OpenAI and ChatGPT for Enterprises – Introducing Azure OpenAI

In this chapter, we'll focus on the enterprise-level applications of OpenAI models and introduce the partnership between OpenAI and Microsoft and **Azure OpenAI** (**AOAI**) Service. We will go through the milestones and developments of Microsoft in the field of **artificial intelligence** (**AI**), highlighting the journey that brought the Azure cloud into the game of OpenAI, and why this is a game-changer for large organizations. Finally, we will consider the topic of responsible AI and how to make sure your AI system complies with ethical standards.

In this chapter, we will discuss the following topics:

- The history of the partnership between Microsoft and OpenAI and the introduction of AOAI Service
- The role of the public cloud in the context of OpenAI models
- Responsible AI

By the end of this chapter, you will have learned about the main features of AOAI Service and how it differs from the OpenAI models we've discussed so far. You will also be familiar with the partnership history between Microsoft and OpenAI, and why there was the need for OpenAI models to be deployed on an enterprise-scale infrastructure. Finally, you will understand Microsoft's continuous and long-lasting commitment toward responsible AI and how it is benefiting AOAI Service.

Technical requirements

The following are the technical requirements for this chapter:

An Azure subscription, which you can create for free here: https://azure.microsoft.com/free/cognitive-services.

Access granted to Azure OpenAI in the desired Azure subscription. At the time of writing, access to this service is granted only by application. You can apply for access to Azure OpenAI by completing the form at https://aka.ms/oai/access.

OpenAI and Microsoft for enterprise-level AI – introducing Azure OpenAI

Microsoft has a long history of investing in AI research and development, with a focus on building AI-powered tools and services that can be used by businesses and individuals to solve complex problems and improve productivity.

It also boasts a series of milestones in terms of achieving human parity in AI domains such as speech recognition (2017), machine translation (2018), conversational Q&A (2019), image captioning (2020), and natural language understanding (2021).

Definition

Human parity in AI refers to the point at which an AI system can perform a task or tasks at a level that is equal to or indistinguishable from a human being. This concept is often used to measure the performance of AI systems, especially in areas such as natural language understanding, speech recognition, and image recognition. Achieving human parity in AI is considered a significant milestone as it demonstrates the AI's ability to effectively match human capabilities in a given domain.

In the next few sections, we are going to explore the research history and background of Microsoft in the domain of AI, to fully understand its journey toward their partnership with OpenAI and, finally, the development of AOAI Service.

Microsoft AI background

Early research in the field of AI traces back to the late 1990s when Microsoft established its **machine learning** (**ML**) and applied statistics groups. Starting with those, Microsoft started researching and experimenting with intelligent agents and

virtual assistants. In this case, the prototype was Clippy, a personal digital assistant for Microsoft Office:

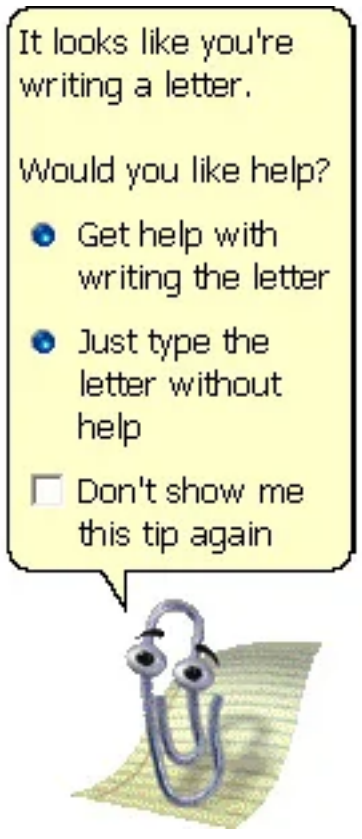

Figure 9.1 – Clippy, the default Office Assistant launched in 2000

Clippy was the forerunner of more sophisticated tools such as Cortana. Launched in 2014, Cortana is a digital assistant that uses **natural language processing** (**NLP**) and ML to provide personalized assistance to users.

Then, in 2016, as an expansion of Microsoft Project Oxford, Microsoft launched Microsoft Cognitive Services in the Azure cloud, a set of APIs that provide AI capabilities to developers without them requiring ML and data science expertise:

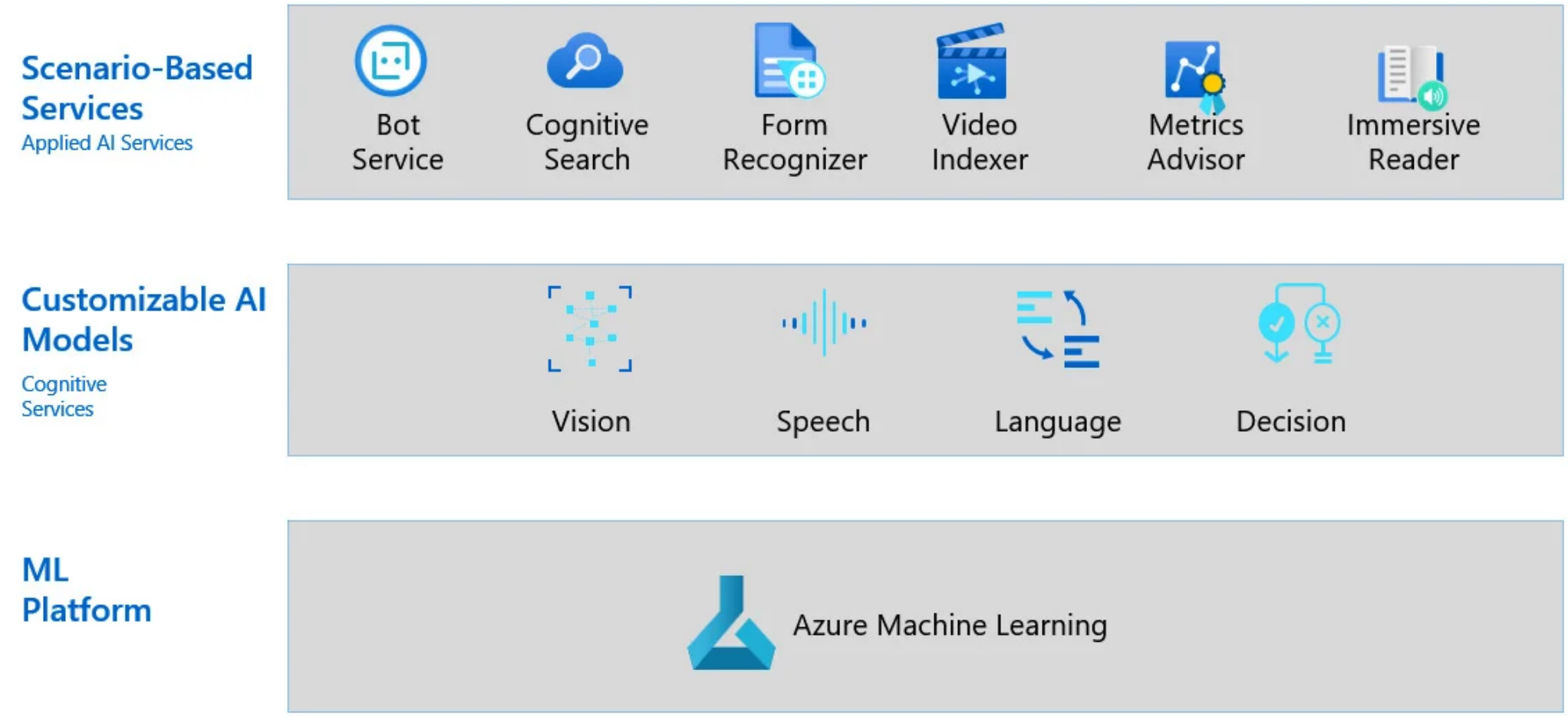

Figure 9.2 – Microsoft Azure AI services

With Cognitive Services, AI could finally be consumed by a wide range of users, from large enterprises to individual developers. From this, we witnessed what we now call **AI democratization:** AI is no longer a privilege for those who have deep technical knowledge and powerful and expensive hardware for model training. Cognitive Services has been developed for the following reasons:

- So that anyone, from data scientists to business users, can leverage Cognitive Services with a no-code approach
- To provide a set of pre-built models that have already been trained – that is, they are ready to use and don't need GPU-powered hardware to run

Microsoft's investments in AI can be seen from its acquisition of AI companies in recent years, including SwiftKey (a predictive keyboard app: https://blogs.microsoft.com/blog/2016/02/03/microsoft-acquires-swiftkey-in-support-of-re-inventing-productivity-ambition/) in 2016, Maluuba (a deep learning startup: https://blogs.microsoft.com/blog/2017/01/13/microsoft-acquires-deep-learning-startup-maluuba-ai-pioneer-yoshua-bengio-advisory-role/) in 2017, and Bonsai (a platform for building AI models: https://blogs.microsoft.com/blog/2018/06/20/microsoft-to-acquire-bonsai-in-move-to-build-brains-for-autonomous-systems/) in 2018.

Among the companies Microsoft invested in and partnered with, there is also OpenAI.

The partnership between the two tech companies began in 2016 when OpenAI agreed to leverage Microsoft's Azure cloud infrastructure to run its AI experiments. Later on, in 2019, Microsoft announced a $1 billion partnership with OpenAI (https://news.microsoft.com/2019/07/22/openai-forms-exclusive-computing-partnership-with-microsoft-to-build-new-azure-ai-supercomputing-technologies/) to develop AI models and technologies that can be used for the benefit of humanity. This partnership is based on the following three main pillars:

- Microsoft and OpenAI will jointly build new Azure supercomputing infrastructure to train AI models
- OpenAI will make its models and technologies consumable from the Azure cloud
- Microsoft will become OpenAI's preferred partner for commercializing new AI solutions to the market

Since then, the two companies kept investing and researching, and finally, in January 2023, a set of OpenAI models was made available in Azure via AOAI Service.

With the general availability of AOAI Service, a new milestone was reached and the Microsoft AI portfolio has been extended with the powerful large language models of OpenAI.

Azure OpenAI Service

AOAI Service is a product of Microsoft that provides REST API access to OpenAI's powerful language models such as GPT-3.5, Codex, and DALL-E. You can use these models for the very same tasks as OpenAI models, such as content generation, summarization, semantic search, natural language, and code translation.

In the context of the Microsoft Azure AI portfolio, AOAI Service is collocated among the following Cognitive Services offerings:

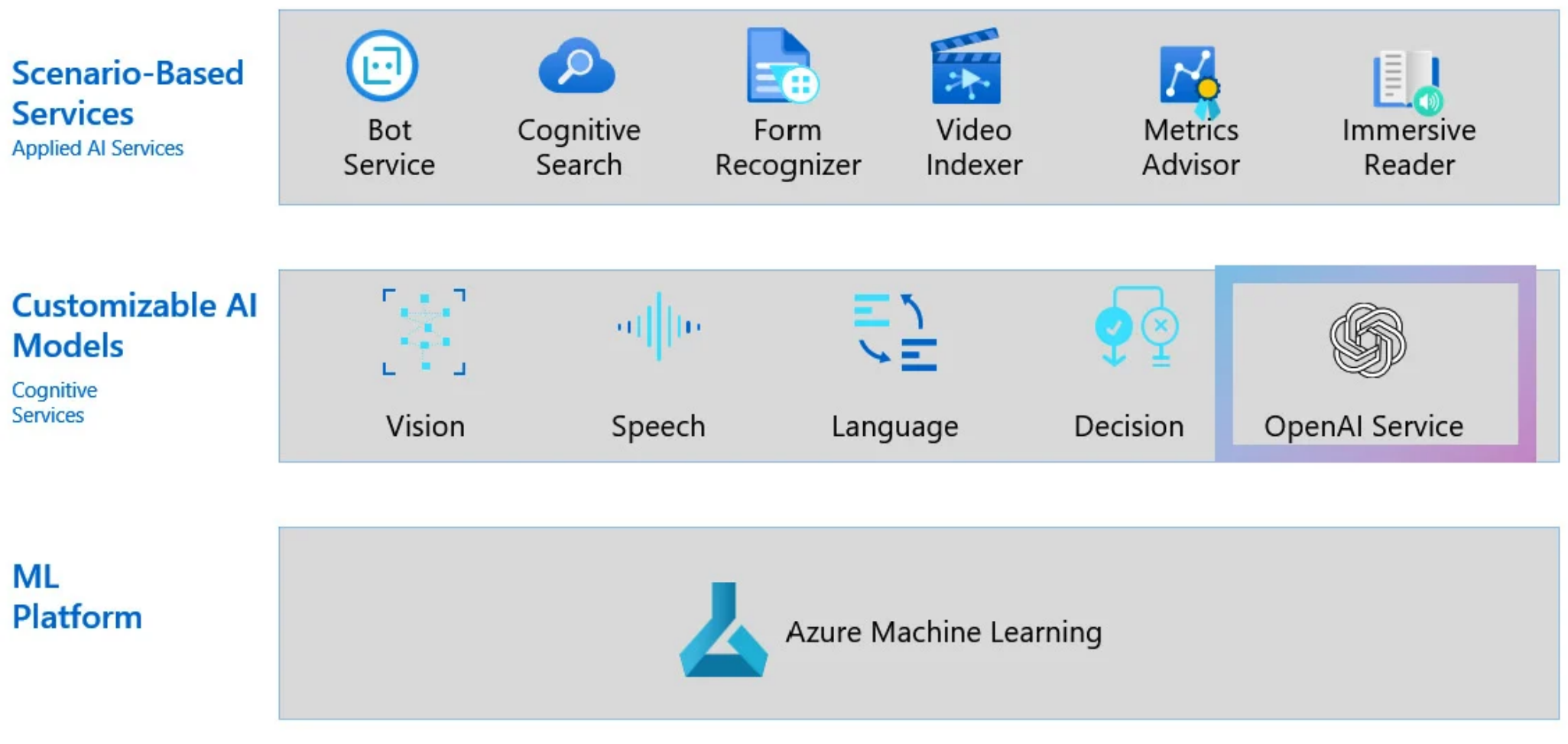

Figure 9.3 – AOAI Service General Availability (GA)

As with any other Cognitive Services offering, AOAI offers models that have already been trained and are ready to be consumed.

To create your AOAI resource, follow these instructions:

1. Navigate to the Azure portal at https://ms.portal.azure.com.
2. Click on **Create a resource**.
3. Type `azure openai` and click on **Create**.
4. Fill in the required information and click on **Review + create**.

This is shown in the following screenshot:

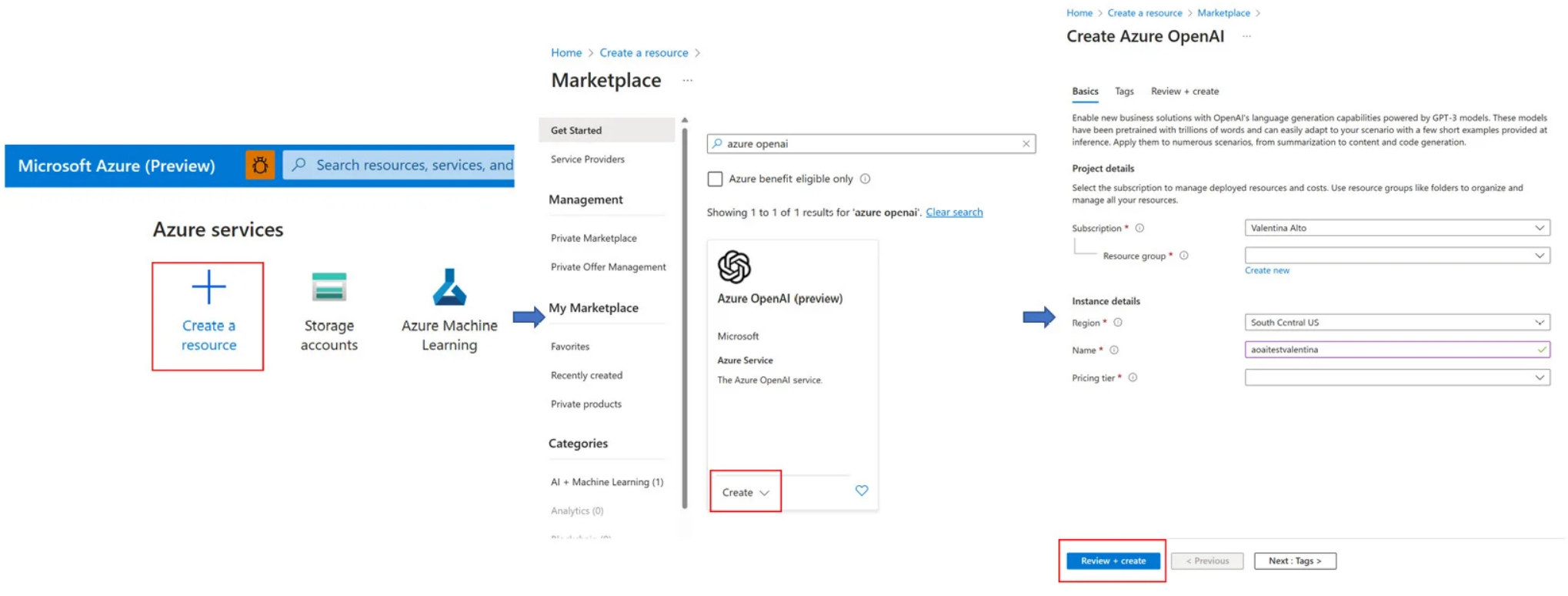

Figure 9.4 – Steps to create an AOAI resource

This process might take a few minutes. Once it is ready, you can directly jump to its user-friendly interface, AOAI Playground, to test your models before deploying them:

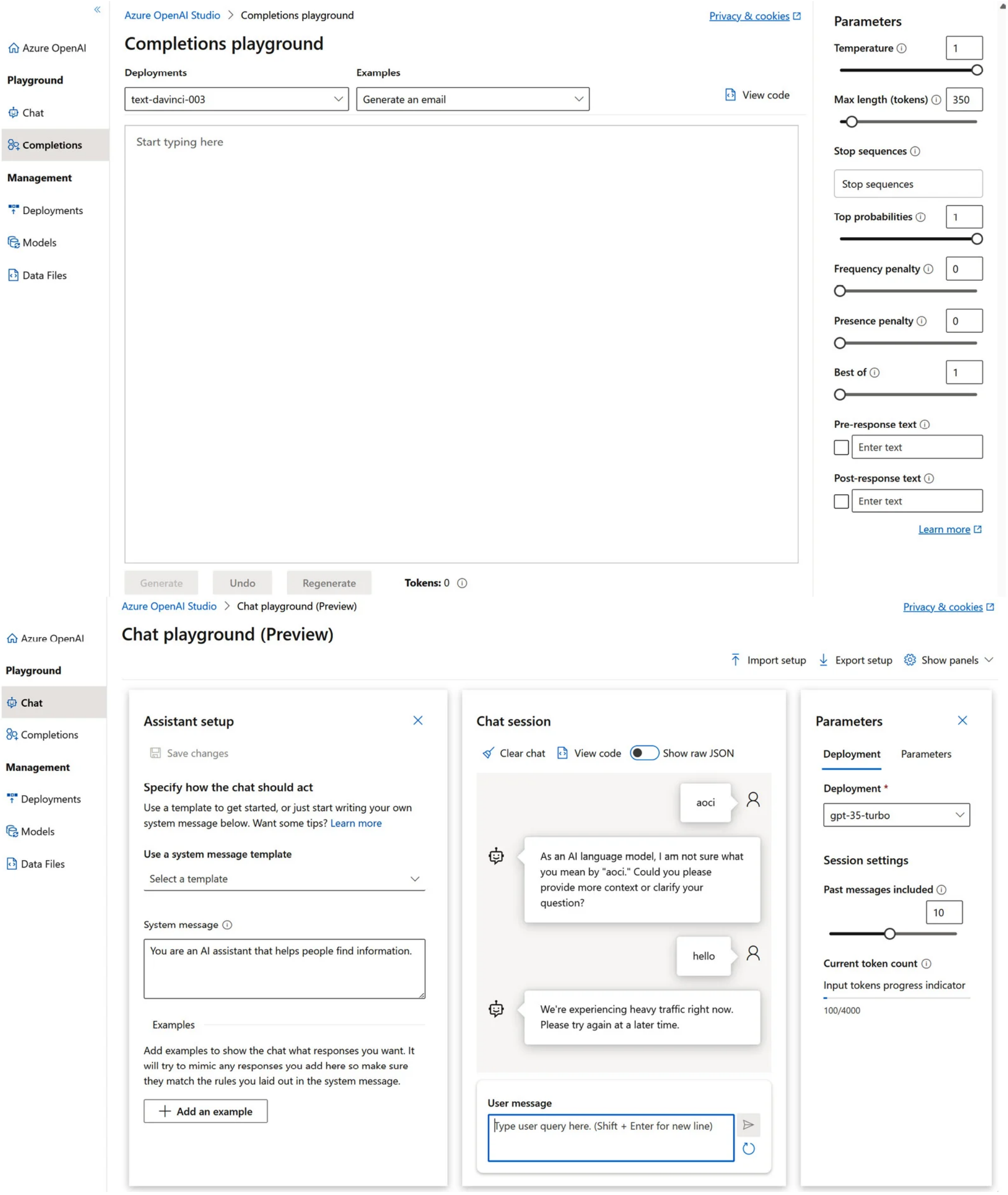

Figure 9.5 – AOAI UI and Playground

Note that AOAI Playground looks almost identical to the OpenAI Playground version we saw in *Chapter 2*. The difference here is that, to use AOAI models, you have to initiate a deployment, which is a serverless compute instance you can attach to a model. You can do so either in Playground or on the resource backend page in the Azure portal:

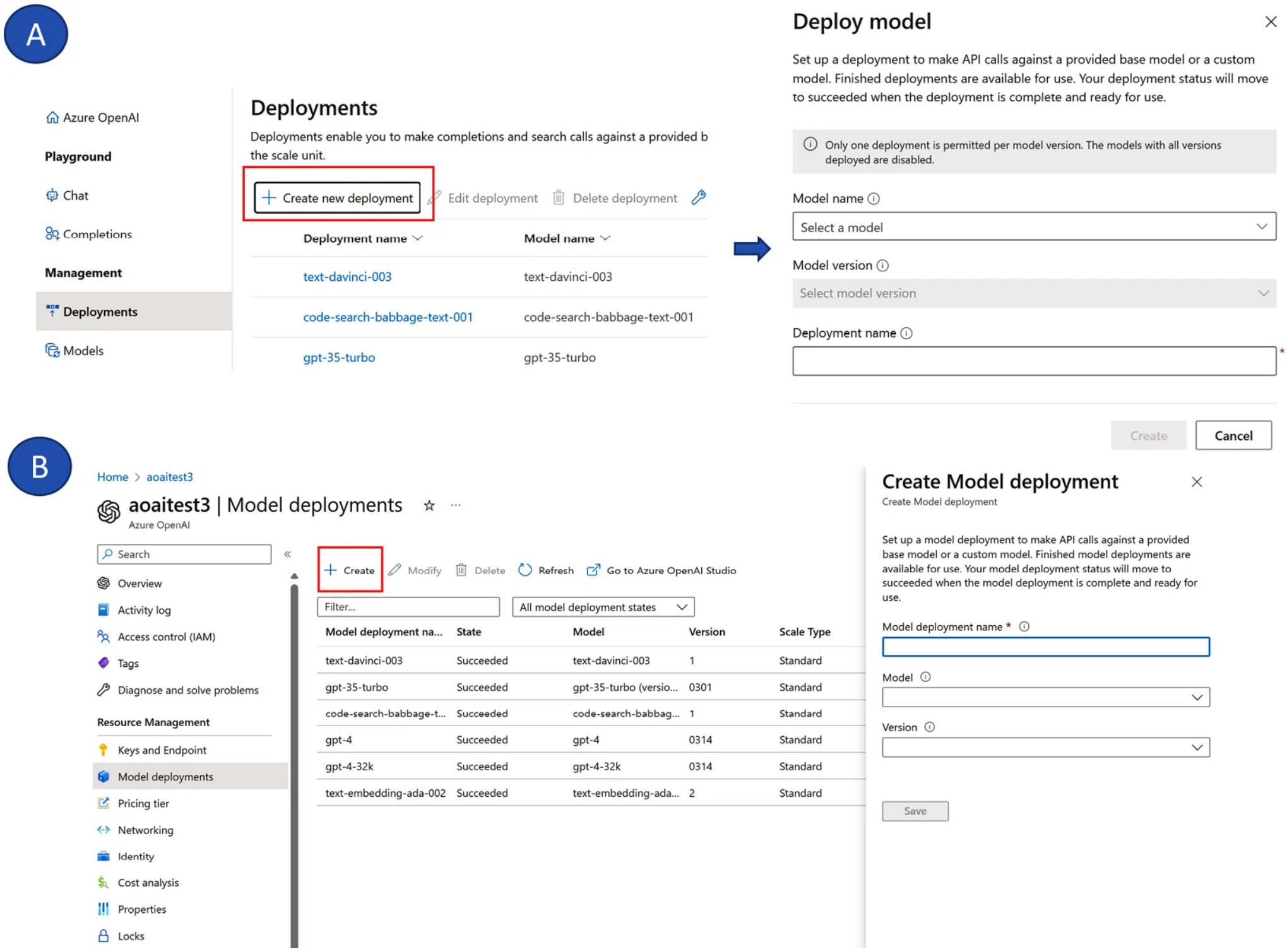

Figure 9.6 – Creating a new AOAI deployment via Playground (A) or in the Azure portal (B)

For example, I created a deployment called `text-davinci-003` with an associated `text-davinci-003` model:

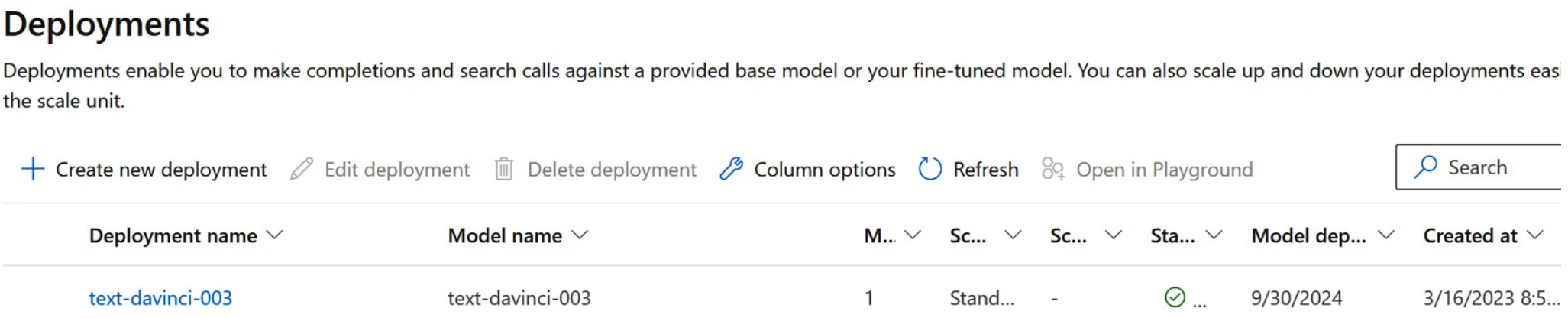

Figure 9.7 – An active deployment of AOAI

In OpenAI Playground, we can test those models either directly via the user interface or by embedding their APIs into our applications. In the next section, we are going to explore how to interact with Playground and try different models' configurations. In *Chapter 10*, we will learn how to integrate AOAI's Models API into enterprise applications.

Exploring Playground

AOAI Playground is the easiest way to get familiar with the underlying models and start planning which model's version is the most suitable for your projects. The user interface presents different tabs and workspaces, as shown in the following screenshot:

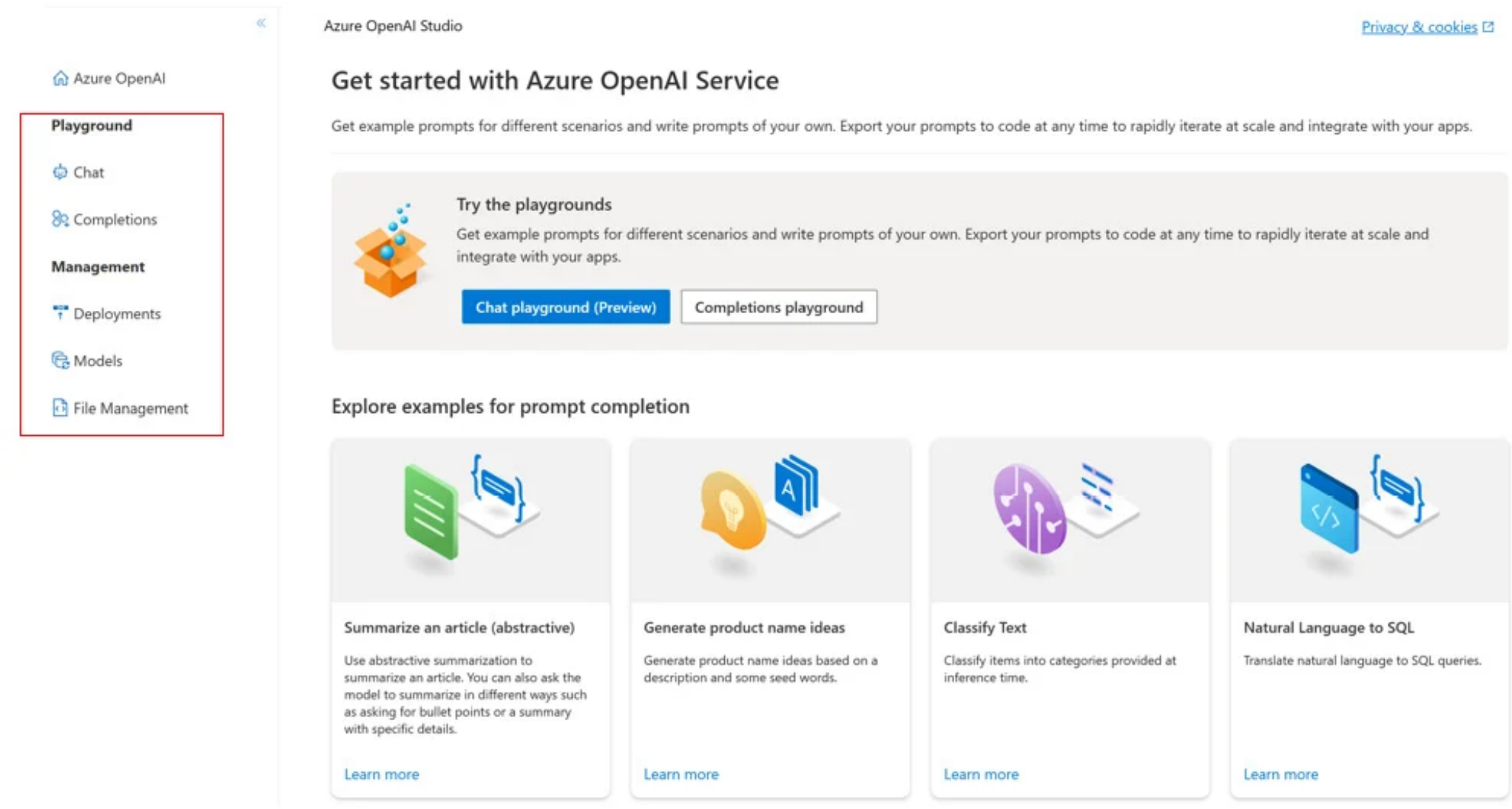

Figure 9.8 – Overview of AOAI Playground

Let's explore each of them:

Playground | Chat: The **Chat** workspace is designed to be only used with conversational models such as GPT-3.5-turbo (the model behind ChatGPT):

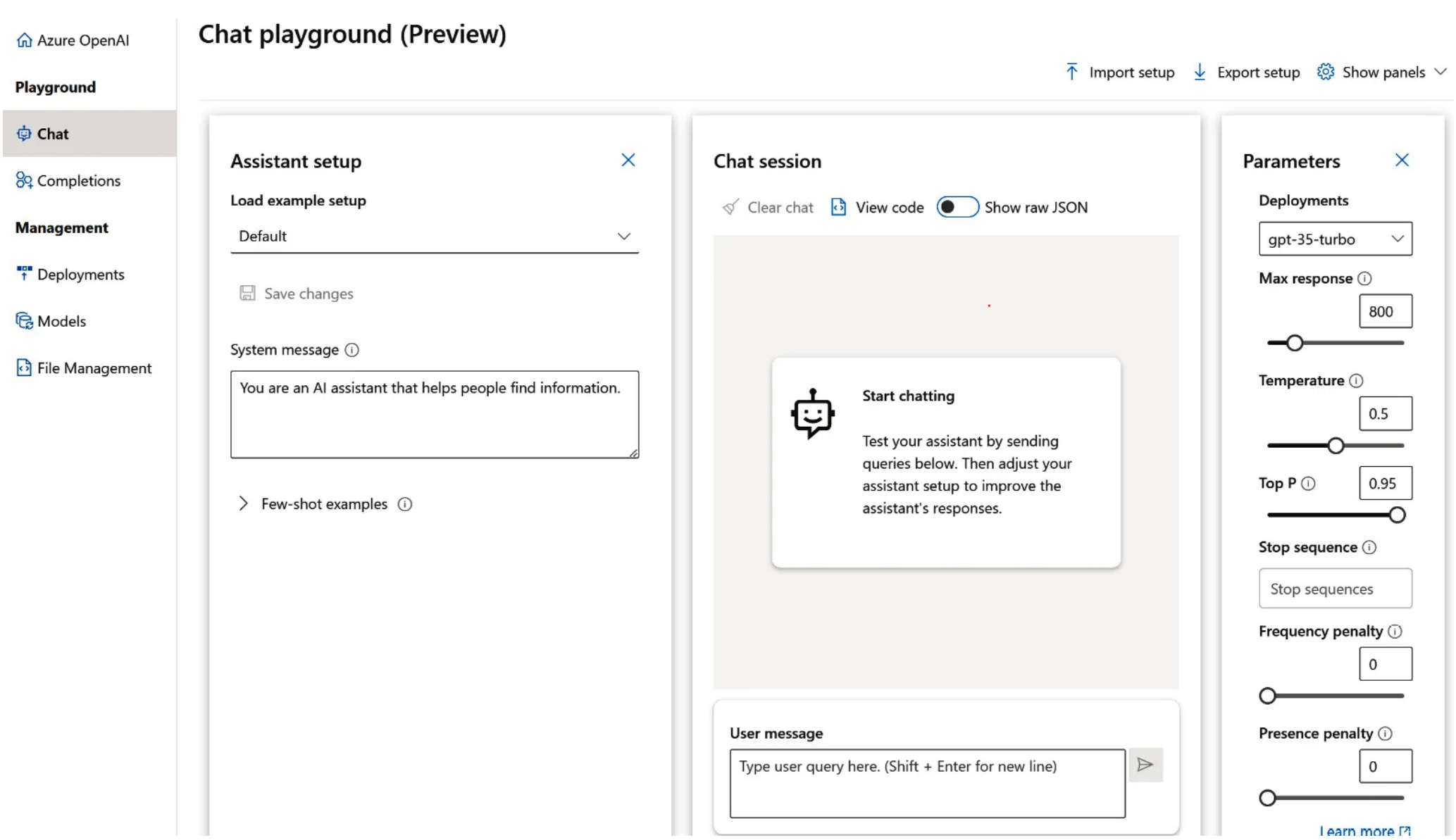

Figure 9.9 – AOAI Chat workspace

It offers a similar experience to ChatGPT itself, with the possibility to configure your model with additional parameters (as we saw in *Chapter 2* with OpenAI Playground). Furthermore, there is an additional feature that makes the **Chat** workspace very interesting, known as **System message:**

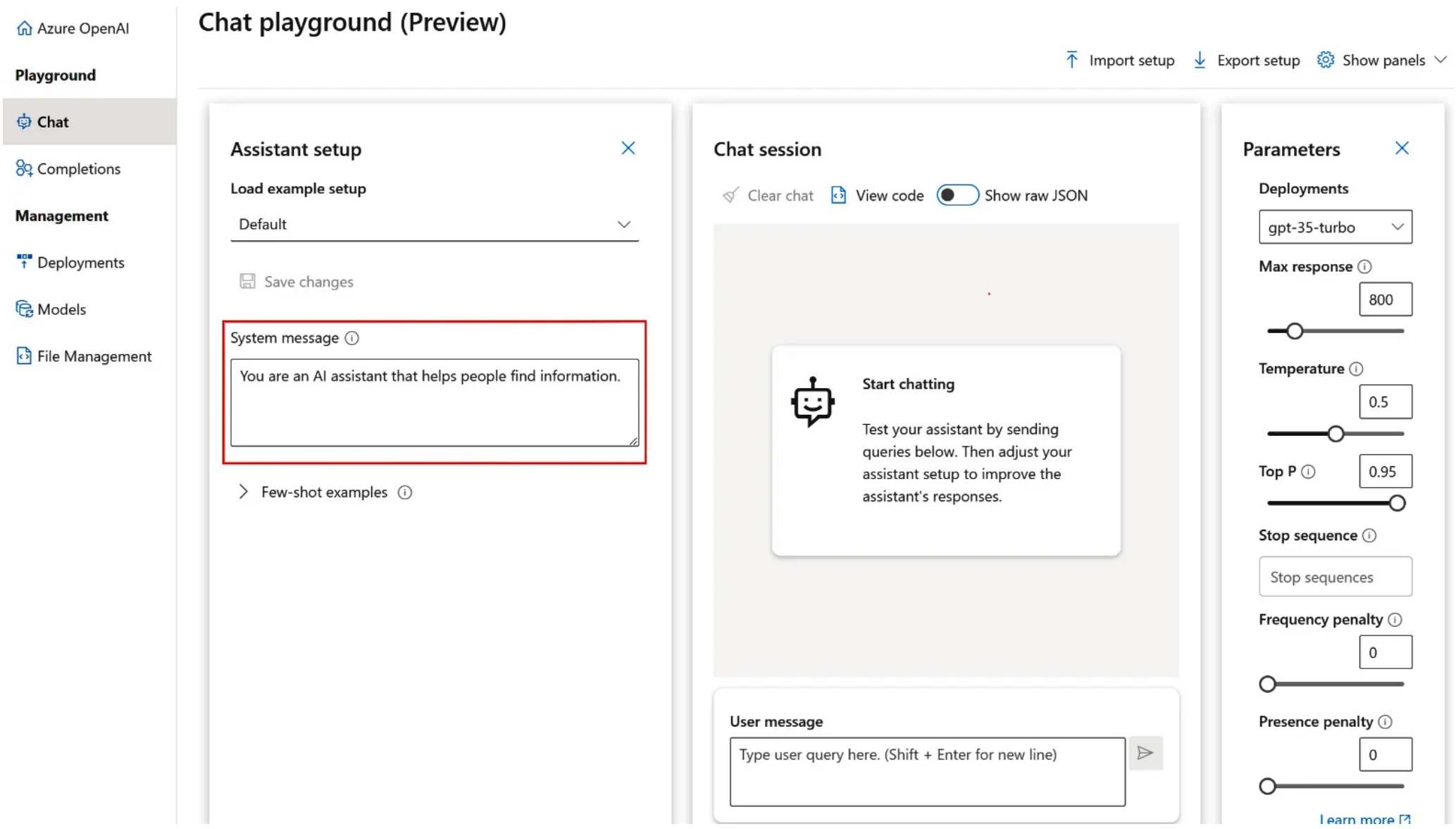

Figure 9.10 – Example of System message

System message is the set of instructions we give the model to tell it how to behave and interact with us. As for the prompt, **System message** represents a key component of a model's configuration since it massively affects model performance.

For example, let's instruct our model to behave as a JSON formatter assistant:

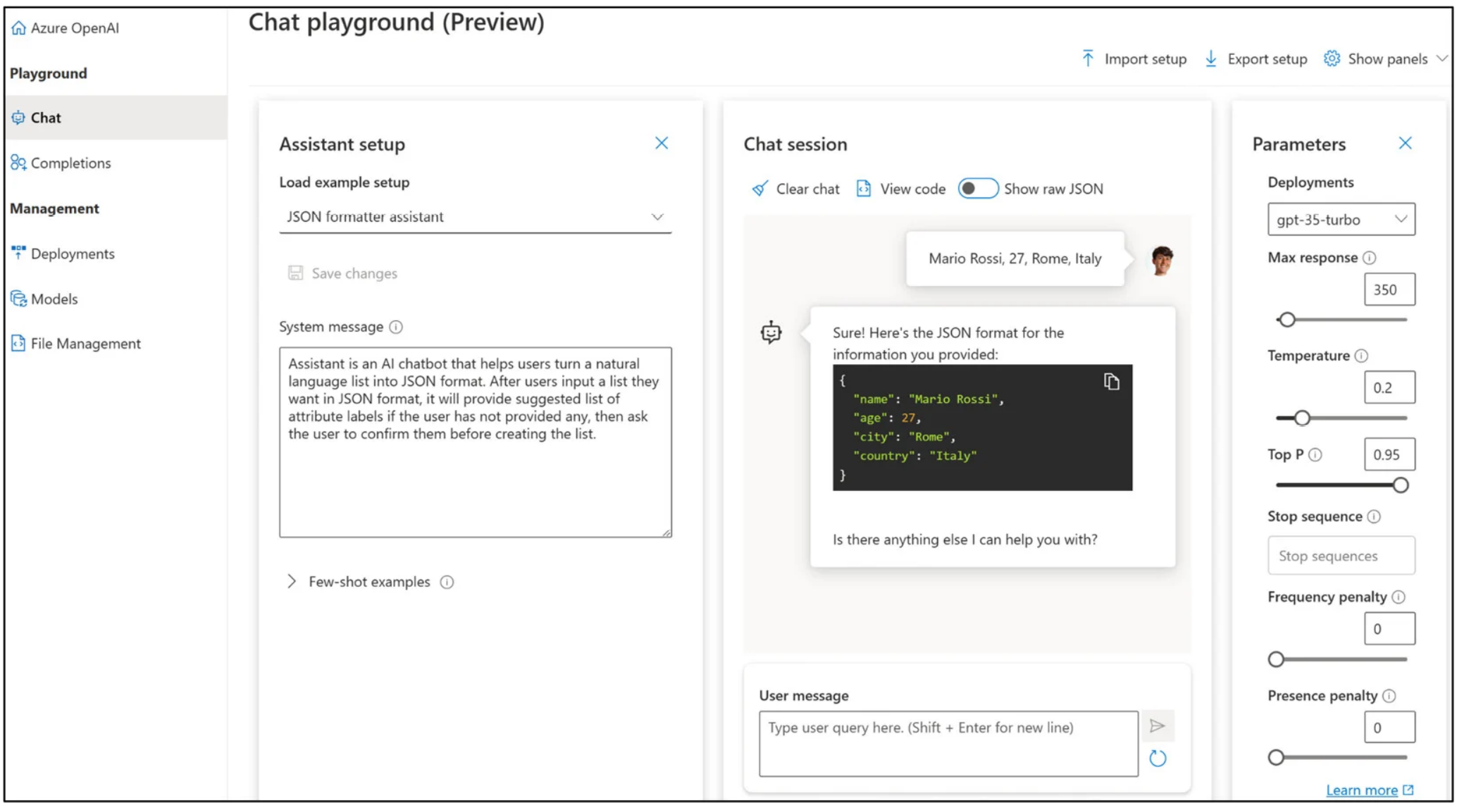

Figure 9.11 – Example of a model acting as a JSON formatter assistant

As you can see from the previous screenshot, the model was able to suggest a JSON file through some simple data, such as name and age, without the need to specify any labels.

Playground | Completions: Different from the previous workspace, the **Completions** workspace offers a sort of *white paper* where you can interact with your models. While GPT-3.5-turbo is designed for conversational tasks (which means it can be consumed via a chatbot-like interface), the GPT-3 series contains more general-purpose models and can be used for a wide range of language tasks, such as content generation, summarization, and so on.

For example, we could ask our model to generate a quiz by giving it a description of the topic and a one-shot example, as shown here:

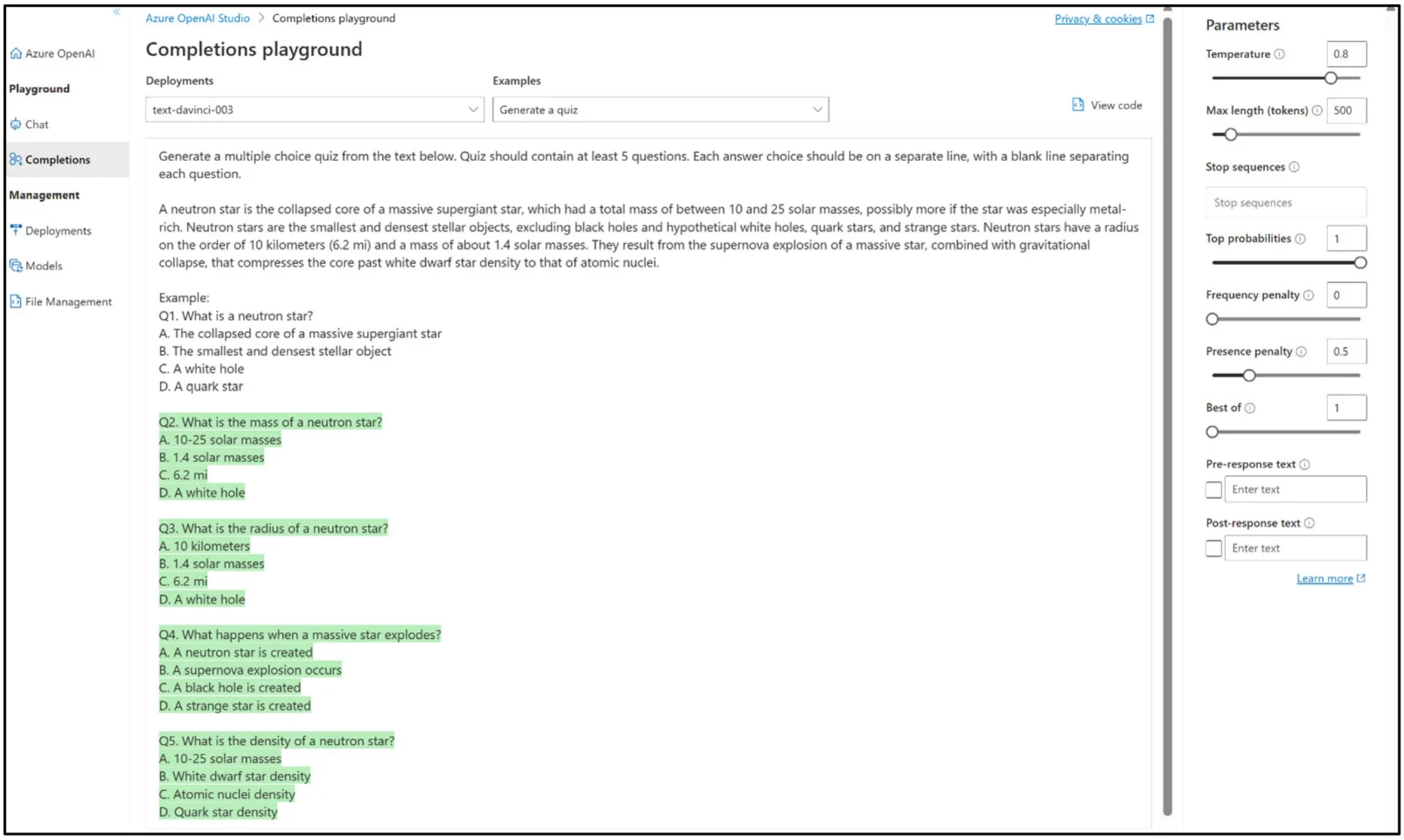

Figure 9.12 – Example of a GPT model generating a quiz

Finally, as per the **Chat** workspace, with **Completions**, you can configure parameters such as the maximum number of tokens or the temperature (refer to *Chapter 2* for a comprehensive list of those parameters and their meanings).

Management | Deployments: Within the **Deployments** tab, you can create and manage new deployments to be associated with AOAI models. They are depicted here:

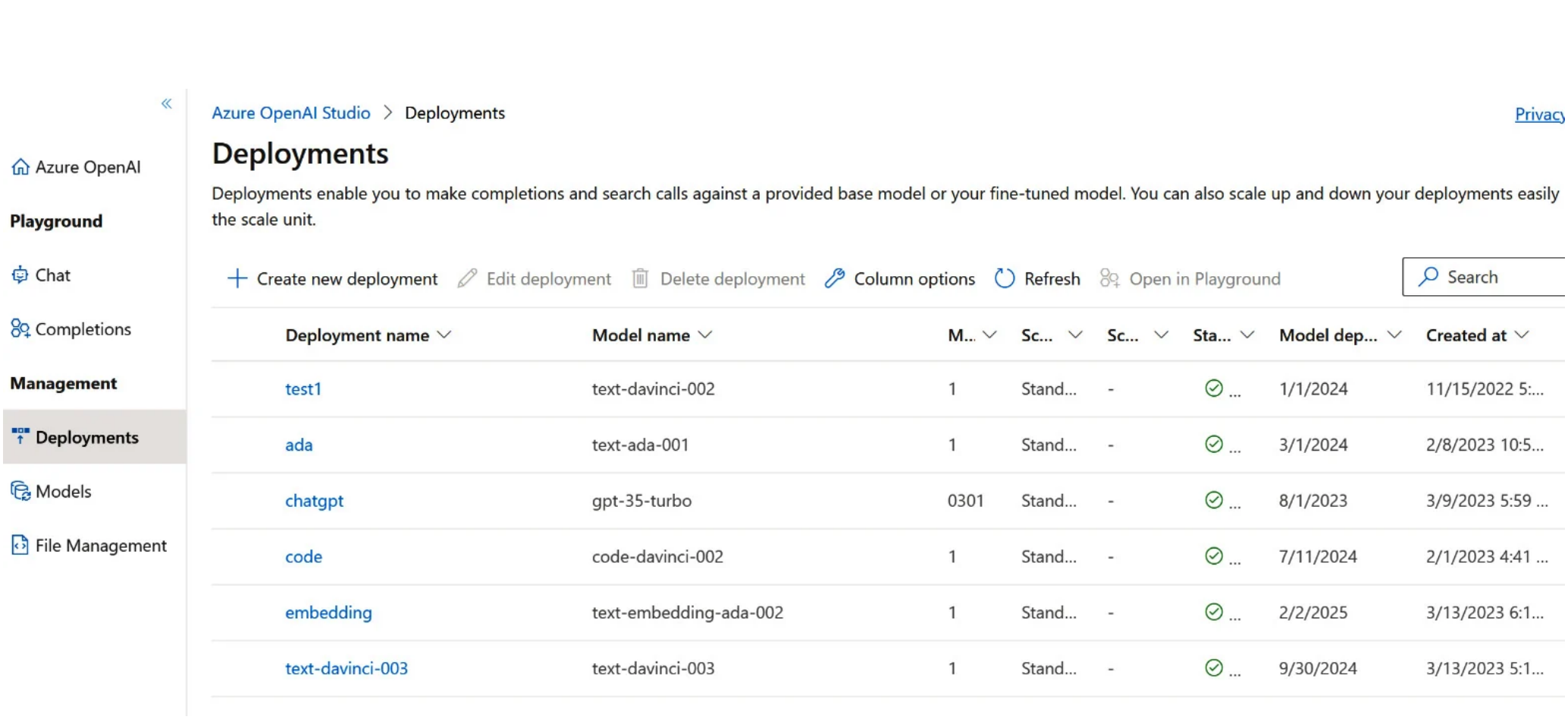

Figure 9.13 – List of AOAI deployments

Each deployment can host only one model. You can edit or delete your deployments at any time. As we mentioned previously, a model deployment is the enabler step for using either the **Completions** or **Chat** workspace within AOAI Service.

Management | Models: Within this tab, you can quickly assess the models that are available within AOAI Service and, among them, those that can be deployed (that is, a model that hasn't been deployed yet). For example, let's consider the following screenshot:

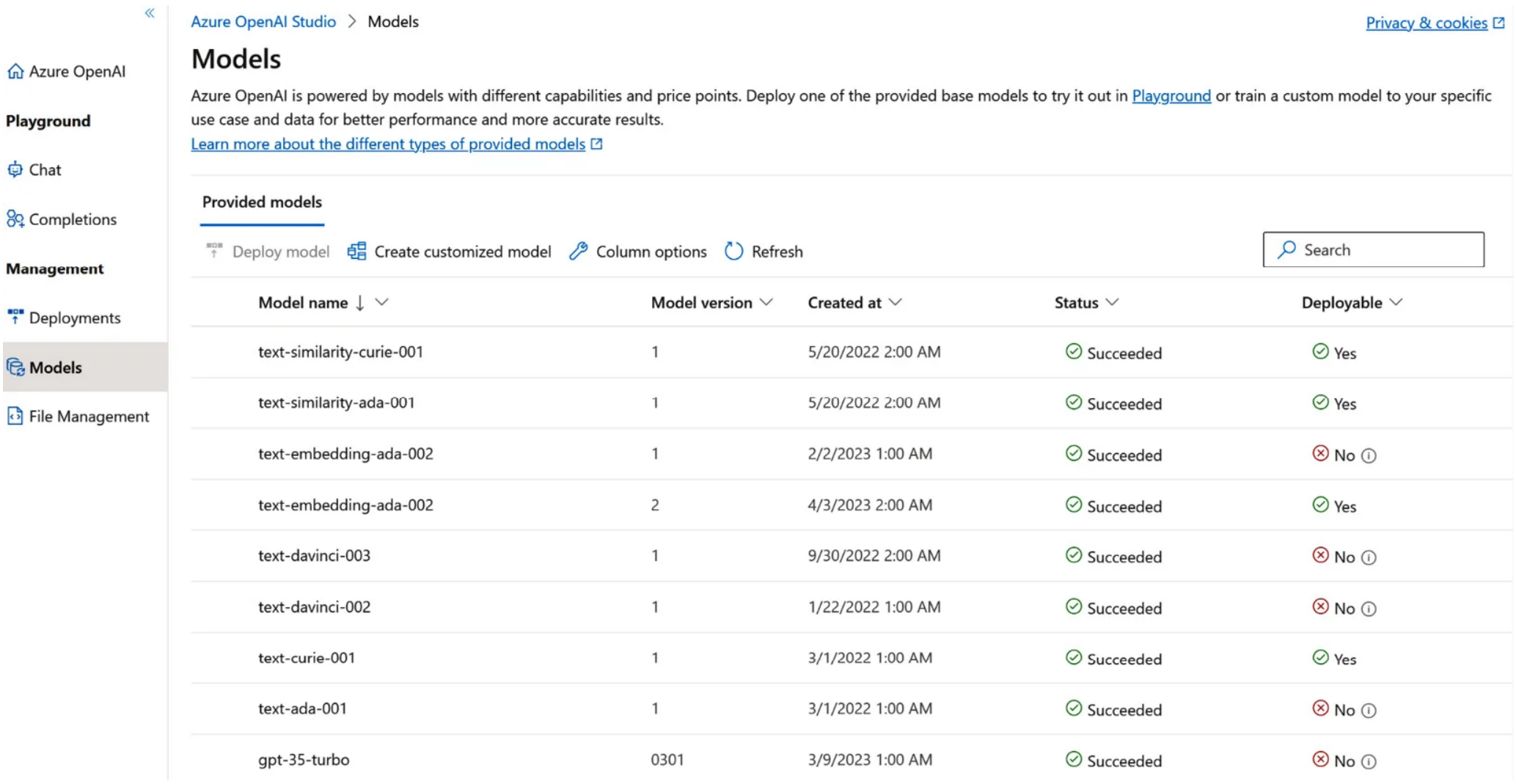

Figure 9.14 – List of AOAI models

Here, we have `text-similarity-curie-001`. It doesn't have an associated deployment, so it can be deployed (as the **Deployable** column shows). On the other hand, `text-similarity-ada-002` already has a deployment, so it is not available anymore.

Within this tab, you can also create a custom model by following a procedure called fine-tuning. We explored this in *Chapter 2*:

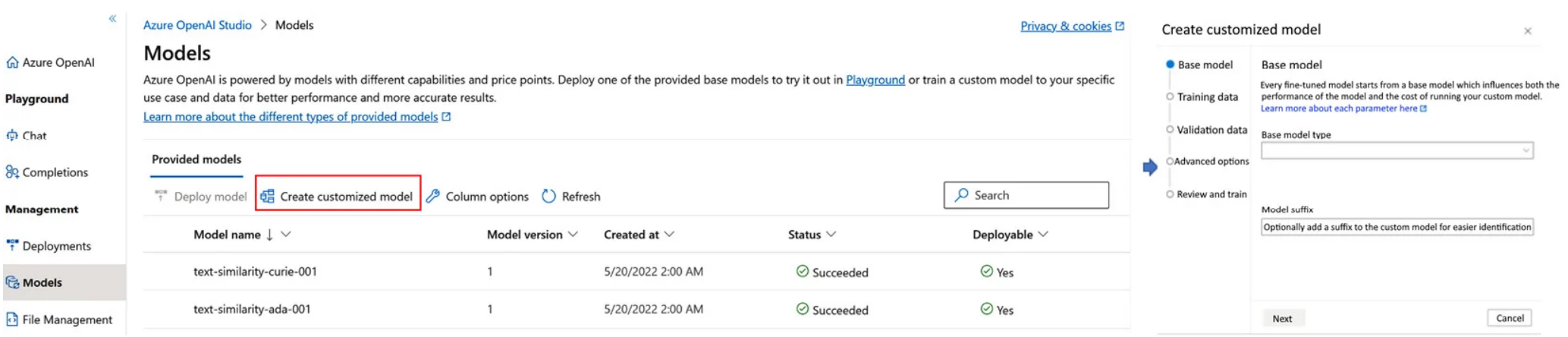

Figure 9.15 – Example of model fine-tuning

Starting from this guided widget, you can upload your training and validation data to produce a customized model, starting from a base model (namely, `text-davinci-002`), which will be hosted on a dedicated deployment.

Note

In *Chapter 2*, we saw that the training dataset should align with a specific format of the following type (called JSONL):

```
{"prompt": "<prompt text>", "completion": "<ideal generated text>"}

{"prompt": "<prompt text>", "completion": "<ideal generated text>"}

{"prompt": "<prompt text>", "completion": "<ideal generated text>"}

...
```

To facilitate this formatting, OpenAI has developed a tool that can format your data into this specific format ready for fine-tuning. It can also provide suggestions on how to modify data so that the tool can be used for fine-tuning. Plus, it accepts various data formats as inputs, including CSV, TXT, and JSON.

To use this tool, you can initialize the OpenAI **command-line interface** (**CLI**) by running the following command:

```
pip install --upgrade openai
```

Once initialized, you can run the tool, as follows:

```
openai tools fine_tunes.prepare_data -f <LOCAL_FILE>
```

Management | **File Management**: Finally, within the **File Management** tab, you can govern and upload your training and test data directly from the user interface, as shown here:

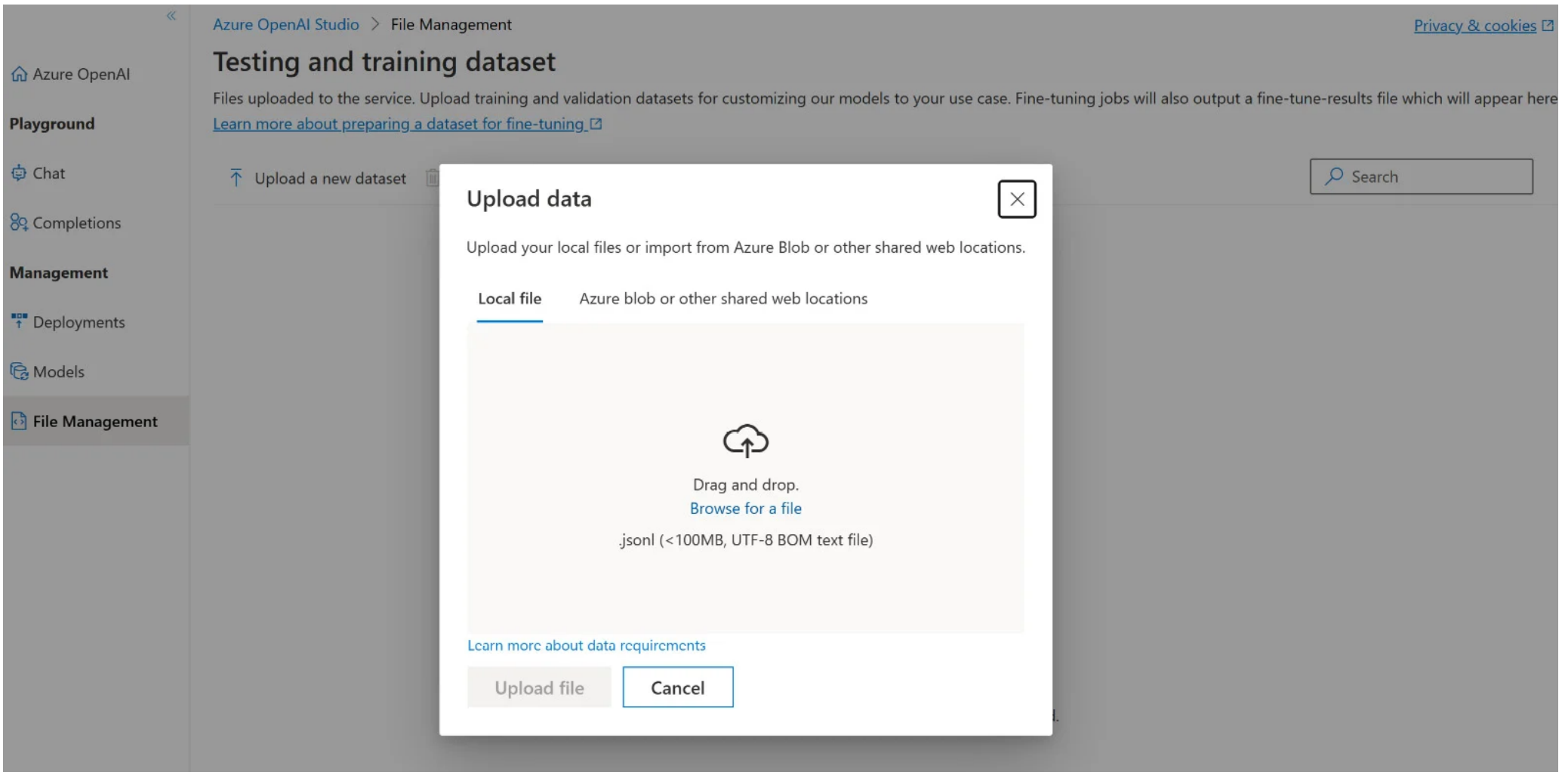

Figure 9.16 – Example of uploading a file within AOAI Service

You can decide to upload files by selecting **Local file** or **Azure blob or other shared web locations.**

Once you've uploaded your files, you will be able to select them while creating customized models, via the **Models** tab.

Finally, as mentioned in the previous section, each model comes with a REST API that can be consumed in your applications.

In the next chapter, we will see many end-to-end implementations of using AOAI's Models API. However, before we jump into that, we need to understand how AOAI differs from the standard OpenAI models and why the Azure cloud became part of the game.

Why introduce a public cloud?

At the beginning of this chapter, we saw how Microsoft and OpenAI have partnered in recent years and how Microsoft's cloud, Azure, became the *gym* for OpenAI model training. However, it also became the cloud infrastructure where OpenAI models can be consumed.

But what is the difference between using models from OpenAI and Azure OpenAI? The difference is the underlying infrastructure: with Azure OpenAI, you are leveraging your *own* infrastructure while living in your *own secured* subscription. This brings a series of advantages:

Scalability and flexibility: You can benefit from the scalability of Azure and accommodate the elastic usage of AOAI models. From small pilots to enterprise-level production projects, AOAI allows you to leverage the required capacity and scale up or down if necessary.

Security and compliance: You can use role-based authentication and private network connectivity to make your deployment more secure and trusted. You can also train your AI model while having full control of your data.

Regional availability: You can run your AI workloads on the Azure global infrastructure that meets your production needs.

Built-in responsible AI: You can use content filtering to ensure that your AI model generates appropriate and ethical output.

With the OpenAI models available in Azure, we can elevate the game to the enterprise and production levels, meeting all security and capacity requirements typical of large organizations.

One of the previously mentioned benefits deserves a particular focus: responsible AI. The rapid development of AI technologies also needs to be addressed in terms of ethical tools. This is what Microsoft has been studying since 2016, as we will explore in the next section.

Understanding responsible AI

We mentioned the built-in responsible AI as one of the key features of AOAI. However, to fully understand it, we first need to understand Microsoft's commitment and journey toward responsible AI.

Microsoft's journey toward responsible AI

Microsoft soon recognized that as AI technologies continue to advance and become more integrated into our lives, there is a growing need to ensure that those systems are developed and used responsibly, ethically, and in ways that benefit everyone.

The beginning of this journey traces back to 2016 when Microsoft's CEO Satya Nadella penned an article exploring how humans and AI can work together to solve society's greatest challenges and introducing the first concepts of responsible AI, among which are transparency, fairness, and that it is designed for privacy and to assist humanity.

Shortly after, in 2017, Microsoft formalized those concepts with the first AI ethics committee – **Aether** (short for **AI, Ethics, and Effects in Engineering and Research**) – formed as an advisory group for the Microsoft senior leadership team.

AETHER spent time listening to customers and internal experts, and then partnered with legal affairs to publish the book titled *The Future Computed: Artificial Intelligence and its role in society* in January 2018. In this book, Microsoft identified six principles meant to guide a company's development of AI systems, as well as to help inform the broader industry and society as a whole about responsible AI practices.

Microsoft's six principles for responsible AI are as follows:

Fairness: Microsoft aims to create AI systems that are unbiased and treat all individuals and groups fairly, without discrimination or prejudice

Reliability and safety: Microsoft seeks to create AI systems that are robust, reliable, and secure, and that do not compromise safety or create unintended harm

Privacy and security: Microsoft values the privacy and security of individuals and their data, and works to protect them through transparency and responsible use of AI technologies

Inclusiveness: Microsoft believes that AI should be designed to empower and include individuals from diverse backgrounds, and to foster equal opportunities for all

Transparency: Microsoft believes in transparency and accountability for the decisions and actions of AI systems and is committed to providing clear explanations for their outcomes

Accountability: Microsoft accepts responsibility for the impact of its AI systems on society and the environment, and seeks to promote ethical and responsible practices in the development and use of AI

Microsoft follows these principles with the help of committees that offer guidance to its leadership, engineering teams, and every other team within the company.

Microsoft also has a **Responsible AI Standard** that provides a framework for building AI systems responsibly.

Following the publication of that book, Microsoft kept investing and researching in the following fields of responsible AI:

From the contribution to government regulation in the field of facial recognition (2018, https://www.geekwire.com/2018/microsoft-calls-government-regulation-facial-recognition-technology/, https://blogs.microsoft.com/on-the-issues/2018/12/17/six-principles-to-guide-microsofts-facial-recognition-work/) to the establishment of responsible AI in systems engineering or RAISE (2020, https://www.microsoft.com/en-us/ai/our-approach?activetab=pivot1%3aprimaryr5)

The development of responsible AI tools in the areas of ML interpretability, unfairness assessment and mitigation, error analysis, causal inference, and counterfactual analysis (2021, https://responsibleaitoolbox.ai/)

The following diagram shows the entire journey for responsible AI:

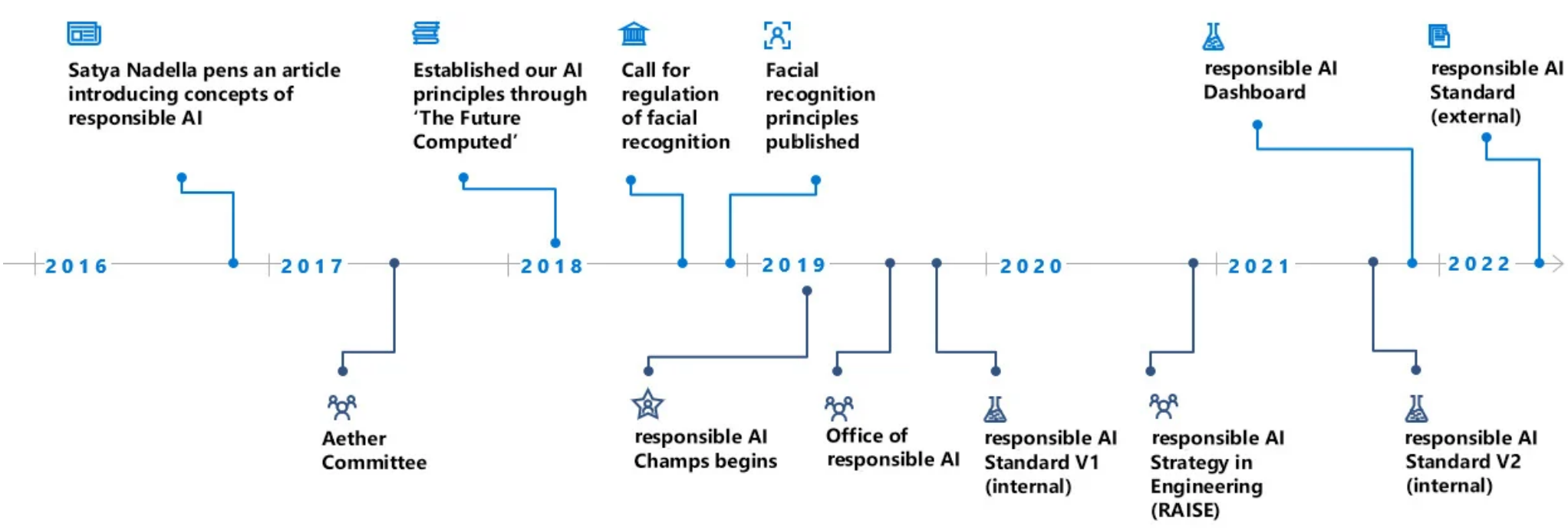

Figure 9.17 – Microsoft's responsible AI journey

Microsoft's commitment to responsible AI is reflected in the way its products are designed and the best practices and guidelines provided.

Of course, this also applies to AOAI Service. As we will see in the next section, AOAI Service comes with a built-in responsible AI at a different level.

Azure OpenAI and responsible AI

When it comes to AOAI Service, we can talk about responsible AI at the following two levels:

Built-in: In this case, we refer to AOAI embedded features of responsible AI that are enforced by a content management system. This system utilizes a series of classification models to detect harmful content. The system works alongside core models to filter content by analyzing both the input prompt and generated content. In cases where harmful content is identified, you'll receive either an error on the API call if the prompt was detected as inappropriate, or see that the `finish_reason` parameter on the response in JSON will be `content_filter` to signify that some of the generation was filtered.

Code of conduct and best practices: As for its other AI services, Microsoft provides **Transparency Notes** for AOAI. This application aims to promote an understanding of how AI technology works, its limitations and capabilities, and the importance of considering the entire system, including people and the environment. These notes can be used by developers and system owners to create AI systems that are fit for their intended purpose and, in the specific case of AOAI, help identify those scenarios that might trigger the built-in content filter.

Both the built-in capabilities and Transparency Notes are manifestations of Microsoft's effort to apply ethical AI practices in real-world scenarios, guided by their AI principles.

In conclusion, as responsible AI for Microsoft signifies the company's unwavering commitment to ethical AI development and deployment, AOAI also benefits from this commitment.

Summary

In this chapter, we saw how the partnership between OpenAI and Microsoft has brought about a powerful and innovative AI solution for enterprise-level organizations: AOAI. This service combines OpenAI's cutting-edge technology with

Microsoft's extensive cloud infrastructure to provide businesses with a scalable and customizable platform for building and deploying advanced AI applications.

We also dwelled on Microsoft's strong focus on responsible AI practices and ethics, and how AOAI Service reflects this commitment to responsible AI, with features such as a content filter built into the platform.

As AI continues to transform industries and shape our future, the collaboration between OpenAI and Microsoft marks an important milestone in the development of enterprise-level AI solutions. AOAI empowers businesses to harness the power of AI to drive growth and innovation while ensuring ethical and responsible practices.

In the next chapter, we will dive deeper into concrete use cases that enterprises are developing with the AOAI Models API. We will also see an end-to-end implementation of a potential use case that uses Python and Streamlit so that you can experience firsthand how AOAI's models can infuse your applications with AI.

References

https://blogs.microsoft.com/blog/2023/01/23/microsoftandopenaiextendpartnership/

https://news.microsoft.com/2019/07/22/openai-forms-exclusive-computing-partnership-with-microsoft-to-build-new-azure-ai-supercomputing-technologies/

https://azure.microsoft.com/en-us/blog/general-availability-of-azure-openai-service-expands-access-to-large-advanced-ai-models-with-added-enterprise-benefits/

https://slate.com/technology/2016/06/microsoft-ceo-satya-nadella-humans-and-a-i-can-work-together-to-solve-societys-challenges.html

https://www.geekwire.com/2018/microsoft-calls-government-regulation-facial-recognition-technology/

https://blogs.microsoft.com/on-the-issues/2018/12/17/six-principles-to-guide-microsofts-facial-recognition-work/

https://www.microsoft.com/en-us/ai/our-approach?activetab=pivot1%3aprimaryr5

https://responsibleaitoolbox.ai/

https://www.microsoft.com/en-us/research/publication/human-parity-on-commonsenseqa-augmenting-self-attention-with-external-attention/

https://learn.microsoft.com/en-gb/azure/cognitive-services/openai/how-to/fine-tuning?pivots=programming-language-studio#openai-cli-data-preparation-tool

Previous Chapter

Next Chapter