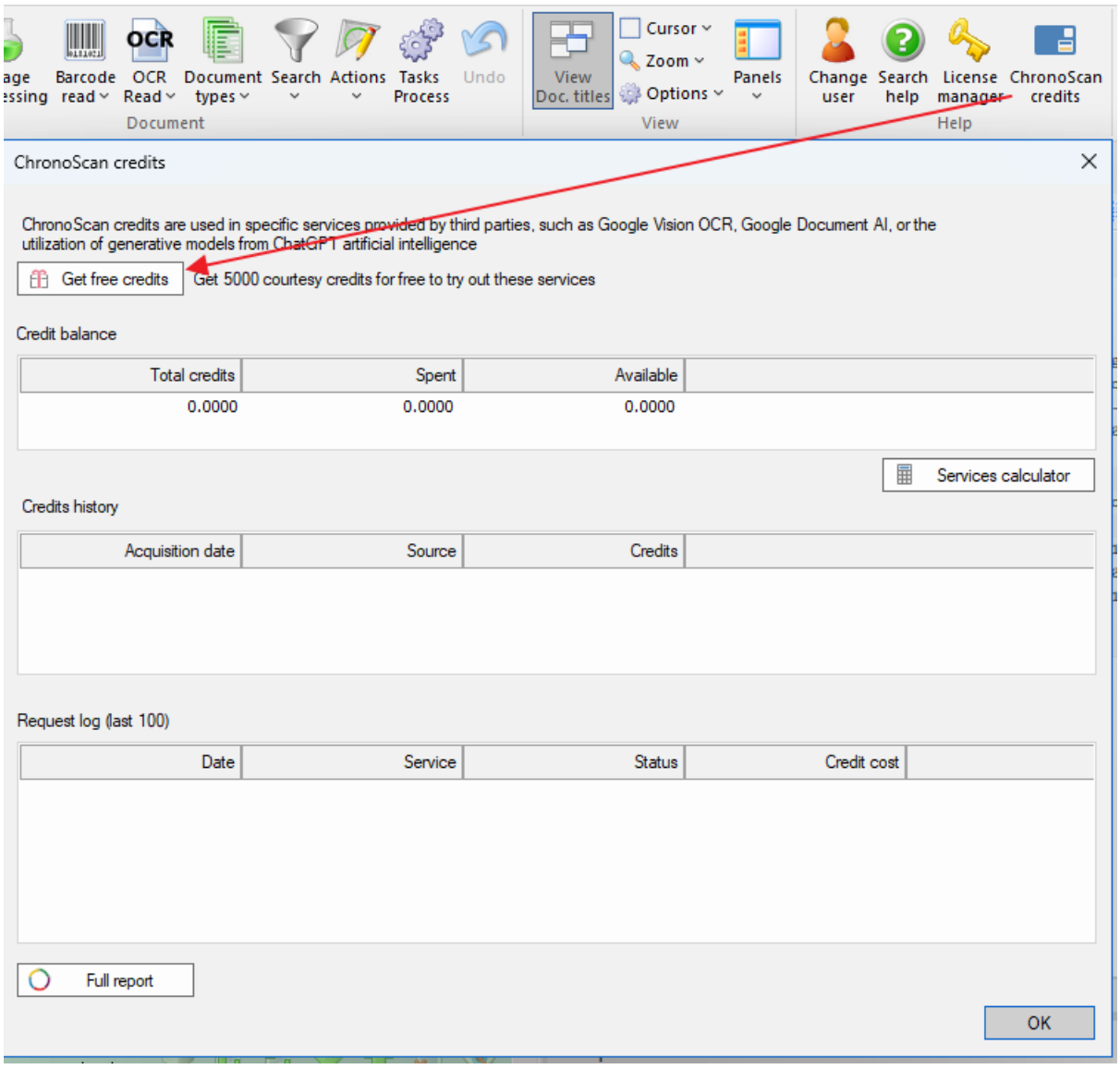
ChronoScan Ollama and GPT Quickstart

Ollama Setup

This document will guide you through the requirements needed to run the “Invoice AI GPT and OS” Configuration.

1) To enable GPT you will need to ensure that you have some demo credits. If you don’t already have some allocated to your licence you can allocate some using the button as shown below.

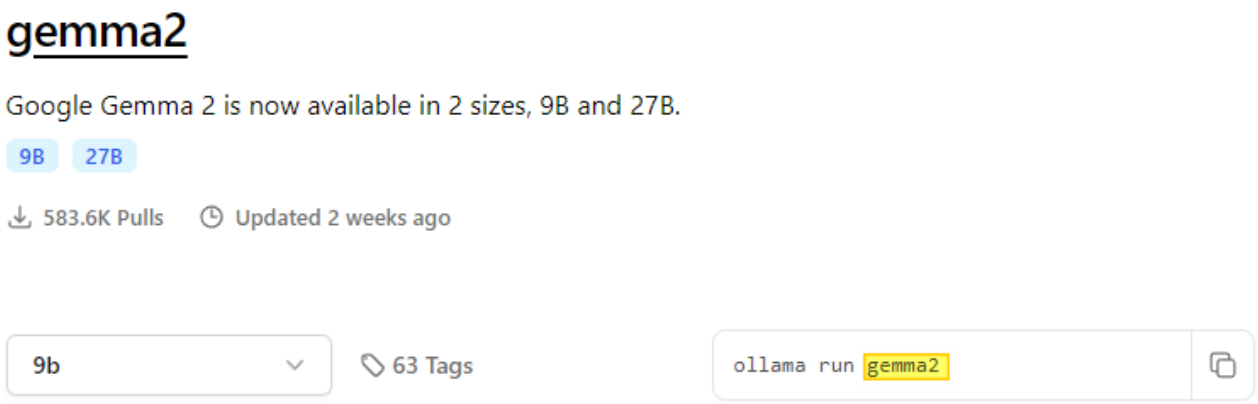


2) To enable the use of Ollama follow the steps below.
Please note that if using Ollama you should have a GPU of a reasonable specification to run at speed, ideally something like an NIVIDIA RTX 4090.

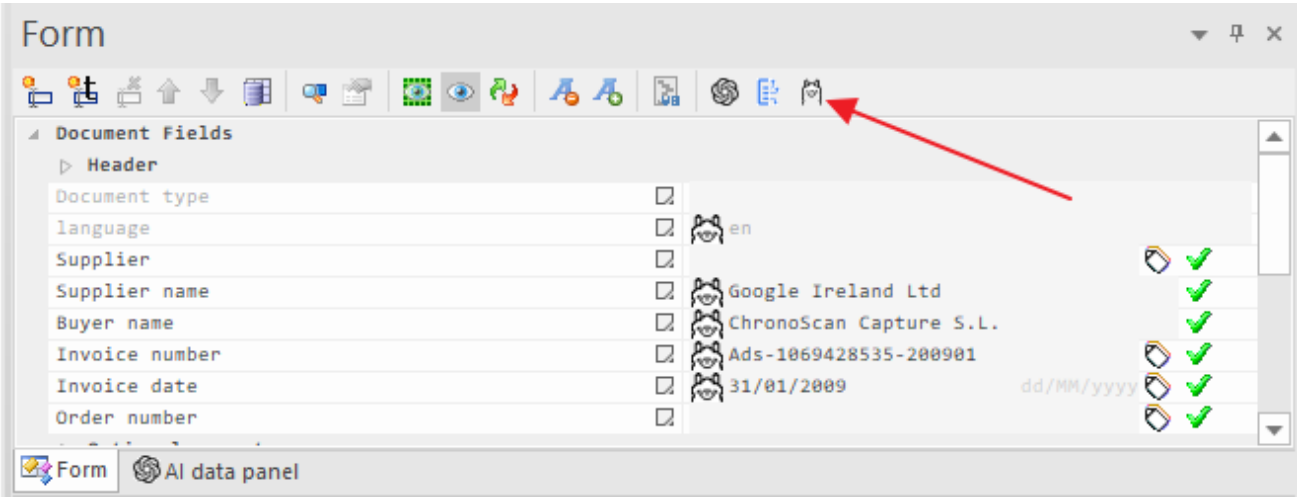
i) Install Ollama by downloading and running the installer here - https://ollama.com/download
ii) You can choose various large language models from the Ollama model library here - https://ollama.com/library

We would recommend to start with Gemma2 9B

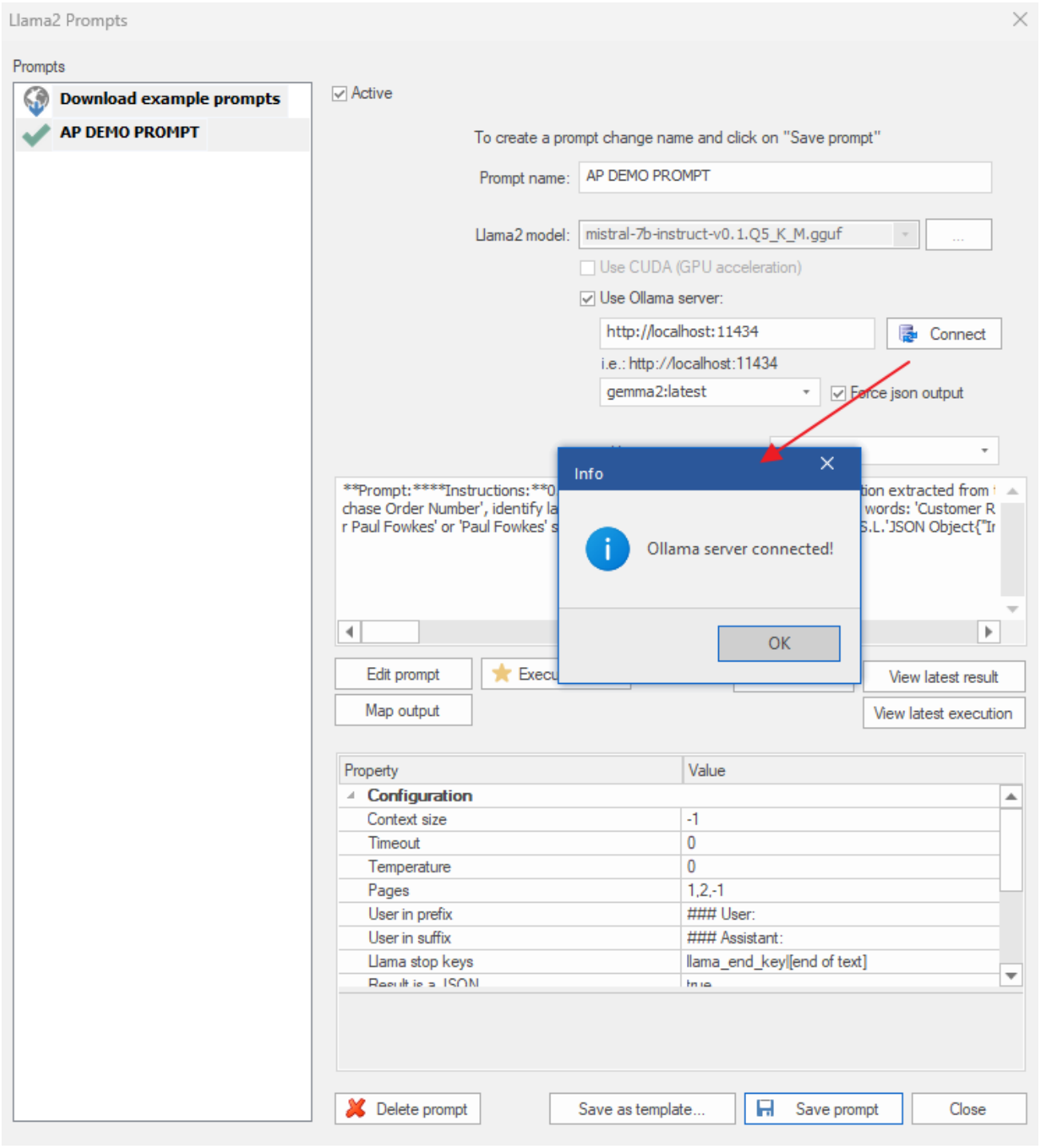
To install a model you need to open a command line and type “Ollama pull MODEL NAME” where MODEL NAME is the one you will find from the Ollama library, as shown below, I recommend using gemma2 9B to start.



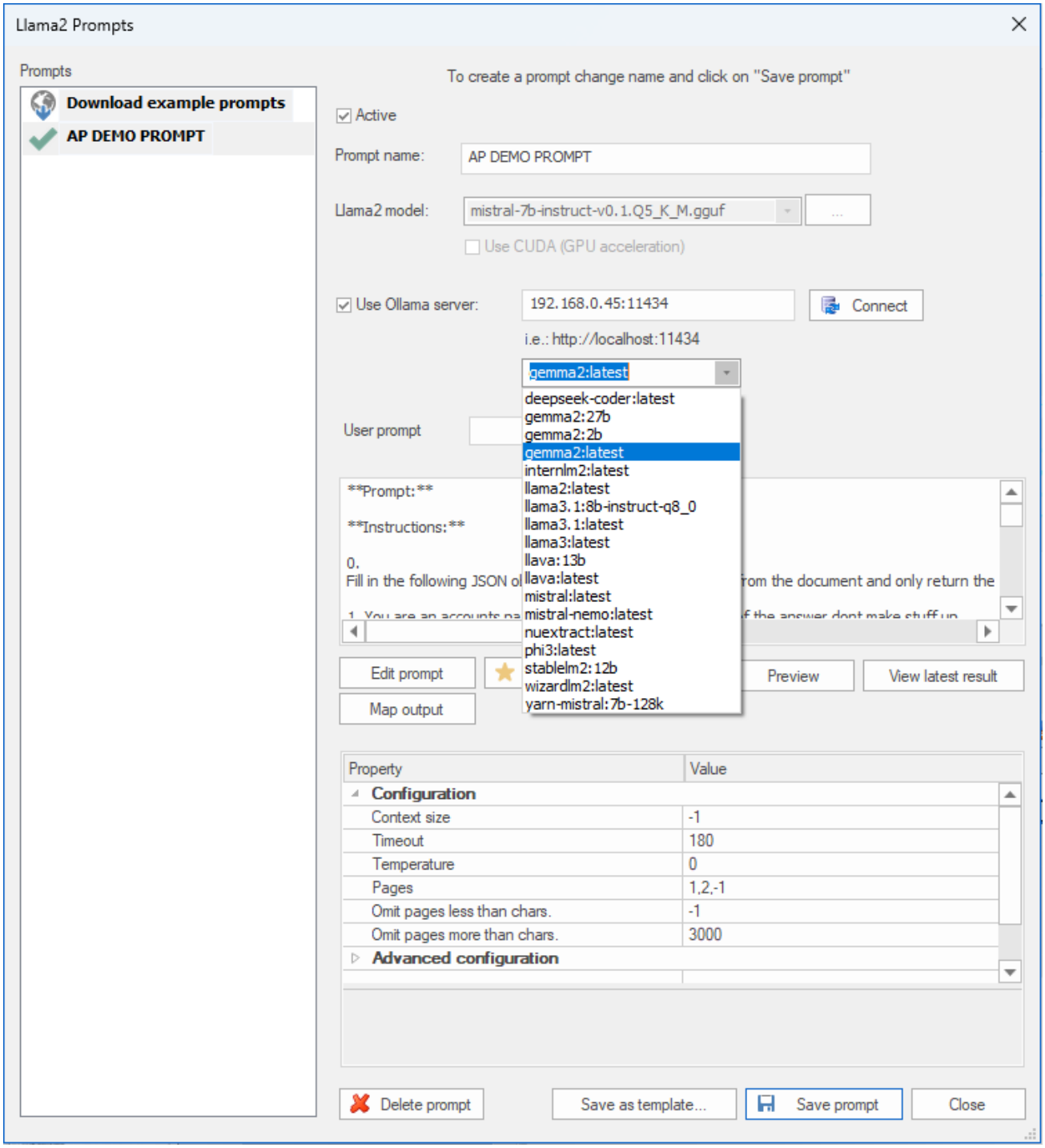
ii) When Ollama and your model have been installed check you can connect by clicking on the Llama icon here,



A new window will open as shown below, click on the “Connect” button where you should see the dialogue indicating a successful connection.

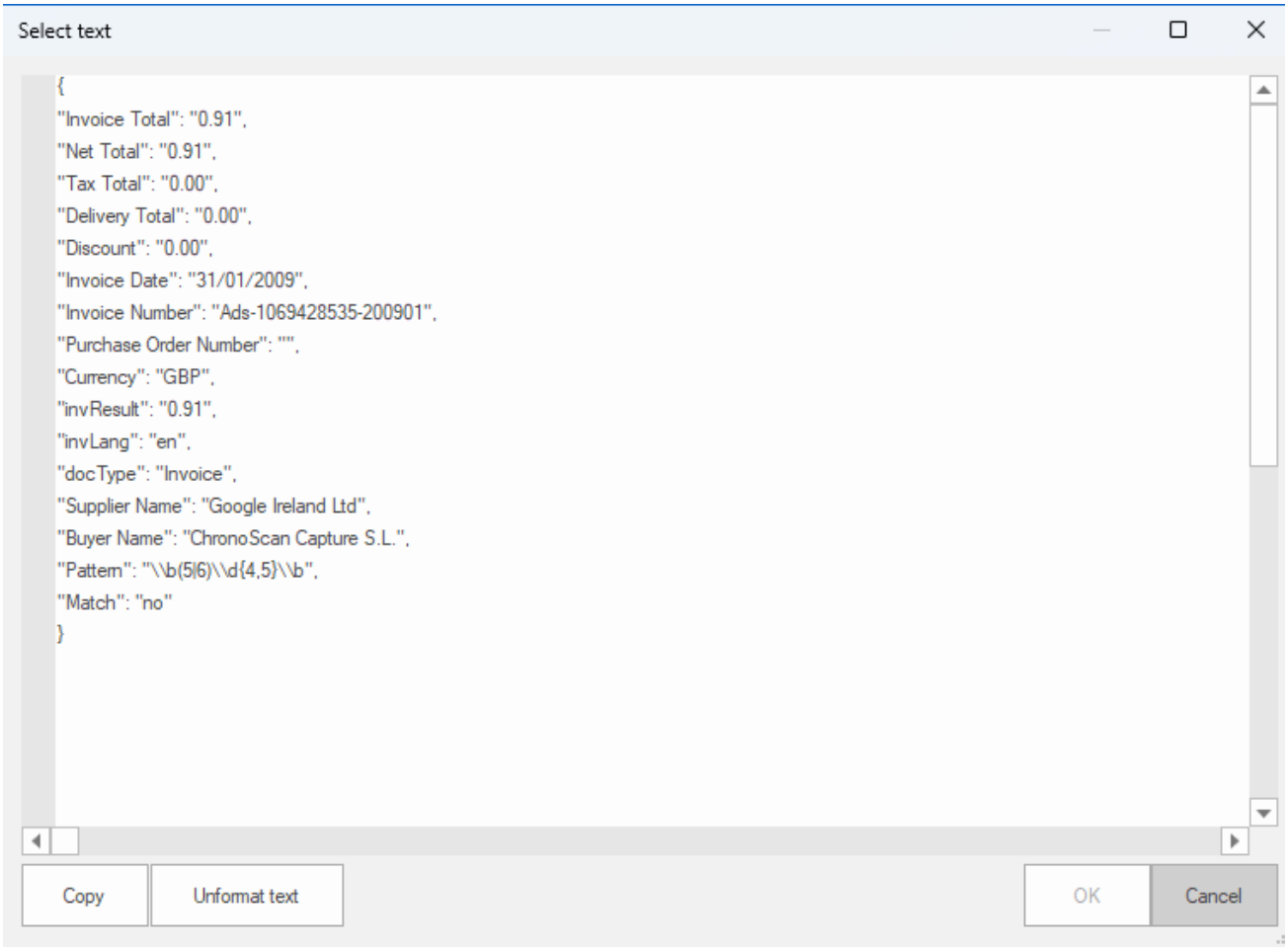


3) Depending which models you have pulled from the Ollama repository you will now be able to select from the dropdown,

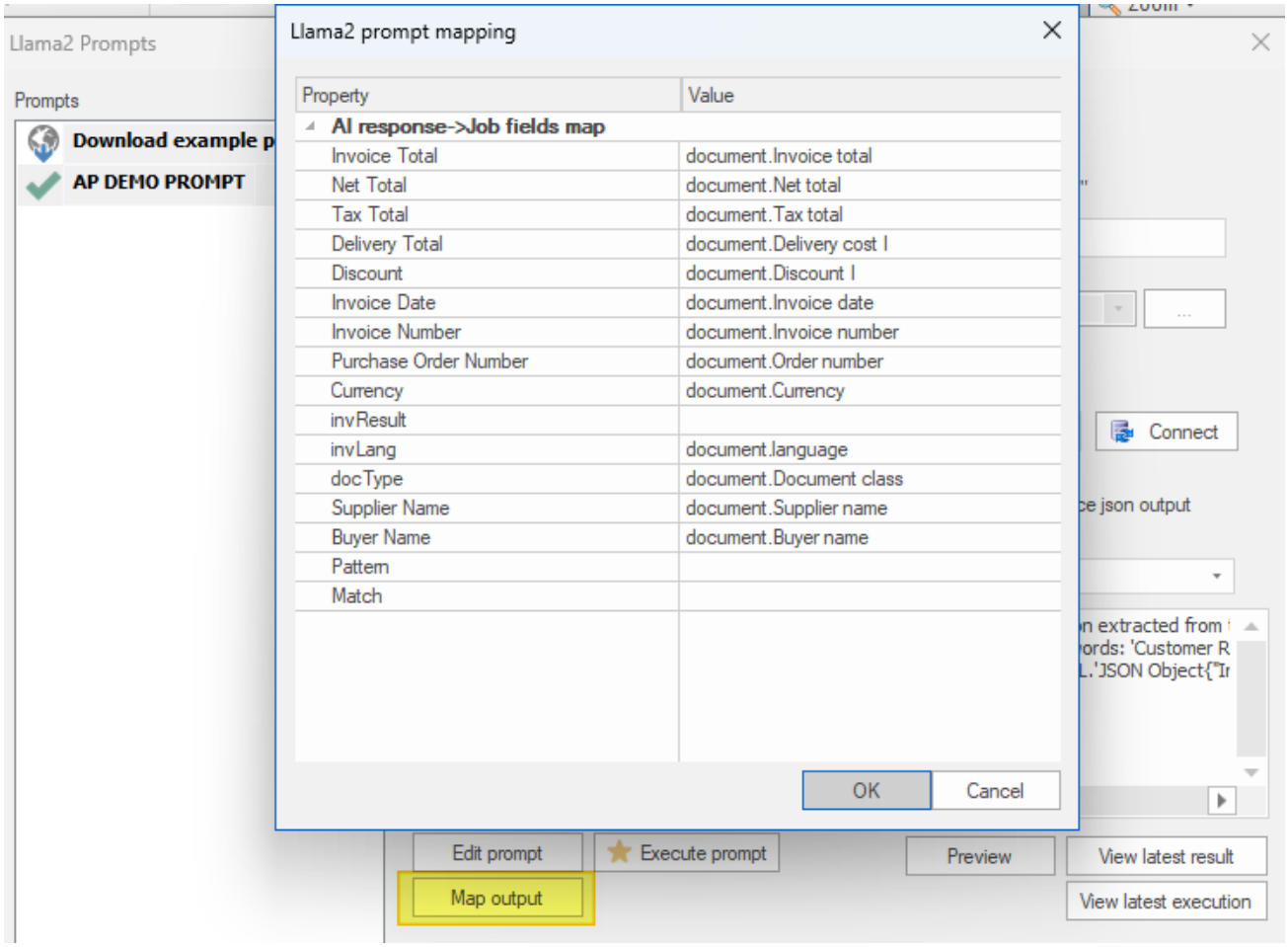


4) With your model selected you can now proceed to execute the prompt using the “Execute Prompt” button.

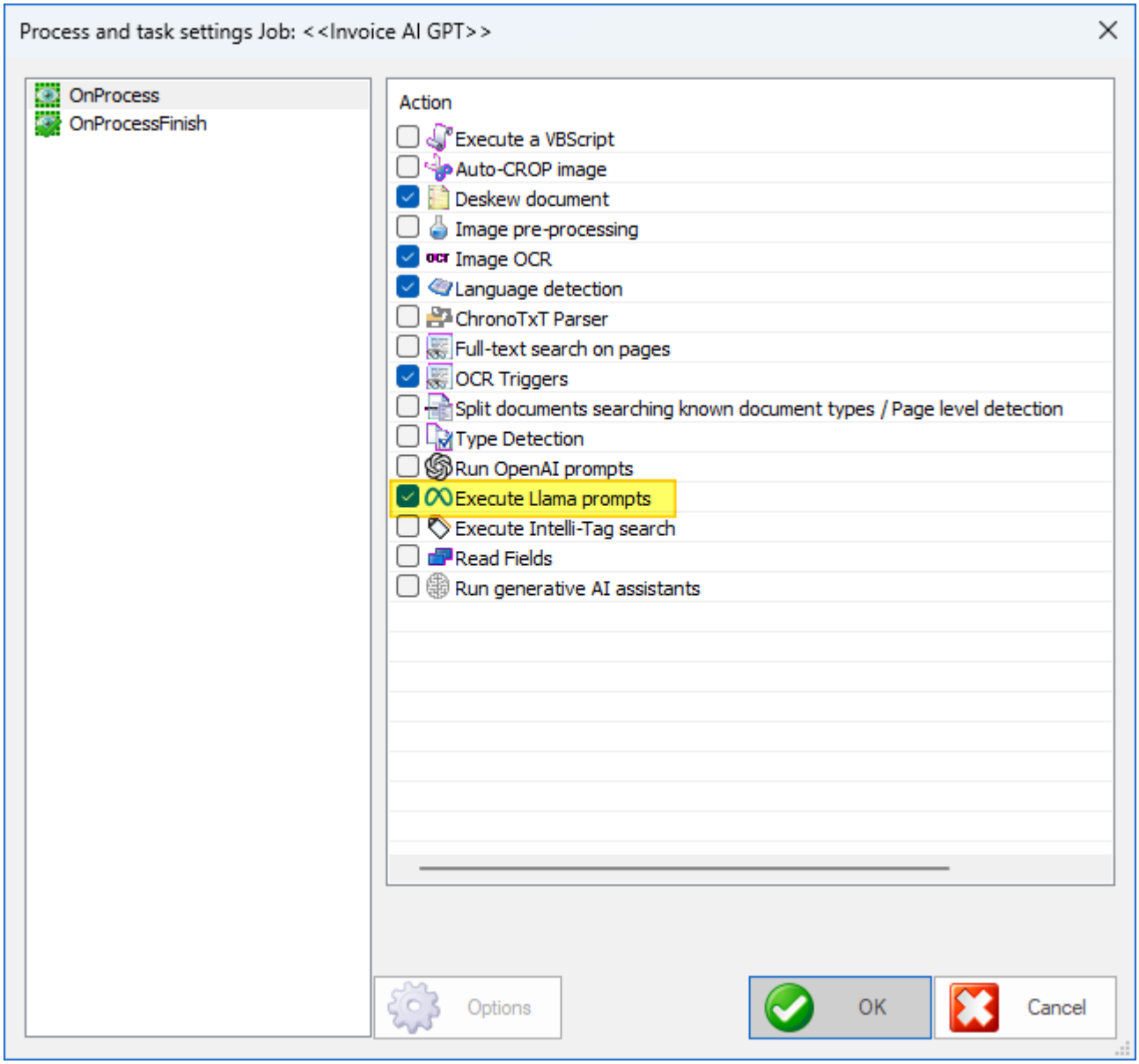
   If your prompt executes successfully you should see something like the window below,



5) Assuming the output from the prompt was successful you can now map the output to ChronoScan form fields by using the “Map output” button and save your prompt.
The sample job is already configured with some standard fields.

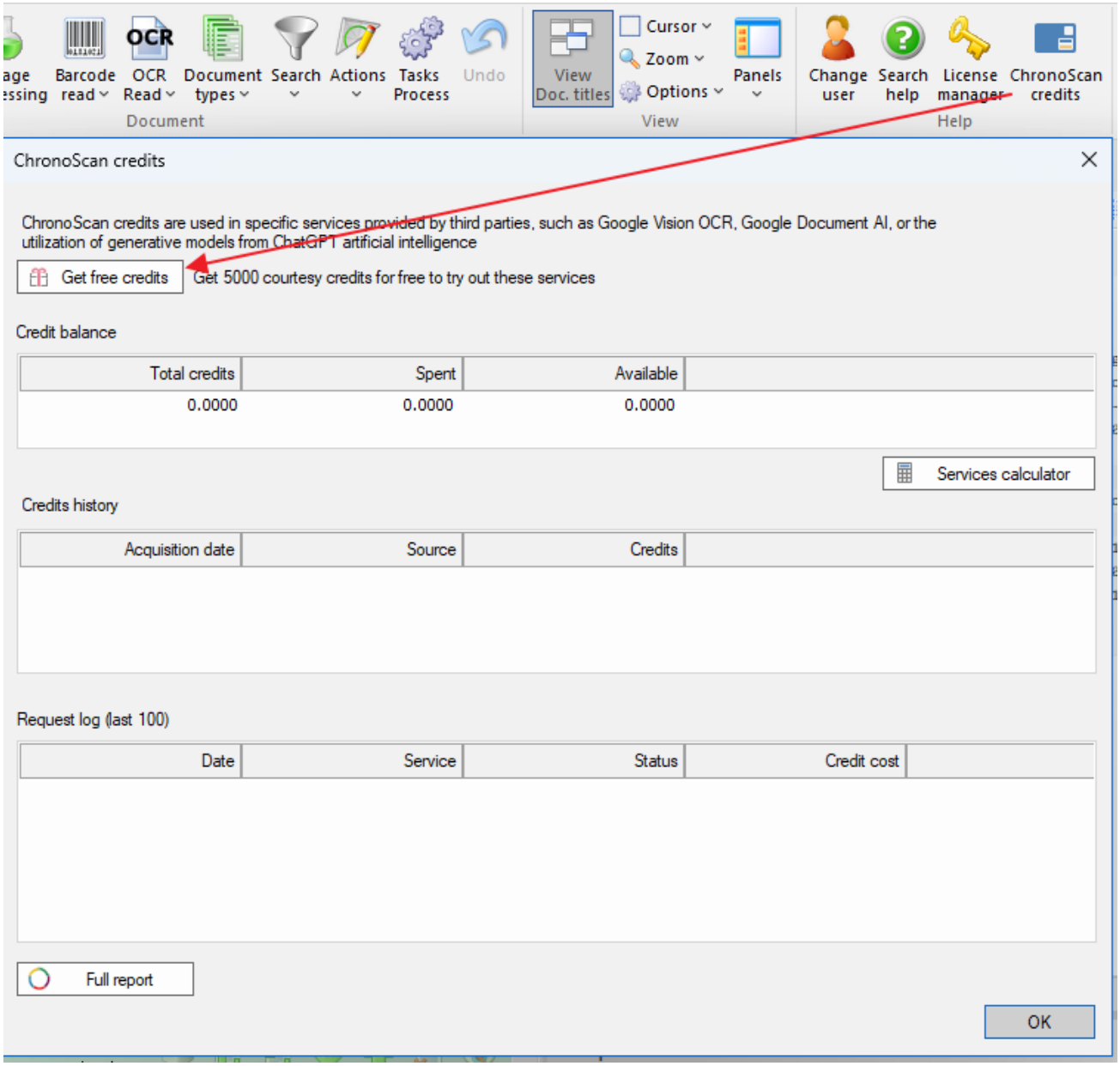


6) With your mapping complete you can start to process documents by enabling "Execute Llama prompts" in your Process and task settings for your job,

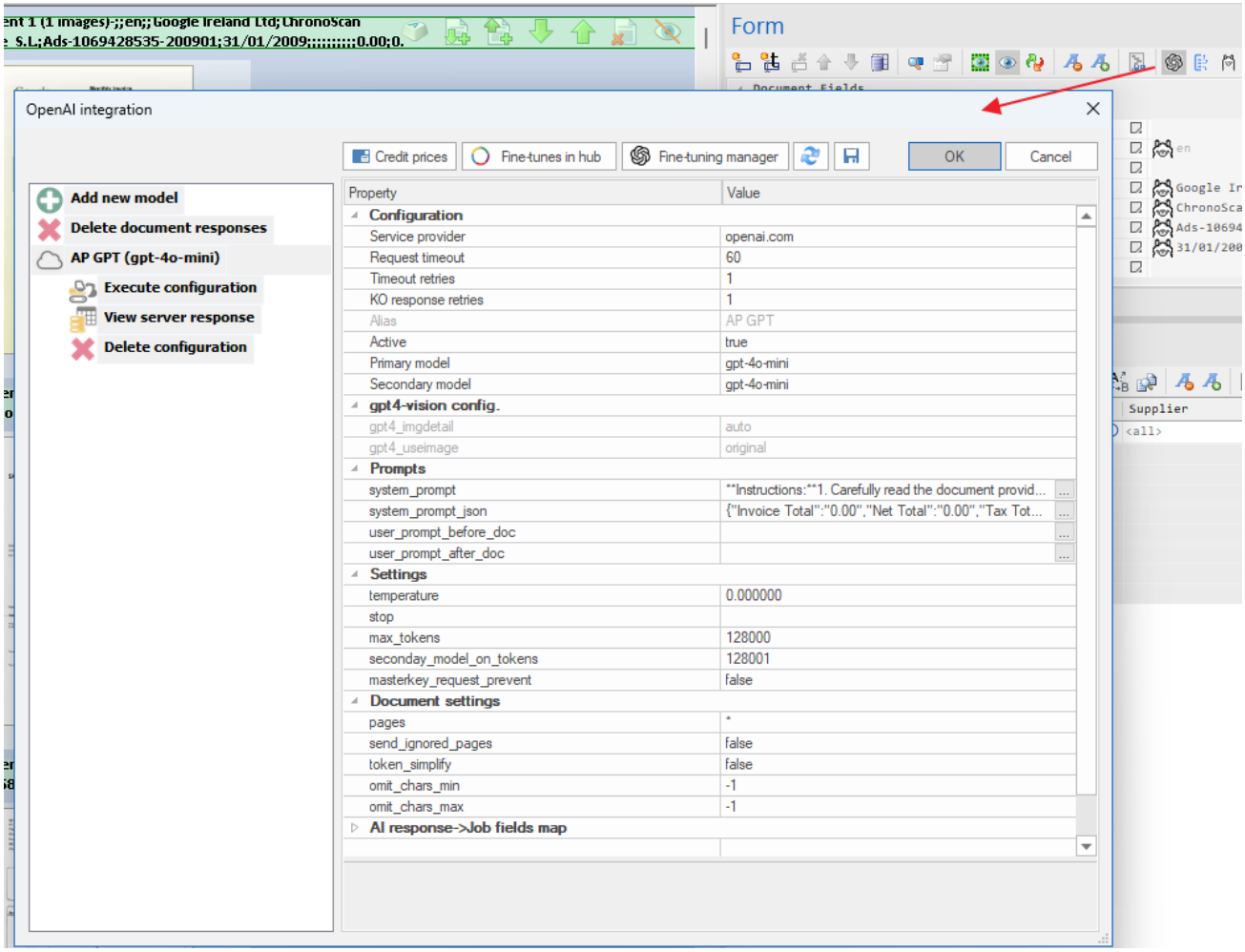

Process and task settings Job: <<Invoice AI GPT>>
OnProcess
OnProcessFinish
Action
Execute a VBScript
Auto-CROP image
Deskew document
Image pre-processing
Image OCR
Language detection
ChronoTxT Parser
Full-text search on pages
OCR Triggers
Split documents searching known document types / Page level detection
Type Detection
Run OpenAI prompts
Execute Llama prompts
Execute Intelli-Tag search
Read Fields
Run generative AI assistants
Options
OK
Cancel

OpenAI GPT Setup

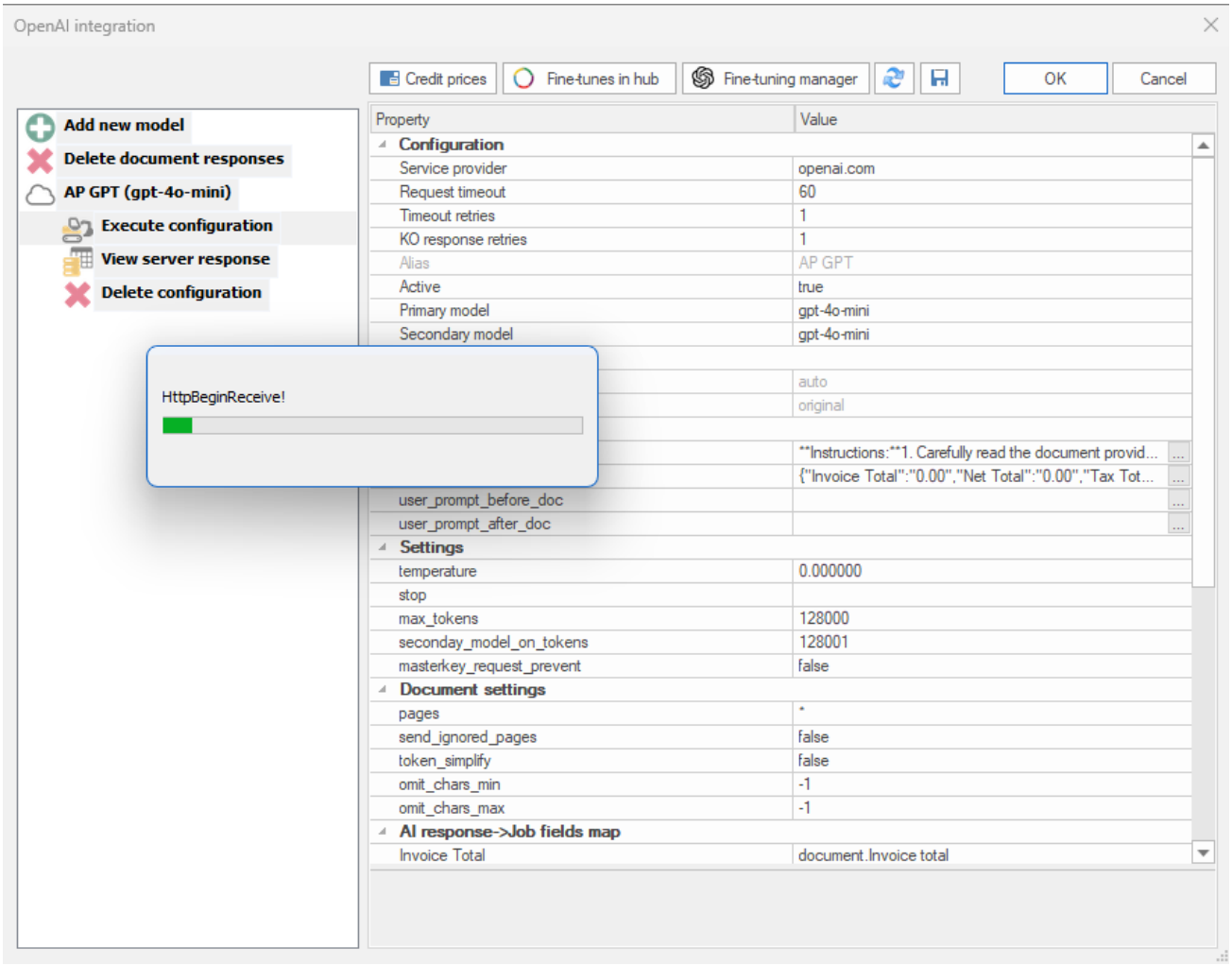
Ensure you have some ChronoScan Credits allocated to your licence, you can get some free credits by clicking on the “ChronoScan credits” button and selecting the “Get free credits” button.



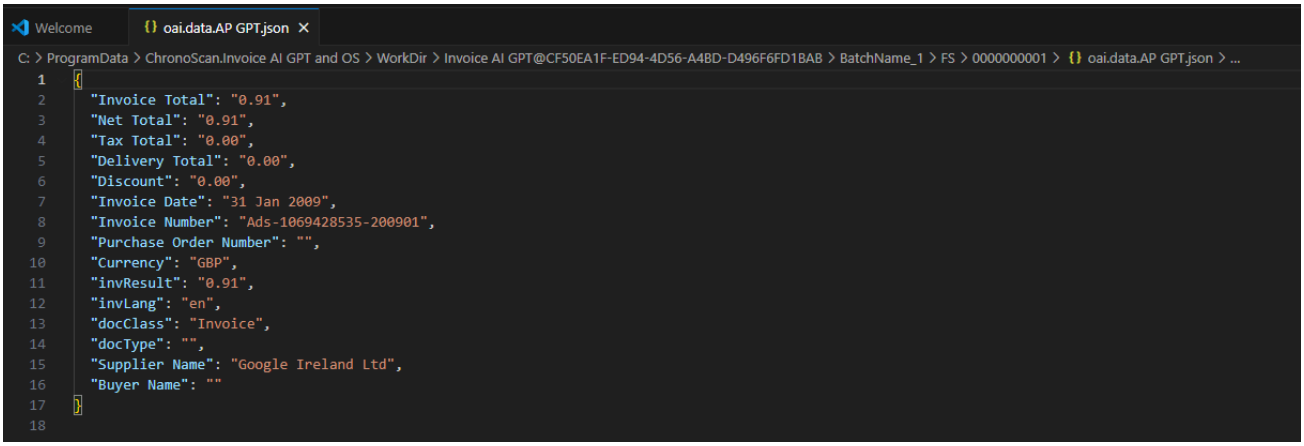
1) Click on the OpenAI GPT icon, the following window will be displayed,



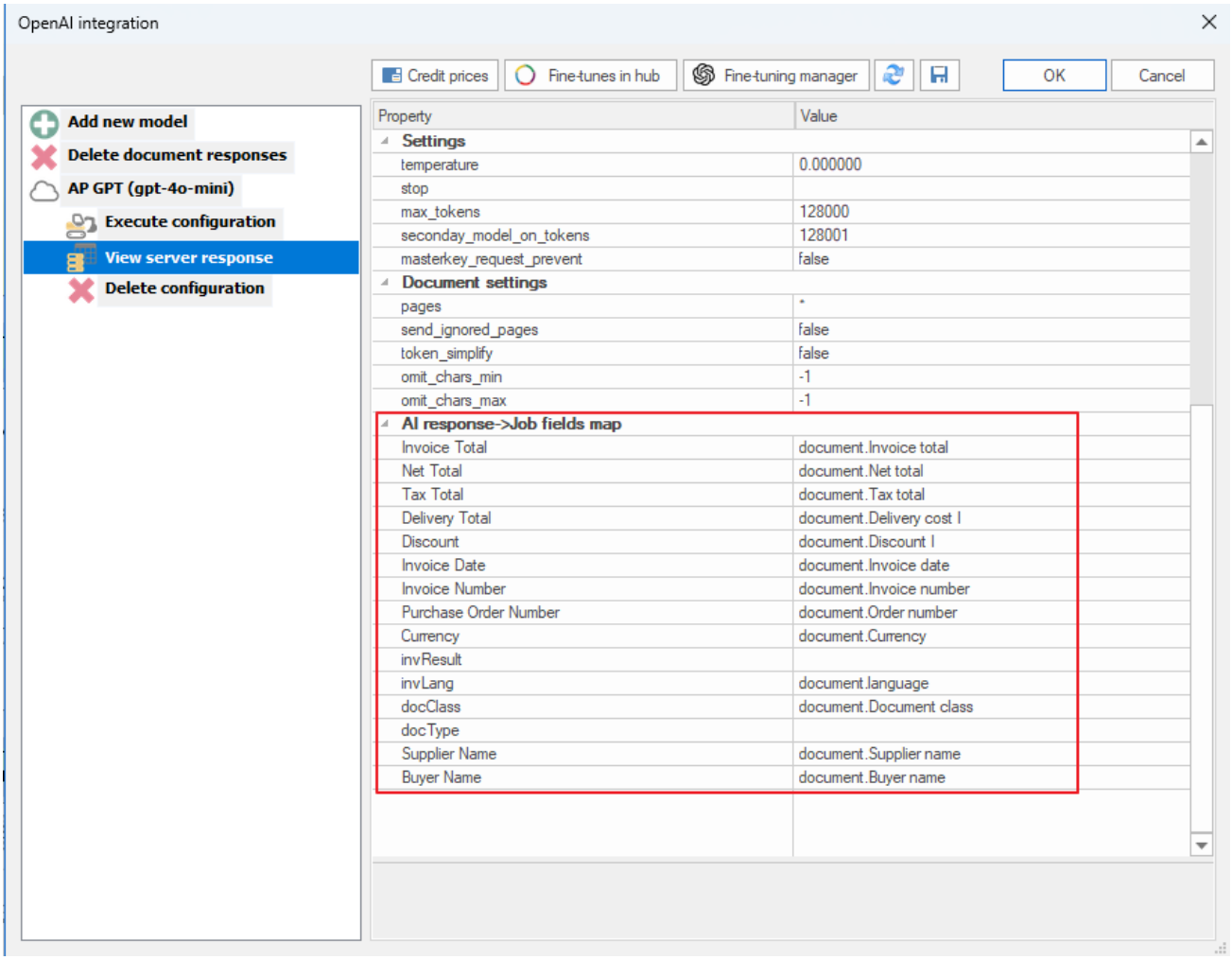
2) Click on the “Execute configuration” option in the left panel, you should briefly see a dialogue with “HttpBeginReceive”,



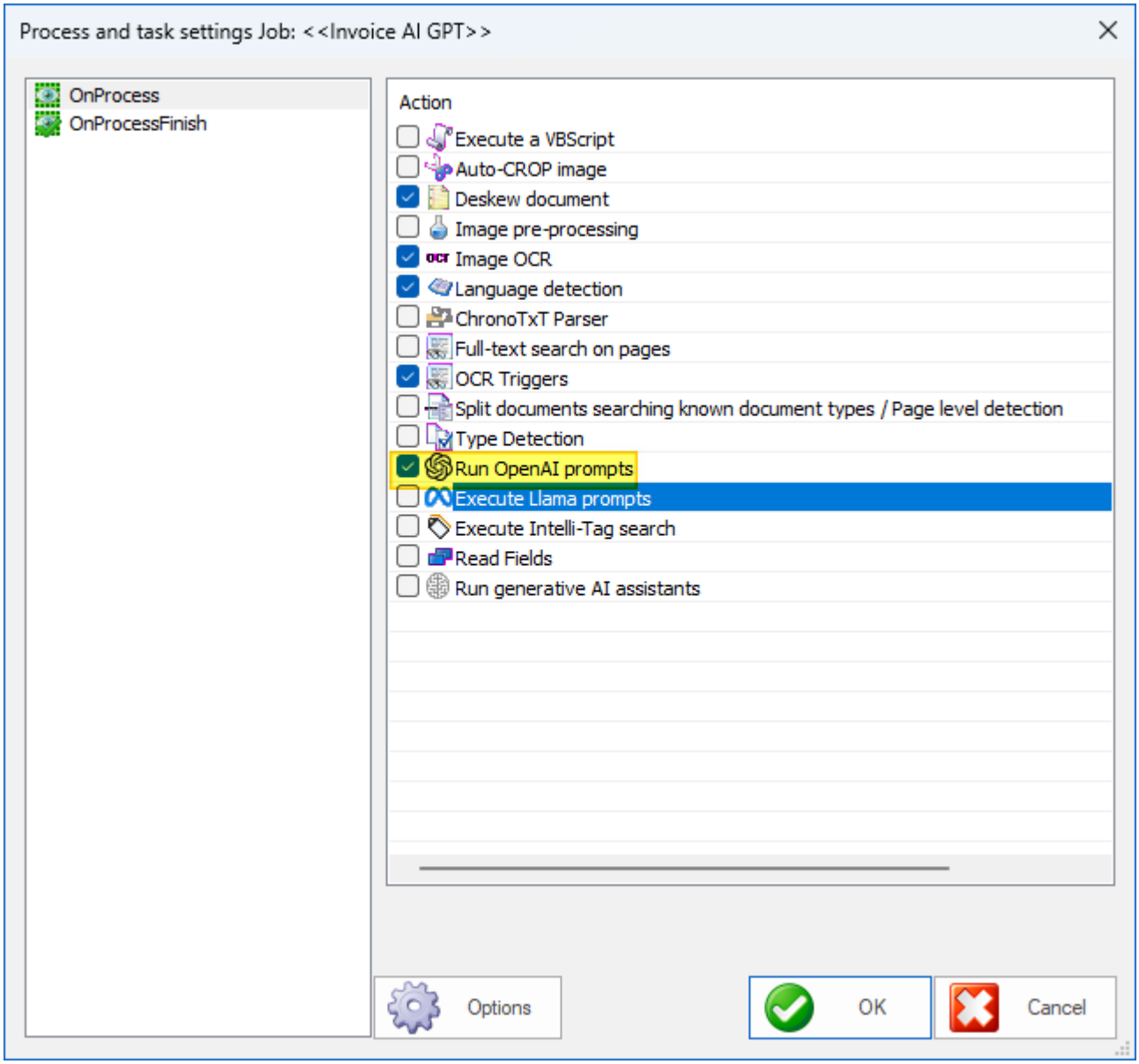
3) Once the dialogue disappears the prompt has executed and you can check the response by clicking on the “View server response” option.
If you have an application configured you will see the JSON response, for example as you see in the VS Code viewer below,



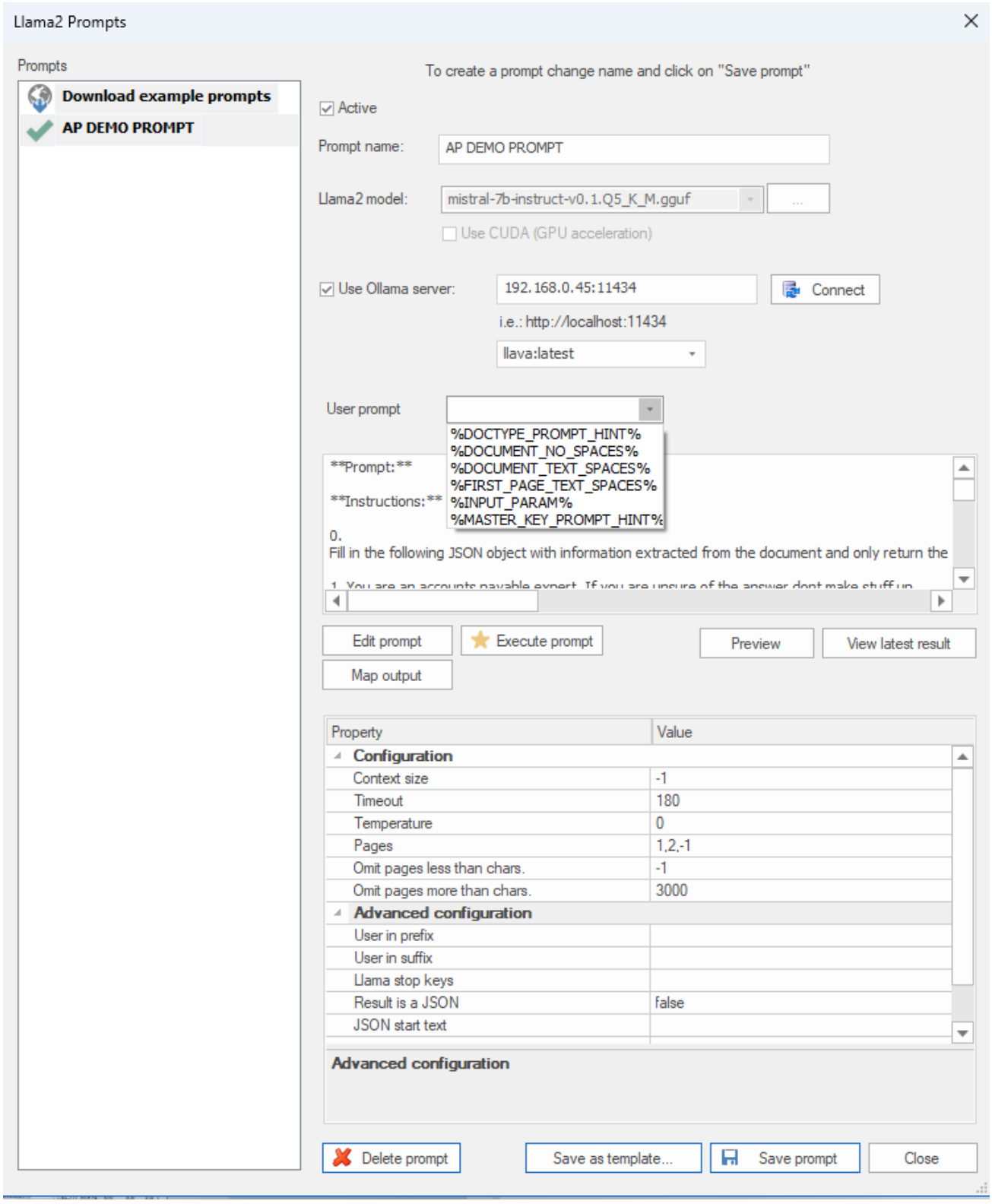
4) You can now map the output to your ChronoScan form fields although the sample configuration is already mapped for you.



Hopefully at this point you now have both Ollama and GPT operational and you can start to test some of your documents by enabling the “Run OpenAI prompts” in the “Process and task” settings.
If you ran into any difficulties, please contact us to assist.


Process and task settings Job: <<Invoice AI GPT>>
OnProcess
OnProcessFinish
Action
Execute a VBScript
Auto-CROP image
Deskew document
Image pre-processing
Image OCR
Language detection
ChronoTxT Parser
Full-text search on pages
OCR Triggers
Split documents searching known document types / Page level detection
Type Detection
Run OpenAI prompts
Execute Llama prompts
Execute Intelli-Tag search
Read Fields
Run generative AI assistants
Options
OK
Cancel

Ollama Parameter Configuration Description



Prompt Name - Name with which the prompt will be saved.
Llama2 model - Used to select the active large language model.

User prompt	-	Used to insert variable data in the prompt, as below.

%DOCTYPE_PROMPT_HINT%	- Use specific prompt for a document type.
%DOCUMENT_NO_SPACES%	- Use document text without multiple spaces.
% DOCUMENT_TEXT_SPACES %	- Use document text with multiple spaces.
%FIRST_PAGE_TEXT_SPACES%	- Use only first page with multiple spaces.
%INPUT_PARAM%	- Use custom variable
%MASTER_KEY_PROMPT_HINT%	- Use specific prompt for a master key.

The parameters %DOCTYPE_PROMPT_HINT% & %MASTER_KEY_PROMPT_HINT% can be utilised to customise the capture over and above the main prompt. This can be useful if for example you want to capture some additional data just for one type of document or you need a custom prompt to adjust the capture.

Edit prompt	-	Used to edit/create prompt instructions.
Execute prompt	-	Used to execute the current prompt for testing/mapping.
Preview	-	Used to display the complete prompt & data sent.
View latest result	-	Shows request duration and returned result.
Map output	-	Opens the mapping utility to map JSON to ChronoScan

Configuration

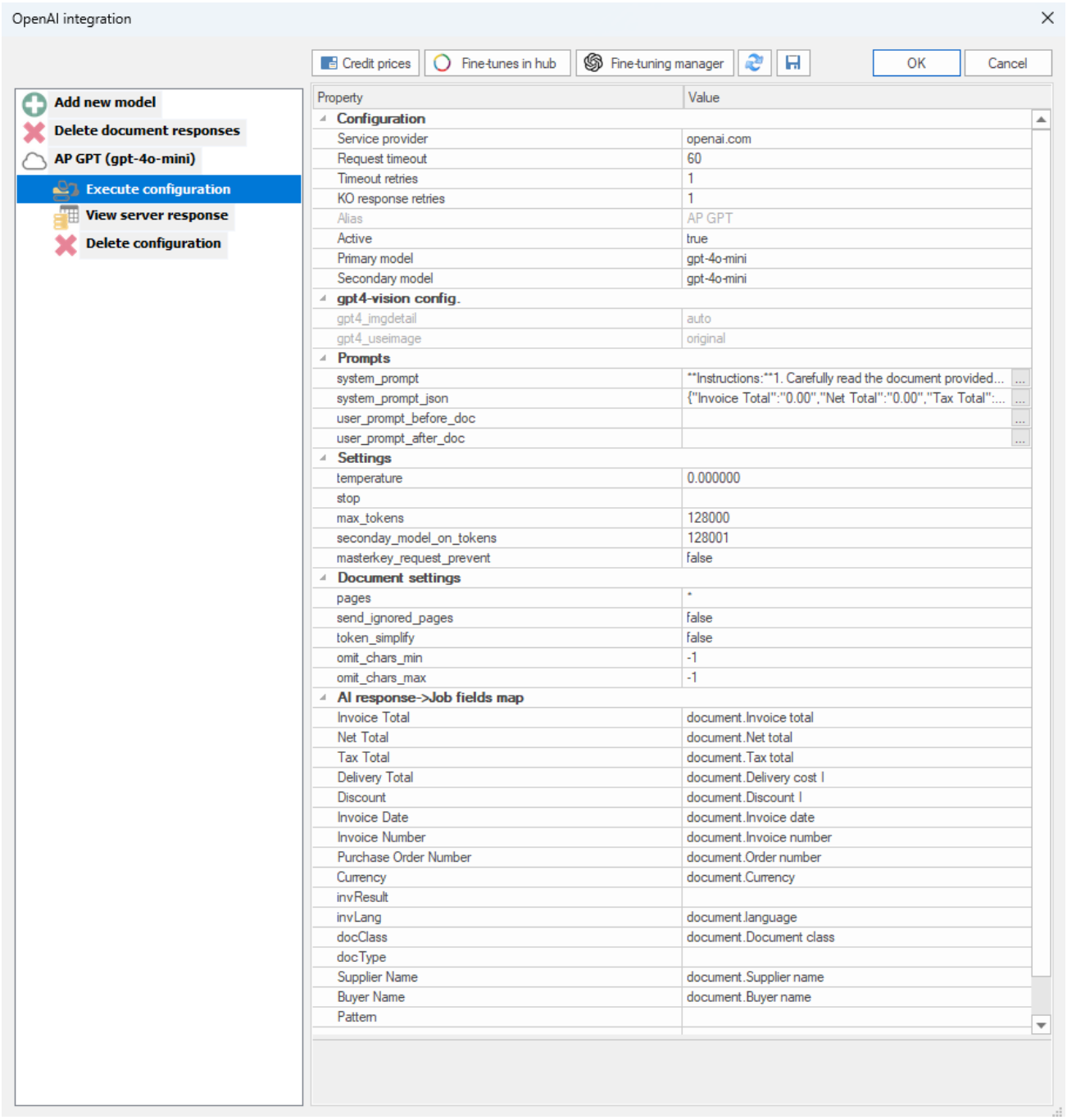
Context size	-	Specifies the LLM context size, -1 for max size by LLM.
Timeout	-	The maximum time for prompt execution.
Temperature	-	Sets the creativity of the LLM, usually this should be ‘0’.
Pages	-	Specifies the pages to include in the prompt.
Omit less than chars	-	Don’t send pages with less than n characters.
Omit pages more than chars	-	Don’t send pages with more than n characters.

Advanced Configuration

It is not necessary when using Ollama to use these parameters

User in prefix	-	Use this text in user prompt prefix.
User in suffix	-	Use this text in user prompt suffix.
Llama stop keys	-	Specifies tokens to stop the return of data.
Result is JSON	-	Forces the result to return a JSON object.
JSON start text	-	Defines the text where JSON starts

GPT Parameter Description



Service provider	-	Option to use OpenAI or Azure.
Request timeout	-	Sets max time allowed for a response return.
Timeout retries	-	Number of times to retry a request when a timeout occurs.
KO response retries	-	Number of times to retry request when no response received.
Alias	-	Configuration alias.
Active	-	Set configuration active/inactive.
Primary model	-	Sets model to use when max tokens not exceeded.
Secondary model	-	Sets model to use when max tokens exceeded.
gpt4_imgdetail	-	Sets the resolution for gpt processing.
gpt4_useimage	-	Sets whether to use the original image or thumbnail.
System_prompt	-	Used to define main system prompt.

System_prompt_json	-	Used to define the JSON prompt.
User_prompt_before_doc	-	Used to define a prompt to run before main prompt.
User_prompt_after_doc	-	Used to define a prompt to run after main prompt.
Temperature	-	Sets response variability (use 0 for most applications).
Stop	-	Sets a stop sequence (usually not required).
Max_tokens	-	Sets max tokens before switching to secondary model.
Secondary_model_on_tokens	-	Sets model to activate when tokens reached.
Masterkey_prevent_request	-	Stop request if no masterkey intellitag found.
Pages	-	Sets which pages to send.
Send_ignored_pages	-	Allows ignored pages to be sent/not sent.
Token_simplify	-	Removes multiple white spaces before sending request.
Omit_chars_min	-	Don’t send request is less than n characters.
Omit_chars_max	-	Don’t send request if n characters exceeded.