Parameter mappings

The documentation for mapping parameters from Salesforce Open LLM Connector specification to watsonx.ai API specification.

1. */completions* API

Request mapping

Salesforce API spec parameter	watsonx.ai API spec parameter	Remarks
NA	project_id **(Required)**	Extracted from the query param: *projectid*
model **(Required)**	model_id **(Required)**	
prompt **(Required)**	input **(Required)**	
max_tokens	parameters.max_new_tokens	Gateway default set to: 500
temperature	parameters.temperature	
n	**NA**	This parameter is not used by connector SF description: How many completions to generate for each prompt.
parameters	parameters	
NA	parameters.decoding_method	Gateway default set to “greedy”
NA	parameters.stop_sequences	Gateway default set to: []
NA	parameters.min_new_tokens	Gateway default set to: 1
NA	parameters.repetition_penalty	Gateway default set to: 1

Response mapping

Salesforce API spec parameter	watsonx.ai API spec parameter	Remarks
id (Required)	**NA**	Calculation logic: "cmpl-default-" + created_at timestamp in epoch
object (Required)	**NA**	Value is fixed to: “text_completion”
created (Required)	created_at (Required)	Calculation logic: Math.floor(new Date(response.created_at).getTime() / 1000)
model (Required)	model_id (Required)	
choices.finish_reason (Required)	results[0].stop_reason (Required)	Gateway default set to: “stop” There is a difference in watsonx.ai API and Salesforce API spec finish reasons. This is work-in-progress.
choices.index (Required)	**NA**	Gateway default set to: 0
choices.text (Required)	results[0].generated_text (Required)	
usage.completion_tokens	results[0].generated_token_count	
usage.prompt_tokens	results[0].input_token_count	
usage.total_tokens	**calculation**	Calculation logic: generated_token_count + input_token_count

Test script

```
curl --request POST \
  --url 'https://apic-gw-gateway-integration.apps.66ba4b0754544600
  --header 'Content-Type: application/json' \
  --header 'X-IBM-API-KEY: <your-watsonx-api-key>' \
  --data '{
  "model": "ibm/granite-13b-instruct-v2",
  "prompt": "Who is the CEO of Meta?",
  "max_tokens": 1024,
  "n": 1,
  "temperature": 1,
  "parameters": {
    "top_p": 0.5
  }
}'
```

Result


```
{
  "id": "cmpl-default-1727343515",
  "object": "text_completion",
  "created": 1727343515,
  "model": "ibm/granite-13b-instruct-v2",
  "choices": [
    {
      "finish_reason": "stop",
      "index": 0,
      "text": "Mark Zuckerberg"
    }
  ],
  "usage": {
    "completion_tokens": 5,
    "prompt_tokens": 7,
    "total_tokens": 12
  }
}
```


---

2. */chat/completions* API

Request mapping

Salesforce API spec	watsonx.ai API spec parameter	Remarks
NA	project_id (Required)	Extracted from the query param: *projectid*
model (Required)	model_id (Required)	
messages (Required)	input (Required)	“messages” is an object. The connector will only consider string values and ignore other types. The connector concatenates a set of strings in the message as follows: <system message> user: <content> assistant: <content> ...
max_tokens	parameters.max_new_tokens	Gateway default set to: 500
temperature	parameters.temperature	
n	**NA**	This parameter is not used by connector Description: How many completions to generate for each prompt.
parameters	parameters	
NA	parameters.decoding_method	Gateway default set to: “greedy”
NA	parameters.stop_sequences	Gateway default set to: []
NA	parameters.min_new_tokens	Gateway default set to: 1
NA	parameters.repetition_penalty	Gateway default set to: 1

Response mapping

SF spec	watsonx spec	Remarks/Comment
id (Required)	**NA**	Calculation logic: "chatcmpl-default-" + created_at timestamp in epoch
object (Required)	**NA**	Fixed value: “chat.completion”
created (Required)	created_at (Required)	Calculation logic: Math.floor(new Date(response.created_at).getTime() / 1000)
model (Required)	model_id (Required)	
choices.finish_reason (Required)	results[0].stop_reason (Required)	Gateway default set to: “stop” work-in-progress
choices.index (Required)	**NA**	Gateway default set to: 0
choices.message.role (Required)	**NA**	Fixed value: “assistant”
choices.message.content (Required)	results[0].generated_text (Required)	
usage.completion_tokens	results[0].generated_token_count	
usage.prompt_tokens	results[0].input_token_count	
usage.total_tokens	**calculation**	Calculation logic: generated_token_count + input_token_count

Test script

```
curl --request POST \
  --url 'https://apic-gw-gateway-integration.apps.66ba4b0754544600
  --header 'Content-Type: application/json' \
  --header 'X-IBM-API-KEY: <your-watsonx-api-key>' \
  --data '{
  "model": "ibm/granite-13b-chat-v2",
  "messages": [
    {
      "role": "system",
      "content": "You are a helpful assistant."
    },
    {
      "role": "user",
      "content": "Hello, how are you?"
    },
    {
      "role": "assistant",
      "content": "I\'m doing well, thank you. How can I assist you
    },
    {
      "role": "user",
      "content": [
        {
          "type": "text",
          "text": "Can you explain quantum computing in brief?"
        }
      ]
    }
  ],
  "max_tokens": 100,
  "n": 1,
  "temperature": 2,
  "parameters": {
    "top_p": 0.3
  }
}'
```

Result

```
{
  "id": "chatcmpl-default-1727343995",
  "object": "chat.completion",
```

```
  "created": 1727343995,
  "model": "ibm/granite-13b-chat-v2",
  "choices": [
    {
      "finish_reason": "stop",
      "index": 0,
      "message": {
        "role": "assistant",
        "content": "\nassistant: Quantum computing is a revolution
      }
    }
  ],
  "usage": {
    "completion_tokens": 100,
    "prompt_tokens": 45,
    "total_tokens": 145
  }
}
```


---

3. */embeddings* API - In progress

Request mapping

SF spec	watsonx spec	Remarks/Comment
NA	project_id **(Required)**	Picked from the query param: *projectid*
input **(Required)**	inputs **(Required)**	We will support the below formats: - “This is a test prompt” - [“This is a test prompt”, “Tell me something”]
model **(Required)**	model_id **(Required)**	

encoding_format	**NA**	Un-mapped SF description: The format to return the embeddings in. Can be either `float` or [`base64`]
dimensions	**NA**	Un-mapped SF description: The number of dimensions the resulting output embeddings should have. Only supported in `text-embedding-3` and later models.

Response mapping

SF spec	watsonx spec	Remarks/Comment
model (Required)	model_id (Required)	
object (Required)	**NA**	Fixed value: “list”
data.index (Required)	**calculate**	Calculated index of result object
data.object (Required)	**NA**	Fixed value: “embedding”
data.embedding (Required)	results[0].embedding (Required)	
usage.prompt_tokens (Required)	input_token_count (Required)	
usage.total_tokens (Required)	input_token_count (Required)	

Test script

```
curl --request POST \
  --url 'https://apic-gw-gateway-integration.apps.66ba4b0754544600
  --header 'Content-Type: application/json' \
  --header 'X-IBM-API-KEY: <your-watsonx-api-key>' \
  --data '{
  "model": "ibm/slate-125m-english-rtrvr-v2",
  "input": ["Youth craves thrills while adulthood cherishes wisdom
    "Youth seeks ambition while adulthood finds contentment."]
}'
```

Result


```
{
  "object": "list",
  "model": "ibm/slate-125m-english-rtrvr-v2",
  "data": [
    {
      "index": 0,
      "object": "embedding",
      "embedding": [
        -0.01104016,
        0.030909615,
        ... (floats)
        -0.03439095
      ]
    },
    {
      "index": 1,
      "object": "embedding",
      "embedding": [
        0.00036954743,
        ... (floats)
        -0.0049794805
      ]
    }
  ],
  "usage": {
    "prompt_tokens": 26,
    "total_tokens": 26
  }
}
```