# Azure OpenAI learning References

Monday, July 24, 2023 8:59 AM

**Github OpenAi**

PowerShellAI/Public/Set-AzureOpenAI.ps1 at master · dfinke/PowerShellAI · GitHub

LZ and Azure AI

https://techcommunity.microsoft.com/t5/azure-architecture-blog/azure-openai-landing-zone-reference-architecture/ba-p/3882102

**How to**

Getting started with Azure OpenAI and PowerShell | A blog about automation and technologies in the cloud (alexholmeset.blog)

**Helpful Videos**

Azure OpenAI 101: An introduction to Building Custom AI Models #python #chatgpt #azure

Connect ChatGPT to your Enterprise Data using Cognitive Search

Train your Own Enterprise Data with Azure OpenAI Service | ChatGPT with Custom Data - PDF, Word, TXT

The difference between OpenAI chat and completion is that the /completions endpoint provides the completion for a single prom pt and takes a single string as an input, whereas the /chat/completions provides the responses for a given dialog and requires the input in a specific for mat corresponding to the message history[1]. There are two different APIs for interacting with Azure OpenAI GPT models: Chat Completion API and Completion API [2][3]. The Chat Completion API is a new dedicated API for interacting with the ChatGPT and GPT-4 models[3].

From <https://www.bing.com/search?pglt=169&q=difference+between+openai+completions+and+chat+use&cvid=397b3b659a8a4fb887585568ed94bc 6a&aqs=edge..69i57j69i11004.17932j0j1&FORM=ANNAB1&PC=LCTS>

| Model types | Type | Description |
|---|---|---|
| **Summarize text** | Prompt Completion | As it says - look at documents and summarize  use **tldr**; before entering text |
| **Classify text** | Prompt Completion | |
| **Natural language to SQL** | Prompt Completion | Prompt to help build queries |
| **Generate new product name** | Prompt Completion | Help create Product names |
| **GPT3 - text-davinci** | model | Most versatile model for all types of tasks (summarize and understands intentions) most computational expensive, can also do codex |
| **GPT3-text-currie** | model | Less strong than davinci (capable of sentiment , classification, summarization) best used for a service chatbox |
| **GPT3 - text-babbage** | model | Best for simple tasks, semantic search , ranking how well documents match up with specific search queries |
| **GPT3 -text-ada** | model | Fastest model , good for parsing text,  address corrections,  things that do not require a lot of nuance and not complicated |
| **GPT3- cushman** | model | Uses codex and able to run a lot faster and cheaper than davinci. |
| **Max response** | parameter | Set a limit on the number of tokens per model response. The API supports a maximum of 4000 tokens shared between the prompt (including system message, examples, message history, and user query) and the model's response. One token is roughly 4 characters for typical English text.<br>From <https://oai.azure.com/portal/b868742ec6994c1a93e1fcd59c1f83bb/chat?tenantid=e594a530-1ec9-4192-a8d4-a9111f8cffa7> |
| **Temperature** | parameter | Controls randomness. Lowering the temperature means that the model will produce more repetitive and deterministic responses. Increasing the temperature will result in more unexpected or creative responses. Try adjusting temperature or Top P but not both. More or less creative<br>From <https://oai.azure.com/portal/b868742ec6994c1a93e1fcd59c1f83bb/chat?tenantid=e594a530-1ec9-4192-a8d4-a9111f8cffa7> |
| **Top P** | parameter | Similar to temperature, this controls randomness but uses a different method. Lowering Top P will narrow the model's token selection to likelier tokens. Increasing Top P will let the model choose from tokens with both high and low likelihood. Try adjusting temperature or Top P but not both.<br>From <https://oai.azure.com/portal/b868742ec6994c1a93e1fcd59c1f83bb/chat?tenantid=e594a530-1ec9-4192-a8d4-a9111f8cffa7> |
| **Stop Sequence** | parameter | Make the model end its response at a desired point. The model response will end before the specified sequence, so it won't contain the stop sequence text. For ChatGPT, using <\|im_end\|> ensures that the model response doesn't generate a follow-up user query. You can include as many as four stop sequences.<br>From <https://oai.azure.com/portal/b868742ec6994c1a93e1fcd59c1f83bb/chat?tenantid=e594a530-1ec9-4192-a8d4-a9111f8cffa7> |
| **Frequency penalty** | parameter | Reduce the chance of repeating a token proportionally based on how often it has appeared in the text so far. This decreases the likelihood of repeating the exact same text in a response.<br>From <https://oai.azure.com/portal/b868742ec6994c1a93e1fcd59c1f83bb/chat?tenantid=e594a530-1ec9-4192-a8d4-a9111f8cffa7> |
| **Presence penalty** | parameter | Reduce the chance of repeating any token that has appeared in the text at all so far. This increases the likelihood of introducing new topics in a response.<br>From <https://oai.azure.com/portal/b868742ec6994c1a93e1fcd59c1f83bb/chat?tenantid=e594a530-1ec9-4192-a8d4-a9111f8cffa7> |
| **Current Token Count** | parameter | This is an estimate of the number of tokens that will be used for the next request.<br>From <https://oai.azure.com/portal/b868742ec6994c1a93e1fcd59c1f83bb/chat?tenantid=e594a530-1ec9-4192-a8d4-a9111f8cffa7> |
| **Pre-response text** | parameter | Insert text after the user's input and before the model's response. This can help prepare the model for a response.<br>From <https://oai.azure.com/portal/b868742ec6994c1a93e1fcd59c1f83bb/playground?tenantid=e594a530-1ec9-4192-a8d4-a9111f8cffa7> |

| | | |
|---|---|---|
| **Post response text** | parameter | Insert text after the model’s generated response to encourage further user input, as when modeling a conversation.<br>From <https://oai.azure.com/portal/b868742ec6994c1a93e1fcd59c1f83bb/playground?tenantid=e594a530-1ec9-4192-a8d4-a9111f8cffa7> |
| **Max length tokens** | Parameter | Set a limit on the number of tokens per model response. The API supports a maximum of 4000 tokens shared between the prompt (including system message, examples, message history, and user query) and the model's response. One token is roughly 4 characters for typical English text.<br>From <https://oai.azure.com/portal/b868742ec6994c1a93e1fcd59c1f83bb/playground?tenantid=e594a530-1ec9-4192-a8d4-a9111f8cffa7> |
| **Top probabilities** | parameter | Similar to temperature, this controls randomness but uses a different method. Lowering Top P will narrow the model’s token selection to likelier tokens. Increasing Top P will let the model choose from tokens with both high and low likelihood. Try adjusting temperature or Top P but not both.<br>From <https://oai.azure.com/portal/b868742ec6994c1a93e1fcd59c1f83bb/playground?tenantid=e594a530-1ec9-4192-a8d4-a9111f8cffa7> |

How to generate text with Azure OpenAI Service - Azure OpenAI | Microsoft Learn