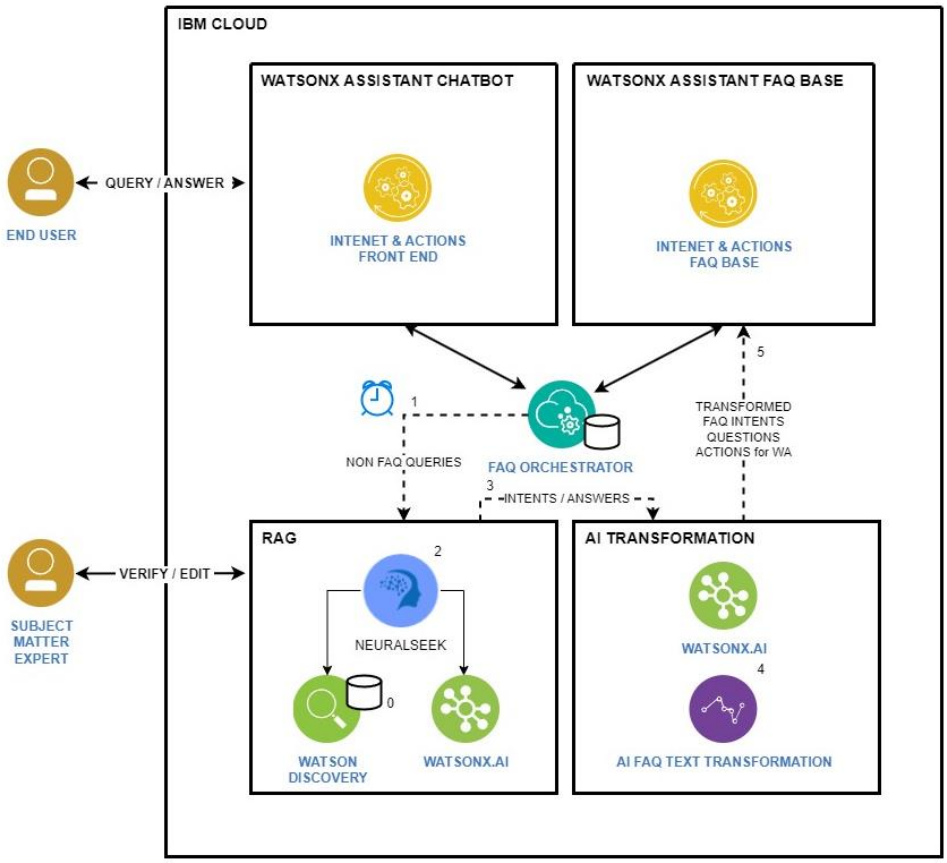
3. We will need to setup two instances of Watsonx Assistant. The first instance will act as front end chatbot processing all users inputs and control communication among the services, the second one will hold Frequently Asking Question (FAQ) knowledge base, will be created from the scratch from Neuralseek export.

   The Watsonx Assistant service is already provisioned, you can find it under watsonx Assistant-YOUR_ID-LAB1531 name on the LAB Resource List page.

4. We will need to configure already provisioned serverless FAQ Orchestrator running in docker and using IBM Cloud Code Engine. The application is available for you with the name ceapp-YOUR_ID-lab1531 under Code Engine project called CE-BAR-LAB1531.

Note: In order to get service dedicated for you, you need to substitute YOUR_ID from previous paragraph with your student number in the format 0xx (e.g. 013).

3.1 FAQ Question Answering System Architecture



- Any use's query goes to Watsonx Assistant Chatbot first
- Chatbot need to display set of most relevant FAQs, so it calls FAQ ORCHESTRATOR, which gets most relevant FAQs from Watsonx Assistant FAQ Base
- The user pickup some of FAQ and display its answer, he/she can repeat the process with additional FAQs from the given set. The user is asked for ranking of the answer for every FAQ.
- FAQ Orchestrator records all queries, answers, FAQ confidence and user answer ranking. This information are used to identify by heuristic algorithm "new questions", i.e. the questions, which were not well answered and need to be newly generate by RAG / Neuralseek from KnowledgeBase using the LLM model.
    1. Based on time event (e.g. daily), there is process executed, which generate these "new questions"
    2. Neuralseek respond internally to "new questions" and update its FAQ base
    3. FAQ intents/answers are exported from Neuralseek to AI Transformation process
    4. The process merge FAQ intents, create better titles of FAQ and additional questions, which identify the intent using watsonx.ai
    5. Generated Watsonx Assistant configuration file is imported into Watsonx Assistant FAQ BASE.

4 NeuralSeek Export

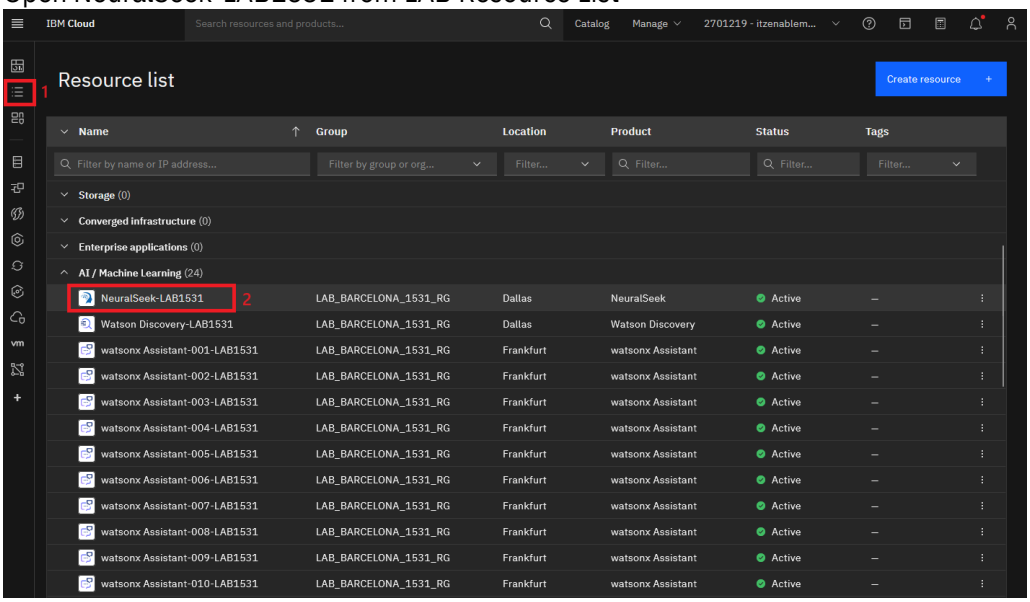
Online querying Neuralseek by question leads to Neuralseek automatic aggregation of vector similar questions into intents (creating FAQs). Every intent has one or several Neuralseek/LLM generated answers. The subject matter expert can browse Neuralseek/LLM generated FAQ answers and if it is required edit them to create verified and corrected answers.

Neuralseek is the ultimate orchestration tool for implementation of RAG, capable of not only to answer user queries online, but also aggregate answers to similar queries into intents. It then allows a subject matter expert to edit intent answers and propagate them as verified and validated answers for given queries.

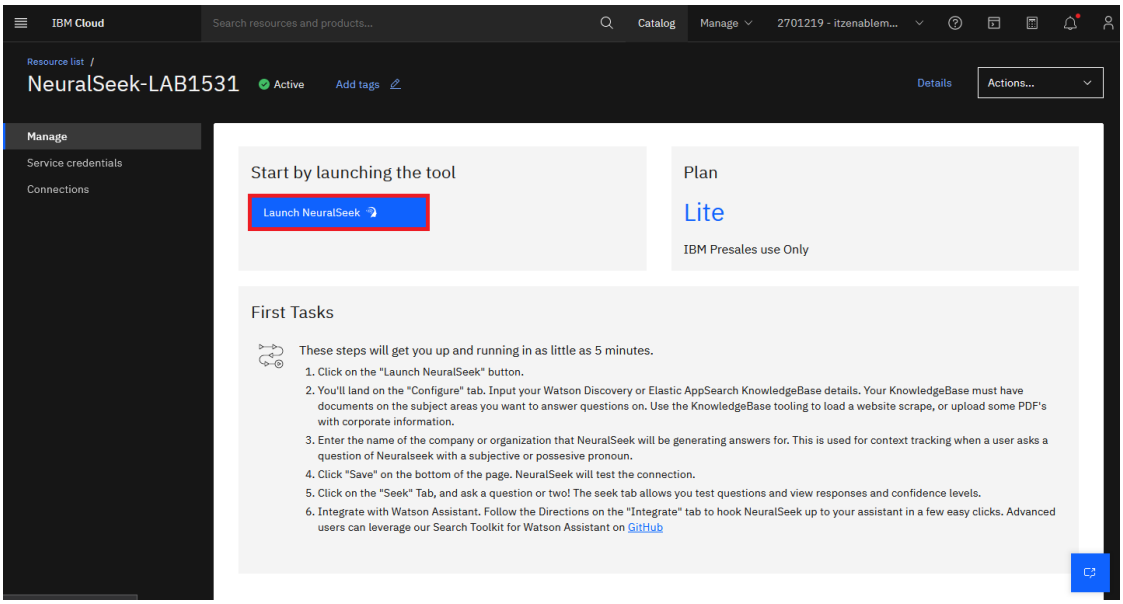
1. Chapter 4.1 and 4.2 let you look to Neuralseek FAQ base and steps done by subject matter expert to get adjust FAQ answers and export configuration file as Watsonx Assistant FAQ Base
2. To simplify required steps to you as student, we make prepared WA FAQ base for you. Go directly to chapter 4.3, where you have instructions to download FAQ base from GITHUB.
3. If you decide to prepare the file by yourself (to learn steps), follow steps in Chapter 4.1 and 4.2:
    a. Subject matter expert need to edit and verify answers using Neuralseek user interface
    b. Export WA action file (Watsonx Assistant configuration file) from Neuralseek
    c. Make extra AI transformation and AI question enrichment to make WA configuration file more natural

4.1 Browse curate FAQ base an edit answers in Neuralseek (optional)

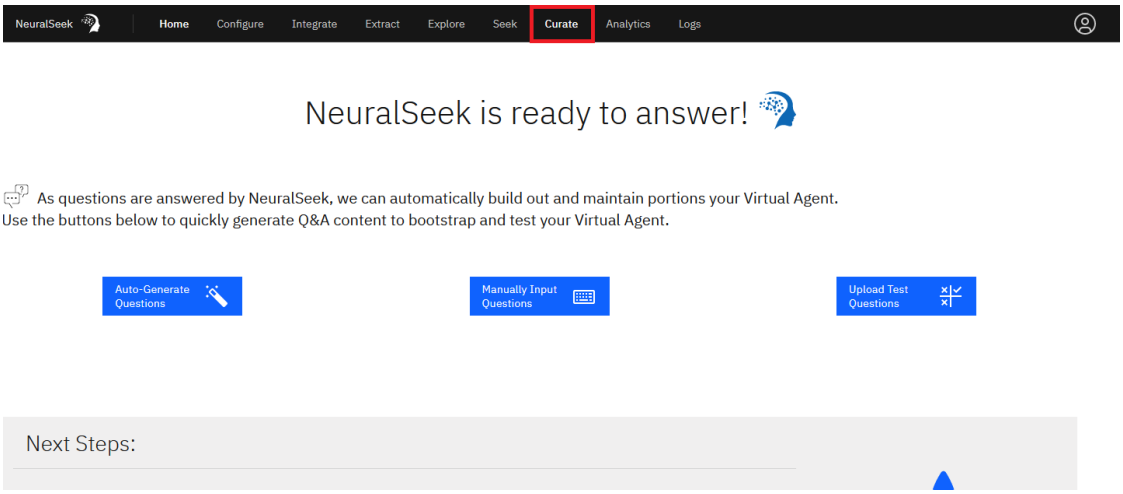
1. Open NeuralSeek-LAB1531 from LAB Resource List



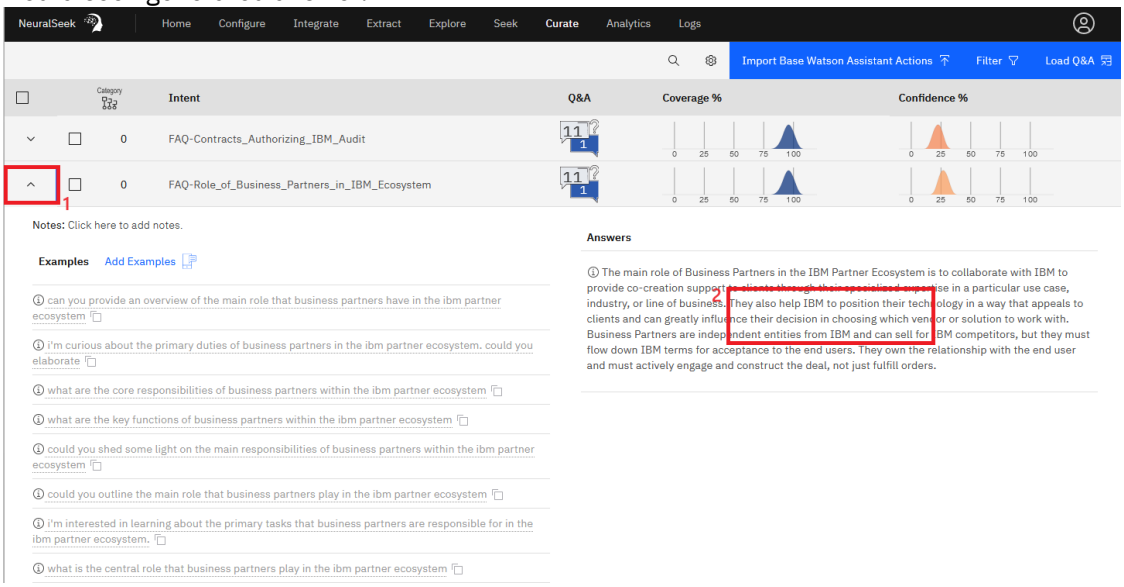
2. Launch NeuralSeek service



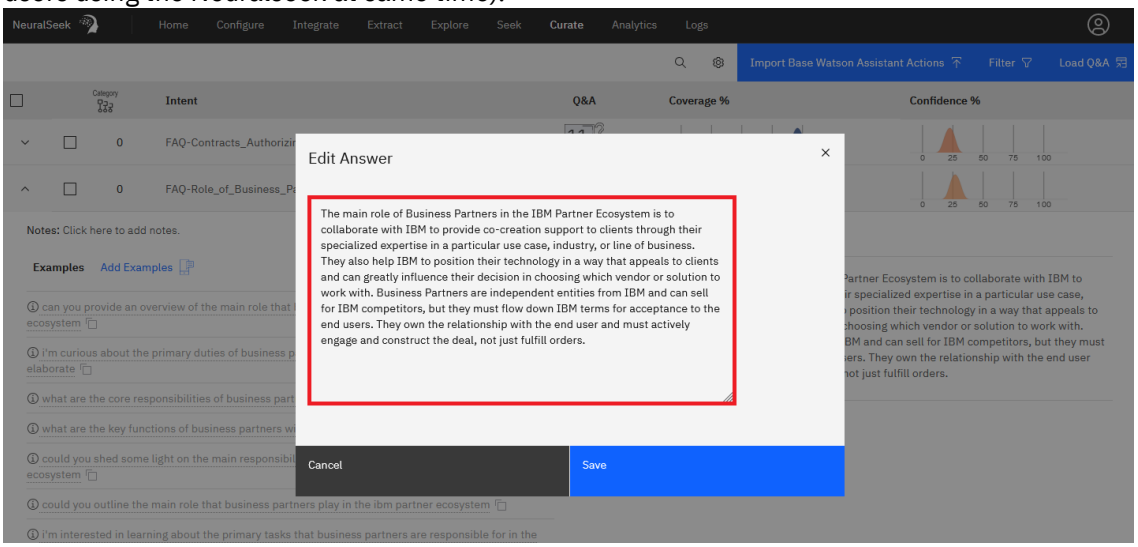
3. Click the Curate tab



4. Open one of the intent (e.g. FAQ-Role_of_Business_Partners_in_IBM_Ecosystem) then click on the Neuralseek generated answer.



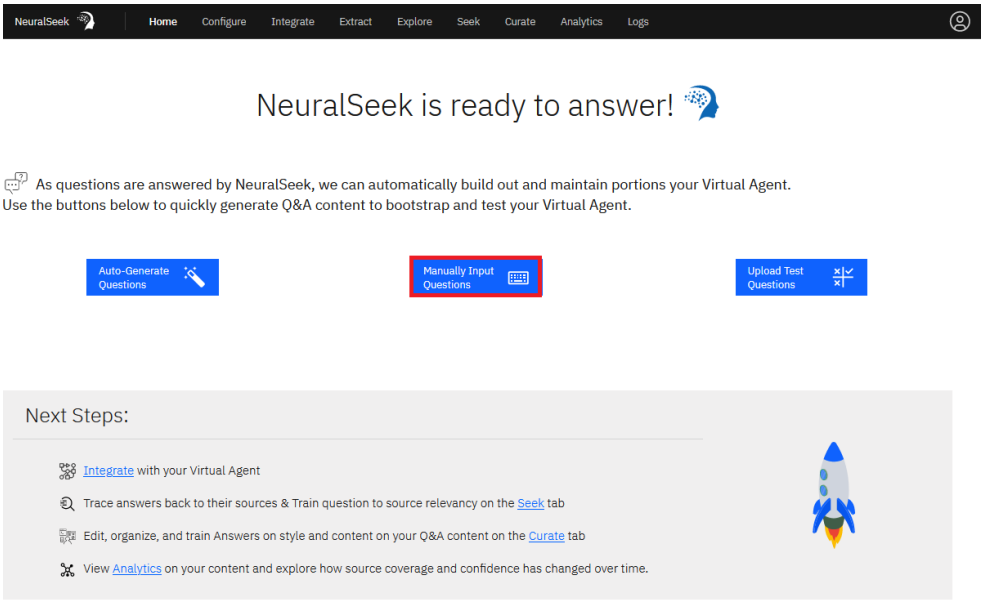
5. Here you can view, verify, and potentially edit LLM generated answers. This is only for your information. (Please, **keep the Neuralseek untouched**, **do not edit it in real**, there are multiple users using the Neuralseek at same time).



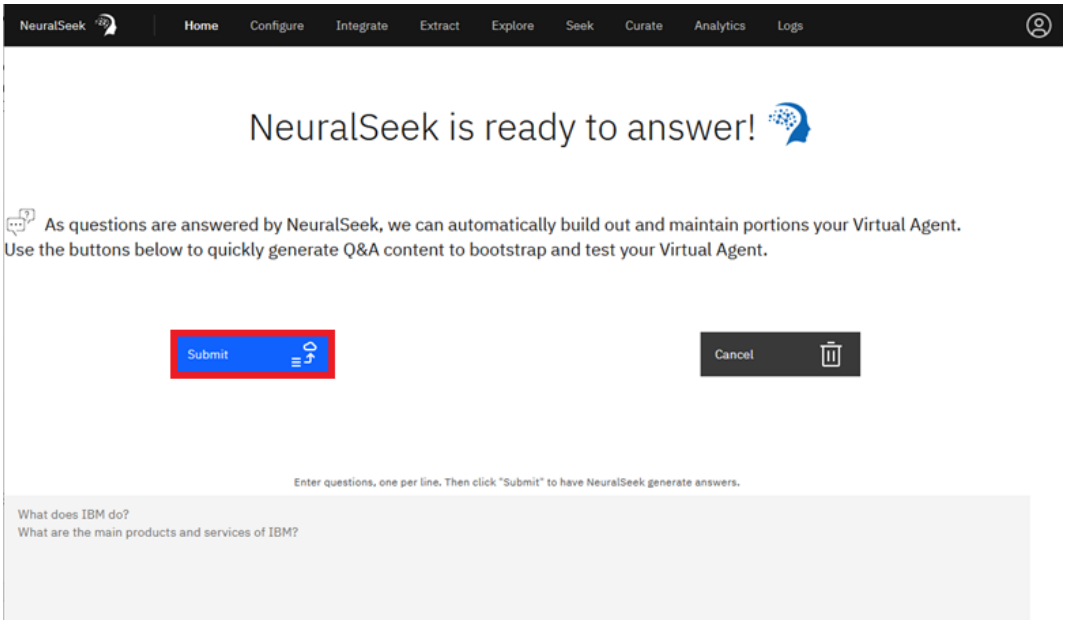
4.2 Export Watsonx Assistant FAQ Based from Neuralseek (optional)

(Optional) In the following steps, you can export Neuralseek curate FAQ base as Watsonx Assistant configuration file by yourself (you are not required). It is recommended to rather use file from GITHUB, it is further AI transformed and prepared file.

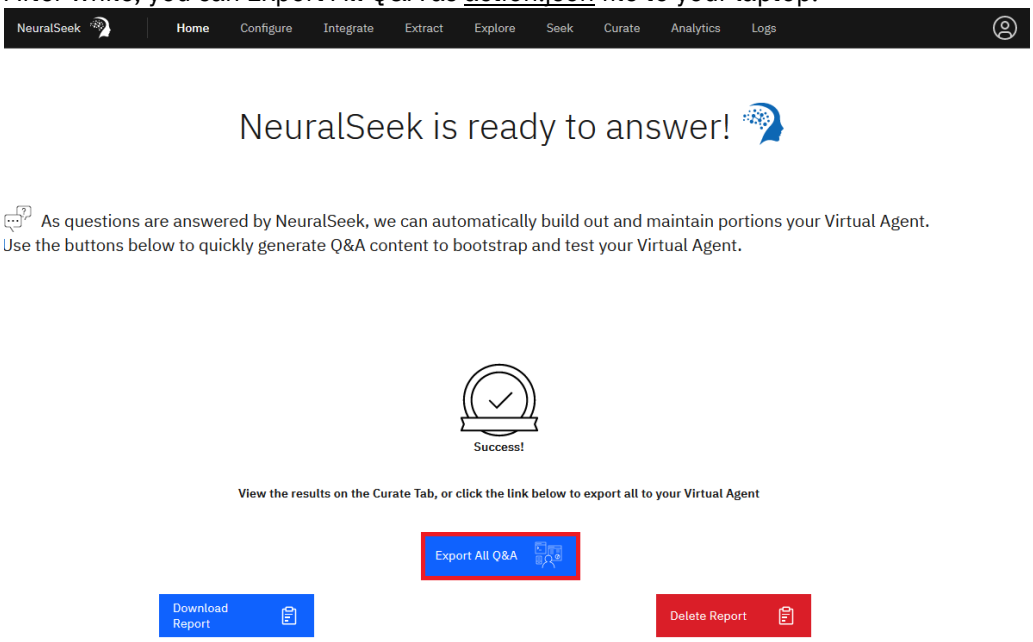
1. You can download Neuralseek FAQ curate base from Neuralseek home screen, click the Manually Input Question



2. Keep form empty and press Submit



3. After while, you can Export All Q&A as action.json file to your laptop.



4.3 Get prepared Watsonx Assistant FAQ Based from GITHUB

1. Download exported FAQ base of intents / questions and answers from Neuralsekk for Watsonx Assistant directly from GITHUB under the name wa-faq-base-action.json.
   https://github.ibm.com/martin-rysanek/BARCELONA-LAB1531/wa-faq-base-action.json

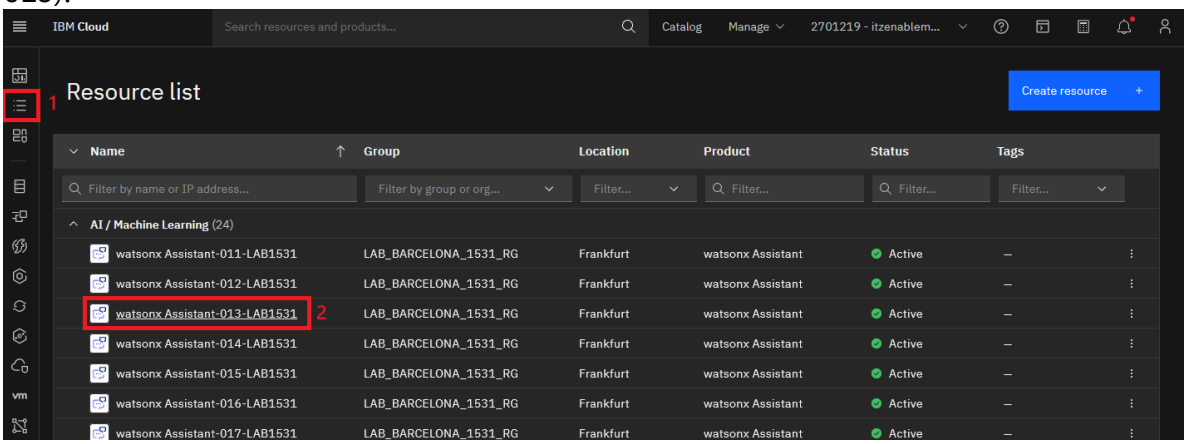
5 Setup Watsonx Assistant - FAQ Base

Watsonx Assistant has NLP classification capability to identify intent for given question. According to NLP benchmarks, Watsonx Assistant is one of the best in intent detection accuracy. There is a linked answer to every intent. Therefore we have decided to use Watsonx Assistant as FAQ base (question and answer tuple) for user's question, which we exported from Neuralseek in previous chapter.

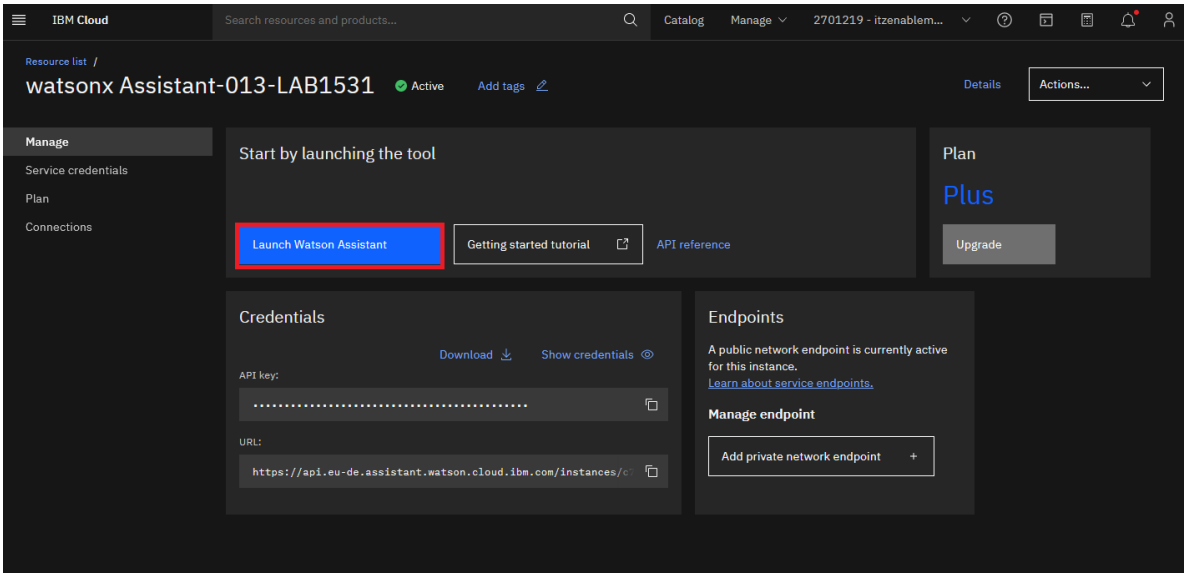
Let us setup new Watsonx Assistant FAQ Base instance and upload it with FAQ Base exported from Neuralseek in the chapter.

5.1 Open and Setup Watsonx Assistant instance

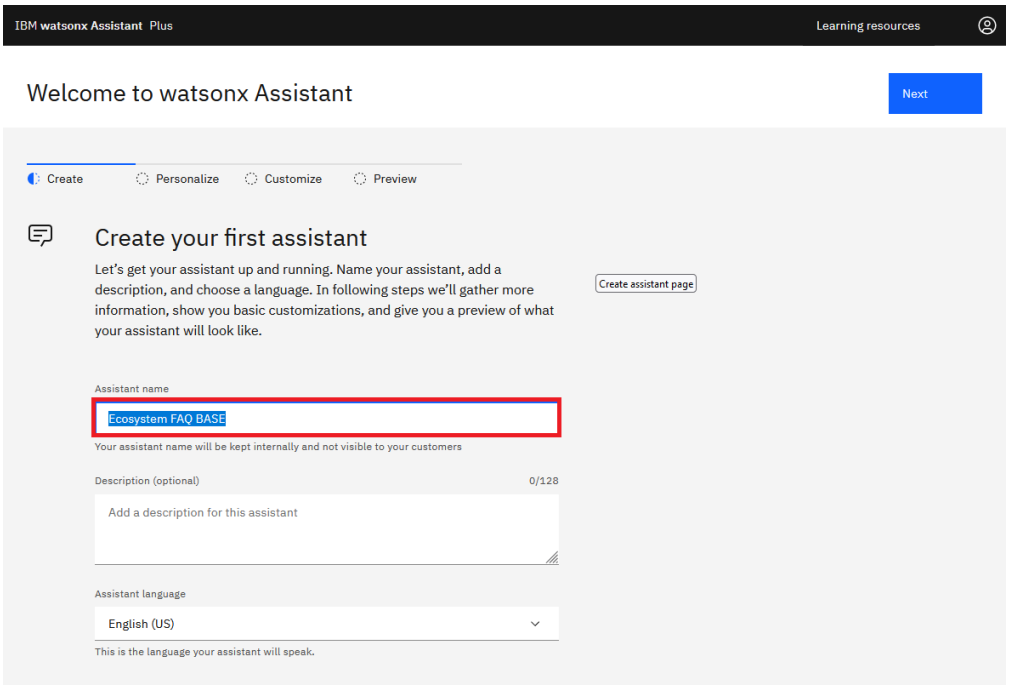
1. Open LAB Resource list, find and click your prepared Watsonx Assistant under the name watsonx Assistant-YOUR_ID-LAB1531, where YOUR_ID is your ID in the LAB in format 0xx (e.g. 013).



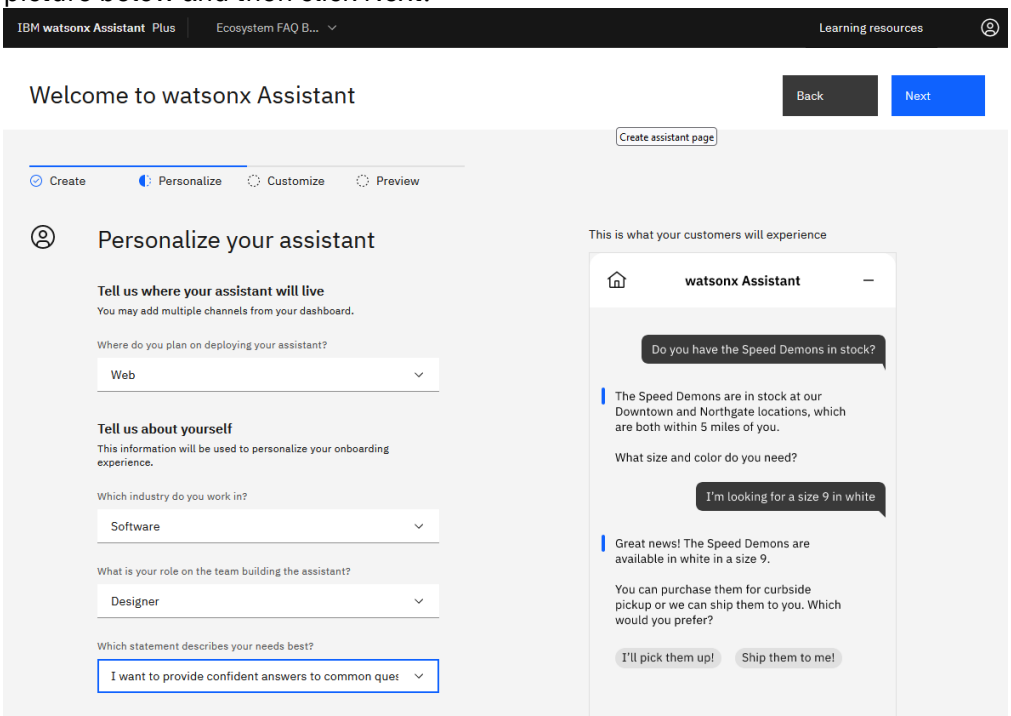
2. Lauch Watsonx Assistant



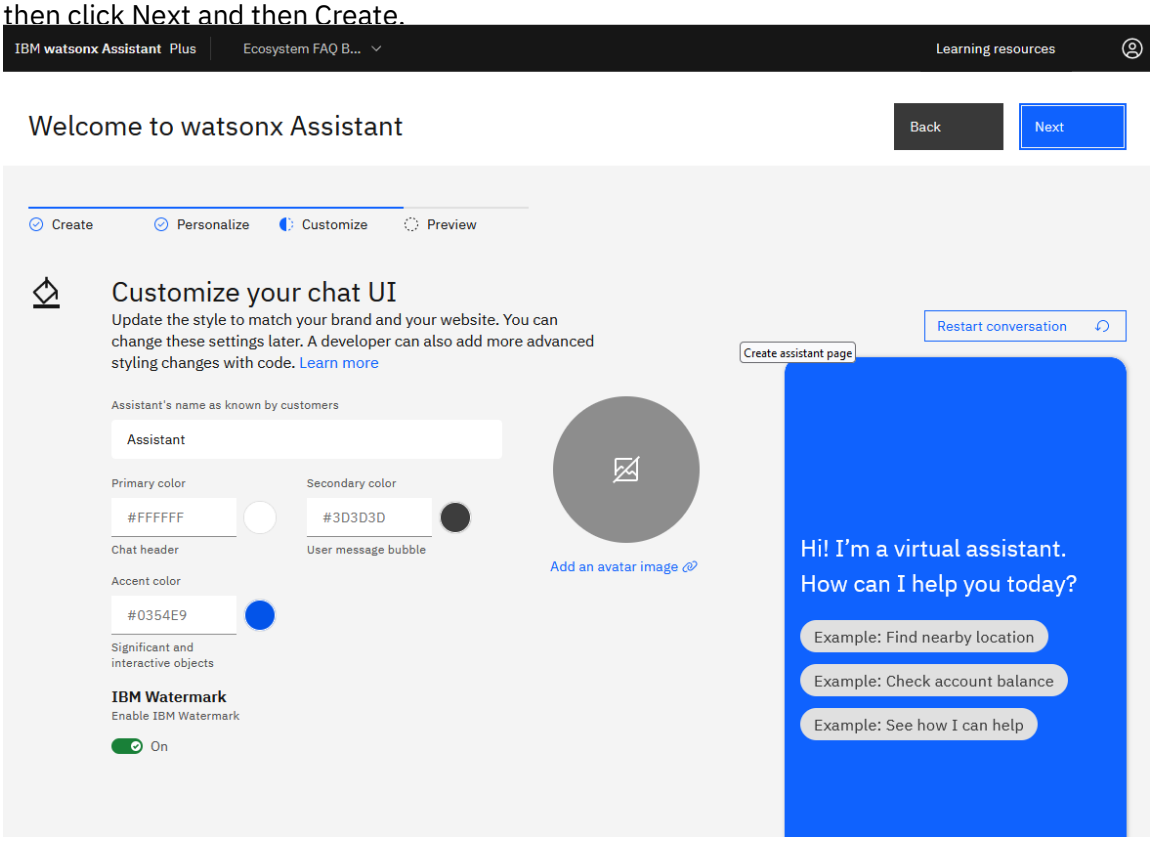
3. Name the Watsonx Assistant instance, e.g. Ecosystem FAQ BASE, click Next.



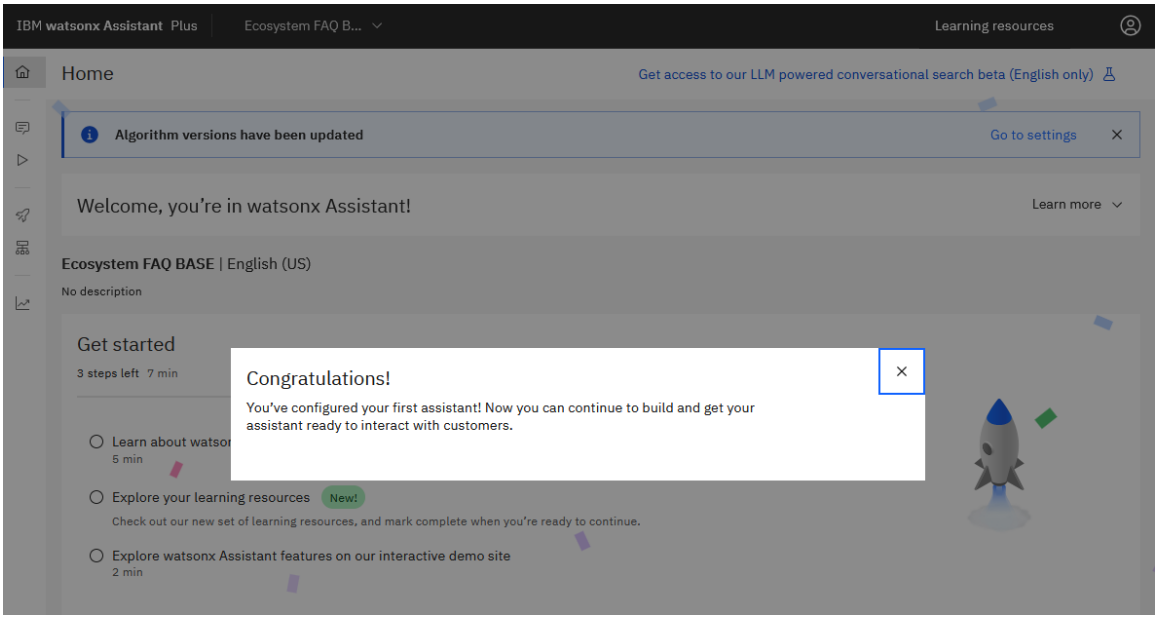
4. Personalize your Watsonx Assistant instance, you can use items similar as you can see on the picture below and then click Next.



5. Give your assistant name, pickup color of your assistant (your can leave them as they are) and then click Next and then Create.

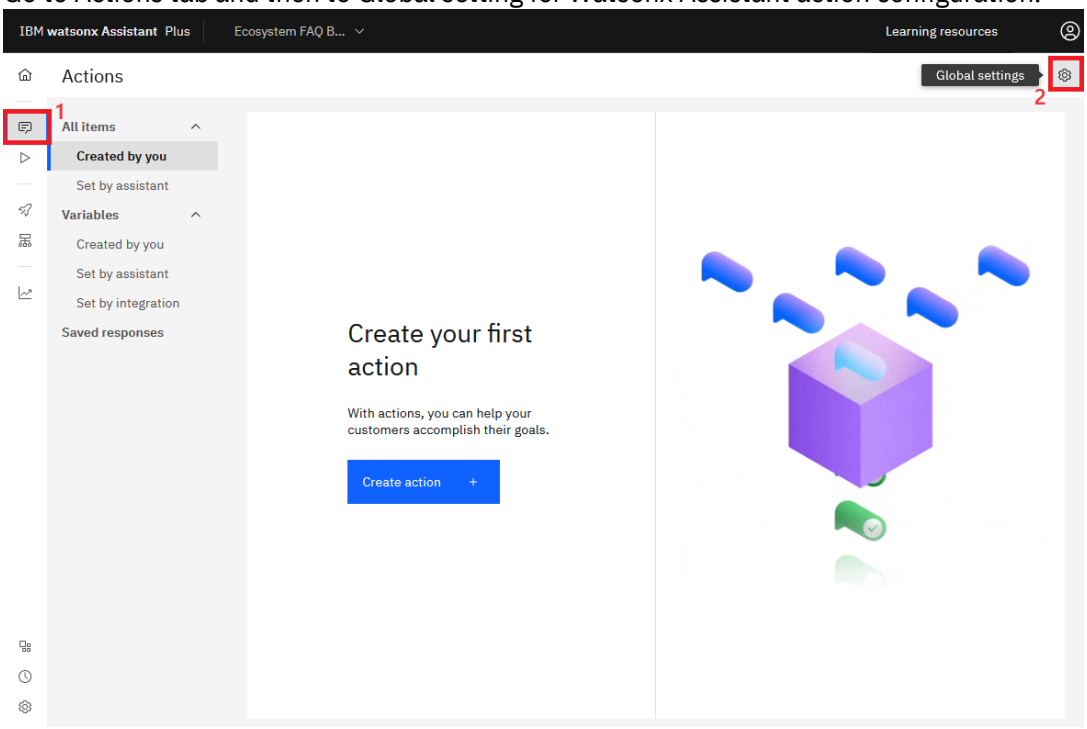


6. You have create a Watsonx Assistant instance.

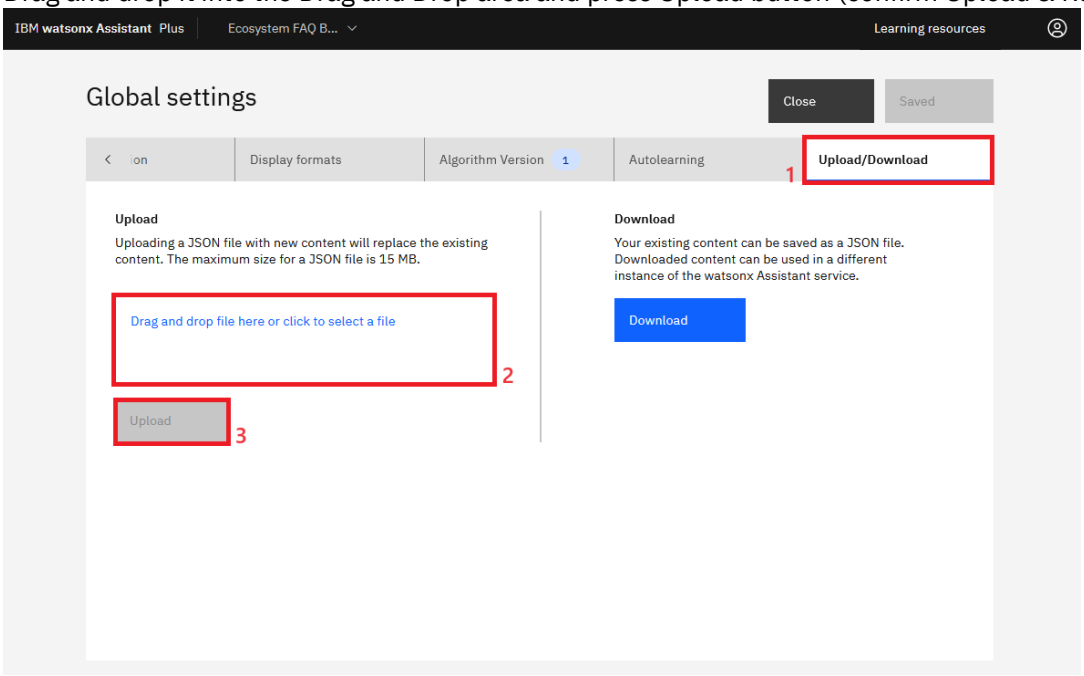


5.2 Upload Watsonx Assistant actions

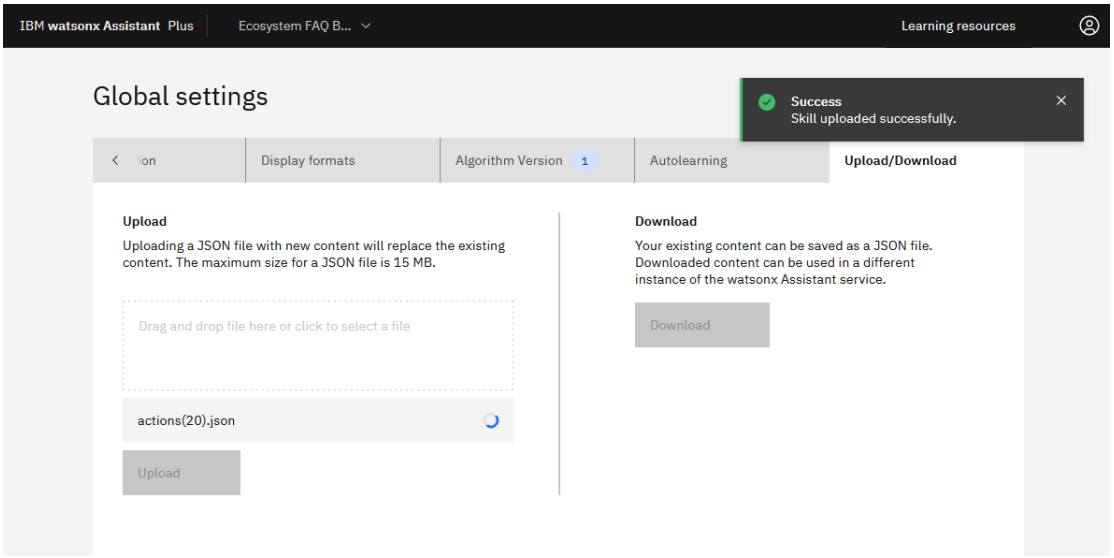
1. Go to Actions tab and then to Global setting for Watsonx Assistant action configuration.



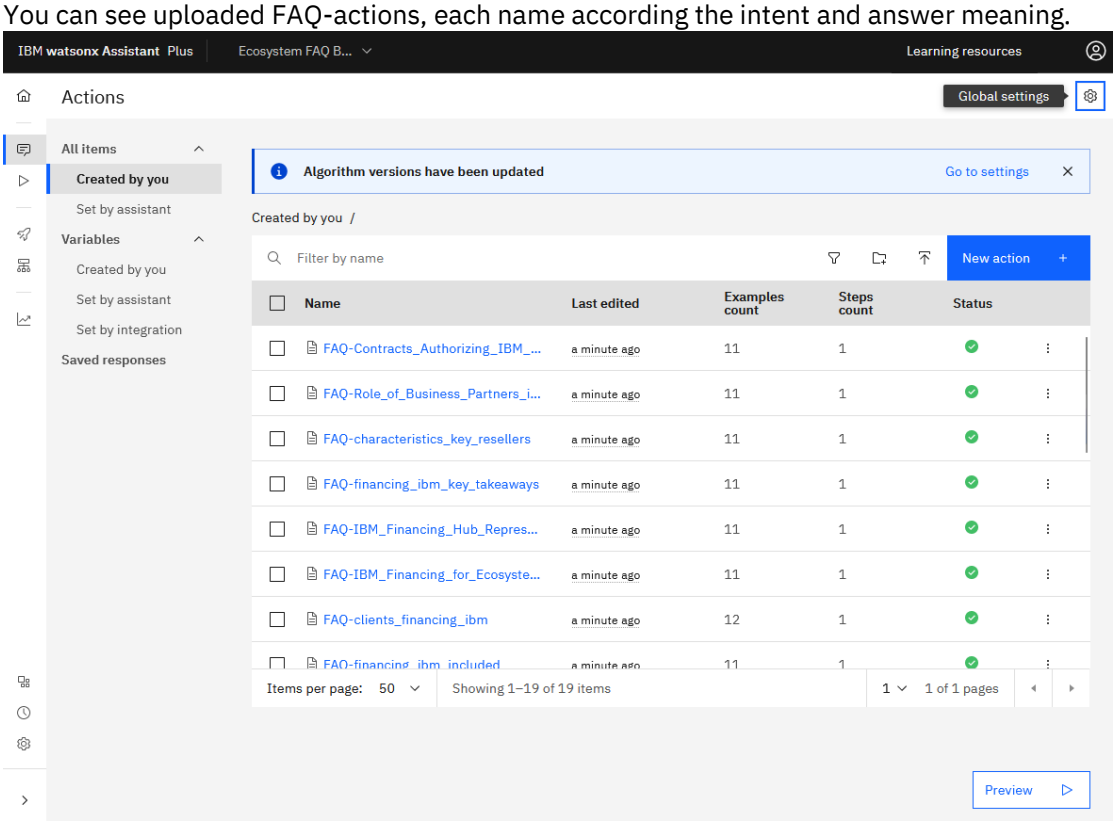
2. Click Upload / Download tab.
3. Take your Watsonx Assistant action file either exported from Neuralseek or the one from GITHUB: https://github.ibm.com/martin-rysanek/BARCELONA-LAB1531/wa-chatbot-action.json
Drag and drop it into the Drag and Drop area and press Upload button (confirm Upload & Replace).



4. After while, the action are uploaded to Watsonx Assistant with new FAQ Base configuration, press Close button.



5. You can see uploaded FAQ-actions, each name according the intent and answer meaning.



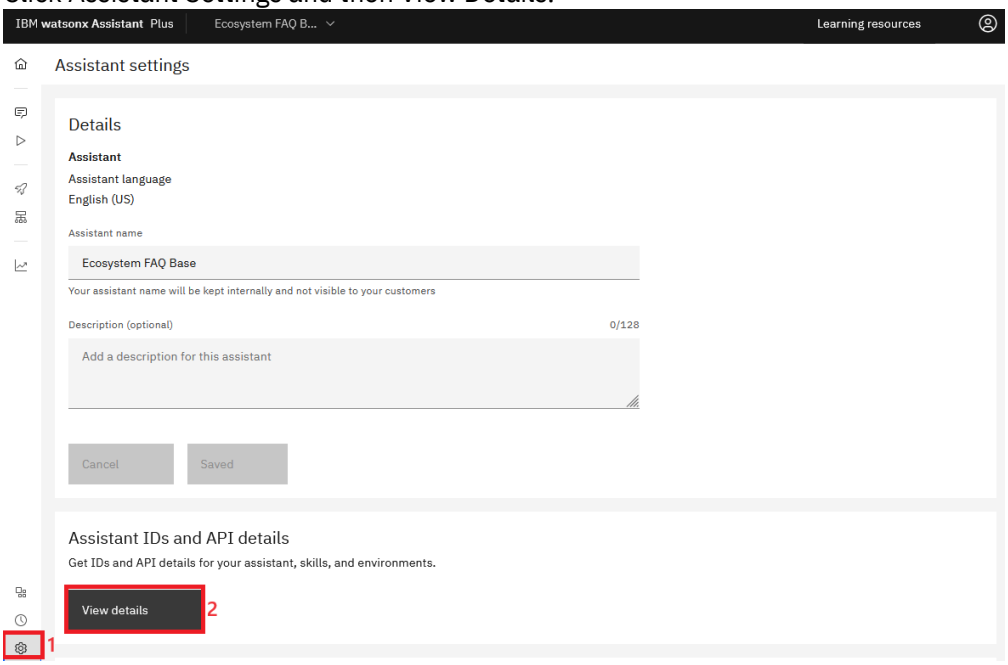
6. Click Preview and put some simple question like “IBM Partner Financing” to test, it works.

5.3 WA API key, URL and Assistant ID for later integration

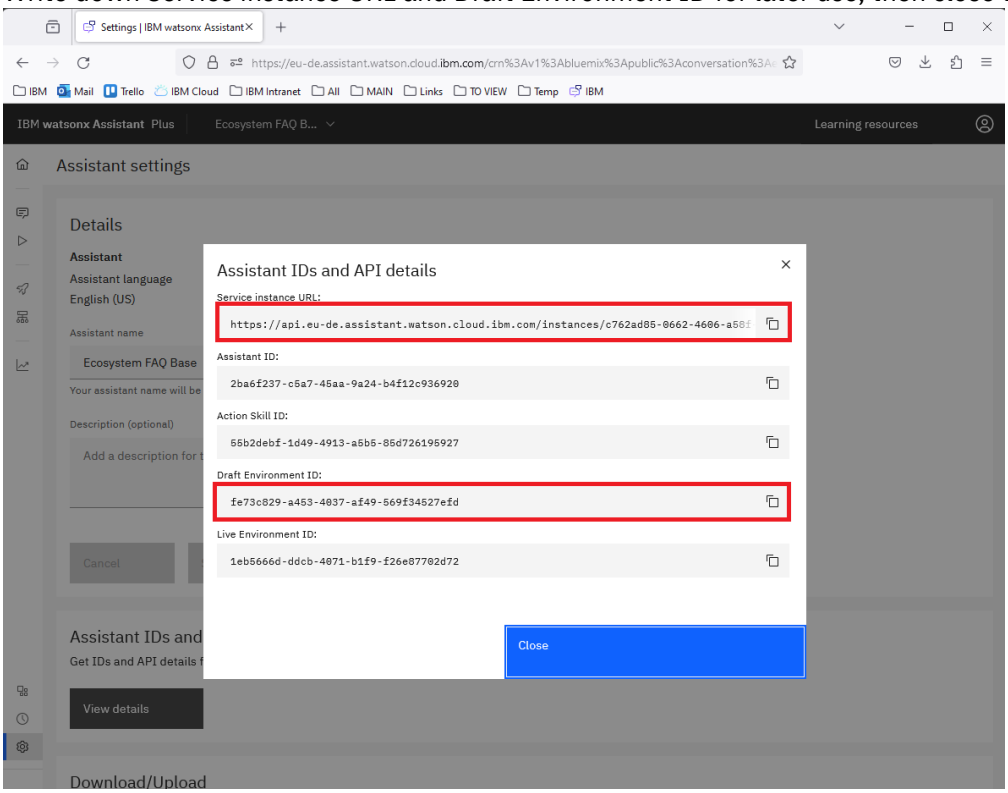
We need to get several integration items for future integration of Watson Assistant with FAQ Orchestrator / Code Engine extension.

The first, we need to get Watsonx Assistant URL and Assistant ID.

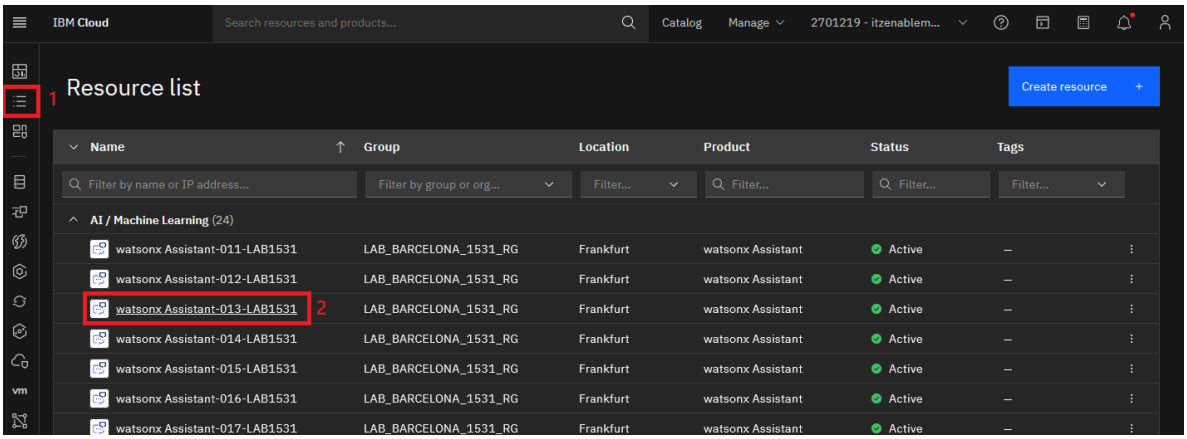
1. Click Assistant Settings and then View Details.



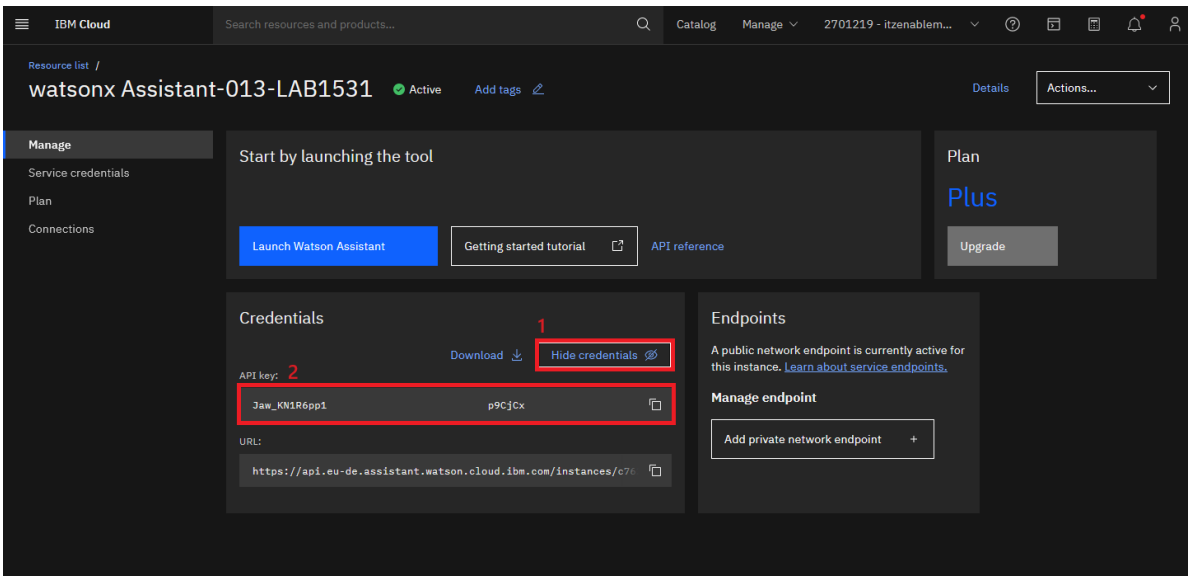
2. Write down Service instance URL and Draft Environment ID for later use, then close the window.



3. Leave Watsonx Assistant (we do not need it anymore) and open again list of LAB services by Open LAB Resource list as you can see below. Find and click prepared Watsonx Assistant service dedicated to you, i.e. service under the name watsonx Assistant-YOUR_ID-LAB1531, where YOUR_ID is your ID in the LAB in format 0xx (e.g. 013) again.



4. Click Show credentials button (now Hide credentials on the picture) and write down **API key** for Watsonx Assistant.



5. Now, you have the API Key, Watsonx Assistant URL and Assistant ID for configuration of FAQ Orchestrator / Code Engine environment variables.

6 Setup of FAQ Orchestrator – WA Custom Extension

FAQ Orchestrator acts as Watsonx Assistant Customer Extension implemented by IBM Cloud Code Engine (serverless container), which helps to orchestrate Watsonx Assistant Front End Chatbot and Watsonx Assistant FAQ Base in getting list of FAQ for given query.

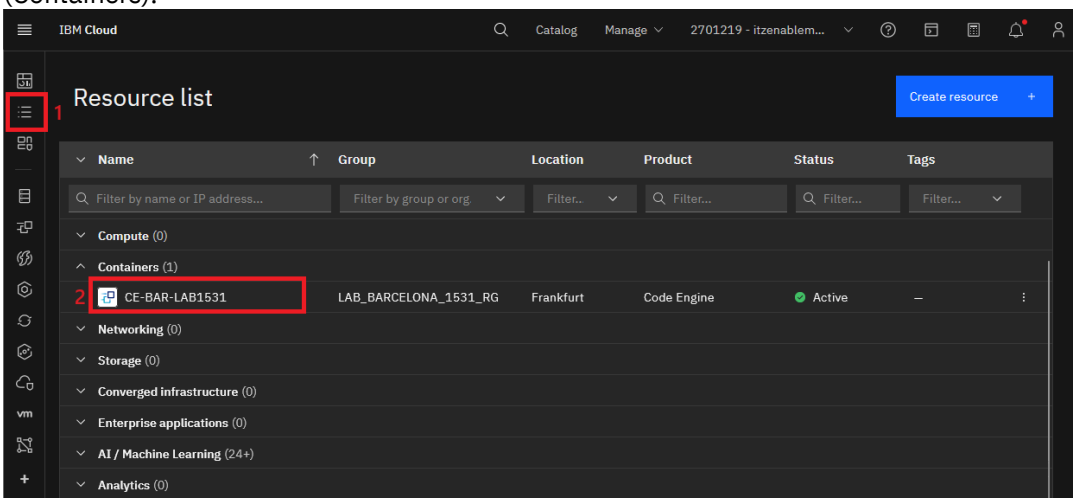
FAQ Orchestrator / Code Engine service is already provisioned in IBM Cloud for you, each LAB student has its own service. The service is running under CE-BAR-LAB1531 project, and the application has following name pattern ceapp-YOUR_ID-lab1531 pattern (e.g. ceapp-013-lab1531 in case of student with 013 ID).

The Code Engine application is configured to be started on demand and runs for 30 minutes from last request (to save IBM Cloud resources a bit). The Application on query request find N most relevant FAQ intents managed by Watsonx Assistant FAQ Base instance, which are upon the request passed back to Watsonx Assistant Chatbot. The Watsonx Assistant Chatbot display specific intent / answer upon user selection. For monitoring purpose, the application records user's query, selected intent, intent confidence and user's ranking of answer to give question evaluated by user.

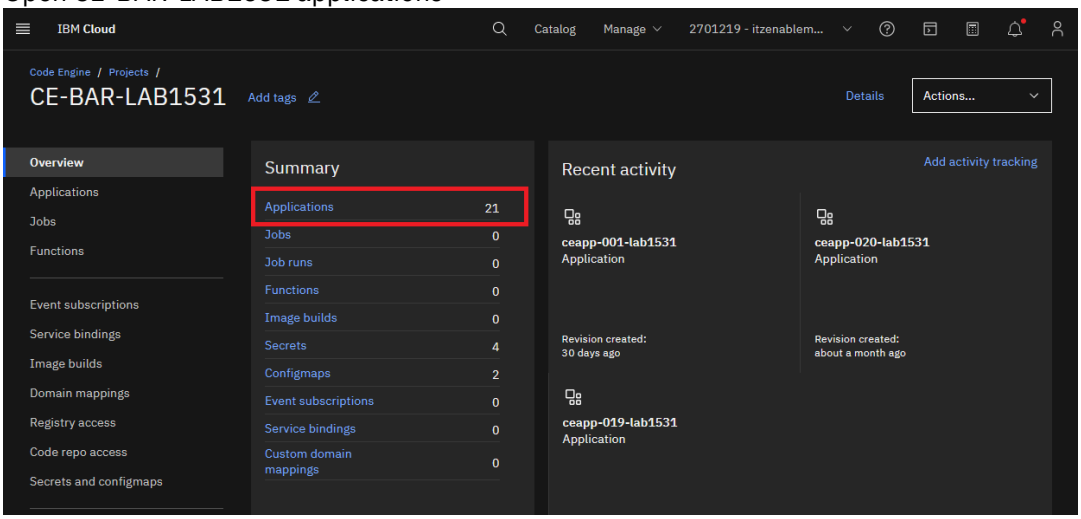
By analyses of user's selections (recorded by FAQ Orchestrator), we know, what queries should be batch/offline evaluated by Neuralseek, verified by subject matter expert and exported as WA FAQ Base configuration, to enable WA FAQ Base to better response to next users questions.

6.1 Setup FAQ Base Orchestrator Environment

1. Open Resource list of your LAB services and select CE-BAR-LAB1531 Code Engine Project (Containers).

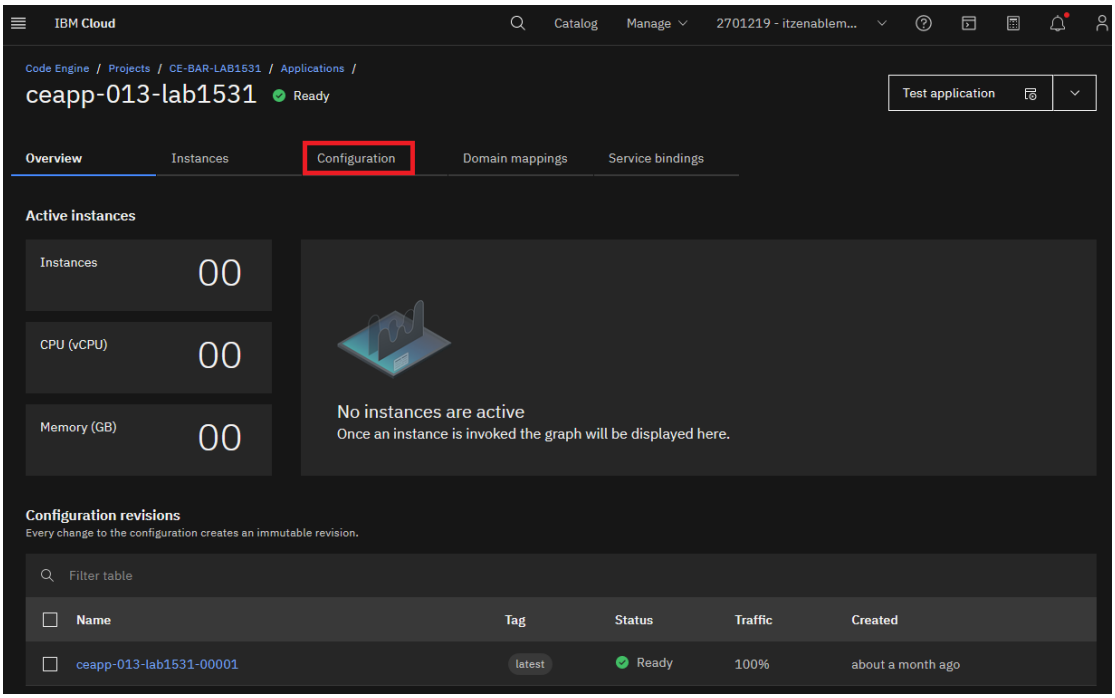


2. Open CE-BAR-LAB1531 applications

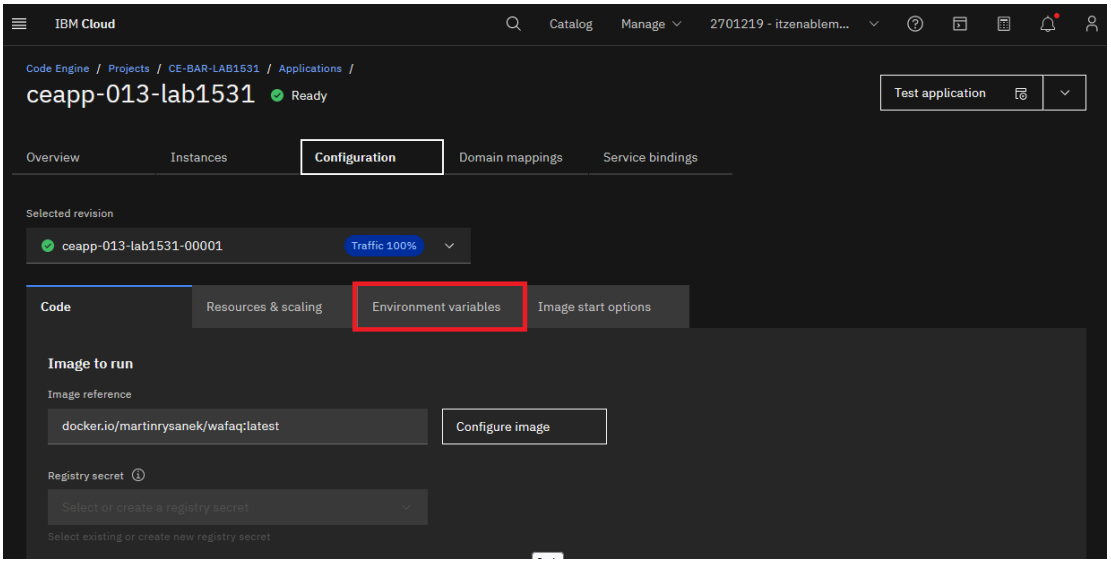


3. Find and click your Code Engine application, the application has following name pattern ceapp-YOUR_ID-lab1531 pattern (e.g. ceapp-013-lab1531 in case of student with 013 ID).

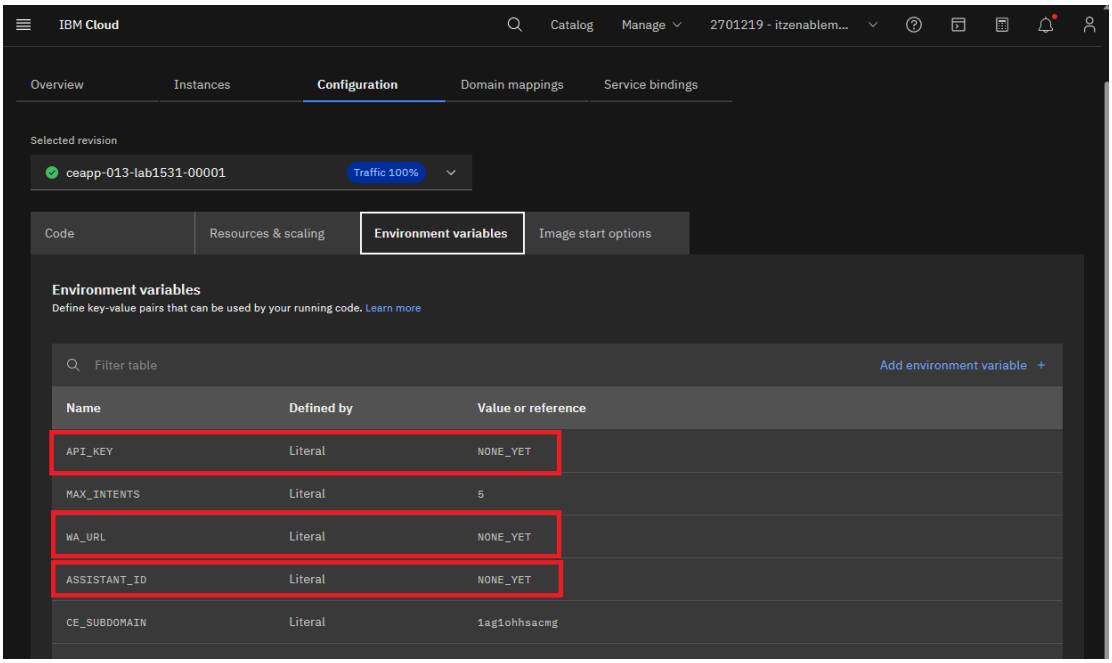
   Click Configuration Tab.



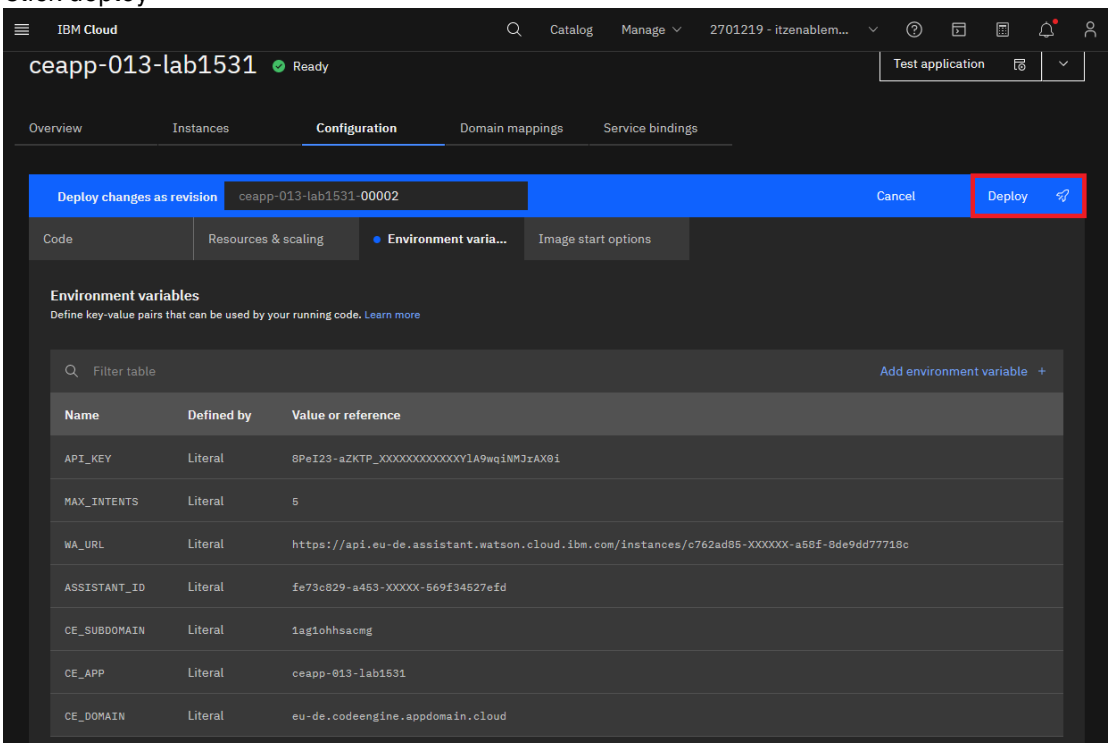
4. Click Environment variables



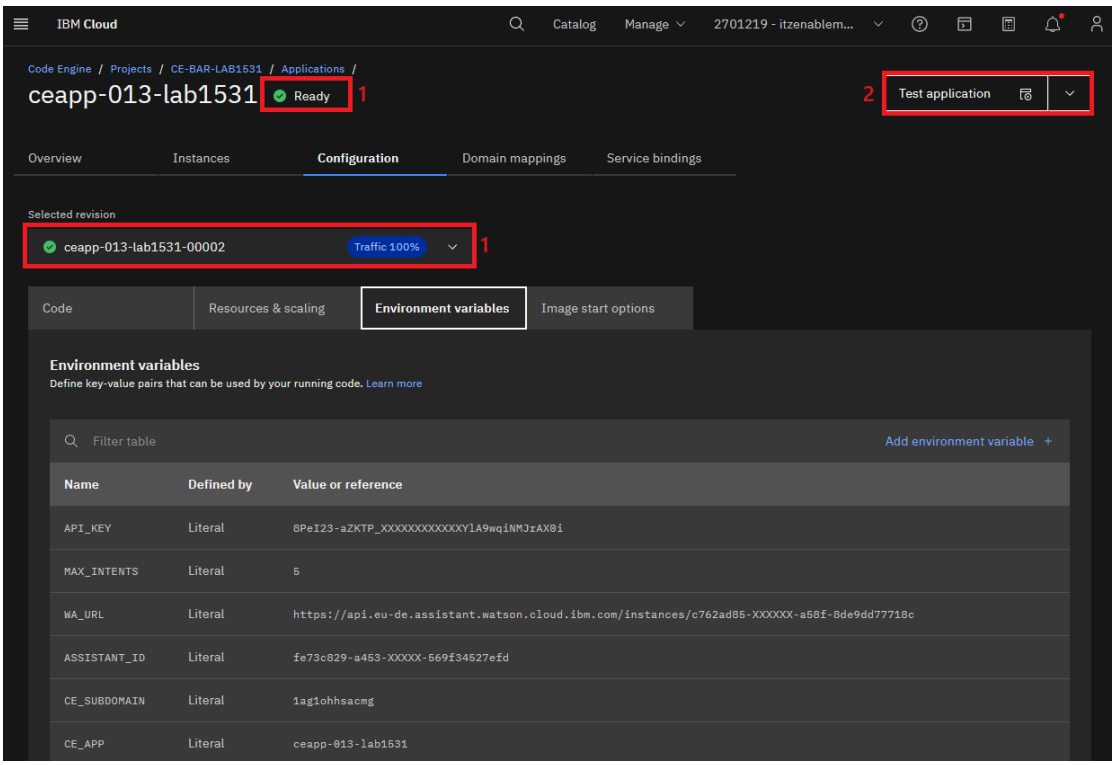
5. Now you need to replace all NONE_YETs by right values, you have recorded in previous chapter. Use recorded API_KEY, WA_URL and ASSISTAND_ID for Watsonx Assistan FAQ Base.



6. Click deploy

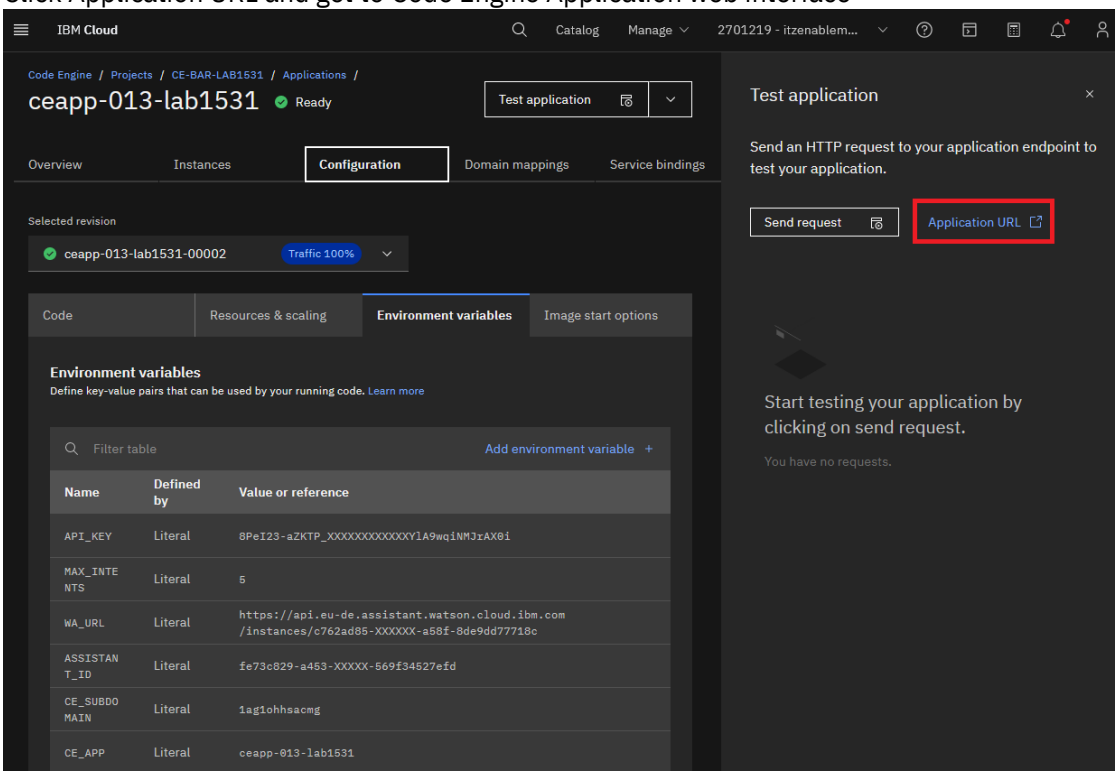


7. It takes like 2 minutes for IBM Cloud to create new Code Engine Application instance and start it. Wait until the Code Engine Application is ready with new environment variables.
Then Click Test Application.



6.2 Open and work with FAQ Orchestrator Web Interface

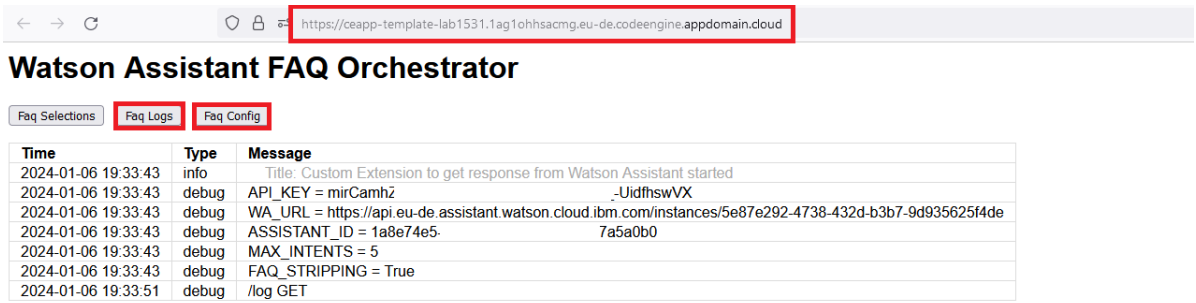
1. Click Application URL and get to Code Engine Application web interface



2. You can see running Code Engine Application and its Faq Logs, which show that Code Engine was just started with configured Environment variables of API_KEY, WA_URL and ASSISTANT_ID.

   You can now configure FAQ Config with number of max intents (MAX_INTENTS) passed from called linked Watsonx Assistant Custom Extension. You can switch back to the screen by clicking Faq Logs.

   Important is, **write down URL of the FAQ Orchestrator / Code Engine** (top rectangle), the URL will be used for configuration of Custom Extension of Watsonx Assistant Chatbot.



Note: be aware, that FAQ Orchestrator / Code Engine Application is up and running for 30 minutes than is automatically destroyed (for production purpose, the FAQ Orchestrator can be permanently up). To start sleeping FAQ Orchestrator / Code Engine Application, it again takes like 1 minute by calling the application either by its web interface or directly from Watsonx Assistant Chatbot.

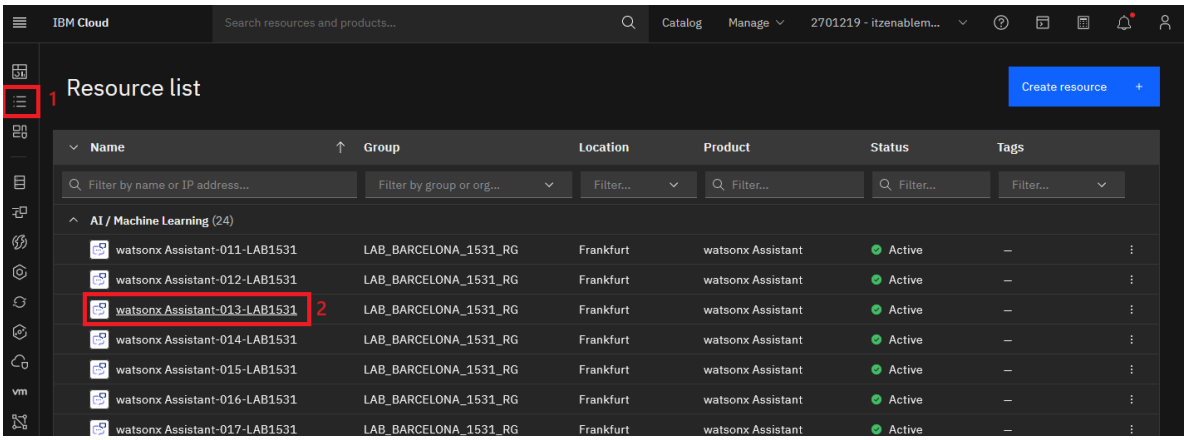
7 Setup Watsonx Assistant Chatbot

The main Watsonx Assistant instance will act as standard chatbot front end being able to respond to any query by predefined dialog action flow. In the case, the intent is not recognized by Watsonx Assistant NLP classification, Watsonx Assistant triggers the system action called “No action matches”. There is a first step system action “No action matches”, which transfer execution our Search action. Search action forward query to Custom Extension, which get for given query list of matching FAQs from the KnowledgeBase represented by Watsonx Analytics FAQ base.

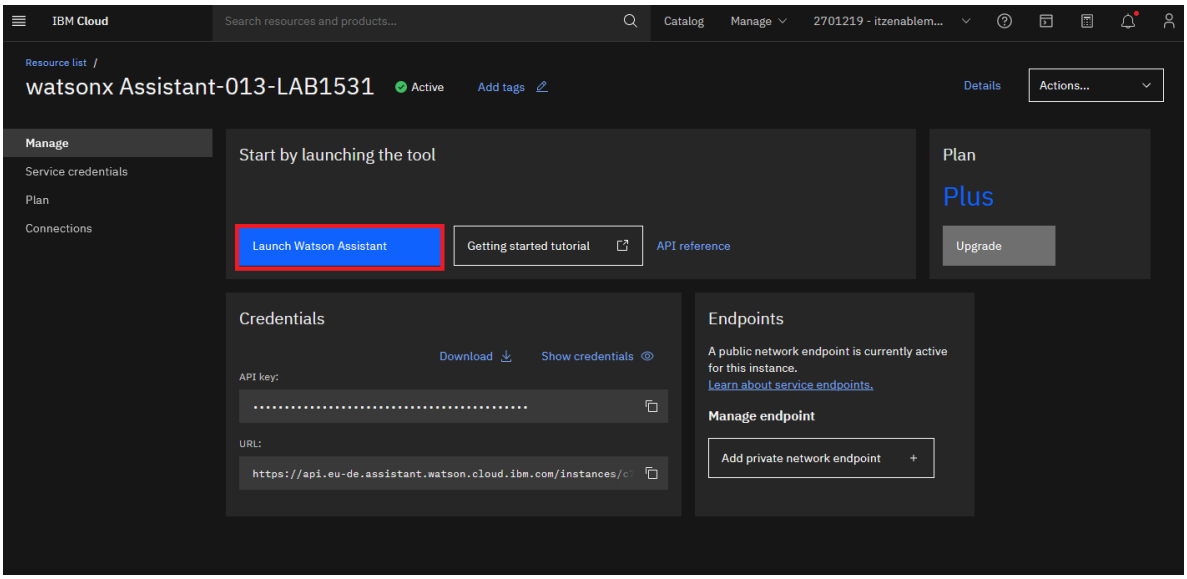
Let us know setup Watsonx Assistant Chatbot and upload it with initial actions.

7.1 Open and Setup Watsonx Assistant Chatbot instance

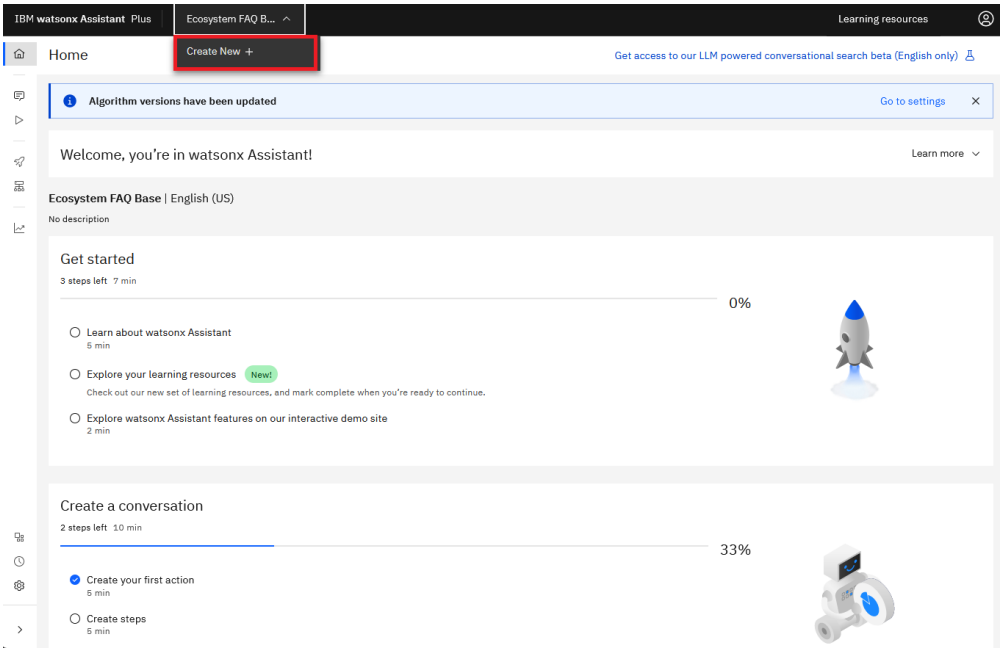
1. Open LAB Resource list, find and click your prepared Watsonx Assistant under the name watsonx Assistant-YOUR_ID-LAB1531, where YOUR_ID is your ID in the LAB in format 0xx (e.g. 013) as we already did for Ecosytem FAQ BASE Watsonx Assistant.

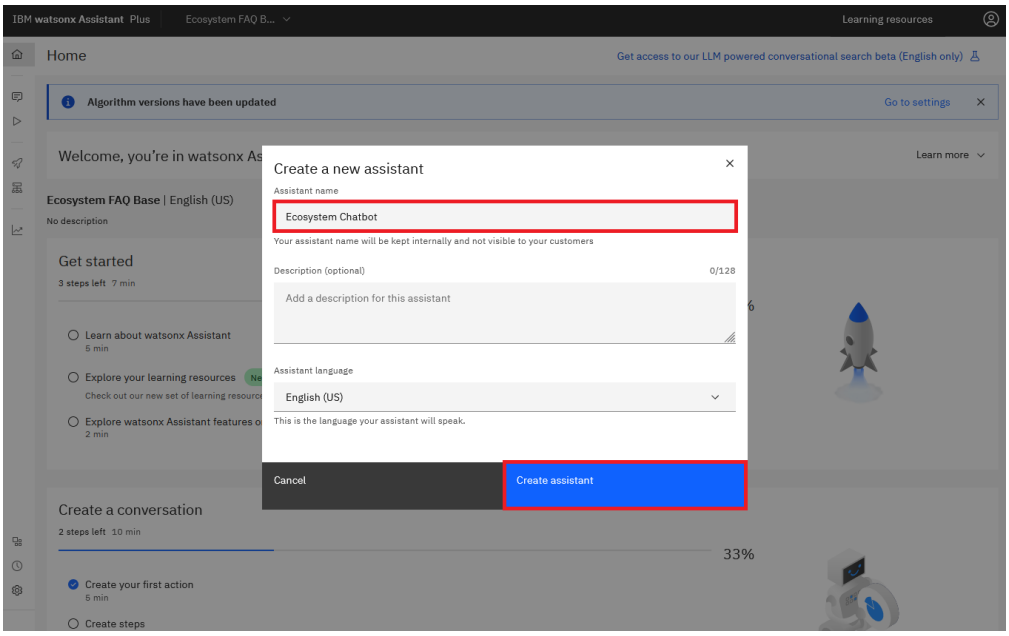


2. Lauch Watsonx Assistant



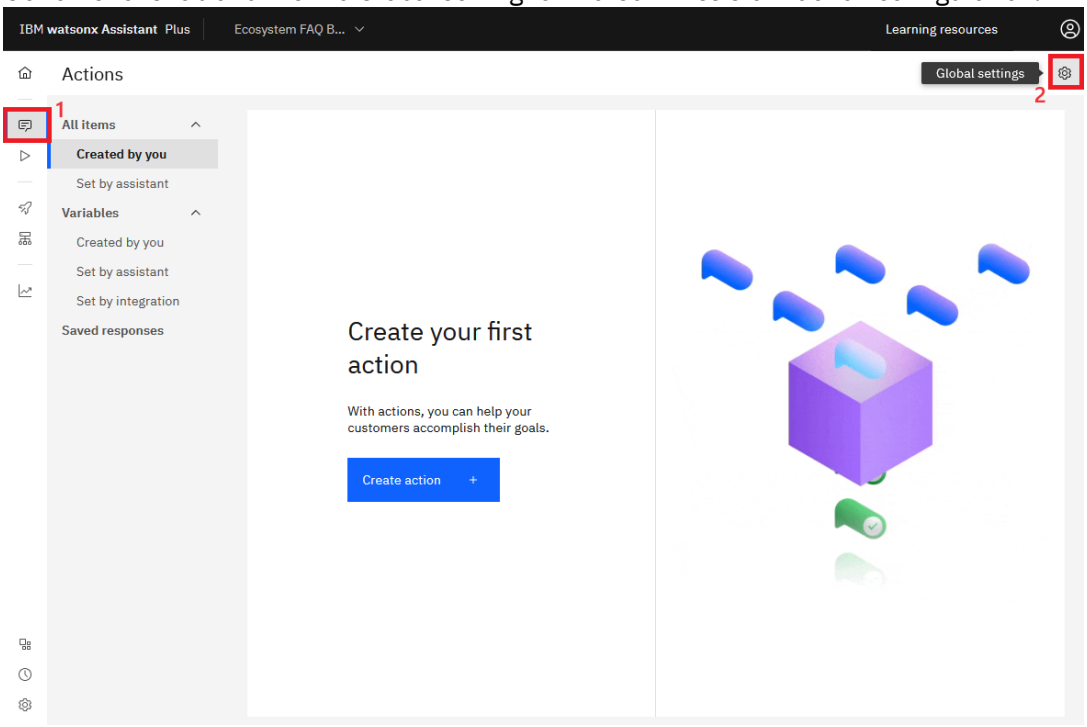
3. From many select Create New+ and name the Watsonx Assistant instance, e.g. Ecosystem Chatbot, click Create assistant.



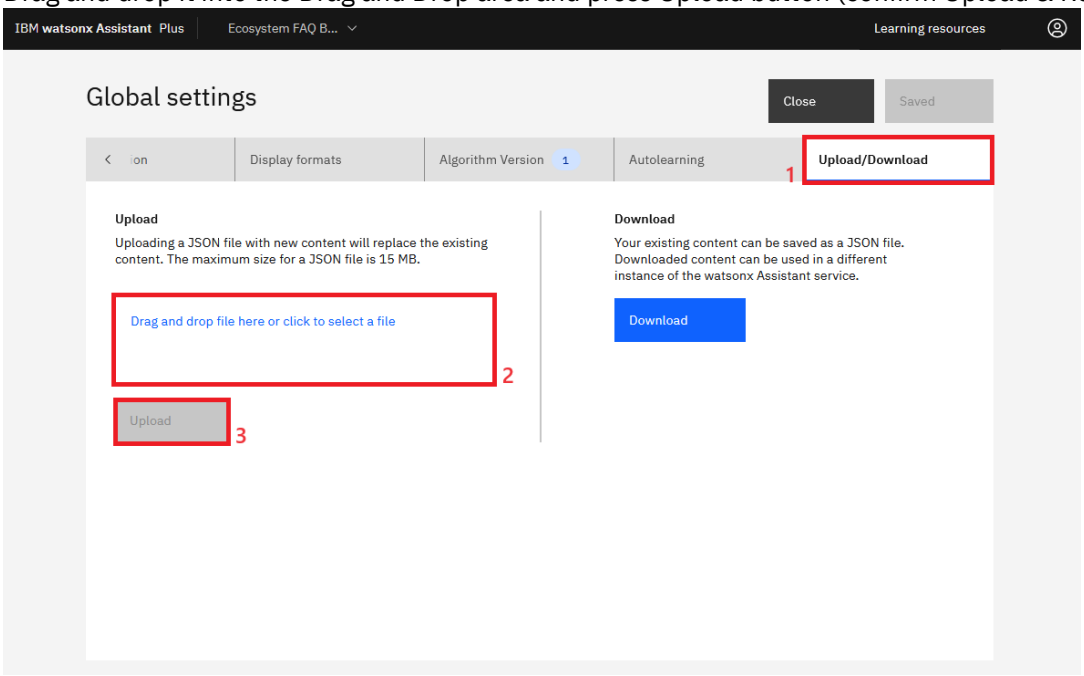


7.2 Upload Watsonx Assistant Chatbot actions

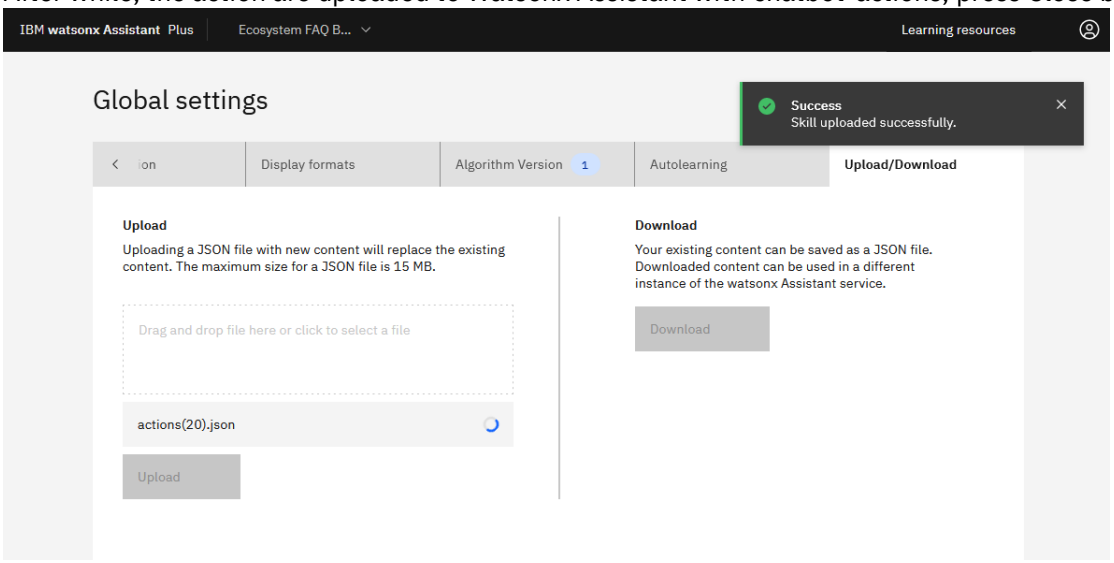
1. Go to Actions tab and then to Global setting for Watsonx Assistant action configuration.



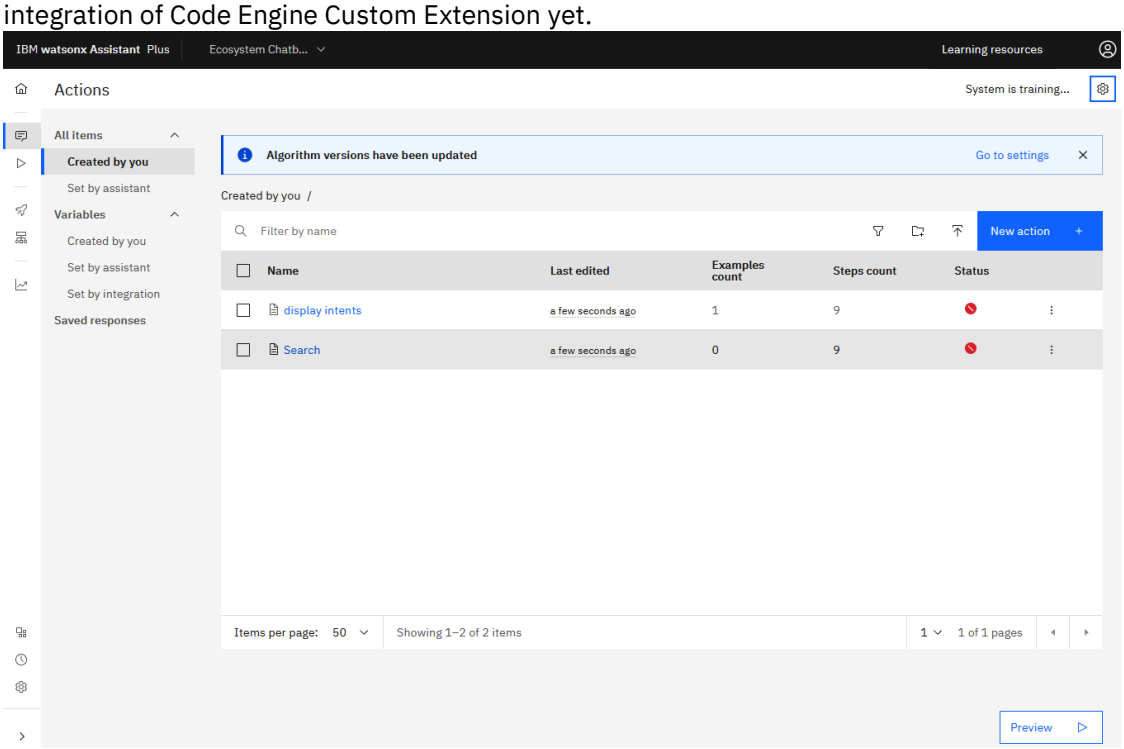
2. Click Upload / Download tab.
3. Take your Watsonx Assistant Chatbot action file from GITHUB: https://github.ibm.com/martin-rysanek/BARCELONA-LAB1531/wa-chatbot -action.json
Drag and drop it into the Drag and Drop area and press Upload button (confirm Upload & Replace).



4. After while, the action are uploaded to Watsonx Assistant with chatbot-actions, press Close button.



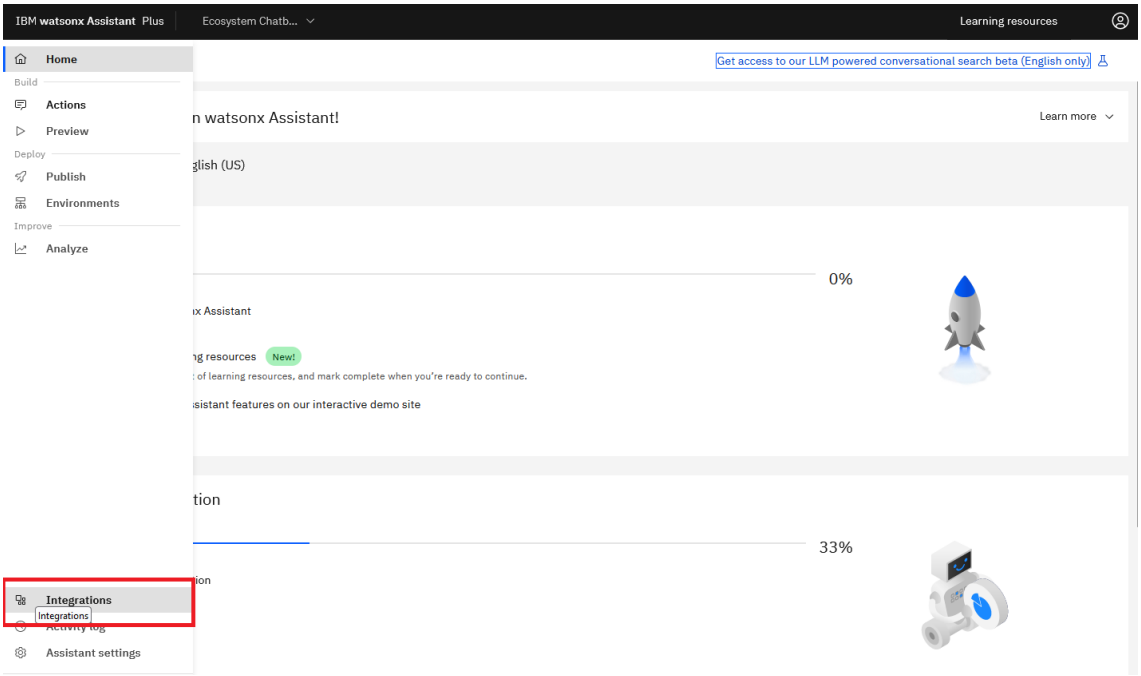
5. You can see uploaded chatbot-actions and each action with red status, because there is missing integration of Code Engine Custom Extension yet.



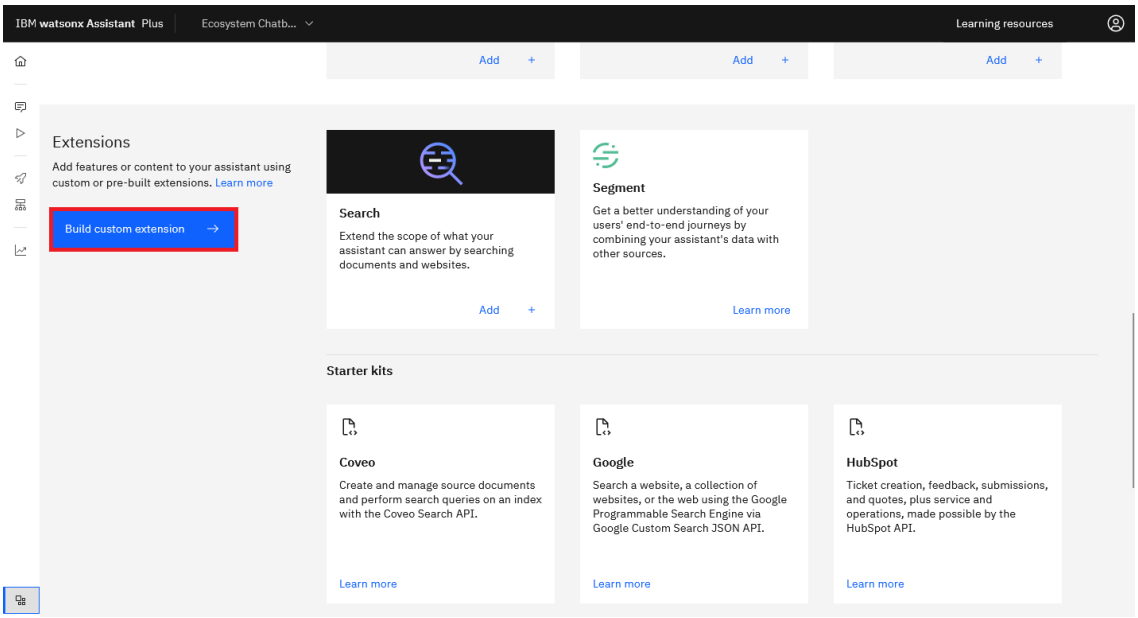
Note: The Watsonx Assistant Chatbot can be further extended by additional intents / actions. Chatbot than responds primarily to these intents / actions and then it searches for intents / answers using FAQ KnowledgeBase. Whenever you update Watsonx Assistant FAQ Base, it does not influence intents / actions stored permanently in Watsonx Assistant Chatbot.

7.3 Integration with FAQ Orchestrator / Code Engine

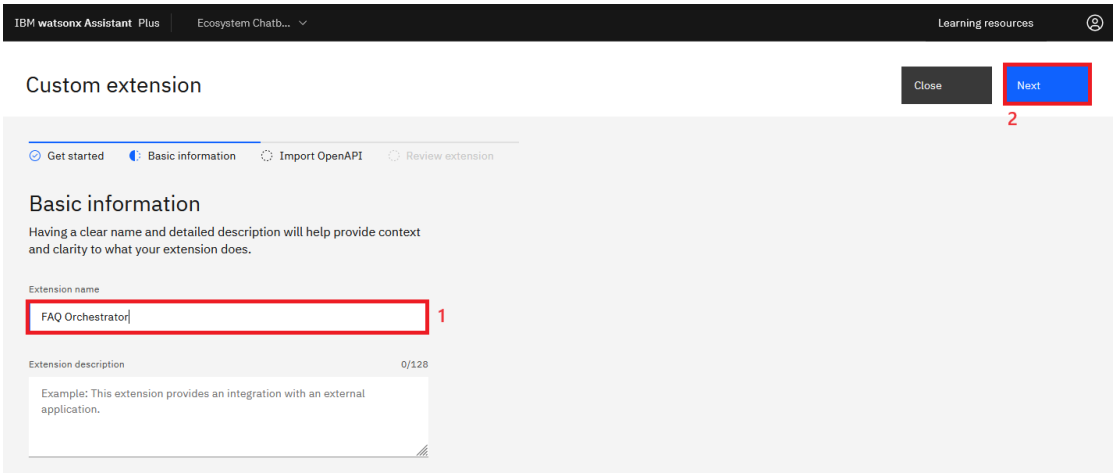
1. Click on Integrations on left side.



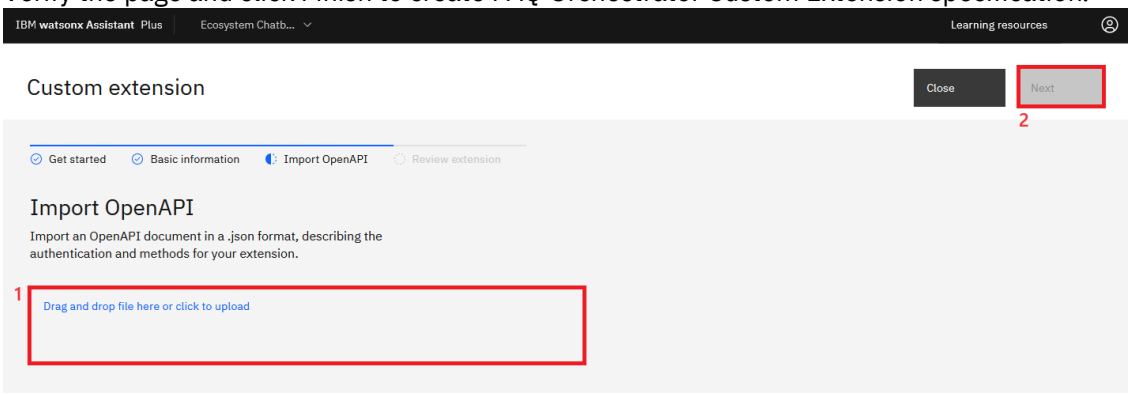
2. Scroll down and click Build custom extension



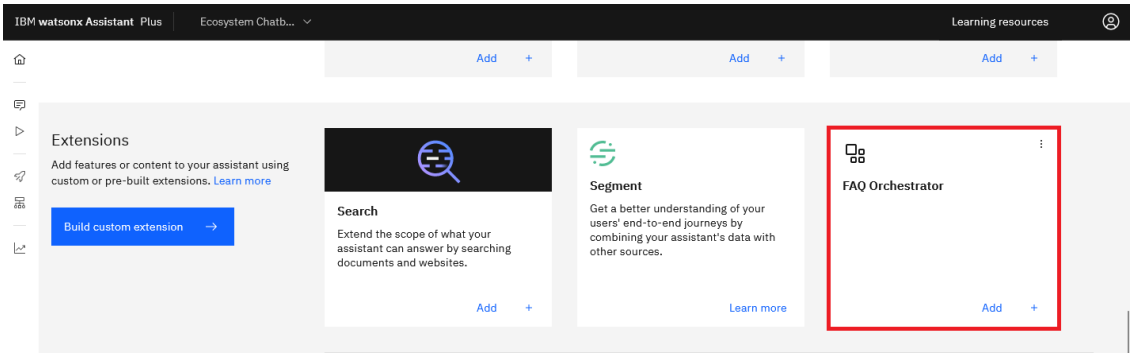
3. In Custom Extension, click Next, name Extension “FAQ Orchestrator” and click Next



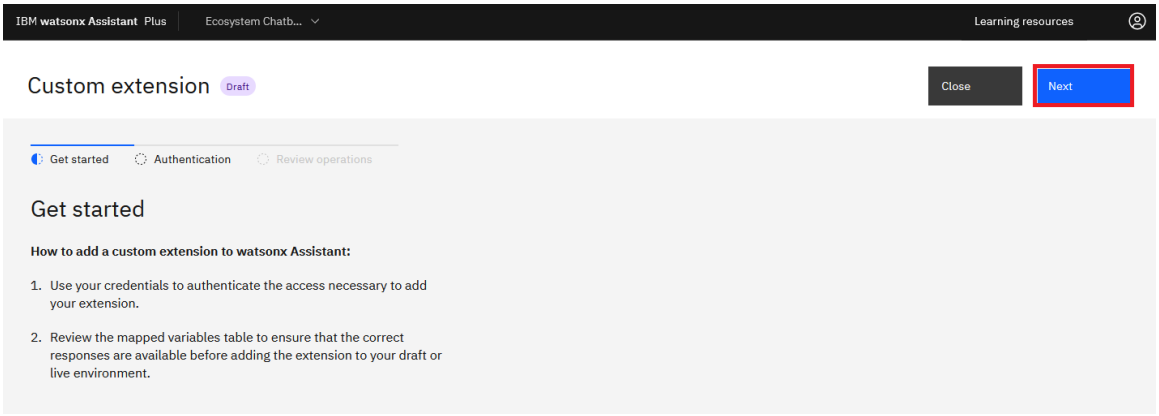
4. Download FAQ-Orchestrator-openapi.json file from GITHUB:
https://github.ibm.com/martin-rysanek/BARCELONA-LAB1531/faq-orchestrator-openapi.json
Drag and Drop file into the area and click Next.
Verify the page and click Finish to create FAQ Orchestrator Custom Extension specification.



5. The new tile appears among Custom Extensions, click Add blue text then click Add button again to create instance of FAQ Orchestrator.

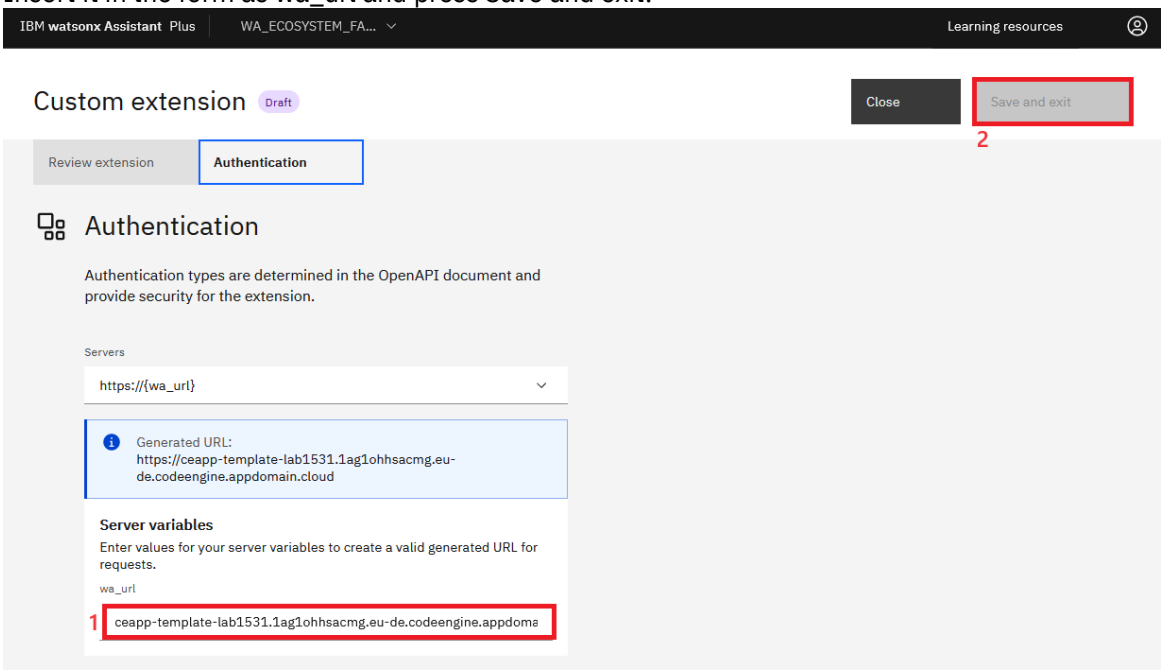


6. Click Next



7. You have written down URL of FAQ Orchestrator in previous chapter. Now extract only domain part of it (e.g. there is an URL https://ceapp-template-lab1531.1ag1ohhsacmg.eu-de.codeengine.appdomain.cloud/ and you need to insert there only domain part: ceapp-template-lab1531.1ag1ohhsacmg.eu-de.codeengine.appdomain.cloud).

   Insert it in the form as wa_url and press Save and exit.



We have FAQ Orchestrator Custom Extension configured and we can start experiment querying FAQ system.

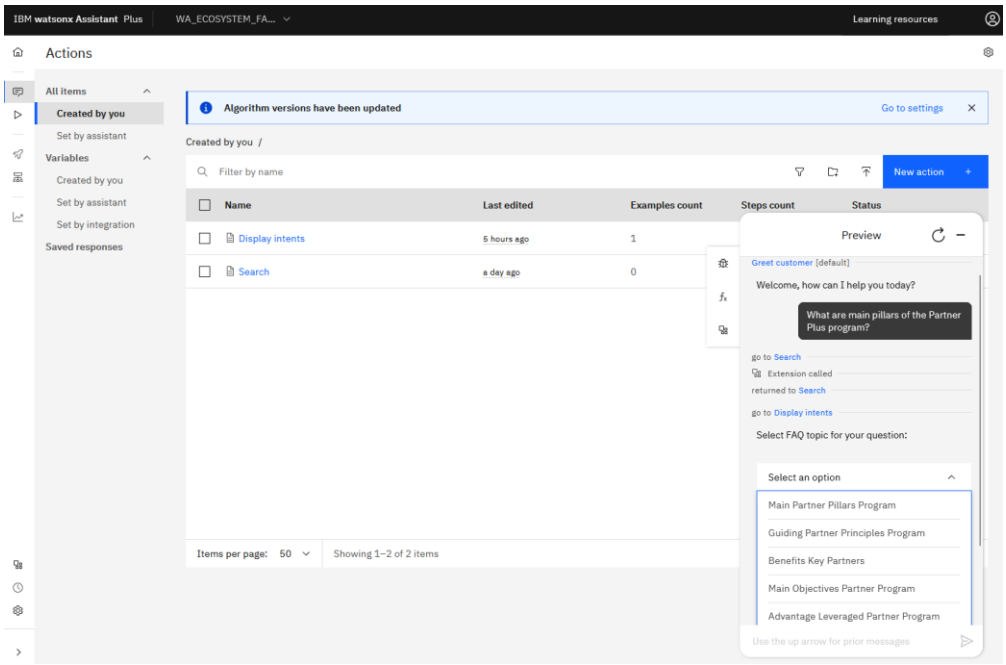
8 Experimenting with Chatbot

Now, we have a complete FAQ system of AI-generated and subject matter expert edited and verified questions + answers system.

Let's start experimenting with the system's ability to answer user queries and learn about the analytical features provided by FAQ Orchestrator. FAQ Orchestrator monitor user's question and FAQ answers, record user's FAQ selections. The FAQ Orchestrator provides a dataset of the user's selection of the most common questions, and by analyzing it we can identify those questions that need to be newly batch generated by Neuralseek and further verified by a subject matter expert.

8.1 Setup Web Chat for Watsonx Assistant Chatbot

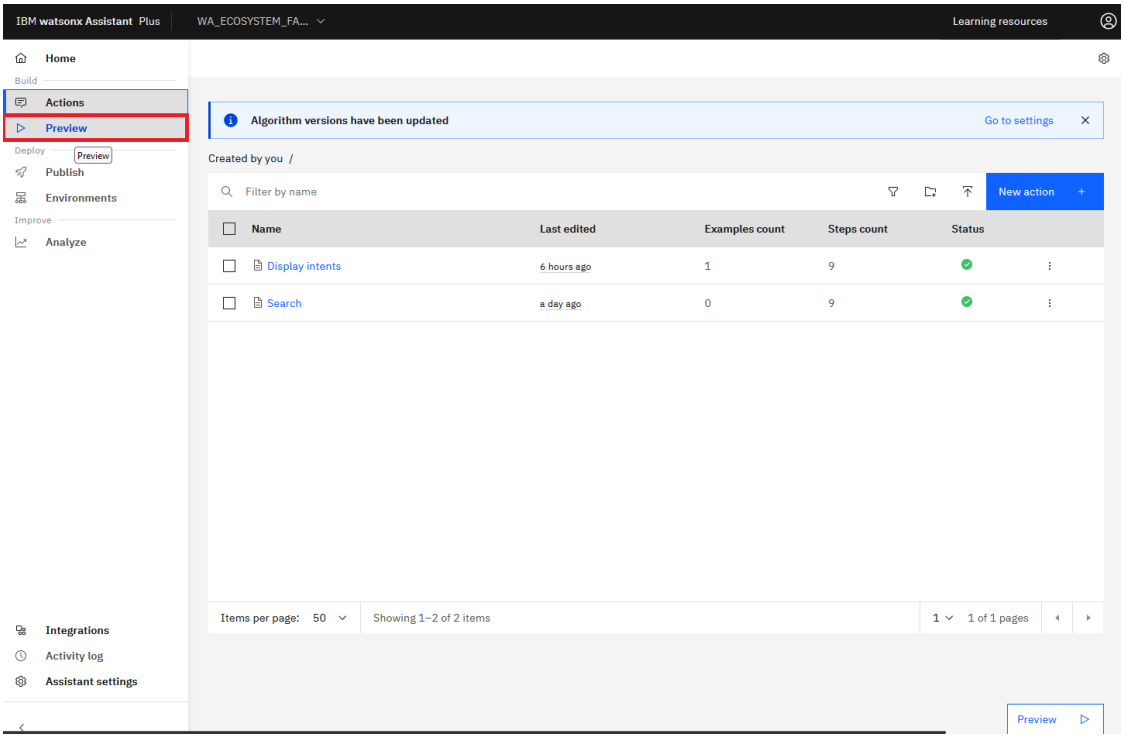
For experimenting with FAQ system, you can use directly preview of Watsonx Assistant Chatbot, but it is not as user friendly for end users as web chat interface.



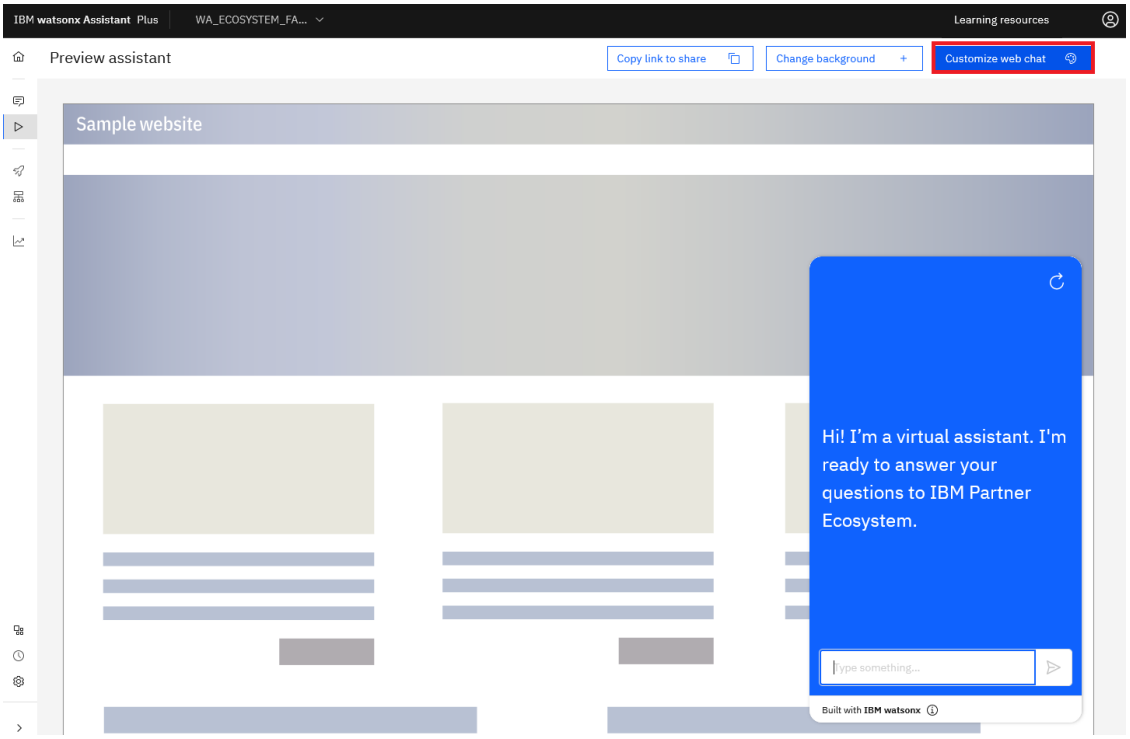
Let us setup Web Chat for Watsonx Assistant, which provides end users with better user interface.

1. Download wa-webchat.zip from GITHUB, which includes 2 files.
   https://github.ibm.com/martin-rysanek/BARCELONA-LAB1531/wa-webchat.zip
   Unzip index.html and .jpg files from wa-webchat.zip

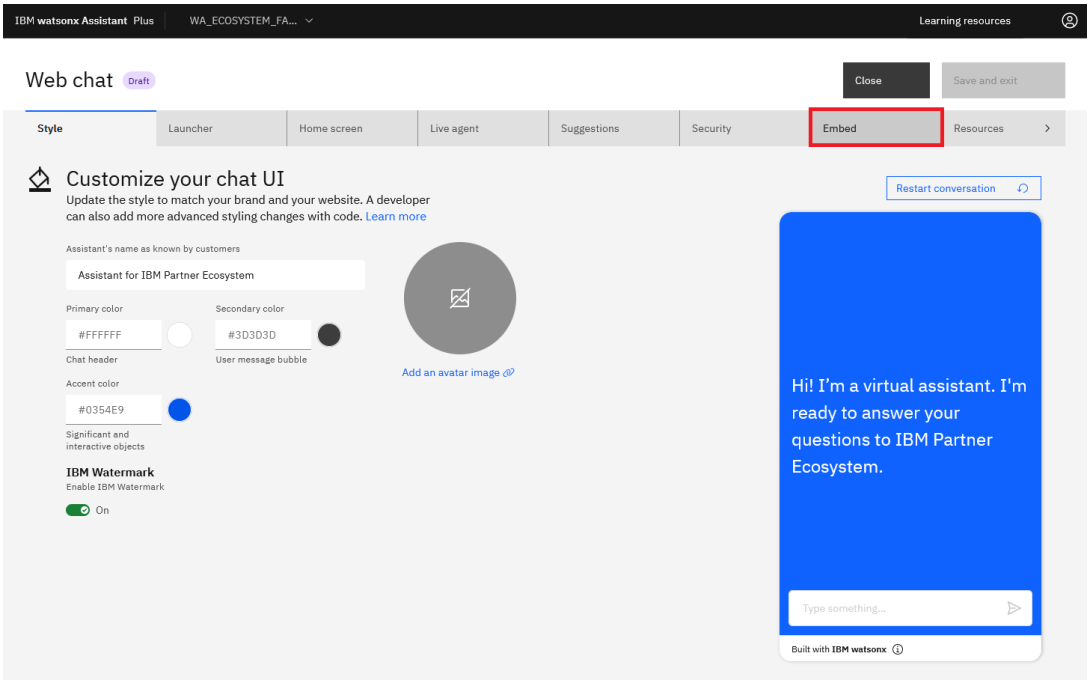
2. Open Preview on the left side menu of Watsonx Assistant.



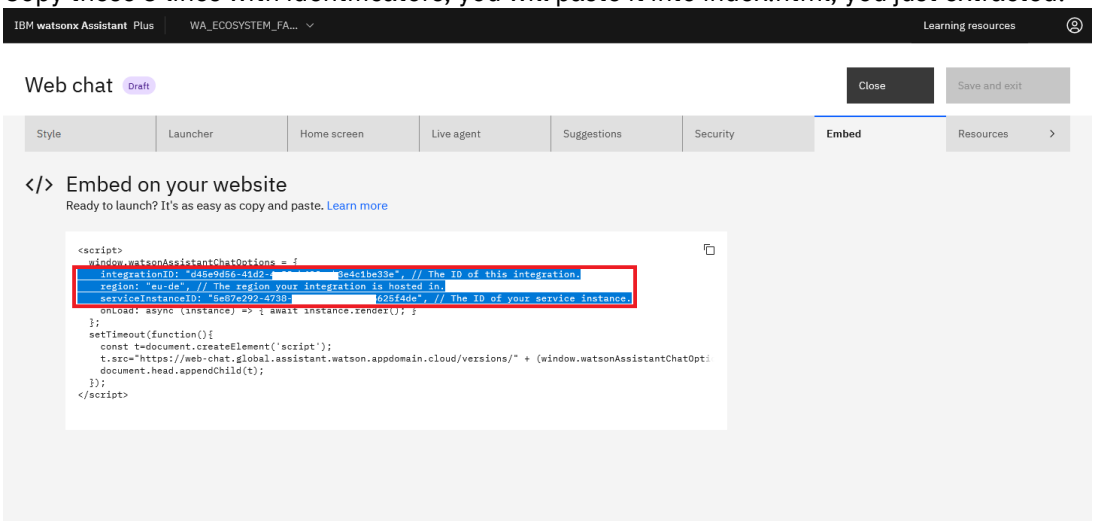
3. Select Customize Web Chat



4. Click Embed



5. Copy these 3 lines with identificators, you will paste it into index.html, you just extracted.



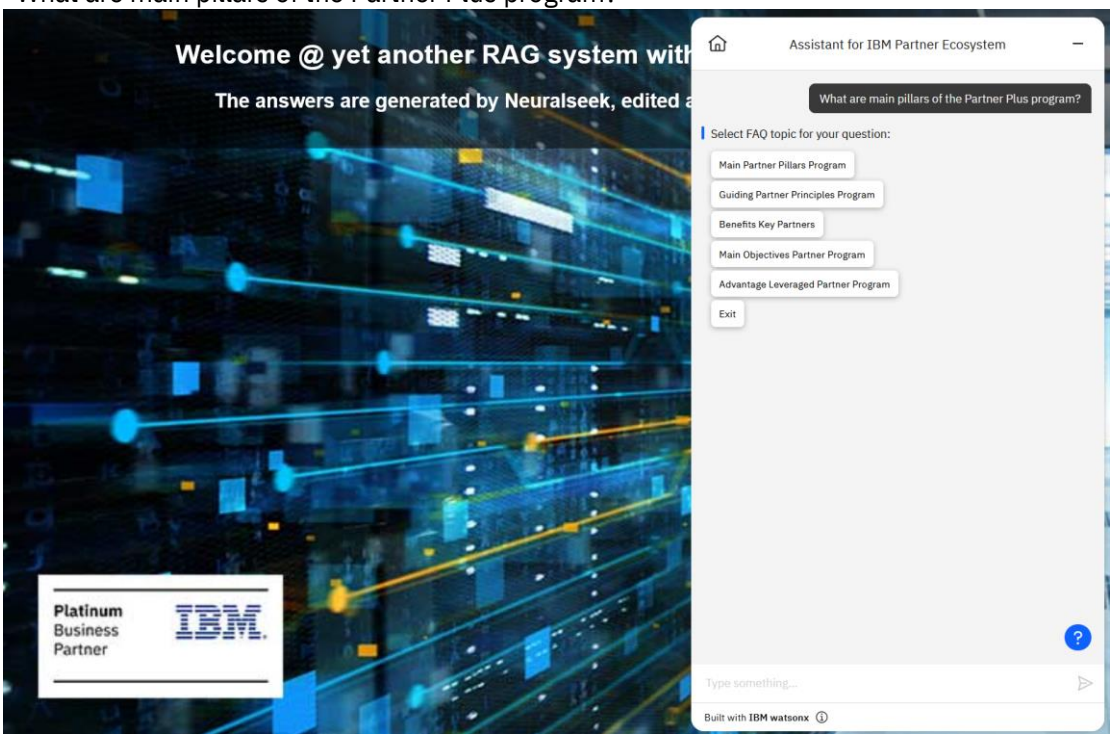
6. Open index.html, you just extracted from .zip file, find similar pattern in index.html (line 285) and replace it with what you just copy in previous step.
   Store the file to you disk and open index.html by your web browser.

```
281     // This is the standard web chat configuration object. You can modify these values with the embed code for your
282     // own assistant if you wish to try this example with your assistant. You can find the documentation for this at
283     // https://web-chat.global.assistant.watson.cloud.ibm.com/docs.html?to=api-configuration#configurationobject.
284     window.watsonAssistantChatOptions = {
285       integrationID: "d45e9d56-                    c1be33e", // The ID of this integration.
286       region: "eu-de", // The r                    on is hosted in.
287       serviceInstanceID: "5e87e                    d935625f4de", // The ID of your service instance.
288       element: customElement,
289       onLoad: onLoad,
290     };
291     setTimeout(function(){const t=document.createElement('script');t.src="https://web-chat.global.assistant.watson.ap
292   </script>
293     <div class="content">
294         <h1>Welcome @ yet another RAG system with verified questions.</h1>
295         <h2><p align="center">The answers are generated by Neuralseek, edited and verified by human.</p></h2>
296     </div>
297 </body>
298 </html>
```

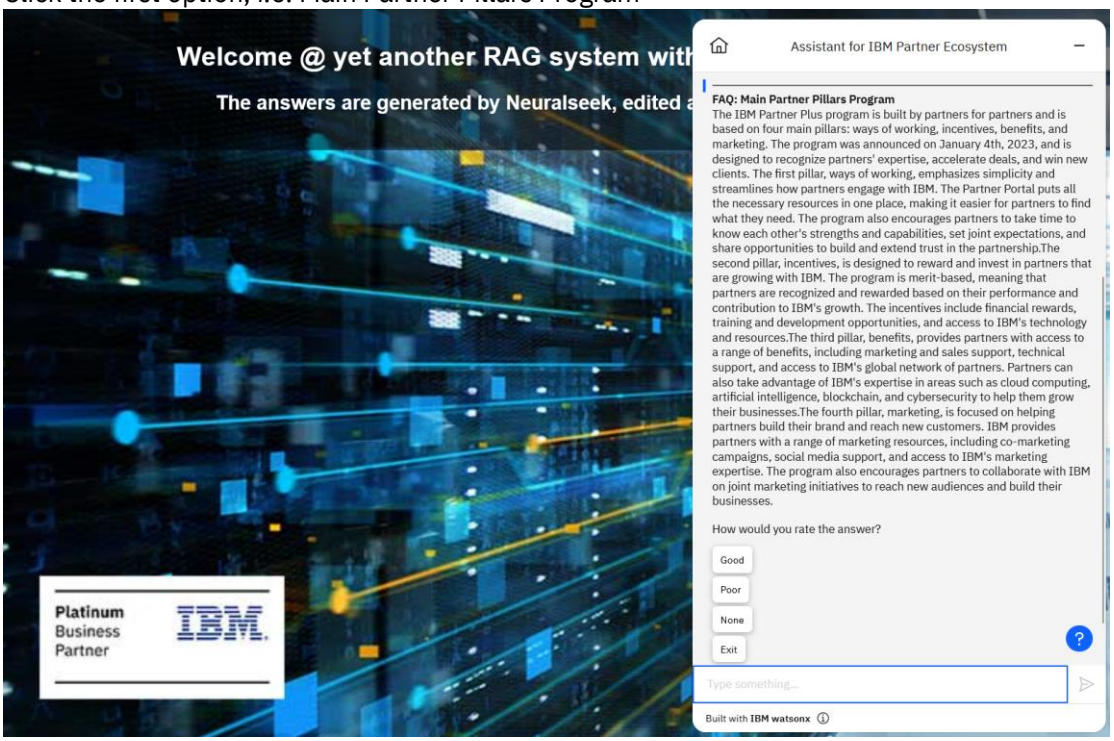
8.2 Chatbot responding by FAQ questions

There is predefined set of questions for our IBM Partner Ecosystem Knowledgebase on the GITHUB:
https://github.ibm.com/martin-rysanek/BARCELONA-LAB1531/ibm-partner-ecosystem-questions.txt

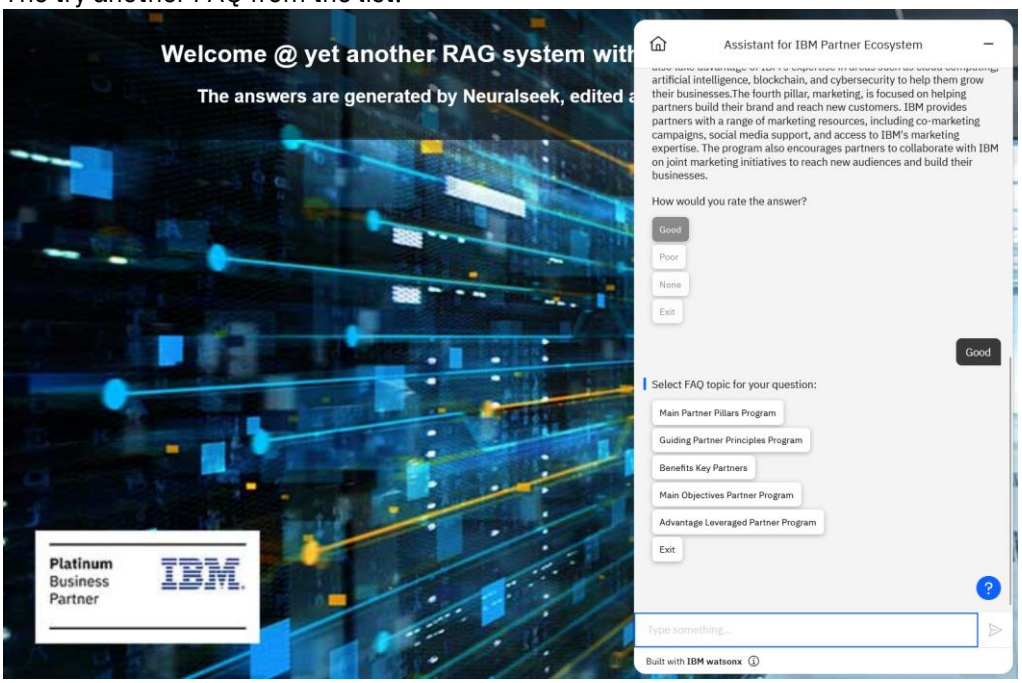
1. Let's try following question:
   "What are main pillars of the Partner Plus program?"



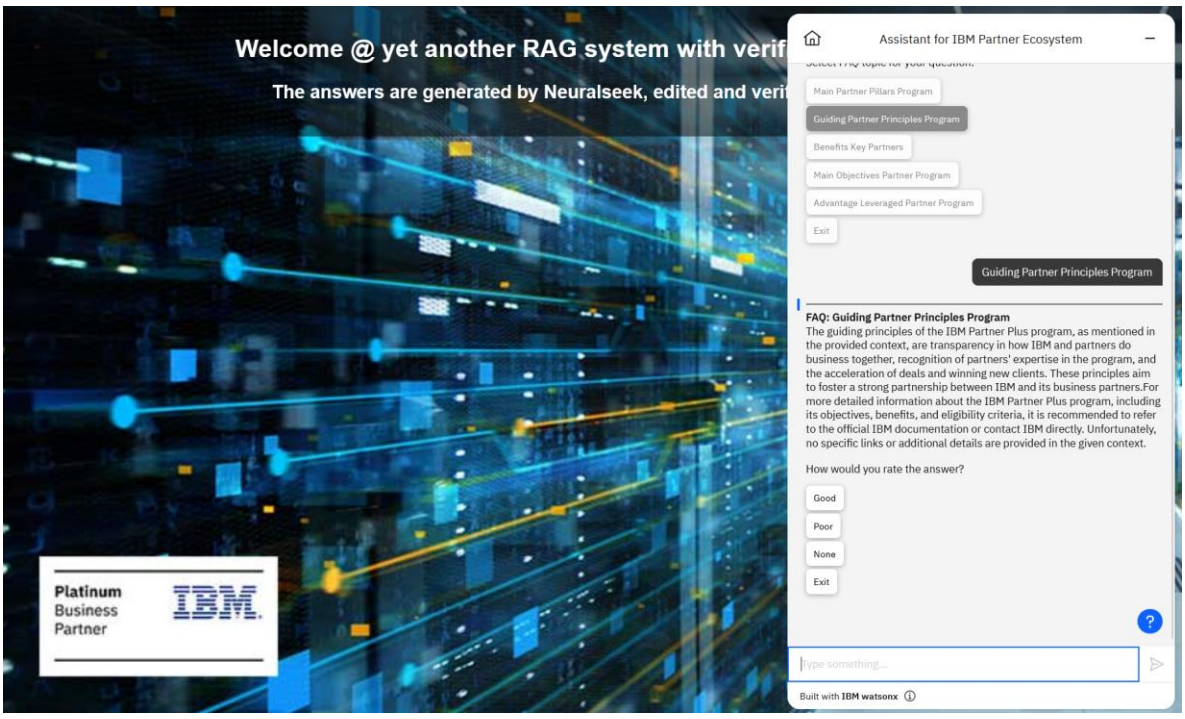
2. Click the first option, i.e. Main Partner Pillars Program



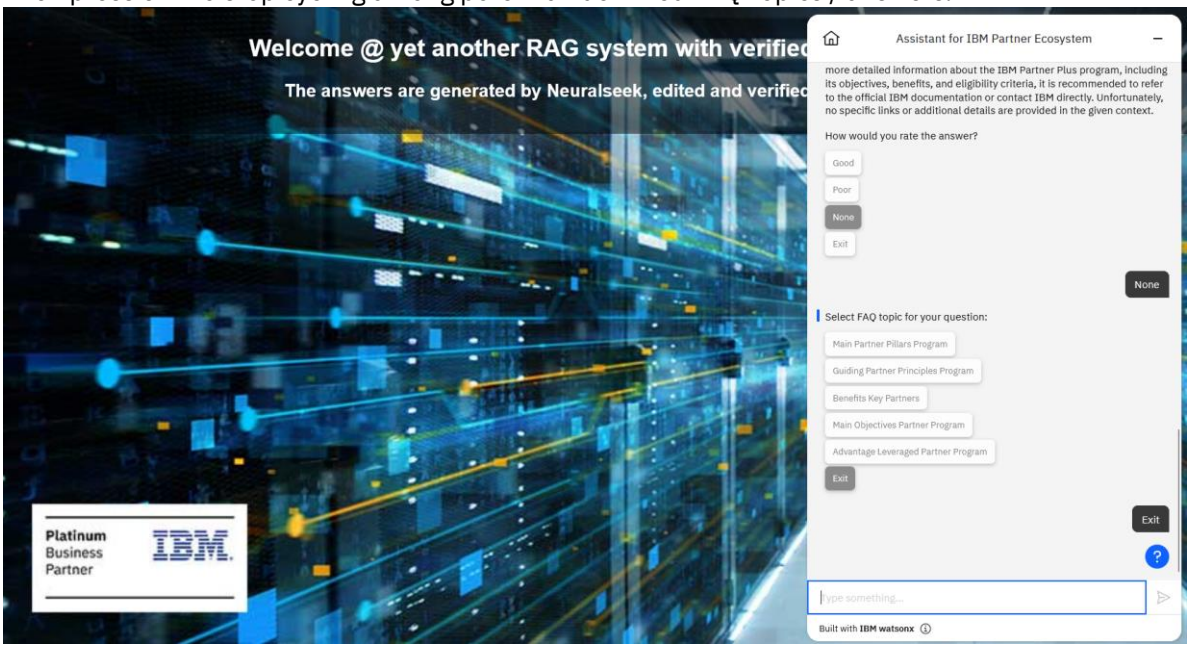
3. Rank the quality of answer Good, for instance.
   The try another FAQ from the list.



4. Select Guiding Partner Principles Program

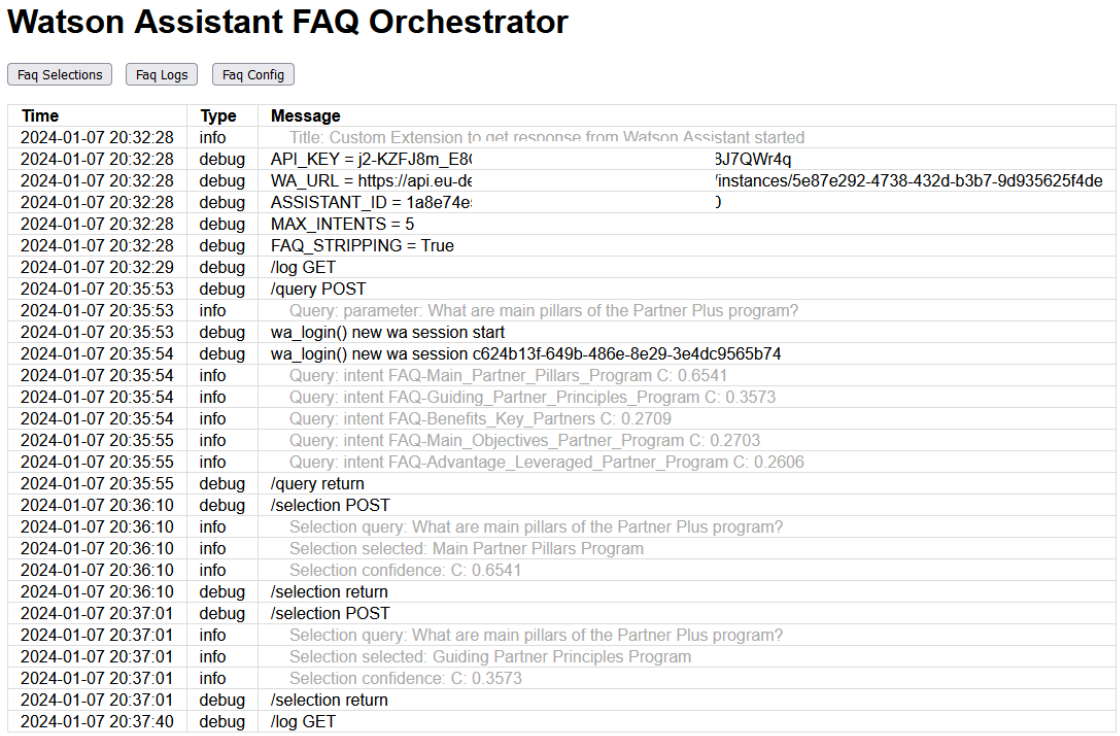


5. Evaluate it as None, since the content seems to be good but not related to your question. Then press exit to stop cycling among potential identified FAQ topics / answers.



8.3 Evaluating users found intents and user's rankings

1. The first, we look, what data are recorded by FAQ Orchestrator.
   Open web page using recorded URL of your FAQ Orchestrator.
   Click Faq Logs and you will get following log page:



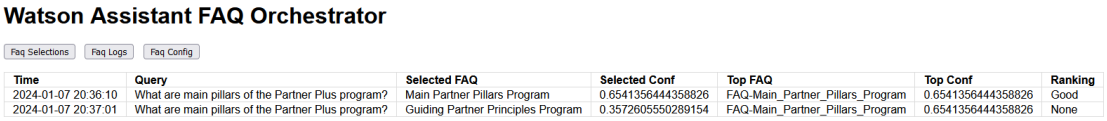
Watson Assistant FAQ Orchestrator

Faq Selections | Faq Logs | Faq Config

Time	Type	Message
2024-01-07 20:32:28	info	Title: Custom Extension to get response from Watson Assistant started
2024-01-07 20:32:28	debug	API_KEY = j2-KZFJ8m_E8(3J7QWr4q
2024-01-07 20:32:28	debug	WA_URL = https://api.eu-de 'instances/5e87e292-4738-432d-b3b7-9d935625f4de
2024-01-07 20:32:28	debug	ASSISTANT_ID = 1a8e74e:)
2024-01-07 20:32:28	debug	MAX_INTENTS = 5
2024-01-07 20:32:28	debug	FAQ_STRIPPING = True
2024-01-07 20:32:29	debug	/log GET
2024-01-07 20:35:53	debug	/query POST
2024-01-07 20:35:53	info	Query: parameter: What are main pillars of the Partner Plus program?
2024-01-07 20:35:53	debug	wa_login() new wa session start
2024-01-07 20:35:54	debug	wa_login() new wa session c624b13f-649b-486e-8e29-3e4dc9565b74
2024-01-07 20:35:54	info	Query: intent FAQ-Main_Partner_Pillars_Program C: 0.6541
2024-01-07 20:35:54	info	Query: intent FAQ-Guiding_Partner_Principles_Program C: 0.3573
2024-01-07 20:35:54	info	Query: intent FAQ-Benefits_Key_Partners C: 0.2709
2024-01-07 20:35:55	info	Query: intent FAQ-Main_Objectives_Partner_Program C: 0.2703
2024-01-07 20:35:55	info	Query: intent FAQ-Advantage_Leveraged_Partner_Program C: 0.2606
2024-01-07 20:35:55	debug	/query return
2024-01-07 20:36:10	debug	/selection POST
2024-01-07 20:36:10	info	Selection query: What are main pillars of the Partner Plus program?
2024-01-07 20:36:10	info	Selection selected: Main Partner Pillars Program
2024-01-07 20:36:10	info	Selection confidence: C: 0.6541
2024-01-07 20:36:10	debug	/selection return
2024-01-07 20:37:01	debug	/selection POST
2024-01-07 20:37:01	info	Selection query: What are main pillars of the Partner Plus program?
2024-01-07 20:37:01	info	Selection selected: Guiding Partner Principles Program
2024-01-07 20:37:01	info	Selection confidence: C: 0.3573
2024-01-07 20:37:01	debug	/selection return
2024-01-07 20:37:40	debug	/log GET

   You can see API_KEY, WA_URL and ASSISTANT_ID assigned to your FAQ Orchestrator.
   There is a Query request "What are main pillars of the Partner Plus program?", which follows opening of WA session and receiving 5 most relevant intent answers with specified machting confidence.
   The user selected "Main Partner Pillars Program" FAQ to display answer.
   Then user selected "Guiding Partner Principles Program" FAQ to display the second answer.

2. Now, click on Faq Selection, you can see that for user's initial question, there are 2 records for 2 FAQ selected with specific FAQ matching confidence, TOP FAQ and its confidence and user's ranking of the FAQ answer.



Time	Query	Selected FAQ	Selected Conf	Top FAQ	Top Conf	Ranking
2024-01-07 20:36:10	What are main pillars of the Partner Plus program?	Main Partner Pillars Program	0.6541356444358826	FAQ-Main_Partner_Pillars_Program	0.6541356444358826	Good
2024-01-07 20:37:01	What are main pillars of the Partner Plus program?	Guiding Partner Principles Program	0.3572605550289154	FAQ-Main_Partner_Pillars_Program	0.6541356444358826	None

3. As you can see, it gives us enough information to identify "new question" i.e. not well-matched FAQ answers for given user's questions (query).

   Consider following situations for "new questions"

   a. TOP FAQ confidence is quite small (less than 25%), interpretation is that WA was not able identified good matching FAQ, so it makes sense to try find new answer by Neuralseek in Knowledgebase
   b. Confidence of selected FAQ is also quite small (less than 25%), interpretation is that user is not satisfied with TOP FAQ and looking for other less confidence answers. If the user is not satisfied with quality of answer (check ranking) than it makes sense to let Neuralseek to generate answer to the question.
   c. Several Poor rankings for given user query can be good signal to batch generate new answer by Neuralseek.
   d. Etc.

   We can use LLM (watsonx.ai) model for FAQ questions/answers enrichment:

   a. to generate other questions which enrich set of questions for the FAQ intent
   b. to merge similar FAQ questions / answers
   c. to generate better titles of FAQ representing intent

9 Conclusion

You have seen how to build FAQ system for given RAG (e.g. Neuralseek), how it is possible to use Watsonx Assistant as Chatbot and the second instance to act as FAQ database responding with set of best matching FAQs intents to user's question.

Typical RAG responds ONLINE to user's queries. Presented solution provide user with best matching verified FAQ answers and being able to identify FAQ intents, which does not well satisfy user. There are called "new question", for which we use BATCH/OFFLINE RAG (Neuralseek) to generate new FAQ answers and update WA FAQ Base (e.g. on daily bases or so).

The entire solution consists of

a. Watsonx Assistant Chatbot (front end)
b. Watsonx Assistant FAQ Base (provide list of FAQ intents)
c. FAQ Orchestrator Code Engine, which integrates these two Watsonx Assistant instances
d. Watson Discovery as document search engine for RAG, source of KnowledgeBase documents
e. Neuralseek, batch generates answers to "new questions", subject matter expert can edit, merge and verify answers using Neuralseek web interface, export set of FAQs to Watsonx Assistant FAQ Base.