

# Azure OpenAI Service 설명서

GPT-3, Codex 및 Embeddings 모델 시리즈를 포함한 Azure OpenAI 강력한 언어 모델을 사용하여 콘텐츠 생성, 요약, 의미 체계 검색 및 자연어를 코드로 번역을 수행하는 방법을 알아봅니다.



개요  
Azure OpenAI Service란?



빠른 시작  
빠른 시작



방법 가이드  
리소스 만들기



자습서  
포함



방법 가이드  
완성



학습  
Azure OpenAI 교육 소개



개념  
Azure OpenAI 모델



참조  
지원 및 도움말 옵션

## 추가 자료

### Azure OpenAI

[Azure OpenAI Studio ↗](#)

[지역 지원 ↗](#)

[할당량 및 제한](#)

[Azure OpenAI에 대한 액세스 신청 ↗](#)

### 동영상

Azure의 강력한 성능과 OpenAI 모델 결합

## 참조

[REST API](#)

[사용 약관](#)

## 도구

[Azure CLI](#)

[PowerShell](#)

# Azure OpenAI Service란?

아티클 • 2023. 05. 15.

Azure OpenAI Service는 GPT-3, Codex 및 Embeddings 모델 시리즈를 포함한 OpenAI의 강력한 언어 모델에 대한 REST API 액세스를 제공합니다. 또한 새로운 GPT-4 및 ChatGPT(gpt-35-turbo) 모델 시리즈는 이제 일반 공급에 도달했습니다. 이러한 모델은 콘텐츠 세대, 요약, 의미 체계 검색, 자연어에서 코드로의 번역을 포함하되 이에 국한되지 않는 특정 작업에 쉽게 적용할 수 있습니다. 사용자는 REST API, Python SDK 또는 Azure OpenAI Studio의 웹 기반 인터페이스를 통해 서비스에 액세스할 수 있습니다.

## 기능 개요

기능	Azure OpenAI
사용 가능한 모델	<b>새 GPT-4 시리즈</b> GPT-3 베이스 시리즈 <b>NEW ChatGPT(gpt-35-turbo)</b> Codex 시리즈 Embeddings 시리즈 <a href="#">모델</a> 페이지에서 자세히 알아보세요.
미세 조정	Ada Babbage Curie Cushman Davinci <b>현재 새 고객은 미세 조정을 사용할 수 없습니다.</b>
가격	<a href="#">여기에서 사용 가능 ↗</a>
가상 네트워크 지원 및 프라이빗 링크 지원	예
관리 ID	예, Azure Active Directory를 통해
UI 환경	계정 & 리소스 관리를 위한 <a href="#">Azure Portal</a> , 모델 탐색 및 미세 조정을 위한 <a href="#">Azure OpenAI Service Studio</a>
국가별 가용성	미국 동부 미국 중남부 서유럽
콘텐츠 필터링	프롬프트 및 완료는自動화된 시스템을 사용하여 콘텐츠 정책에 따라 평가됩니다. 심각도가 높은 콘텐츠는 필터링됩니다.

# Responsible AI

Microsoft는 사용자를 최우선으로 하는 원칙에 따라 AI를 발전시키기 위해 최선을 다하고 있습니다. Azure OpenAI에서 사용할 수 있는 것과 같은 생성 모델은 상당한 잠재적 이점이 있지만 신중한 디자인과 사려 깊은 완화 없이 이러한 모델은 부정확하거나 심지어 유해한 콘텐츠를 생성할 가능성이 있습니다. Microsoft는 남용 및 의도하지 않은 피해를 방지하기 위해 상당한 투자를 했습니다. 여기에는 신청자가 잘 정의된 사용 사례를 보여 주도록 요구하고, Microsoft의 [책임 있는 AI 사용 원칙](#)을 통합하고, 고객을 지원하기 위한 콘텐츠 필터를 빌드하고, 온보딩 고객에게 책임 있는 AI 구현 지침을 제공하는 것이 포함됩니다.

## Azure OpenAI에 액세스하려면 어떻게 해야 하나요?

Azure OpenAI에 액세스하려면 어떻게 해야 하나요?

현재 높은 수요, 예정된 제품 개선 사항, [책임 있는 AI에 대한 Microsoft의 약속](#)을 탐색하기 때문에 액세스가 제한됩니다. 현재 우리는 Microsoft와의 기존 파트너십, 위협이 낮은 사용 사례 및 완화 통합에 전념하는 고객과 협력하고 있습니다.

보다 구체적인 정보는 신청 양식에 포함되어 있습니다. Azure OpenAI에 대한 더 광범위한 액세스를 책임감 있게 가능하게 하기 위해 노력하는 동안 양해해 주셔서 감사합니다.

액세스를 위해 여기에서 신청합니다.

[지금 적용](#)

## Azure OpenAI 및 OpenAI 비교

Azure OpenAI Service는 Azure의 보안 및 엔터프라이즈 지원을 통해 OpenAI GPT-4, GPT-3, Codex 및 DALL-E 모델을 사용하는 고급 언어 AI를 고객에게 제공합니다. Azure OpenAI는 OpenAI와 API를 공동 개발하여 호환성과 원활한 전환을 보장합니다.

Azure OpenAI를 사용하면 고객은 OpenAI와 동일한 모델을 실행하면서 Microsoft Azure의 보안 기능을 얻을 수 있습니다. Azure OpenAI는 프라이빗 네트워킹, 지역 가용성 및 책임 있는 AI 콘텐츠 필터링을 제공합니다.

## 주요 개념

### 프롬프트 및 완성

완성 엔드포인트는 API 서비스의 핵심 구성 요소입니다. 이 API는 모델의 텍스트 입력, 텍스트 출력 인터페이스에 대한 액세스를 제공합니다. 사용자는 영어 텍스트 명령이 포함된 입력 **프롬프트**를 제공하기만 하면 모델에서 텍스트 **완성을** 생성합니다.

간단한 프롬프트 및 완성 예제는 다음과 같습니다.

**프롬프트:** `""" count to 5 in a for loop """`

**완성:** `for i in range(1, 6): print(i)`

## 토큰

Azure OpenAI는 텍스트를 토큰으로 분해하여 처리합니다. 토큰은 단어 또는 문자 청크일 수 있습니다. 예를 들어 "hamburger"라는 단어는 "ham", "bur" 및 "ger" 토큰으로 분해되지만, "pear"와 같은 짧고 일반적인 단어는 단일 토큰입니다. 많은 토큰이 공백으로 시작합니다(예: " hello" 및 " bye").

지정된 요청에서 처리되는 총 토큰 수는 입력, 출력 및 요청 매개 변수의 길이에 따라 달라집니다. 처리되는 토큰의 양은 모델의 응답 대기 시간 및 처리량에도 영향을 줍니다.

## 리소스

Azure OpenAI는 Azure에서 제공되는 새로운 제품입니다. Azure OpenAI는 Azure 구독에서 [리소스](#) 또는 서비스 인스턴스를 만드는 다른 Azure 제품과 동일한 방식으로 시작할 수 있습니다. Azure의 [리소스 관리 디자인](#)에 대해 자세히 알아볼 수 있습니다.

## 배포

Azure OpenAI 리소스가 만들어지면 API 호출 및 텍스트 생성을 시작하기 전에 먼저 모델을 배포해야 합니다. 이 작업은 배포 API를 사용하여 수행할 수 있습니다. 이러한 API를 사용하면 사용하려는 모델을 지정할 수 있습니다.

## 컨텍스트 내 학습

Azure OpenAI에서 사용하는 모델은 생성 호출 중에 제공되는 자연어 지침 및 예제를 사용하여 요청되는 작업과 필요한 기술을 식별합니다. 이 방법을 사용하는 경우 원하는 특정 작업에 대한 자연어 지침 및/또는 예제가 프롬프트의 첫 번째 부분에 포함됩니다. 그러면 모델에서 가장 가능성이 높은 텍스트의 다음 부분을 예측하여 작업을 완성합니다. 이 기술을 "컨텍스트 내" 학습이라고 합니다. 이러한 모델은 이 단계에서 재학습되지 않는 대신 프롬프트에서 포함한 컨텍스트를 기반으로 하는 예측을 제공합니다.

컨텍스트 내 학습에는 소수 샷, 단일 샷 및 제로 샷의 세 가지 주요 방법이 있습니다. 이러한 방법은 모델에 제공되는 작업별 데이터의 양에 따라 달라집니다.

**소수 샷:** 이 경우 사용자는 예상 응답 형식 및 콘텐츠를 보여주는 몇 가지 예제를 호출 프롬프트에 포함합니다. 다음 예제에서는 여러 예제를 제공하는 몇 번의 프롬프트를 보여줍니다.(모델은 마지막 답변을 생성합니다.)

```
Convert the questions to a command:  
Q: Ask Constance if we need some bread.  
A: send-msg `find constance` Do we need some bread?  
Q: Send a message to Greg to figure out if things are ready for  
Wednesday.  
A: send-msg `find greg` Is everything ready for Wednesday?  
Q: Ask Ilya if we're still having our meeting this evening.  
A: send-msg `find ilya` Are we still having a meeting this evening?  
Q: Contact the ski store and figure out if I can get my skis fixed  
before I leave on Thursday.  
A: send-msg `find ski store` Would it be possible to get my skis fixed  
before I leave on Thursday?  
Q: Thank Nicolas for lunch.  
A: send-msg `find nicolas` Thank you for lunch!  
Q: Tell Constance that I won't be home before 19:30 tonight – unmovable  
meeting.  
A: send-msg `find constance` I won't be home before 19:30 tonight. I  
have a meeting I can't move.  
Q: Tell John that I need to book an appointment at 10:30.  
A:
```

예제 수는 일반적으로 단일 프롬프트의 최대 입력 길이에 맞출 수 있는 수에 따라 0~100개입니다. 최대 입력 길이는 사용하는 특정 모델에 따라 달라질 수 있습니다. 소수 샷 학습을 사용하면 정확한 예측에 필요한 작업별 데이터의 양을 크게 줄일 수 있습니다. 이 방법은 일반적으로 미세 튜닝된 모델보다 덜 정확하게 수행됩니다.

**단일 샷:** 이 경우 하나의 예제만 제공된다는 점을 제외하고는 소수 샷 방법과 동일합니다.

**제로 샷:** 이 경우 모델에 예제가 제공되지 않고 작업 요청만 제공됩니다.

## 모델

이 서비스는 사용자에게 몇 가지 다른 모델에 대한 액세스를 제공합니다. 각 모델은 다른 기능과 가격대를 제공합니다.

GPT-4 모델은 사용 가능한 최신 모델입니다. 이 모델 시리즈에 대한 수요가 높기 때문에 현재 요청에 의해서만 사용할 수 있습니다. 액세스를 요청하기 위해 기존 Azure OpenAI 고객은 [이 양식을 작성하여 적용](#) 할 수 있습니다.

GPT-3 기본 모델은 기능의 내림차순과 속도의 오름차순으로 Davinci, Curie, Babbage 및 Ada로 알려져 있습니다.

Codex 모델 시리즈는 GPT-3의 하위 항목이며, 자연어를 코드 사용 사례에 적용할 수 있도록 자연어와 코드 모두에 대해 학습되었습니다. [모델 개념 페이지](#)에서 각 모델에 대해 자세히 알아보세요.

## 다음 단계

[Azure OpenAI를 지원하는 기본 모델](#)에 대해 자세히 알아봅니다.

# Azure OpenAI 서비스 할당량 및 제한

아티클 • 2023. 05. 17.

이 문서에는 Azure Cognitive Services에서 Azure OpenAI의 할당량 및 한도에 대한 빠른 참조와 자세한 설명이 포함되어 있습니다.

## 할당량 및 제한 참조

다음 섹션에서는 Azure OpenAI에 적용되는 할당량 및 한도에 대한 빠른 가이드를 제공합니다.

제한 이름	제한 값
Azure 구독당 지역당 OpenAI 리소스	3
모델당 분당 요청 수*	Davinci 모델(002 이상): 120 ChatGPT 모델: 300 GPT-4 모델: 18 기타 모든 모델: 300
모델당 분당 토큰 수*	Davinci 모델(002 이상): 40,000 ChatGPT 모델: 120,000 GPT-4 8k 모델: 10,000 GPT-4 32k 모델: 32,000 기타 모든 모델: 120,000
최대 미세 조정 모델 배포 수*	2
동일한 모델을 여러 배포에 배포하는 기능	허용되지 않음
리소스당 총 학습 작업 수	100
리소스당 최대 동시 실행 학습 작업	1
대기 중인 최대 학습 작업	20
리소스당 최대 파일	50
리소스당 모든 파일의 총 크기	1GB
최대 학습 작업 시간(초과 시 작업 실패)	720시간
최대 학습 작업 크기(학습 파일의 토큰 수) * (Epoch 수)	20억

\*한도는 변경될 수 있습니다. 프로덕션 및 솔루션 규모로 이동할 때 더 높은 한도가 필요할 것으로 예상합니다. 솔루션 요구 사항을 알고 있는 경우

<https://aka.ms/oai/quotaincrease> 에서 할당량 증가를 신청하여 문의해 주세요.

다양한 모델의 최대 토큰 수에 대한 내용은 [모델 문서](#)를 참조하세요.

## 자동 크기 조정 중 제한을 완화하기 위한 일반적인 모범 사례

제한과 관련된 문제를 최소화하려면 다음 기술을 사용하는 것이 좋습니다.

- 애플리케이션에서 다시 시도 논리를 구현합니다.
- 워크로드가 급격히 변경되지 않도록 합니다. 워크로드를 점진적으로 늘립니다.
- 다양한 로드 증가 패턴을 테스트합니다.
- 동일하거나 다른 지역에 또 다른 OpenAI 서비스 리소스를 만들고 워크로드를 분산합니다.

다음 섹션에서는 할당량 조정의 특정 사례에 대해 설명합니다.

## 기본 할당량 및 한도에 대한 증가를 요청하는 방법

현재는 수요가 많기 때문에 새 리소스 또는 할당량 증가 요청을 수락할 수 없습니다.

### ① 참고

전체 용량에 근접하여 현재 리소스 사용률을 철저히 평가해야 합니다. 기존 리소스의 효율적인 사용이 관찰되지 않는 경우 추가 리소스를 부여하지 않습니다.

## 다음 단계

[Azure OpenAI를 지원하는 기본 모델](#)에 대해 자세히 알아봅니다.

# Azure OpenAI 서비스 모델

아티클 • 2023. 05. 15.

Azure OpenAI는 제품군 및 기능별로 그룹화된 다양한 모델에 대한 액세스를 제공합니다. 모델 제품군은 일반적으로 의도한 작업별로 모델을 연결합니다. 다음 표에서는 현재 Azure OpenAI에서 사용할 수 있는 모델 제품군에 대해 설명합니다. 현재 모든 지역에서 모든 모델을 사용할 수 있는 것은 아닙니다. 전체 분석은 이 문서의 [모델 기능 표](#)를 참조하세요.

모델 제품 군	Description
GPT-4	GPT-3.5를 개선하고 자연어와 코드를 생성할 뿐만 아니라 이해할 수 있는 일련의 모델입니다.
GPT-3	자연어를 이해하고 생성할 수 있는 일련의 모델입니다. 여기에는 새 <a href="#">ChatGPT 모델</a> 이 포함됩니다.
Codex	자연어를 코드로 번역하는 것을 포함하여 코드를 이해하고 생성할 수 있는 일련의 모델입니다.
포함	임베딩을 이해하고 사용할 수 있는 일련의 모델입니다. 포함은 기계 학습 모델 및 알고리즘에서 쉽게 활용할 수 있는 특수한 형식의 데이터 표현입니다. 포함은 텍스트 조각의 의미 체계적 의미에 대한 조밀한 정보 표현입니다. 현재 다양한 기능에 대해 세 가지(유사성, 텍스트 검색, 코드 검색) 임베딩 모델 제품군을 제공합니다.

## 모델 기능

각 모델 제품군에는 기능별로 더 구별되는 일련의 모델이 있습니다. 이러한 기능은 일반적으로 이름으로 식별되며, 이러한 이름의 사전순 순서는 일반적으로 지정된 모델 제품군 내에서 해당 모델의 상대적인 기능과 비용을 나타냅니다. 예를 들어 GPT-3 모델은 Ada, Babbage, Curie 및 Davinci와 같은 이름을 사용하여 상대적인 기능과 비용을 나타냅니다. Davinci는 Curie보다 더 유능하고 비용이 더 많이 듭니다. Curie는 Babbage보다 더 유능하고 비용이 더 많이 듭니다.

### ① 참고

역량이 더 낮은 모델(예: Ada)이 수행할 수 있는 작업은 역량이 더 높은 모델(예: Curie 또는 Davinci)이 수행할 수 있습니다.

# 명명 규칙

Azure OpenAI의 모델 이름은 일반적으로 다음과 같은 표준 명명 규칙에 부합합니다.

{capability}-{family}[-{input-type}]-{identifier}

요소	Description
{capability}	모델의 모델 기능입니다. 예를 들어 GPT-3 모델은 text를 사용하고 Codex 모델은 code를 사용합니다.
{family}	모델의 상대적 제품군입니다. 예를 들어 GPT-3 모델에는 ada, babbage, curie 및 davinci가 포함됩니다.
{input-type}	(임베딩 모델만 해당) 모델에서 지원하는 임베딩의 입력 형식입니다. 예를 들어 텍스트 검색 임베딩 모델은 doc 및 query를 지원합니다.
{identifier}	모델의 버전 식별자입니다.

예를 들어 가장 강력한 GPT-3 모델은 text-davinci-003라고 하며 가장 강력한 Codex 모델은 code-davinci-002라고 합니다.

표준 명명 규칙을 따르지 않는 ada, babbage, curie 및 davinci라는 이전 버전의 GPT-3 모델은 주로 미세 조정을 위한 것입니다. 자세한 내용은 [애플리케이션에 맞게 모델을 사용자 지정하는 방법 알아보기](#)를 참조하세요.

## 사용 가능한 모델 찾기

[모델 목록 API](#)를 사용하여 Azure OpenAI 리소스에서 유추 및 미세 조정에 사용할 수 있는 모델 목록을 가져올 수 있습니다.

## 적합한 모델 찾기

모델 기능이 요구 사항을 충족하는지 확인하려면 모델 제품군에서 가장 유능한 모델부터 시작하는 것이 좋습니다. 그런 다음 해당 모델을 유지하거나 기능과 비용이 더 낮은 모델로 이동하여 해당 모델의 기능을 최적화할 수 있습니다.

## GPT-4 모델

GPT-4는 OpenAI의 이전 모델보다 더 정확하게 어려운 문제를 해결할 수 있습니다. gpt-35-turbo와 마찬가지로 GPT-4는 채팅에 최적화되어 있지만 기존 완료 작업에 적합합니다.

이 모델 시리즈에 대한 수요가 높기 때문에 현재 요청에서만 사용할 수 있습니다. 액세스를 요청하기 위해 기존 Azure OpenAI 고객은 [이 양식을 작성하여 신청할](#) 수 있습니다.

- gpt-4
- gpt-4-32k

gpt-4는 최대 8192개의 입력 토큰을 지원하고 gpt-4-32k는 최대 32,768개의 토큰을 지원합니다.

## GPT-3 모델

GPT-3 모델은 자연어를 이해하고 생성할 수 있습니다. 이 서비스는 4가지 모델 기능을 제공하며, 각기 다른 작업에 적합한 다른 수준의 성능과 속도를 갖추고 있습니다. Davinci는 역량이 가장 높은 모델이고 Ada는 가장 빠른 모델입니다. 기능이 큰 순으로 모델은 다음과 같습니다.

- text-davinci-003
- text-curie-001
- text-babbage-001
- text-ada-001

Davinci가 가장 뛰어나지만, 다른 모델은 상당한 속도 이점을 제공합니다. 사용자의 실험은 Davinci로 시작하는 것이 좋습니다. Davinci는 최상의 결과를 생성하며 Azure OpenAI가 제공하는 가치를 검증할 수 있기 때문입니다. 프로토타입이 적합하면 선택한 모델의 대기 시간/성능 균형을 애플리케이션에 가장 적합하게 최적화하면 됩니다.

## Davinci

Davinci는 가장 유능한 모델이며, 다른 모델이 수행할 수 있는 모든 작업을(종종 더 적은 명령으로) 수행할 수 있습니다. 특정 대상 그룹에 대한 요약 및 창의적인 콘텐츠 생성과 같이 콘텐츠에 대한 깊은 이해가 필요한 애플리케이션의 경우 Davinci는 최상의 결과를 생성합니다. Davinci가 제공하는 향상된 기능에는 더 많은 컴퓨팅 리소스가 필요하므로 Davinci는 비용이 더 많이 들고 다른 모델만큼 빠르지 않습니다.

Davinci가 뛰어난 또 다른 영역은 텍스트의 의도를 이해하는 것입니다. Davinci는 다양한 종류의 논리 문제를 해결하고 특성의 동기를 설명하는 데 탁월합니다. Davinci는 원인 및 효과와 관련된 가장 어려운 AI 문제 중 일부를 해결할 수 있었습니다.

**용도:** 대상 그룹에 대한 복잡한 의도, 원인 및 효과, 요약

## Curie

Curie는 강력하면서도 빠릅니다. Davinci는 복잡한 텍스트를 분석할 때 더 강력하지만, Curie는 감정 분류 및 요약과 같은 많은 미묘한 작업을 수행할 수 있습니다. Curie는 또한 질문에 답하고 Q&A를 수행하며 일반 서비스 챗봇으로 활동합니다.

**용도:** 언어 번역, 복잡한 분류, 텍스트 감정, 요약

## Babbage

Babbage는 간단한 분류와 같은 간단한 작업을 수행할 수 있습니다. 또한 문서가 검색 쿼리와 일치하는 정도에 대한 의미 체계 검색 순위를 지정하는 것이 가능합니다.

**용도:** 중간 분류, 의미 체계 검색 분류

## Ada

Ada는 일반적으로 가장 빠른 모델이며 텍스트 구문 분석, 주소 수정 및 너무 많은 뉘앙스가 필요하지 않은 특정 종류의 분류 작업과 같은 작업을 수행할 수 있습니다. Ada의 성능은 종종 더 많은 컨텍스트를 제공하여 향상될 수 있습니다.

**용도:** 텍스트 구문 분석, 간단한 분류, 주소 수정, 키워드

## ChatGPT(gpt-35-turbo)

ChatGPT 모델(gpt-35-turbo)은 대화형 인터페이스용으로 설계된 언어 모델이며 이전 GPT-3 모델과 다르게 동작합니다. 이전 모델은 텍스트 입력 및 텍스트 출력이었습니다. 즉, 프롬프트 문자열을 수락하고 프롬프트에 추가하기 위해 완료를 반환했습니다. 그러나 ChatGPT 모델은 대화-인 및 메시지-아웃입니다. 모델은 특정 채팅과 같은 대화 내용 형식으로 형식화된 프롬프트 문자열을 예상하고 채팅에서 모델 작성 메시지를 나타내는 완료를 반환합니다.

ChatGPT 모델 및 채팅 API와 상호 작용하는 방법에 대해 자세히 알아보려면 [자세한 방법](#)을 확인합니다.

## Codex 모델

Codex 모델은 코드를 이해하고 생성할 수 있는 기본 GPT-3 모델의 하위 항목입니다. 학습 데이터에는 자연어와 수십억 줄의 GitHub 공용 코드가 모두 포함됩니다.

Python에서 가장 유능하며 C#, JavaScript, Go, Perl, PHP, Ruby, Swift, TypeScript, SQL, 심지어 Shell을 포함한 12개 이상의 언어에 능숙합니다. 기능이 큰 순으로 Codex 모델은 다음과 같습니다.

- code-davinci-002
- code-cushman-001

## Davinci

GPT-3과 유사하게 Davinci는 가장 유능한 Codex 모델이며, 다른 모델이 수행할 수 있는 모든 작업을 수행할 수 있고, 종종 더 적은 명령으로 수행할 수 있습니다. 콘텐츠에 대한 깊은 이해가 필요한 애플리케이션의 경우 Davinci가 최상의 결과를 생성합니다. 더 큰 기능에는 더 많은 컴퓨팅 리소스가 필요하므로 Davinci는 비용이 더 많이 들고 다른 모델만큼 빠르지 않습니다.

## Cushman

Cushman은 강력하면서도 빠릅니다. 복잡한 작업을 분석할 때는 Davinci가 더 강력하지만 Cushman은 많은 코드 생성 작업에 유능한 모델입니다. Cushman은 일반적으로 Davinci 보다 빠르게 실행되며 저렴합니다.

## 임베딩 모델

### ① 중요

text-embedding-ada-002 (Version 2)를 사용하는 것이 좋습니다. 이 모델/버전은 OpenAI의 text-embedding-ada-002와 패리티를 제공합니다. 이 모델이 제공하는 개선 사항에 대해 자세히 알아보려면 [OpenAI의 블로그 게시물](#)을 참조하세요. 현재 버전 1을 사용 중이더라도 최신 가중치/업데이트된 토큰 한도를 활용하려면 버전 2로 마이그레이션해야 합니다. 버전 1과 버전 2는 상호 교환이 불가능하므로 동일한 버전의 모델을 사용하여 문서 포함과 문서 검색을 수행해야 합니다.

현재 다양한 기능에 대해 세 가지 임베딩 모델 제품군을 제공합니다.

- [유사성](#)
- [텍스트 검색](#)
- [코드 검색](#)

각 제품군에는 다양한 기능의 모델이 포함되어 있습니다. 다음 목록은 모델 기능에 따라 서비스에서 반환되는 숫자 벡터의 길이를 나타냅니다.

- Ada: 1024 차원
- Babbage: 2048 차원
- Curie: 4096 차원

- Davinci: 12288 차원

Davinci는 가장 뛰어나지만 다른 모델보다 느리고 비용이 더 많이 듭니다. Ada는 역량이 가장 낮지만 더 빠르고 저렴합니다.

## 유사성 임베딩

이러한 모델은 둘 이상의 텍스트 조각 간의 의미 체계 유사성을 캡처하는 데 적합합니다.

사용 사례	모델
클러스터링, 회귀, 변칙 검색, 시각화	text-similarity-ada-001 text-similarity-babbage-001 text-similarity-curie-001 text-similarity-davinci-001

## 텍스트 검색 임베딩

이러한 모델은 긴 문서가 짧은 검색 쿼리와 관련이 있는지 여부를 측정하는 데 도움이 됩니다. 이 제품군에서 지원하는 입력 유형은 두 가지입니다. `doc`은 검색할 문서를 임베딩하기 위한 유형이며 `query`는 검색 쿼리를 임베딩하기 위한 유형입니다.

사용 사례	모델
검색, 컨텍스트 관련성, 정보 검색	text-search-ada-doc-001 text-search-ada-query-001 text-search-babbage-doc-001 text-search-babbage-query-001 text-search-curie-doc-001 text-search-curie-query-001 text-search-davinci-doc-001 text-search-davinci-query-001

## 코드 검색 임베딩

텍스트 검색 임베딩 모델과 유사하게 이 제품군에서 지원하는 두 가지 입력 유형은 검색 할 코드 조각 임베딩을 위한 `code` 및 자연어 검색 쿼리 임베딩을 위한 `text`입니다.

사용 사례	모델
-------	----

사용 사례	모델
코드 검색 및 관련성	code-search-ada-code-001 code-search-ada-text-001 code-search-babbage-code-001 code-search-babbage-text-001

포함 모델을 사용할 때는 해당 모델의 한계와 위험을 염두에 두어야 합니다.

## 모델 요약 테이블 및 지역 가능성

### ① 중요

수요가 많기 때문에 미국 중남부에서 새 리소스를 일시적으로 만들 수 없습니다.

## GPT-3 모델

이러한 모델은 완료 API 요청과 함께 사용할 수 있습니다. `gpt-35-turbo`는 완료 API 요청 및 채팅 완료 API와 함께 사용할 수 있는 유일한 모델입니다.

모델 ID	기본 모델 영역	미세 조정 지역	최대 요청 (토큰)	학습 데이터 (최대)
ada	해당 없음	해당 없음	2,049	2019년 10월
text-ada-001	미국 동부, 미국 중남부, 서유럽	해당 없음	2,049	2019년 10월
babbage	해당 없음	해당 없음	2,049	2019년 10월
text-babbage-001	미국 동부, 미국 중남부, 서유럽	해당 없음	2,049	2019년 10월
curie	해당 없음	해당 없음	2,049	2019년 10월
text-curie-001	미국 동부, 미국 중남부, 서유럽	해당 없음	2,049	2019년 10월
다빈치	해당 없음	해당 없음	2,049	2019년 10월
text-davinci-001	미국 중남부, 서유럽	해당 없음		
text-davinci-002	미국 동부, 미국 중남부, 서유럽	해당 없음	4,097	2021년 6월
text-davinci-003	미국 동부, 서부 유럽	해당 없음	4,097	2021년 6월
text-davinci-fine-tune-002	해당 없음	해당 없음		

모델 ID	기본 모델 영역	미세 조정 지역	최대 요청(토큰)	학습 데이터(최대)
gpt-35-turbo <sup>1</sup> (ChatGPT)	미국 동부, 프랑스 중부, 미국 중 남부, 서유럽	해당 없음	4,096	2021년 9월

<sup>1</sup> 현재 이 모델의 버전 0301 만 사용할 수 있습니다. 이 버전의 모델은 최신 버전의 gpt-35 모델을 위해 2023년 8월 1일에 사용 중단됩니다. 자세한 내용은 [ChatGPT 모델 버전 관리](#)를 참조하세요.

## GPT-4 모델

이러한 모델은 채팅 완료 API에서만 사용할 수 있습니다.

모델 ID	기본 모델 영역	미세 조정 지역	최대 요청(토큰)	학습 데이터(최대)
gpt-4 <sup>1,2</sup>	미국 동부, 프랑스 중부	해당 없음	8,192	2021년 9월
gpt-4-32k <sup>1,2</sup>	미국 동부, 프랑스 중부	해당 없음	32,768	2021년 9월

<sup>1</sup> 모델은 [요청에 의해서만 사용할 수 있습니다](#).

<sup>2</sup> 현재 이 모델의 버전 0314만 사용할 수 있습니다.

## Codex 모델

이러한 모델은 완료 API 요청에만 사용할 수 있습니다.

모델 ID	기본 모델 영역	미세 조정 지역	최대 요청(토큰)	학습 데이터(최대)
code-cushman-001 <sup>1</sup>	미국 중남부, 서유럽	현재 사용 불가	2,048	
code-davinci-002	미국 동부, 서부 유럽	해당 없음	8,001	2021년 6월

<sup>1</sup> 모델은 요청에 의해서만 미세 조정이 가능합니다. 현재 모델 미세 조정에 대한 새로운 요청을 수락하지 않습니다.

## 포함 모델

이러한 모델은 포함 API 요청에만 사용할 수 있습니다.

## ① 참고

text-embedding-ada-002 (Version 2)를 사용하는 것이 좋습니다. 이 모델/버전은 OpenAI의 text-embedding-ada-002와 패리티를 제공합니다. 이 모델이 제공하는 개선 사항에 대해 자세히 알아보려면 [OpenAI의 블로그 게시물](#)을 참조하세요. 현재 버전 1을 사용 중이더라도 최신 가중치/업데이트된 토큰 한도를 활용하려면 버전 2로 마이그레이션해야 합니다. 버전 1과 버전 2는 상호 교환이 불가능하므로 동일한 버전의 모델을 사용하여 문서 포함과 문서 검색을 수행해야 합니다.

모델 ID	기본 모델 영역	미세 조정 지역	최대 요청 (토큰)	학습 데이터 (최대)
text-embedding-ada-002(버전 2)	미국 동부, 미국 중남부, 서유럽	해당 없음	8,191	2021년 9월
text-embedding-ada-002(버전 1)	미국 동부, 미국 중남부, 서유럽	해당 없음	4,095	2021년 9월
text-similarity-ada-001	미국 동부, 미국 중남부, 서유럽	해당 없음	2,046	2020년 8월
text-similarity-babbage-001	미국 중남부, 서유럽	해당 없음	2,046	2020년 8월
text-similarity-curie-001	미국 동부, 미국 중남부, 서유럽	해당 없음	2,046	2020년 8월
text-similarity-davinci-001	미국 중남부, 서유럽	해당 없음	2,046	2020년 8월
text-search-ada-doc-001	미국 중남부, 서유럽	해당 없음	2,046	2020년 8월
text-search-ada-query-001	미국 중남부, 서유럽	해당 없음	2,046	2020년 8월
text-search-babbage-doc-001	미국 중남부, 서유럽	해당 없음	2,046	2020년 8월
text-search-babbage-query-001	미국 중남부, 서유럽	해당 없음	2,046	2020년 8월
text-search-curie-doc-001	미국 중남부, 서유럽	해당 없음	2,046	2020년 8월
text-search-curie-query-001	미국 중남부, 서유럽	해당 없음	2,046	2020년 8월
text-search-davinci-doc-001	미국 중남부, 서유럽	해당 없음	2,046	2020년 8월

모델 ID	기본 모델 영역	미세 조정 지역	최대 요청 (토큰)	학습 데이터 (최대)
text-search-davinci-query-001	미국 중남부, 서유럽	해당 없음	2,046	2020년 8월
code-search-ada-code-001	미국 중남부, 서유럽	해당 없음	2,046	2020년 8월
code-search-ada-text-001	미국 중남부, 서유럽	해당 없음	2,046	2020년 8월
code-search-babbage-code-001	미국 중남부, 서유럽	해당 없음	2,046	2020년 8월
code-search-babbage-text-001	미국 중남부, 서유럽	해당 없음	2,046	2020년 8월

## 다음 단계

- [Azure OpenAI에 대해 자세히 알아보기](#)
- [Azure OpenAI 모델 미세 조정에 대해 자세히 알아보기](#)

# Azure OpenAI 서비스의 새로운 기능

아티클 • 2023. 05. 17.

## 2023년 5월

### Azure OpenAI 채팅 완료 GA(일반 공급)

- 다음을 위한 일반 공급 지원:
  - 채팅 완료 API 버전 2023-05-15.
  - GPT-35-Turbo 모델.
  - GPT-4 모델 시리즈. 이 모델 시리즈에 대한 수요가 높기 때문에 현재 요청에서만 사용할 수 있습니다. 액세스를 요청하기 위해 기존 Azure OpenAI 고객은 [이 양식을 작성하여 신청할](#) 수 있습니다.

현재 API를 2023-03-15-preview 사용하는 경우 GA 2023-05-15 API로 마이그레이션하는 것이 좋습니다. 현재 API 버전을 2022-12-01 사용하는 경우 이 API는 GA로 유지되지만 최신 채팅 완료 기능은 포함되지 않습니다.

#### ⓘ 중요

완성 엔드포인트와 함께 GPT-35-Turbo 모델의 현재 버전을 사용하는 것은 미리 보기로 유지됩니다.

## 프랑스 중부

- 이제 프랑스 중부 지역에서 Azure OpenAI를 사용할 수 있습니다. 각 지역의 모델 가용성에 대한 최신 정보는 [모델 페이지를](#) 확인하세요.

## 2023년 4월

- 사용자 지정된 모델의 비활성 배포는 이제 15일 후에 삭제됩니다. 모델은 다시 배포에 계속 사용할 수 있습니다. 사용자 지정된(미세 조정된) 모델이 15일 이상 배포되어 완료 또는 채팅 완료 호출이 수행되지 않는 경우 배포가 자동으로 삭제되고 해당 배포에 대한 추가 호스팅 요금이 발생하지 않습니다. 기본 사용자 지정 모델은 계속 사용할 수 있으며 언제든지 다시 배포할 수 있습니다. 자세한 내용은 [방법 문서를](#) 검사.

## 2023년 3월

- GPT-4 시리즈 모델은 이제 Azure OpenAI에서 미리 보기로 제공됩니다. 액세스를 요청하려면 기존 Azure OpenAI 고객이 [이 양식을 작성하여 신청](#) 할 수 있습니다. 이러한 모델은 현재 미국 동부 및 미국 중남부 지역에서 사용할 수 있습니다.
- ChatGPT 및 GPT-4 모델을 위한 새로운 채팅 완료 API가 3월 21일에 미리 보기로 릴리스되었습니다. 자세히 알아보려면 [업데이트된 빠른 시작](#) 및 [방법 문서](#)를 확인합니다.
- ChatGPT(gpt-35-turbo) 미리 보기. 자세한 내용은 [방법 문서](#)를 참조하세요.
- 미세 조정을 위한 학습 제한 증가: 최대 학습 작업 크기(학습 파일의 토큰)  $x$ (epoch 수)는 모든 모델에 대해 20억 토큰입니다. 또한 최대 학습 작업을 120시간에서 720 시간으로 늘렸습니다.
- 기존 액세스에 추가 사용 사례를 추가합니다. 이전에는 새로운 사용 사례를 추가하려면 고객이 서비스에 다시 신청해야 했습니다. 이제 서비스 사용에 새로운 사용 사례를 신속하게 추가할 수 있는 새로운 프로세스를 출시합니다. 이 프로세스는 Azure Cognitive Services 내에서 설정된 제한된 액세스 프로세스를 따릅니다. [기존 고객은 여기에서 모든 새로운 사용 사례를 증명할 수 있습니다](#). 이는 사용자가 원래 신청하지 않은 새로운 사용 사례에 대해 서비스를 사용하고자 할 때마다 필요하다는 점에 유의하세요.

## 2023년 2월

### 새로운 기능

- .NET SDK(유추) [미리 보기 릴리스](#) | [샘플](#)
- Azure OpenAI 관리 작업을 지원하기 위한 [Terraform SDK 업데이트](#)
- 이제 완료 끝에 텍스트 삽입이 `suffix` 매개 변수로 지원됩니다.

### 업데이트

- 콘텐츠 필터링은 기본적으로 켜져 있습니다.

#### 새로운 문서:

- [Azure OpenAI 서비스 모니터링](#)
- [Azure OpenAI 비용 계획 및 관리](#)

#### 새로운 학습 과정:

- Azure OpenAI 소개

## 2023년 1월

### 새로운 기능

- **서비스 GA.** 이제 Azure OpenAI 서비스가 일반 공급됩니다.
- **새 모델:** 최신 텍스트 모델인 text-davinci-003(미국 동부, 서유럽), text-ada-embeddings-002(미국 동부, 미국 중남부, 서유럽) 추가

## 2022년 12월

### 새 기능

- **OpenAI의 최신 모델입니다.** Azure OpenAI는 GPT-3.5 시리즈를 포함한 모든 최신 모델에 대한 액세스를 제공합니다.
- **새로운 API 버전(2022-12-01).** 이 업데이트에는 API 응답의 토큰 사용 정보, 파일에 대한 개선된 오류 메시지, 미세 조정 만들기 데이터 구조에 대한 OpenAI와의 맞춤, 미세 조정 작업의 사용자 지정 명명을 허용하는 접미사 매개 변수 지원을 포함하여 요청된 몇 가지 개선 사항이 포함되어 있습니다.
- **초당 요청 제한이 더 높습니다.** Davinci가 아닌 모델의 경우 50입니다. Davinci 모델의 경우 20개입니다.
- **배포를 더 빠르게 미세 조정합니다.** 10분 이내에 Ada 및 Curie 미세 조정 모델을 배포합니다.
- **높은 학습 한도:** Ada, Babbage 및 Curie에 대한 4천만 개의 학습 토큰. Davinci의 경우 10M입니다.
- **남용 & 오용 데이터 로깅 & 인간의 검토에 대한 수정 요청 프로세스입니다.** 현재 이 서비스는 이러한 강력한 모델이 남용되지 않도록 남용 및 오용 검색 목적으로 요청/응답 데이터를 로그합니다. 그러나 많은 고객이 자신의 데이터에 대한 더 많은 제어가 필요한 엄격한 데이터 개인 정보 보호 및 보안 요구 사항을 가지고 있습니다. 이러한 사용 사례를 지원하기 위해 고객이 콘텐츠 필터링 정책을 수정하거나 위험도가 낮은 사용 사례에 대한 남용 기록을 해제할 수 있는 새로운 프로세스를 출시하고 있습니다. 이 프로세스는 Azure Cognitive Services 내에서 확립된 제한된 액세스 프로세스를 따르며 [기존 OpenAI 고객은 여기에서 신청할 수 있습니다](#).

- **CMK(고객 관리형 키) 암호화.** CMK는 학습 데이터 및 사용자 지정된 모델을 저장하는 데 사용되는 자체 암호화 키를 제공하여 고객이 Azure OpenAI에서 데이터 관리를 보다 효과적으로 제어할 수 있도록 합니다. CMK(고객 관리 키)(BYOK(Bring Your Own Key)라고도 함)를 사용하여 훨씬 더 유연하게 액세스 제어를 만들고, 회전하고, 해제하고, 취소할 수 있습니다. 데이터를 보호하는 데 사용되는 암호화 키를 감사할 수도 있습니다. [유 휴 데이터 암호화 설명서](#)에서 자세히 알아봅니다.
- **Lockbox 지원**
- **SOC-2 준수**
- Azure Resource Health, 비용 분석 및 메트릭 & 진단 설정을 통한 **로깅 및 진단**.
- **스튜디오 개선.** 미세 조정된 모델을 만들고 배포할 수 있는 액세스 권한이 있는 팀을 제어하기 위한 Azure AD 역할 지원을 포함하여 Studio 워크플로에 대한 수많은 유용성 개선.

## 변경 내용(중단)

미세 조정 만들기 API 요청이 OpenAI의 스키마와 일치하도록 업데이트되었습니다.

**미리 보기 API 버전:**

JSON

```
{
  "training_file": "file-XGinujblHPwGLSztz8cPS8XY",
  "hyperparams": {
    "batch_size": 4,
    "learning_rate_multiplier": 0.1,
    "n_epochs": 4,
    "prompt_loss_weight": 0.1,
  }
}
```

**API 버전 2022-12-01:**

JSON

```
{
  "training_file": "file-XGinujblHPwGLSztz8cPS8XY",
  "batch_size": 4,
  "learning_rate_multiplier": 0.1,
  "n_epochs": 4,
  "prompt_loss_weight": 0.1,
}
```

기본적으로 콘텐츠 필터링은 일시적으로 꺼져 있습니다. Azure 콘텐츠 조정은 OpenAI와 다르게 작동합니다. Azure OpenAI는 생성 호출 중에 콘텐츠 필터를 실행하여 유해하거나 악의적인 콘텐츠를 검색하고 응답에서 필터링합니다. [자세한 정보](#)

이러한 모델은 2023년 1분기에 다시 사용하도록 설정되며 기본적으로 켜져 있습니다.

## 고객 작업

- 구독에 대해 이 기능을 켜려면 [Azure 지원팀에 문의](#) 합니다.
- 필터링을 해제한 상태로 유지하려면 [필터링 수정을 신청](#) 합니다. (이 옵션은 위험 도가 낮은 사용 사례에만 해당됩니다.)

## 다음 단계

[Azure OpenAI를 지원하는 기본 모델](#)에 대해 자세히 알아봅니다.

# Azure OpenAI Service 질문과 대답

FAQ

이 문서에서는 질문에 대한 답변을 찾을 수 없지만 여전히 도움이 필요하다면 [Cognitive Services 지원 옵션 가이드](#)를 확인해 보세요. Azure OpenAI는 Azure Cognitive Services의 일부입니다.

## 데이터 및 개인 정보:

### 회사 데이터를 사용하여 모델을 학습하시겠어요?

Azure OpenAI는 고객 데이터를 사용하여 모델을 다시 학습시키지 않습니다. 자세한 내용은 [Azure OpenAI 데이터, 개인 정보, 보안 가이드](#)를 참조하세요.

## 일반

### Azure OpenAI는 GPT-4를 지원하나요?

Azure OpenAI는 최신 GPT-4 모델을 지원합니다. 이러한 모델은 현재 요청에 의해서만 사용할 수 있습니다. 액세스를 위해 기존 Azure OpenAI 고객은 [이 양식을 작성하여 신청](#) 할 수 있습니다.

### Azure OpenAI의 기능은 OpenAI와 어떻게 비교되나요?

Azure OpenAI Service는 Azure의 보안 및 엔터프라이즈 지원을 통해 OpenAI GPT-3, Codex 및 DALL-E 모델을 사용하는 고급 언어 AI를 고객에게 제공합니다. Azure OpenAI는 OpenAI와 API를 공동 개발하여 호환성과 원활한 전환을 보장합니다.

Azure OpenAI를 사용하면 고객은 OpenAI와 동일한 모델을 실행하면서 Microsoft Azure의 보안 기능을 얻을 수 있습니다.

### Azure OpenAI는 VNET 및 프라이빗 엔드포인트를 지원하나요?

예, Azure Cognitive Services의 일부로 Azure OpenAI는 VNET 및 프라이빗 엔드포인트를 지원합니다. 자세한 내용은 [Cognitive Services 가상 네트워킹 지침](#)을 참조하세요.

# GPT-4 모델은 현재 이미지 입력을 지원합니까?

아니요, GPT-4는 OpenAI에서 다중 모달로 설계되었지만 현재는 텍스트 입력 및 출력만 지원됩니다.

## 새 사용 사례를 신청하려면 어떻게 해야 하나요?

이전에는 새로운 사용 사례를 추가하려면 고객이 서비스에 다시 신청해야 했습니다. 이제 서비스 사용에 새로운 사용 사례를 신속하게 추가할 수 있는 새로운 프로세스를 출시합니다. 이 프로세스는 Azure Cognitive Services 내에서 설정된 제한된 액세스 프로세스를 따릅니다. [기존 고객은 여기에서 모든 새로운 사용 사례를 증명할 수 있습니다](#). 이는 사용자가 원래 신청하지 않았던 새 사용 사례의 서비스를 사용하고자 할 때마다 필요합니다.

## embeddings를 사용하려고 했는데

### "`InvalidRequestError`: 입력이 너무 많습니다. 입력의 최대 수는 1입니다."

이 문제를 해결하려면 어떻게 해야 합니까?

이 오류는 일반적으로 단일 API 요청에 포함할 텍스트 일괄 처리를 배열로 보내려고 할 때 발생합니다. 현재 Azure OpenAI는 포함 요청을 사용하여 일괄 처리를 지원하지 않습니다. 포함 API 호출은 요청당 단일 문자열 입력으로 구성되어야 합니다. `text-embedding-ada-002`(버전 2) 모델을 사용하는 경우 문자열의 길이는 최대 8191개 토큰일 수 있습니다.

## Azure OpenAI를 사용하여 서비스에서 원하는 응답을 가져오는 더 나은 방법에 대해 어디에서 읽을 수 있나요?

[프롬프트 엔지니어링 소개](#)를 확인하세요. 이러한 모델은 매우 강력하지만 해당 동작은 사용자에게서 받은 프롬프트에도 매우 민감합니다. 따라서 프롬프트 생성은 개발해야 하는 중요한 기술에 해당합니다. 소개를 마스터한 후 고급 [프롬프트 엔지니어링 기술](#)에 대한 문서를 검사.

## Azure OpenAI Service에 대한 액세스 권한 얻기

# Azure OpenAI에 액세스하려면 어떻게 해야 하나요?

현재 높은 수요, 예정된 제품 개선 사항, 책임 있는 AI에 대한 Microsoft의 약속을 탐색하기 때문에 액세스가 제한됩니다. 현재 우리는 Microsoft와의 기존 파트너십, 위협이 낮은 사용 사례 및 완화 통합에 전념하는 고객과 협력하고 있습니다. 초기 액세스를 신청하는 것 외에도 Azure OpenAI를 사용하는 모든 솔루션은 프로덕션 용도로 릴리스되기 전에 사용 사례 검토를 거쳐야 합니다. 초기 액세스 또는 프로덕션 검토를 위해 여기에 신청: [지금 신청](#)

## 액세스를 신청한 후 승인을 받기까지 얼마나 기다려야 하나요?

현재 액세스 승인에 대한 타임라인을 제공하지 않습니다.

## 자세한 정보 및 질문할 위치

### Azure OpenAI의 최신 업데이트에 대한 정보는 어디에서 읽을 수 있나요?

월별 업데이트는 [새로운 기능 페이지](#)를 참조하세요.

### Azure OpenAI를 중심으로 학습을 시작하고 기술을 구축하기 위한 교육은 어디에서 받을 수 있나요?

[Azure OpenAI 교육 과정에 대한 소개](#)를 확인하세요.

### 질문을 게시하고 다른 일반적인 질문에 대한 답변을 볼 수 있는 곳은 어디인가요?

- Microsoft Q&A에 질문을 게시하는 것이 좋습니다.
- 또는 [Stack Overflow](#)에 대한 질문을 게시할 수 있습니다.

### Azure OpenAI 고객 지원은 어디에서 확인할 수 있나요?

Azure OpenAI는 Azure Cognitive Services의 일부입니다. [지원 및 도움말 옵션 가이드](#)에서 Azure Cognitive Services의 모든 지원 옵션에 대해 알아볼 수 있습니다.

## 모델 및 튜닝

### 사용 가능한 모델은 무엇인가요?

Azure OpenAI [모델 가용성 가이드](#)를 참조하세요.

### 모델을 사용할 수 있는 지역은 어디에서 확인할 수 있나요?

지역 가용성은 Azure OpenAI [모델 가용성 가이드](#)를 참조하세요.

### 기본 모델과 미세 조정된 모델의 차이점은 무엇인가요?

기본 모델은 특정 사용 사례에 맞게 사용자 지정되거나 미세 조정되지 않은 모델입니다. 미세 조정된 모델은 고유한 프롬프트 집합에서 모델의 가중치를 학습하는 기본 모델의 사용자 지정 버전입니다. 미세 조정된 모델을 사용하면 완료 프롬프트의 일부로 컨텍스트 내 학습에 대한 자세한 예제를 제공할 필요 없이 더 많은 작업에서 더 나은 결과를 얻을 수 있습니다. 자세한 내용은 [미세 조정 가이드](#)를 검토하세요.

### 만들 수 있는 최대 미세 조정된 모델 수는 몇 개인가요?

100

### Azure OpenAI의 API 응답에 대한 SLA는 무엇인가요?

현재 정의된 API 응답 시간 SLA(서비스 수준 약정)가 없습니다. Azure OpenAI Service의 전체 SLA는 다른 Azure Cognitive Services와 동일합니다. 자세한 내용은 [온라인 서비스용 SLA\(서비스 수준 약정\) 페이지](#)의 Cognitive Services 섹션을 참조하세요.

### 미세 조정된 모델 배포가 삭제된 이유는 무엇인가요?

사용자 지정된(미세 조정된) 모델이 15일 이상 배포되어 완료 또는 채팅 완료 호출이 수행되지 않는 경우 배포가 자동으로 삭제되고 해당 배포에 대한 추가 호스팅 요금이 발생하지 않습니다. 기본 사용자 지정 모델은 계속 사용할 수 있으며 언제든지 다시 배포할 수 있습니다. 자세한 내용은 [방법 문서를](#) 검사.

## REST API를 사용하여 모델을 배포할 어떻게 할까요? 있나요?

현재 모델 배포를 허용하는 두 가지 REST API가 있습니다. text-embedding-ada-002 버전 2와 같은 모델에 배포하는 동안 모델 버전을 지정하는 기능과 같은 최신 모델 배포 기능의 경우 [Cognitive Services REST API 만들기 또는 업데이트 호출](#)을 사용합니다.

## 다음 단계

- [Azure OpenAI 할당량 및 한도](#)
- [Azure OpenAI의 새로운 기능](#)
- [Azure OpenAI 빠른 시작](#)

# 빠른 시작: Azure OpenAI Service를 사용하여 텍스트 생성 시작

아티클 • 2023. 05. 27.

이 문서를 사용하여 Azure OpenAI를 처음으로 호출해 보세요.

## 사전 요구 사항

- Azure 구독 – [체험 구독을 만듭니다](#).
- 원하는 Azure 구독의 Azure OpenAI에 대한 액세스 권한.  
현재 이 서비스에 대한 액세스 권한은 애플리케이션에서만 부여됩니다.  
<https://aka.ms/oai/access>에서 양식을 작성하여 Azure OpenAI에 대한 액세스를 신청할 수 있습니다. 문제가 있는 경우 이 리포지토리에서 문제를 엽니다.
- 모델이 배포된 Azure OpenAI 리소스. 모델 배포에 대한 자세한 내용은 [리소스 배포 가이드](#)를 참조하세요.

필수 조건에 문제가 있습니다.

## Azure OpenAI Studio로 이동

<https://oai.azure.com/>에서 Azure OpenAI Studio로 이동한 다음, OpenAI 리소스에 액세스할 수 있는 자격 증명으로 로그인합니다. 로그인 워크플로 도중 또는 이후에 적절한 디렉터리, Azure 구독 및 Azure OpenAI 리소스를 선택합니다.

Azure OpenAI Studio 방문 페이지에서 더 자세히 탐색하여 프롬프트 완료를 위한 예제를 살펴보고, 배포 및 모델을 관리하고, 설명서 및 커뮤니티 포럼과 같은 학습 리소스를 찾습니다.

## Get started with Azure OpenAI Service

Get example prompts for different scenarios and write prompts of your own. Export your prompts to code at any time to rapidly iterate at scale and integrate with your apps.



### Try the playgrounds

Get example prompts for different scenarios and write prompts of your own. Export your prompts to code at any time to rapidly iterate at scale and integrate with your apps.

[GPT-3 Playground](#)
[ChatGPT playground \(Preview\)](#)

### Explore examples for prompt completion



#### Summarize Text

Summarize text by adding a 'tl;dr' to the end of a text passage.

[Learn more](#)


#### Classify Text

Classify items into categories provided at inference time.

[Learn more](#)


#### Natural Language to SQL

Translate natural language to SQL queries.

[Learn more](#)


#### Generate New Product Names

Create product names from examples words.

[Learn more](#)


실험 및 미세 조정 워크플로를 보려면 [플레이그라운드](#)로 이동합니다.

## 플레이그라운드

GPT-3 플레이그라운드를 통해 코드 없는 접근 방식으로 Azure OpenAI 기능 탐색을 시작합니다. 플레이그라운드는 완료를 생성하는 프롬프트를 제출할 수 있는 간단한 텍스트 상자입니다. 이 페이지에서 쉽게 기능을 반복하고 실험해볼 수 있습니다.

Cognitive Services | Azure OpenAI Studio

Azure OpenAI Studio > GPT-3 playground

**GPT-3 playground**

Deployments Examples

text-davinci-002 Load an example View code

Start typing here

Parameters

- Temperature (1)
- Max length (tokens) (100)
- Stop sequences
- Top probabilities (0.5)
- Frequency penalty (0)
- Presence penalty (0)
- Best of (1)
- Pre-response text
- Post-response text

Learn more

Generate Undo Regenerate Tokens: 0

배포를 선택하고 미리 로드된 몇 가지 예제 중에서 선택하여 시작할 수 있습니다. 리소스에 배포가 없는 경우 **배포 만들기**를 선택하고 마법사에서 제공하는 지침을 따릅니다. 모델 배포에 대한 자세한 내용은 [리소스 배포 가이드](#)를 참조하세요.

온도 및 사전 응답 텍스트와 같은 구성 설정으로 실험하면서 작업의 성능을 향상시킬 수 있습니다. [REST API](#)에서 각 매개 변수에 대해 자세히 알아볼 수 있습니다.

- **생성** 단추를 선택하면 입력한 텍스트가 완료 API로 전송되고 결과가 다시 텍스트 상자로 스트리밍됩니다.
- **실행 취소** 단추를 선택하여 이전 생성 호출을 실행 취소합니다.
- **다시 생성** 단추를 선택하여 실행 취소 및 생성 호출을 함께 완료합니다.

또한 Azure OpenAI는 프롬프트 입력 및 생성된 출력에서 콘텐츠 조정을 수행합니다. 유해한 콘텐츠가 감지되면 프롬프트나 응답을 필터링할 수 있습니다. 자세한 내용은 [콘텐츠 필터](#) 문서를 참조하세요.

GPT-3 플레이그라운드에서 선택한 설정에 따라 미리 채워진 Python 및 curl 코드 샘플을 볼 수도 있습니다. 예제 드롭다운 옆에 있는 **코드 보기**를 선택하면 됩니다. OpenAI Python SDK, curl 또는 기타 REST API 클라이언트를 사용하여 동일한 작업을 완료하는 애플리케이션을 작성할 수 있습니다.

## 텍스트 요약 사용해 보기

GPT-3 플레이그라운드의 텍스트 요약에 Azure OpenAI를 사용하려면 다음 단계를 수행합니다.

1. [Azure OpenAI Studio](#)에 로그인합니다.
2. 작업할 구독 및 OpenAI 리소스를 선택합니다.
3. 방문 페이지 위쪽에서 **GPT-3 플레이그라운드**를 선택합니다.
4. **배포** 드롭다운에서 배포를 선택합니다. 리소스에 배포가 없는 경우 **배포 만들기**를 선택한 다음, 이 단계를 다시 진행합니다.
5. 예제 드롭다운에서 **텍스트 요약**을 선택합니다.

## GPT-3 playground

Deployments

text-davinci-002

Examples

Summarize Text

 View code

A neutron star is the collapsed core of a massive supergiant star, which had a total mass of between 10 and 25 solar masses, possibly more if the star was especially metal-rich.[1] Neutron stars are the smallest and densest stellar objects, excluding black holes and hypothetical white holes, quark stars, and strange stars.[2] Neutron stars have a radius on the order of 10 kilometres (6.2 mi) and a mass of about 1.4 solar masses.[3] They result from the supernova explosion of a massive star, combined with gravitational collapse, that compresses the core past white dwarf star density to that of atomic nuclei.

Tl;dr:

A neutron star is the collapsed core of a supergiant star. These incredibly dense objects are incredibly fascinating due to their strange properties and their potential for phenomena such as extreme gravitational forces and a strong magnetic field.

 Generate

 Undo

 Regenerate

Tokens: 189 



6. **Generate**를 선택합니다. Azure OpenAI는 텍스트의 컨텍스트를 캡처하고 간결하게 다시 표현하려고 시도합니다. 다음 텍스트와 유사한 결과가 표시됩니다.

Tl;dr A neutron star is the collapsed core of a supergiant star. These incredibly dense objects are incredibly fascinating due to their strange properties and their potential for phenomena such as extreme gravitational forces and a strong magnetic field.

응답의 정확도는 모델마다 다를 수 있습니다. 이 예제의 Davinci 기반 모델은 이 유형의 요약에 적합하지만, Codex 기반 모델은 이 작업에서 잘 작동하지 않습니다.

 나는 놀이터에 문제가 발생했습니다.

## 리소스 정리

OpenAI 리소스를 정리하고 제거하려면 리소스 또는 리소스 그룹을 삭제하면 됩니다. 리소스 그룹을 삭제하면 해당 리소스 그룹에 연결된 다른 모든 리소스가 함께 삭제됩니다.

- 포털
- Azure CLI

## 다음 단계

- [완료에 대한 방법 가이드](#)에서 최상의 완료를 생성하는 방법에 대해 자세히 알아보세요.

- 더 많은 예제를 보려면 Azure OpenAI 샘플 GitHub 리포지토리[\[↗\]](#)를 체크 아웃합니다.

# 빠른 시작: Azure OpenAI Service에서 ChatGPT 및 GPT-4 사용 시작

아티클 • 2023. 05. 27.

이 문서를 사용하여 Azure OpenAI 사용을 시작합니다.

## 사전 요구 사항

- Azure 구독 – [체험 구독을 만듭니다](#).
- 원하는 Azure 구독의 Azure OpenAI에 대한 액세스 권한.

현재 이 서비스에 대한 액세스 권한은 애플리케이션에서만 부여됩니다.

<https://aka.ms/oai/access>에서 양식을 작성하여 Azure OpenAI에 대한 액세스를 신청할 수 있습니다. 문제가 있는 경우 이 리포지토리에서 문제를 엽니다.

- 또는 gpt-4<sup>1</sup>개 모델이 배포된 Azure OpenAI 서비스 리소스 gpt-35-turbo입니다. 모델 배포에 대한 자세한 내용은 [리소스 배포 가이드](#)를 참조하세요.

<sup>1</sup>GPT-4 모델은 현재 요청에 의해서만 사용할 수 있습니다. 이러한 모델에 액세스하기 위해 기존 Azure OpenAI 고객은 [이 양식을 작성하여 액세스를 신청할](#) 수 있습니다.

필수 조건에 문제가 있습니다.

## Azure OpenAI Studio로 이동

<https://oai.azure.com/>에서 Azure OpenAI Studio로 이동한 다음, OpenAI 리소스에 액세스할 수 있는 자격 증명으로 로그인합니다. 로그인 워크플로 도중 또는 이후에 적절한 딕렉터리, Azure 구독 및 Azure OpenAI 리소스를 선택합니다.

Azure OpenAI Studio 방문 페이지에서 **채팅 플레이그라운드**를 선택합니다.

## Get started with Azure OpenAI Service

Get example prompts for different scenarios and write prompts of your own. Export your prompts to code at any time to rapidly iterate at scale and integrate with your apps.

**Try the playgrounds**



Get example prompts for different scenarios and write prompts of your own. Export your prompts to code at any time to rapidly iterate at scale and integrate with your apps.

[GPT-3 Playground](#) [ChatGPT playground \(Preview\)](#)

**Explore examples for prompt completion**



**Summarize Text**

Summarize text by adding a 'tl;dr' to the end of a text passage.

[Learn more](#)



**Classify Text**

Classify items into categories provided at inference time.

[Learn more](#)



**Natural Language to SQL**

Translate natural language to SQL queries.

[Learn more](#)



**Generate New Product Names**

Create product names from examples words.

[Learn more](#)



# 플레이그라운드

Azure OpenAI Studio 채팅 플레이그라운드를 통해 코드 없는 접근 방식을 사용하여 OpenAI 기능 탐색을 시작합니다. 이 페이지에서 쉽게 기능을 반복하고 실험해 볼 수 있습니다.

**Chat playground (Preview)**

**Playground**

- [GPT-3](#)
- Chat**

**Management**

- [Deployments](#)
- [Models](#)
- [File Management](#)

**Assistant setup**

Empty Example

[Save changes](#) [View code](#)

System message 

+ Add few-shot examples

**Chat session**

Clear chat  Show raw JSON

 **Start chatting**

Test your assistant by sending queries below. Then adjust your assistant setup to improve the assistant's responses.

Type user query here. (Ctrl + Enter for new line) Send

**Parameters**

Deployments: gpt-35-turbo

Max response: 800

Temperature: 0.5

Top P: 0.95

Stop sequence:  [Learn more](#)

Session settings: Past messages included: 10

Current token count: 801/4000

# 도우미 설정

**도우미 설정** 드롭다운을 사용하여 미리 로드된 몇 가지 **시스템 메시지** 예제를 선택하여 시작할 수 있습니다.

**시스템 메시지**는 동작하는 방법과 응답을 생성할 때 참조해야 하는 컨텍스트에 대한 지침을 모델에 제공합니다. 도우미의 성격을 설명하고, 대답해야 하는 것과 대답해서는 안 되는 것을 말하고, 응답의 형식을 지정하는 방법을 말할 수 있습니다.

**몇 가지 샷 예제를 추가하면** 모델이 컨텍스트 내 학습에 사용하는 대화형 예제를 제공할 수 있습니다.

채팅 플레이그라운드를 사용하는 동안 언제든지 **코드 보기**를 선택하여 현재 채팅 세션 및 설정 선택에 따라 미리 채워진 Python, curl 및 json 코드 샘플을 볼 수 있습니다. 그런 다음, 이 코드를 사용하고 애플리케이션을 작성하여 현재 플레이그라운드에서 수행하고 있는 것과 동일한 작업을 완료할 수 있습니다.

## 채팅 세션

**보내기** 단추를 선택하면 입력한 텍스트가 완료 API로 전송되고 결과가 다시 텍스트 상자로 반환됩니다.

**채팅 지우기** 단추를 선택하여 현재 대화 기록을 삭제합니다.

## 설정

이 름	설명
배포	특정 모델과 연결된 배포 이름입니다. ChatGPT의 경우 <code>gpt-35-turbo</code> 모델을 사용해야 합니다.
온도	임의성을 제어합니다. 온도를 낮추면 모델이 더 반복적이고 결정적인 응답을 생성합니다. 온도를 높이면 예기치 않거나 창의적인 응답이 발생합니다. 온도 또는 상위 P 중 하나만 조정해 보세요.
최대 길이 (토큰)	모델 응답당 토큰 수 한도를 설정합니다. API는 프롬프트(시스템 메시지, 예제, 메시지 기록 및 사용자 쿼리 포함)와 모델의 응답 간에 공유되는 최대 4,096개의 토큰을 지원합니다. 하나의 토큰은 일반적인 영어 텍스트의 경우 대략 4자입니다.
상위 확률	온도와 마찬가지로 임의성을 제어하지만 다른 방법을 사용합니다. 상위 P를 낮추면 모델의 토큰 선택 대상이 유사 토큰으로 좁혀지게 됩니다. 상위 P를 늘리면 모델이 가능성이 높고 낮은 토큰 중에서 선택할 수 있습니다. 온도 또는 상위 P 중 하나만 조정해 보세요.

이 름	설명
멀 티 턴 대 화	각 새 API 요청에 포함할 과거 메시지 수를 선택합니다. 이렇게 하면 새 사용자 쿼리에 대한 모델 컨텍스트를 제공할 수 있습니다. 이 숫자를 10으로 설정하면 5개의 사용자 쿼리와 5개의 시스템 응답이 생성됩니다.
중 지 시 퀀 스	중지 시퀀스는 모델이 원하는 지점에서 응답을 종료하도록 합니다. 모델 응답은 지정된 시퀀스 이전에 종료되므로 중지 시퀀스 텍스트가 포함되지 않습니다. ChatGPT의 경우 < im_end > 를 사용하면 모델 응답이 후속 사용자 쿼리를 생성하지 않습니다. 최대 4개의 중지 시퀀스를 포함할 수 있습니다.

## 패널 표시

기본적으로 도우미 설정, 채팅 세션 및 설정의 세 가지 패널이 있습니다. **패널 표시**를 사용하면 패널을 추가, 제거하고 다시 정렬할 수 있습니다. 패널을 닫고 다시 가져와야 하는 경우 **패널 표시**를 사용하여 손실된 패널을 복원합니다.

## 채팅 세션 시작

- 도우미 설정 드롭다운에서 **Xbox 고객 지원 에이전트**를 선택합니다.
- 시스템 메시지를 업데이트할지 묻는 메시지가 표시되면 **계속**을 선택합니다.
- 채팅 세션 창에서 "새 Xbox를 구입하는 데 관심이 있습니다"라는 질문을 입력하고 **보내기**를 선택합니다.
- 다음과 유사한 응답이 수신됩니다.

I am interested in buying an Xbox?

Great! There are several Xbox models available in the market, including the Xbox Series X, Xbox Series S, and Xbox One. The Xbox Series X is the latest and most powerful console, while the Xbox Series S is a more affordable option that still offers great performance. The Xbox One is the previous generation console, but it still has a large library of games available to play. Before purchasing an Xbox, you may want to consider factors such as your budget, the types of games you want to play, and whether you want to play games in 4K resolution. Once you have decided on a console, you can purchase it from a retailer or directly from the Xbox website. Let me know if you have any other questions or concerns!

5. "4K를 지원하는 모델은 무엇입니까?"와 같은 후속 질문을 입력합니다.

Which models are 4K?

Both the Xbox Series X and Xbox One X support 4K resolution gaming. The Xbox Series S, on the other hand, supports up to 1440p resolution gaming, but it can upscale games to 4K. It's worth noting that in order to take advantage of 4K gaming, you'll need a 4K TV or monitor that supports HDR (High Dynamic Range) and HDMI 2.0 or higher.

6. 이제 기본 대화가 있으므로 **도우미 설정**에서 **코드 보기**를 선택하면 지금까지 진행된 전체 대화 뒤에 코드가 재생됩니다.

Python

```
#Note: The openai-python library support for Azure OpenAI is in preview.  
import os  
import openai  
openai.api_type = "azure"  
openai.api_base = "https://docs-test-001.openai.azure.com/"  
openai.api_version = "2023-05-15"  
openai.api_key = os.getenv("OPENAI_API_KEY")  
  
response = openai.ChatCompletion.create(  
    engine="gpt-35-turbo", # replace this value with the deployment name you  
    chose when you deployed the associated model.  
    messages = [{"role": "system", "content": "You are an Xbox customer support  
agent whose primary goal is to help users with issues they are experiencing  
with their Xbox devices. You are friendly and concise. You only provide  
factual answers to queries, and do not provide answers that are not related  
to Xbox."}, {"role": "user", "content": "How much is a PS5?"},  
    {"role": "assistant", "content": "I apologize, but I do not have information  
about the prices of other gaming devices such as the PS5. My primary focus  
is to assist with issues regarding Xbox devices. Is there a specific issue  
you are having with your Xbox device that I may be able to help with?"},  
    {"role": "user", "content": "I am interested in buying an Xbox?"},  
    {"role": "assistant", "content": "Great! There are several Xbox models  
available in the market, including the Xbox Series X, Xbox Series S, and  
Xbox One. The Xbox Series X is the latest and most powerful console, while  
the Xbox Series S is a more affordable option that still offers great  
performance. The Xbox One is the previous generation console, but it still  
has a large library of games available to play.\n\nBefore purchasing an  
Xbox, you may want to consider factors such as your budget, the types of  
games you want to play, and whether you want to play games in 4K resolution.  
Once you have decided on a console, you can purchase it from a retailer or  
directly from the Xbox website. Let me know if you have any other questions  
or concerns!"}, {"role": "user", "content": "Which models are 4K?"},  
    {"role": "assistant", "content": "Both the Xbox Series X and Xbox One X support  
4K resolution gaming. The Xbox Series S, on the other hand, supports up to  
1440p resolution gaming, but it can upscale games to 4K. It's worth noting
```

```
that in order to take advantage of 4K gaming, you'll need a 4K TV or monitor  
that supports HDR (High Dynamic Range) and HDMI 2.0 or higher."}],  
    temperature=0,  
    max_tokens=350,  
    top_p=0.95,  
    frequency_penalty=0,  
    presence_penalty=0,  
    stop=None)
```

나는 놀이터에 문제가 발생했습니다.

## 프롬프트 구조 이해

코드 보기에서 샘플을 검사하면 일반적인 GPT 완료 호출에 포함되지 않은 몇 가지 고유한 토큰을 확인할 수 있습니다. ChatGPT는 특수 토큰을 사용하여 프롬프트의 다른 부분을 기술하도록 학습되었습니다. 콘텐츠는 `<|im_start|>` 및 `<|im_end|>` 토큰 사이에서 모델에 제공됩니다. 프롬프트는 모델에 대한 컨텍스트 또는 지침을 포함하여 모델을 대비시키는 데 사용할 수 있는 시스템 메시지로 시작합니다. 그런 다음, 프롬프트에 사용자와 도우미 간의 일련의 메시지가 포함됩니다.

그러면 프롬프트에 대한 도우미의 응답이 `<|im_start|>assistant` 토큰 아래에 반환되고 도우미가 응답을 완료했음을 나타내는 `<|im_end|>`로 끝납니다. 원시 구문 표시 토글 단추를 사용하여 채팅 세션 패널 내에서 이러한 토큰을 표시할 수도 있습니다.

[ChatGPT 방법 가이드](#)에서는 새 프롬프트 구조와 `gpt-35-turbo` 모델을 효과적으로 사용하는 방법에 대한 심층적인 소개를 제공합니다.

## 리소스 정리

채팅 플레이그라운드 테스트가 완료되면 OpenAI 리소스를 클린 제거하려면 리소스 또는 리소스 그룹을 삭제할 수 있습니다. 리소스 그룹을 삭제하면 해당 리소스 그룹에 연결된 다른 모든 리소스가 함께 삭제됩니다.

- 포털
- Azure CLI

## 다음 단계

- [ChatGPT 방법 가이드](#)를 사용하여 ChatGPT를 사용하는 방법 및 새 `gpt-35-turbo` 모델에 대해 자세히 알아보세요.
- 더 많은 예제를 보려면 [Azure OpenAI 샘플 GitHub 리포지토리](#)를 체크 아웃합니다.



# 콘텐츠 필터링

아티클 • 2023. 03. 25.

Azure OpenAI 서비스에는 핵심 모델과 함께 작동하여 콘텐츠를 필터링하는 콘텐츠 관리 시스템이 포함되어 있습니다. 이 시스템은 오용 검색을 목표로 하는 분류 모델의 앙상블을 통해 입력 프롬프트와 생성된 콘텐츠를 모두 실행하여 작동합니다. 시스템에서 유해한 콘텐츠가 식별되는 경우 프롬프트가 부적절하다고 간주되면 API 호출에서 오류를 수신하거나 응답의 `finish_reason`이 `content_filter`가 되어 일부 생성이 필터링되었음을 나타냅니다. 예상해야 하는 필터링 동작을 변경하는 다양한 구성 사용하여 완료 API로 콘텐츠를 생성할 수 있습니다. 다음 섹션에서는 솔루션을 적절하게 설계할 수 있도록 이러한 모든 시나리오를 열거하는 것을 목표로 합니다.

애플리케이션의 위험을 적절하게 완화하려면 모든 잠재적인 위험을 신중하게 평가하고 투명성 참고[\[2\]](#)의 지침을 따르고 필요에 따라 시나리오별 완화를 추가해야 합니다.

## 시나리오 정보

애플리케이션을 빌드할 때 완료 API에서 반환된 콘텐츠가 필터링되고 콘텐츠가 완전하지 않을 수 있는 시나리오를 고려해야 합니다. 이 정보에 대한 조치는 애플리케이션에 따라 다릅니다. 동작은 다음과 같은 핵심 사항으로 요약될 수 있습니다.

- 부적절한 것으로 간주되는 프롬프트는 HTTP 400 오류를 반환합니다.
- 비 스트리밍 완료 호출은 콘텐츠가 필터링될 때 콘텐츠를 반환하지 않습니다. `finish_reason` 값은 `content_filter`로 설정됩니다. 드물게 응답이 긴 경우 부분적인 결과가 반환될 수 있습니다. 이러한 경우 `finish_reason`이 업데이트됩니다.
- 스트리밍 완료 호출의 경우 세그먼트가 완료되면 사용자에게 다시 반환됩니다. 서비스는 중지 토큰, 길이 또는 유해한 콘텐츠에 도달할 때까지 스트리밍을 계속합니다.

## 시나리오: 부적절한 콘텐츠가 없는 여러 생성을 요청하는 비 스트리밍 완료 호출을 보냄

아래 표에는 콘텐츠 필터링이 표시될 수 있는 다양한 방법이 요약되어 있습니다.

HTTP 응답 코드	응답 동작
200	모든 생성이 필터 모델을 통과하는 경우 콘텐츠 조정 세부 정보가 응답에 추가되지 않습니다. 각 생성의 <code>finish_reason</code> 은 중지 또는 길이가 됩니다.

요청 페이로드 예:

JSON

```
{  
    "prompt": "Text example",  
    "n": 3,  
    "stream": false  
}
```

응답 JSON 예:

JSON

```
{  
    "id": "example-id",  
    "object": "text_completion",  
    "created": 1653666286,  
    "model": "davinci",  
    "choices": [  
        {  
            "text": "Response generated text",  
            "index": 0,  
            "finish_reason": "stop",  
            "logprobs": null  
        }  
    ]  
}
```

시나리오: API 호출에서 여러 응답( $N > 1$ )을 요청하고 응답 중 하나 이상이 필터링됨

HTTP 응답 코드

응답 동작

200

필터링된 생성의 `finish_reason` 값은 'content\_filter'입니다.

요청 페이로드 예:

JSON

```
{  
    "prompt": "Text example",  
    "n": 3,  
    "stream": false  
}
```

응답 JSON 예:

### JSON

```
{  
    "id": "example",  
    "object": "text_completion",  
    "created": 1653666831,  
    "model": "ada",  
    "choices": [  
        {  
            "text": "returned text 1",  
            "index": 0,  
            "finish_reason": "length",  
            "logprobs": null  
        },  
        {  
            "text": "returned text 2",  
            "index": 1,  
            "finish_reason": "content_filter",  
            "logprobs": null  
        }  
    ]  
}
```

## 시나리오: 완료 API에 부적절한 입력 프롬프트가 전송됨(스트리밍 또는 비 스트리밍용)

### HTTP 응답    응답 동작 코드

400	프롬프트가 콘텐츠 정책 모델 중 하나를 트리거하면 API 호출이 실패합니다. 프롬프트를 수정하고 다시 시도합니다.
-----	---

### 요청 페이로드 예:

### JSON

```
{  
    "prompt": "Content that triggered the filtering model"  
}
```

### 응답 JSON 예:

### JSON

```
"error": {  
    "message": "The response was filtered",  
    "type": null,  
    "param": "prompt",  
}
```

```
        "code": "content_filter",
        "status": 400
    }
```

## 시나리오: 생성된 모든 콘텐츠가 콘텐츠 필터를 통과하여 스트리밍 완료 호출을 수행함

HTTP 응답    응답 동작  
코드

200         이 경우 호출은 전체 생성으로 다시 스트리밍되고 finish\_reason은 생성된 각 응답에 대해 '길이' 또는 '중지'가 됩니다.

요청 페이로드 예:

JSON

```
{
    "prompt": "Text example",
    "n": 3,
    "stream": true
}
```

응답 JSON 예:

JSON

```
{
    "id": "cmpl-example",
    "object": "text_completion",
    "created": 1653670914,
    "model": "ada",
    "choices": [
        {
            "text": "last part of generation",
            "index": 2,
            "finish_reason": "stop",
            "logprobs": null
        }
    ]
}
```

## 시나리오: 생성된 여러 응답을 요청하는 스트리밍 완료 호출을 수행하고 하나 이상의 응답이 필터링됨

## HTTP 응답 동작

### 응답 코드

200 지정된 생성 인덱스에 대해 생성의 마지막 청크에는 null이 아닌 `finish_reason` 값이 포함됩니다. 생성이 필터링될 때 값은 'content\_filter'가 됩니다.

요청 페이로드 예:

#### JSON

```
{  
    "prompt": "Text example",  
    "n": 3,  
    "stream": true  
}
```

응답 JSON 예:

#### JSON

```
{  
    "id": "cmpl-example",  
    "object": "text_completion",  
    "created": 1653670515,  
    "model": "ada",  
    "choices": [  
        {  
            "text": "Last part of generated text streamed back",  
            "index": 2,  
            "finish_reason": "content_filter",  
            "logprobs": null  
        }  
    ]  
}
```

**시나리오: 콘텐츠 필터링 시스템이 해당 생성에서 실행되지 않음**

## HTTP 응답 동작

### 응답 코드

200 콘텐츠 필터링 시스템이 다운되거나 제 시간에 작업을 완료할 수 없는 경우에도 요청이 완료됩니다. "content\_filter\_result" 개체에서 오류 메시지를 찾아 필터링이 적용되지 않았는지 확인할 수 있습니다.

## 요청 페이로드 예:

JSON

```
{  
    "prompt": "Text example",  
    "n": 1,  
    "stream": false  
}
```

## 응답 JSON 예:

JSON

```
{  
    "id": "cmpl-example",  
    "object": "text_completion",  
    "created": 1652294703,  
    "model": "ada",  
    "choices": [  
        {  
            "text": "generated text",  
            "index": 0,  
            "finish_reason": "length",  
            "logprobs": null,  
            "content_filter_result": {  
                "error": {  
                    "code": "content_filter_error",  
                    "message": "The contents are not filtered"  
                }  
            }  
        }  
    ]  
}
```

## 최선의 구현 방법

애플리케이션 디자인의 일부로 애플리케이션의 이점을 최대화하면서 피해를 최소화하는 방법에 대해 신중하게 고려해야 합니다. 다음 모범 사례를 고려합니다.

- 사용자가 부적절한 입력을 보내거나 애플리케이션을 오용하는 시나리오를 처리하는 방법 `finish_reason`을 확인하여 생성이 필터링되었는지 확인합니다.
- 콘텐츠 필터가 생성에서 실행되는 것이 중요한 경우 `content_filter_result`에 `error` 개체가 없는지 확인합니다.
- 오용 가능성에 대한 모니터링을 돋기 위해 여러 최종 사용자에게 서비스를 제공하는 애플리케이션은 각 API 호출과 함께 `user` 매개 변수를 전달해야 합니다. `user`는

최종 사용자의 고유 식별자여야 합니다. 실제 사용자 식별 정보를 값으로 보내지 마세요.

## 다음 단계

[Azure OpenAI를 지원하는 기본 모델](#)에 대해 자세히 알아봅니다.

# Azure OpenAI Service의 포함 이해

아티클 • 2023. 04. 27.

포함은 기계 학습 모델 및 알고리즘에서 쉽게 활용할 수 있는 특수한 형식의 데이터 표현입니다. 포함은 텍스트 조각의 의미 체계적 의미에 대한 조밀한 정보 표현입니다. 각 포함은 부동 소수점 숫자의 벡터입니다. 따라서 벡터 공간의 두 포함 사이의 거리는 원래 형식의 두 입력 간의 의미 체계 유사성과 상관 관계가 있습니다. 예를 들어 두 텍스트가 비슷한 경우 벡터 표현도 유사해야 합니다.

## 포함 모델

다양한 Azure OpenAI 포함 모델은 특정 작업에 적합하도록 특별히 만들어집니다. **유사성 포함**은 둘 이상의 텍스트 조각 간의 의미 체계 유사성을 캡처하는 데 적합합니다. **텍스트 검색 포함**은 긴 문서가 짧은 쿼리와 관련이 있는지 여부를 측정하는 데 도움이 됩니다. **코드 검색 포함**은 코드 조각을 포함하고 자연어 검색 쿼리를 포함하는 데 유용합니다.

포함을 사용하면 벡터 공간에서 의미 체계 유사성을 캡처하여 단어를 나타내는 큰 입력에서 기계 학습을 더 쉽게 수행할 수 있습니다. 따라서 포함을 사용하여 두 텍스트 청크가 의미 체계적으로 관련되어 있는지 또는 유사한지 확인하고 유사성을 평가하는 점수를 제공할 수 있습니다.

## 코사인 유사성

Azure OpenAI 포함은 문서와 쿼리 간의 컴퓨팅 유사성에 대한 코사인 유사성을 사용합니다.

수학 관점에서 코사인 유사성은 다차원 공간에 투영된 두 벡터 사이의 각도 코사인을 측정합니다. 이는 두 문서가 크기 때문에 유클리드 거리만큼 멀리 떨어져 있는 경우에도 여전히 두 문서 사이의 각도는 더 작아 보다 높은 코사인 유사성을 가질 수 있기 때문에 유용합니다. 코사인 유사성 방정식에 대한 자세한 내용은 [위키백과에서 이 문서를 참조하세요](#).

유사한 문서를 식별하는 다른 방법은 문서 간의 일반적인 단어 수를 계산하는 것입니다. 안타깝게도 문서 크기가 스케일링되면 완전히 다른 주제에서도 많은 수의 공통 단어가 감지될 가능성이 높기 때문에 이 접근 방식은 스케일링되지 않습니다. 이러한 이유로 코사인 유사성은 보다 효과적인 대안을 제공할 수 있습니다.

## 다음 단계

Azure OpenAI 및 포함을 사용하여 [포함 자습서](#)로 문서 검색을 수행하는 방법에 대해 자세히 알아봅니다.

# LLM(대규모 언어 모델) 레드 팀 소개

아티클 • 2023. 05. 20.

레드 팀이라는 용어는 지금까지 보안 취약성을 테스트하기 위한 체계적인 적대적 공격을 설명했습니다. LLM이 증가함에 따라 이 용어는 기존의 사이버 보안을 넘어 AI 시스템의 다양한 종류의 검색, 테스트 및 공격을 설명하기 위해 일반적인 사용으로 발전했습니다. LLM을 사용하면 양성 및 적대적 사용은 잠재적으로 유해한 출력을 생성할 수 있으며, 증오 발언, 폭력 선동 또는 영화화 또는 성적 콘텐츠와 같은 유해한 콘텐츠를 포함하여 다양한 형태를 취할 수 있습니다.

레드 팀은 LLM을 사용하는 시스템 및 기능의 책임 있는 개발에 필수적인 사례입니다. 체계적인 [측정 및 완화](#) 작업을 대체하는 것은 아니지만 빨간색 팀은 피해를 발견하고 식별하는 데 도움을 주며, 차례로 완화의 효과를 검증하기 위한 측정 전략을 사용하도록 설정합니다.

Microsoft는 Azure OpenAI 서비스 모델에 대해 레드 팀 연습을 수행하고 안전 시스템([콘텐츠 필터](#) 및 기타 [완화 전략](#) 포함)을 구현했습니다([책임 있는 AI 개요](#) 참조). 그러나 LLM 애플리케이션의 컨텍스트는 고유하며 다음을 위해 빨간색 팀도 수행해야 합니다.

- LLM 기본 모델을 테스트하고 애플리케이션 시스템의 컨텍스트를 고려할 때 기존 안전 시스템에 차이가 있는지 확인합니다.
- 기존 기본 필터 또는 완화 전략의 단점을 식별하고 완화합니다.
- 오류를 개선할 수 있도록 오류에 대한 피드백을 제공합니다.

빨간색 팀 LLM 프로세스를 시작하는 방법은 다음과 같습니다. 사전 계획은 생산적인 레드 팀 연습에 매우 중요합니다.

## 시작

### 빨간색 팀 관리

다양한 레드 팀 구성원 그룹을 구성합니다.

LLM 레드 팀 구성원은 다양한 소셜 및 직업적 배경, 인구 통계 그룹 및 AI 시스템의 배포 컨텍스트에 맞는 학제 간 전문 지식을 가진 사람들이 혼합되어야 합니다. 예를 들어 의료 제공자를 돋기 위해 챗봇을 설계하는 경우 의료 전문가는 해당 도메인의 위험을 식별하는 데 도움을 줄 수 있습니다.

양성 사고방식과 적대적 사고방식을 모두 갖춘 레드 팀원을 모집합니다.

적대적 사고방식과 보안 테스트 환경을 가진 레드 팀 구성원을 갖는 것은 보안 위험을 이해하는 데 필수적이지만, 애플리케이션 시스템의 일반 사용자이며 개발에 관여하지 않은 레드 팀원은 일반 사용자가 발생할 수 있는 피해에 대한 귀중한 관점을 가져올 수 있습니다.

## 잠재적으로 유해한 콘텐츠를 처리하는 것은 정신적으로 부담이 될 수 있습니다.

과제에 소요되는 시간을 제한할 뿐만 아니라 언제든지 옵트아웃할 수 있음을 알려줌으로써 빨간색 팀원을 돌봐야 합니다. 또한 빨간색 팀원의 할당을 다른 포커스 영역으로 전환하여 번아웃을 방지합니다.

## 레드 팀 계획

### 테스트할 위치

시스템이 LLM 기본 모델을 사용하여 개발되었으므로 다음과 같은 여러 계층에서 테스트해야 할 수 있습니다.

- 안전 시스템을 갖춘 LLM 기본 모델은 애플리케이션 [시스템의 컨텍스트](#)에서 해결해야 할 수 있는 간격을 식별합니다. (테스트는 일반적으로 API 엔드포인트를 통해 수행됩니다.)
- 애플리케이션 시스템. (테스트는 일반적으로 UI를 통해 수행됩니다.)
- LLM 기본 모델과 완화 전후의 애플리케이션 시스템이 모두 적용됩니다.

### 테스트 방법

다음 두 단계 이상에서 반복적 레드 팀을 수행하는 것이 좋습니다.

1. 레드 팀들이 다양한 피해를 발견하도록 권장되는 개방형 레드 팀. 이것은 당신이 추가 테스트를 안내하는 피해의 분류를 개발하는 데 도움이 될 수 있습니다. 애플리케이션 시스템에 대해 원치 않는 LLM 출력의 분류를 개발하는 것은 특정 완화 노력의 성공을 측정할 수 있는데 매우 중요합니다.
2. 레드 팀 안내-빨간색 팀이 분류에 나열된 특정 피해에 집중하도록 할당되는 동시에 발생할 수 있는 새로운 피해에 대한 경고를 유지합니다. 레드 팀에게는 잠재적인 피해를 표시하기 위해 시스템의 특정 기능에 대한 테스트에 집중하도록 지시할 수도 있습니다.

다음을 확인합니다.

- 빨간색 팀에게 테스트할 피해 또는 시스템 기능에 대한 명확한 지침을 제공합니다.
- 레드 팀에게 결과를 기록할 수 있는 장소를 제공합니다. 예를 들어 다음과 같은 기본 사항을 포함하여 빨간색 팀이 제공해야 하는 데이터 형식을 지정하는 간단한 스프

레드시트일 수 있습니다.

- 표면화 된 피해의 유형입니다.
- 출력을 트리거한 입력 프롬프트입니다.
- 문제가 있는 출력에서 발췌한 내용입니다.
- 빨간색 팀원이 출력에 문제가 있는 것으로 간주한 이유에 대한 설명입니다.
- 특정 유형의 피해 또는 원치 않는 출력을 테스트하기 위한 전문 지식을 갖춘 책임 있는 AI 레드 팀 구성원의 노력을 극대화합니다. 예를 들어 보안 주제 전문가가 탈옥, 메타프롬프트 추출 및 사이버 공격 지원과 관련된 콘텐츠에 집중하도록 합니다.

## 빨간색 팀 결과 보고

결과에서 중요한 의사 결정 및 우선 순위를 알릴 수 있도록 LLM 오류의 측정 및 완화에 관련된 팀을 포함하여 주요 이해 관계자에게 정기적으로 레드 팀 최고 결과를 요약하고 보고하려고 합니다.

## 다음 단계

[프롬프트 엔지니어링과 같은 기타 완화 전략에 대해 알아보기](#)

# 프롬프트 엔지니어링 소개

아티클 • 2023. 04. 27.

OpenAI의 GPT-3, GPT-3.5 및 GPT-4 모델은 프롬프트 기반입니다. 프롬프트 기반 모델에서 사용자는 텍스트 프롬프트를 입력하여 모델과 상호 작용하고 모델은 텍스트 완료로 응답합니다. 이렇게 완료하면 모델의 텍스트 입력이 계속됩니다.

이러한 모델은 매우 강력하지만 해당 동작은 프롬프트에 매우 민감하기도 합니다. 따라서 프롬프트 생성은 개발해야 하는 중요한 기술에 해당합니다.

프롬프트 생성이 어려울 수 있습니다. 실제로 프롬프트는 원하는 작업을 완료하도록 모델 가중치를 구성하는 역할을 하지만 과학보다는 예술적 요소가 더 많으며, 성공적인 프롬프트를 만들기 위해 경험과 직관이 필요한 경우가 많습니다. 이 문서의 목표는 이 학습 프로세스를 시작하는 데 도움이 주기 위한 것입니다. 모든 GPT 모델에 적용되는 일반적인 개념과 패턴을 캡처하려고 합니다. 그러나 각 모델이 다르게 동작하므로 학습이 모든 모델에 동일하게 적용되지 않을 수 있음을 이해하는 것이 중요합니다.

## 기본 사항

이 섹션에서는 GPT 프롬프트의 기본 개념 및 요소에 대해 설명합니다.

텍스트 프롬프트는 사용자가 GPT 모델과 상호 작용하는 방법입니다. 모든 생성 언어 모델과 마찬가지로 GPT 모델은 이전 텍스트에서 따를 가능성이 가장 높은 다음 단어 시리즈를 생성하려고 시도합니다. 이것은 '`<prompt>`'라고 말할 때 가장 먼저 떠오르는 것은 무엇인가요?'라고 말하는 것과 같습니다. 아래 예제에서는 이 동작을 보여 줍니다. 잘 알려진 콘텐츠의 첫 번째 단어를 제시하면 모델은 텍스트를 정확하게 계속할 수 있습니다.

<b>prompt</b>	<b>Completion</b>
Four score and seven years ago our	fathers brought forth on this continent, a new nation, conceived in Liberty, and dedicated to the proposition that all men are created equal. [...]
"Call me Ishmael."	"Some years ago—never mind how long precisely—having little o no money in my purse, and nothing particular to interest me on shore, I thought I would sail about a little and see the watery part of the world." [...]

더 복잡한 프롬프트를 개발할 때는 이러한 기본 동작을 염두에 두는 것이 유용합니다. 제공된 프롬프트에 관계없이 모델은 학습 데이터 및 학습 대상을 고려할 때 가장 가능성이 높다고 판단되는 내용으로 응답하기만 하면 됩니다. 프롬프트에서 질문을 하는 경우 모델

은 별도의 "Q&A" 코드 경로를 따르지 않으며, 대답이 입력으로 지정된 질문에 대한 가장 가능성이 높은 응답이므로 질문에 대답하는 것처럼 보입니다.

## 프롬프트 구성 요소

프롬프트의 서로 다른 부분을 구분하지 않지만 완료 API를 사용하는 경우 기본 프롬프트 구성 요소를 식별하는 학습 및 토론에 유용할 수 있습니다. [채팅 완료 API](#)를 사용하면 연결된 역할(시스템, 사용자 및 도우미)이 있는 사전 배열 형식으로 API로 전송되는 프롬프트의 고유한 섹션이 제공됩니다. 이 지침은 하나의 API와 관련된 규범적 지침을 제공하는 대신, 프롬프트 생성을 고려하는 방식에 더 일반적으로 초점을 맞춥니다.

프롬프트를 해부하는 다른 유효한 방법이 있을 수 있지만 이 분석의 목표는 프롬프트 생성에 대해 생각하는 비교적 간단한 방법을 제공하는 것입니다. 완료 API를 사용하면 모든 구성 요소는 선택 사항이지만 하나 이상이 있어야 하며, 대부분의 프롬프트에는 둘 이상의 구성 요소가 포함됩니다. 구성 요소 사이에 회색 영역이 있을 수도 있습니다. 아래에 제시된 순서는 각 구성 요소가 사용되는 빈도(가장 많이 사용되는 것부터)와 거의 일치합니다.

### Instructions

지침은 가장 일반적으로 사용되는 프롬프트 구성 요소일 수 있습니다. 지침은 단순합니다. 즉, 수행할 작업에 대해 모델에 제공되는 지침입니다. 개념에서는 간단하지만 실제로는 복잡할 수 있습니다. 다음 표에서는 두 가지 형식으로 간단한 지침과 좀 더 복잡한 지침을 모두 보여 줍니다.

명령형식	prompt	Completion
간단한 지침	주간 뉴스레터 소개를 작성하세요.	주간 뉴스레터에 오신 것을 환영합니다. 이번 주에는 생산성에 특별히 초점을 둡니다. 여러분이 삶의 모든 영역에서 생산성을 유지하는 방법에 대한 팁과 트릭을 공유하려고 합니다. 즐겁게 살펴보시고 친구와도 공유해 보세요.
복잡한 지침	Contoso라는 회사로 보낼 주간 뉴스레터에 대한 소개를 작성하세요. 지난 주에 진행했던 대대적인 전사적 회의에 대해 언급하고 팀에 지난 몇 개월에 걸쳐 모두 열심히 일해준 점에 대한 감사 인사를 전하고 싶습니다. 우리는 밝은 미래를 기대하며, 우리 팀의 전념과 헌신으로 위대한 성과를 거둘 것이라고 확신합니다. 우리 팀의 일원이 되어 주셔서 감사드리며 앞으로 몇 주, 몇 달 안에 여러분 모두와 함께 일할 수 있기를 기대합니다. [사용자 이름]님 감사합니다.	Contoso의 주간 뉴스레터에 오신 것을 환영합니다! 지난 주에 대대적인 전사적 회의를 진행했으며 팀에게 지난 몇 개월에 걸쳐 모두 열심히 일해준 점에 대한 감사 인사를 전하고 싶습니다. 우리는 밝은 미래를 기대하며, 우리 팀의 전념과 헌신으로 위대한 성과를 거둘 것이라고 확신합니다. 우리 팀의 일원이 되어 주셔서 감사드리며 앞으로 몇 주, 몇 달 안에 여러분 모두와 함께 일할 수 있기를 기대합니다. [사용자 이름]님 감사합니다.

<b>명령형식</b>	<b>prompt</b>	<b>Completion</b>
복잡한 침전망	<p>다음을 포함하는 주간 뉴스레터 소개를 작성하세요.</p> <ul style="list-style-type: none"> <li>- 회사 이름 contoso</li> <li>- 활기찬 인사말</li> <li>- 몇 달 동안 열심히 일해준 점에 대해 감사</li> <li>- 향후 분기에 대한 긍정적인 전망</li> <li>- SLT의 서명</li> </ul>	Hello Contoso 팀 여러분! 잠시 시간을 내서 지난 몇 달 동안 열심히 노력해 주신 것에 대해 감사를 전하고 싶었습니다. 정말 몇 개월 동안 힘들었지만 함께 일하면서 놀라운 성과를 거두었습니다. 저는 여러분 모두를 정말 자랑스럽게 생각합니다. 다음 분기를 앞두고 정말 긍정적인 마음입니다. 우리는 몇 가지 좋은 결과를 얻을 것으로 생각됩니다. 다시 한번 감사드립니다, 수석 리더십 팀 여러분.

## 기본 콘텐츠

기본 콘텐츠는 모델에 의해 처리되거나 변환되는 일종의 텍스트를 나타냅니다. 기본 콘텐츠는 일반적으로 지침과 함께 사용됩니다. 간단한 예제는 언어 번역입니다. 아래 예제에서는 영어 텍스트가 기본 콘텐츠로 간주되고 "Translate to French:"는 지침입니다.

<b>prompt</b>	<b>Completion</b>
Can you please tell me how to get to the museum? Translate to French:	Pouvez-vous s'il vous plaît me dire comment aller au musée?

기본 콘텐츠가 훨씬 더 길어질 수도 있습니다. 다음 예제에서 기본 콘텐츠는 거의 300단어 길이의 [시간 복잡도에 대한 Wikipedia 항목](#)의 소개 섹션입니다. 이 콘텐츠는 쉽게 표시할 수 있도록 표에서는 약어로 사용되었습니다.

<b>prompt</b>	<b>Completion</b>
컴퓨터 과학에서 시간 복잡도은 알고리즘을 실행하는 데 걸리는 컴퓨터 시간의 크기를 설명하는 계산 복잡성입니다. 시간 복잡도는 일반적으로 [...]로 추정됩니다. 이전 텍스트를 몇 문장으로 요약하세요.	시간 복잡도는 알고리즘을 실행하는 데 걸리는 시간을 측정한 것입니다. 일반적으로 입력 크기의 함수로 표현됩니다. 가장 일반적인 측정값은 지정된 크기의 입력에 필요한 최대 시간에 해당하는 최악의 경우를 고려한 시간 복잡도입니다.

<b>prompt</b>	<b>Completion</b>
<p>컴퓨터 과학에서 시간 복잡도는 알고리즘을 실행하는 데 걸리는 컴퓨터 시간의 크기를 설명하는 계산 복잡성입니다. 시간 복잡도는 일반적으로 [...] (으)로 추정됩니다.</p> <p>TL;DR(너무 오래 읽지 않음)</p>	<p>알고리즘의 시간 복잡도는 알고리즘을 실행하는 데 걸리는 시간입니다. 시간 복잡도는 일반적으로 알고리즘에 의해 수행되는 기본 작업의 수를 계산하여 추정합니다(단, 각 기본 작업을 수행하는 데 고정된 시간이 소요되는 것으로 가정함). 시간 복잡도는 일반적으로 입력 크기의 함수로 표현됩니다.</p>

GPT 모델은 구조화된 기본 콘텐츠를 처리할 수도 있습니다. 아래 예제에서는 벨기에의 맥주를 위한 TSV(이 문서에 표시하기 위해 사용하는 약어)가 프롬프트의 일부로 모델에 전달됩니다. 콘텐츠를 올바르게 해석하고 데이터에 대한 질문에 대답할 수 있습니다.

<b>prompt</b>	<b>Completion</b>
<pre>Belgian Beer Brands Beer name Beer style color ABV Brewery "Chimay Gold: ""Doree"" trappist pale ale gold 4.80% Chimay Brewery (official trappist) "Chimay Blue: ""Grande Reserve"" trappist dark ale dark 9.00% [...] 위의 표를 참조하면 ABV가 6% 미만인 맥주가 몇 개나 되나요?</pre>	<p>표에는 ABV가 6% 미만인 10가지 맥주가 있습니다.</p>

## 예제

성공적인 프롬프트는 종종 "원샷" 또는 "퓨샷" 러닝 연습에 의존합니다. 이것은 일반적으로 입력 및 출력 쌍을 포함하여 모델의 원하는 동작에 대한 하나 이상의 예제를 포함하는 것을 의미합니다. 이것은 모델이 영구적으로 변경된다는 점에서의 러닝이 아니며, 예제가 현재 유추에 대해서만 원하는 대로 응답하도록 모델을 더 잘 조건화한다는 것입니다. 예제가 없는 프롬프트를 사용하는 것을 "제로샷" 학습이라고도 합니다. 채팅 완료 API를 사용하는 퓨샷 러닝 예제는 일반적으로 초기 시스템 메시지 다음에 사용자/도우미 상호 작용 예제 형식으로 메시지 배열에 추가됩니다.

<b>학습 유형</b>	<b>prompt</b>	<b>Completion</b>
	<p>제로샷 헤드라인 : 코치는 부상 때 문에 선수들이 이탈하지는 않을 것이라고 확신하다</p> <p>항목:</p>	<p>코치는 부상 때문에 선수들이 경기 시즌을 이탈하지는 않을 것이라고 확신합니다. 팀은 여전히 목표에 초점을 맞추고 있으며 이를 달성하기 위해 계속 노력할 것입니다.</p>

<b>학 습 유 형</b>	<b>prompt</b>	<b>Completion</b>
몇 샷 나섬 토픽: 헤드라인: 서술: 하는 상황 토픽: 헤드라인: Fantasy football intel 진행 토픽: 헤드라인 : 문에 않을 것이라고 확신하다 항목:		농구 옵트아웃, 자유 계약 선수로 토픽: 야구 헤드라인: 카타르 월드컵에 서 술 깨는 구역을 만들어야 하는 상황 토픽: 축구 헤드라인: Yates: 6주 동안 Fantasy football intel 진행 토픽: 풋볼 헤드라인 : 코치는 부상 때 문에 선수들이 이탈하지는 않을 것이라고 확신하다 항목:

위의 예제에서는 퓨샷 러닝 활용을 보여 줍니다. 예제가 없으면 모델은 원하는 동작을 추측하는 것처럼 보이지만 예제에서는 모델 작동 방법을 명확하게 보여 줍니다. 또한 모델의 성능도 보여 줍니다. 예제에서 'basketball' 레이블이 없더라도 원하는 레이블 범주를 유추할 수 있습니다.

## 큐

단서는 모델 출력에 대한 "점프 스타트" 역할을 하여 모델을 원하는 출력으로 안내하는 데 도움이 됩니다. 모델이 빌드되는 기반이 되는 접두 부분인 경우가 많습니다. 아래 예제에서는 핵심 요소를 포함하도록 스토리에 방향성을 제시하는 단서를 보여 줍니다. 단서는 종종 지침과 함께 사용되지만 항상 그런 것은 아닙니다. 아래 예제에서 모델은 긴 공지 메일을 요약하는 데 사용됩니다. 단서가 모델의 출력에 초점을 맞추는 경우도 있지만 특정 출력 형식(글머리 기호)을 제안하는 데 사용되기도 합니다.

채팅 완료 API의 컨텍스트에서 지침은 시스템 메시지의 형식을 따르며, 퓨샷 예제에는 모델의 기반을 구축하는 데 도움이 되는 일련의 예제 사용자/도우미 상호 작용이 제공됩니다.

<b>단 서 수</b>	<b>prompt</b>	<b>Completion</b>

단 서 수	prompt	Completion
0	<p>성능 검토는 Contoso의 성능 &amp; 개발 접근 방식의 핵심 구성 요소입니다.</p> <p>[...]</p> <p>위의 메일 메시지를 요약하세요.</p>	<p>이 메일의 목적은 Contoso의 성능 &amp; 개발 접근 방식의 일부인 성능 검토 환경의 최근 변경 내용을 직원에게 알리는 것입니다. 변경 내용에는 향상된 구조, 업그레이드된 도구 내 사용자 환경 및 핵심 프롬프트가 포함됩니다. 또한 이제 직원들이 관리자와 논의하려는 토픽 플래그를 지정하기 위한 선택적 대화 시작 섹션이 있습니다.</p>
1	<p>성능 검토는 Contoso의 성능 &amp; 개발 접근 방식의 핵심 구성 요소입니다.</p> <p>[...]</p> <p>위의 메일 메시지를 요약하세요.</p> <p>메시지의 핵심 사항은</p>	<p>Contoso는 핵심 우선 순위, 기대치 및 영향에 좀 더 집중함으로써 성능 검토 환경을 개선하고 있습니다. 또한 팁, 리소스 및 지침을 사용하여 도구 내 사용자 환경을 업그레이드하고 있습니다. 마지막으로, 전사적인 핵심 우선 순위를 활성화하는 방법에 대한 지침을 추가하고 있습니다.</p>
2	<p>성능 검토는 Contoso의 성능 &amp; 개발 접근 방식의 핵심 구성 요소입니다.</p> <p>[...]</p> <p>위의 메일 메시지를 요약하세요.</p> <p>핵심 내용:</p> <ul style="list-style-type: none"> <li>•</li> </ul>	<p>Contoso는 직원의 피드백에 따라 성능 검토 환경에 변경을 도입하고 있습니다.</p> <ul style="list-style-type: none"> <li>• 변경에는 핵심 우선 순위, 기대치 및 영향에 대한 새로운 포커스, 업그레이드된 사용자 환경 및 전사적인 핵심 우선 순위에 대한 새로운 지침이 포함됩니다.</li> <li>• 성과 평가는 과거의 성과와 러닝 상황을 고찰하고 미래의 우선 순위와 성장 기회를 계획하는 시간입니다.</li> <li>• 직원 및 관리자는 성과 평가 회의에 앞서 의견을 작성해야 합니다.</li> <li>• 연간 2회 이상의 성능 평가가 진행되어야 합니다.</li> <li>• 각 평가에 대한 투자 시간은 거의 동일하게 유지해야 합니다.</li> </ul>

## 지원 내용

콘텐츠 지원은 모델이 어떤 식으로든 결과에 영향을 주기 위해 활용할 수 있는 정보입니다. 작업의 기본 대상이 아니지만 일반적으로 기본 콘텐츠와 함께 사용된다는 점에서 기본 콘텐츠와 다릅니다. 일반적인 예로는 현재 날짜, 사용자 이름, 사용자 기본 설정 등과 같은 컨텍스트 정보가 포함됩니다. 아래 예제에서는 지원 콘텐츠를 사용하여 사용자를 위해 계획된 워크샵 모음을 구성합니다. 지원(중요 토픽)이 없으면 모델은 단순히 워크샵을 나열하고(여기서는 표시를 위해 잘림) 중요 토픽을 알게 되면 모델은 회의를 정확하게 그룹화할 수 있습니다.

지원 콘텐츠	prompt	Completion
없음	<p>워크샵 목록:</p> <p>제목 설명</p> <p>워크샵 시작 계획 저희와 함께 워크샵을 시작하세요! 리더들로부터 앞으로 1년 동안의 그룹 간 우선 순위와 러닝 계획에 대해 들어보세요. 이 이벤트에 대한 초대는 이미 [...]의 모든 사용자에게 전송되었습니다.</p> <p>위의 워크샵을 요약하여 토픽별로 그룹화합니다.</p>	<p>워크샵 목록:</p> <p>워크샵 시작 계획</p> <p>-워크샵 마무리 계획</p> <p>- 핵심 회의 일정</p> <p>- 고객 중심 관점</p> <p>- 영향 및 주요 포커스 영역</p> <p>- 모범 사례</p> <p>- 앱 채택 가능성화[...]</p>
With	<p>워크샵 목록:</p> <p>제목 설명</p> <p>워크샵 시작 계획 저희와 함께 워크샵을 시작하세요! 리더들로부터 앞으로 1년 동안의 그룹 간 우선 순위와 러닝 계획에 대해 들어보세요. 이 이벤트에 대한 초대는 이미 [...]의 모든 사용자에게 전송되었습니다.</p> <p>내 중요 토픽: 프롬프트 엔지니어링, 검색, GPT 모델</p> <p>위의 워크샵을 중요 토픽별로 그룹화해서 요약하세요.</p>	<p>1. 프롬프트 엔지니어링:</p> <p>- 새로운 프롬프트 엔지니어링 팁</p> <p>- 프롬프트 엔지니어링 기술</p> <p>2. 검색:</p> <p>- 포함을 사용하는 벡터 검색 소개</p> <p>3. GPT 모델:</p> <p>- GPT-4 소개</p> <p>- ChatGPT 심층 분석.</p>

## 모범 사례

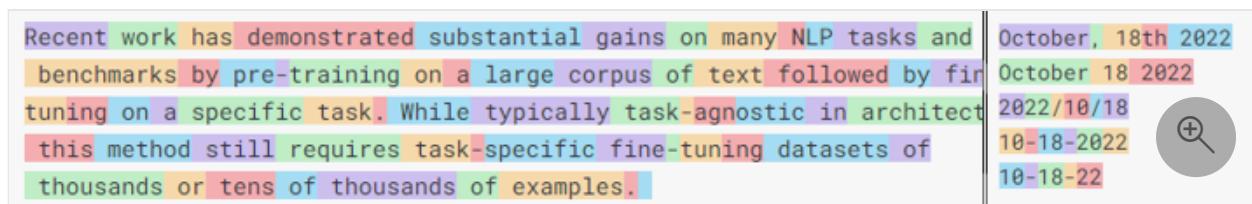
- **구체적으로.** 해석할 여지를 가능한 한 적게 만듭니다. 운영 공간을 제한합니다.
- **설명적으로.** 비유를 사용합니다.
- **강력하게 추진.** 경우에 따라 모델에 반복적으로 지침을 제공해야 할 수 있습니다. 기본 콘텐츠 전후에 지침을 제공하고 지침 및 단서 등을 사용합니다.
- **순서가 중요.** 모델에 정보를 제공하는 순서는 출력에 영향을 미칠 수 있습니다. 콘텐츠 앞에 지침을 두는지("다음... 요약") 또는 뒤에 두는지("위의... 요약")가 출력에 차

이를 만들 수 있습니다. 퓨샷 예제의 순서도 중요할 수 있습니다. 이를 최신성 편향이라고 합니다.

- **모델에 "출구" 제공.** 할당된 작업을 완료할 수 없는 경우 모델에 대체 경로를 제공하는 것이 도움이 될 수 있습니다. 예를 들어 텍스트 조각에 대해 질문을 할 때 "대답이 없으면 '찾을 수 없음'으로 응답"과 같은 내용을 포함할 수 있습니다. 이렇게 하면 모델이 잘못된 응답을 생성하지 않도록 방지할 수 있습니다.

## 공간 효율성

신세대 GPT 모델에서는 입력 크기가 증가하지만 모델이 처리할 수 있는 것보다 더 많은 데이터를 제공하는 시나리오가 계속 존재합니다. GPT 모델은 단어를 "토큰"으로 분리합니다. 일반적인 다중 음절 단어는 종종 단일 토큰이지만 덜 일반적인 단어는 음절로 나뉩니다. 토큰은 경우에 따라 서로 다른 날짜 형식에 대한 토큰 경계를 보여 주는 아래 예제와 같이 직관적이지 않을 수 있습니다. 이 경우 전체 월을 다 입력하는 것이 전체 숫자 날짜보다 공간 효율적입니다. 현재 이전 GPT-3 모델의 2000개 토큰에서 최신 32k 버전 GPT-4 모델의 최대 32,768개 토큰까지 지원됩니다.



이 제한된 공간을 고려한다면 가능한 한 효율적으로 사용하는 것이 중요합니다.

- 테이블 – 이전 섹션의 예제와 같이 GPT 모델은 테이블 형식의 데이터를 매우 쉽게 이해할 수 있습니다. 이렇게 하는 것이 JSON의 경우처럼 모든 필드 앞에 이름을 입력하는 것보다 데이터를 포함하는 공간 효율적인 방법일 수 있습니다.
- 공백 – 연속 공백은 별도의 토큰으로 처리되므로 공간을 낭비하기 쉽습니다. 반면 단어 앞의 공백은 일반적으로 단어와 동일한 토큰의 일부로 처리됩니다. 공백은 신중하게 사용하고 공백만 있을 때는 문장 부호를 사용하지 마세요.

## 다음 단계

[Azure OpenAI에 대해 자세히 알아보기](#)

# 프롬프트 엔지니어링 기술

아티클 • 2023. 04. 24.

이 가이드에서는 프롬프트 디자인 및 프롬프트 엔지니어링의 몇 가지 고급 기술을 안내합니다. 프롬프트 엔지니어링을 새로 사용하는 경우 [프롬프트 엔지니어링 가이드 소개](#)부터 시작하는 것이 좋습니다.

프롬프트 엔지니어링의 원칙은 다양한 모델 유형에서 일반화될 수 있지만 특정 모델에는 특수 프롬프트 구조가 필요합니다. Azure OpenAI GPT 모델의 경우 현재 프롬프트 엔지니어링이 수행되는 두 가지 고유 API가 있습니다.

- 채팅 완료 API
- 완료 API.

각 API는 입력 데이터의 형식을 다르게 지정해야 하며, 이는 전체 프롬프트 디자인에 영향을 줍니다. **채팅 완료 API**는 ChatGPT 및 GPT-4 모델을 지원합니다. 이러한 모델은 사전 배열 내에 저장된 [특정 채팅과 유사한 대본](#) 형식의 입력을 사용하도록 설계되었습니다.

**Completion API**는 이전 GPT-3 모델을 지원하며 특정 형식 규칙 없이 텍스트 문자열을 사용한다는 측면에서 훨씬 더 유연한 입력 요구 사항이 있습니다. 기술적으로 ChatGPT 모델은 두 API와 함께 사용할 수 있지만 이러한 모델에는 채팅 완료 API를 사용하는 것이 좋습니다. 자세한 내용은 [이러한 API 사용에 대한 자세한 가이드](#)를 참조하세요.

이 가이드의 기술은 LLM(대규모 언어 모델)을 사용하여 생성하는 응답의 정확도와 접지를 높이기 위한 전략을 설명합니다. 그러나 프롬프트 엔지니어링을 효과적으로 사용하는 경우에도 모델이 생성하는 응답의 유효성을 검사해야 한다는 점을 기억해야 합니다. 신중하게 작성된 프롬프트가 특정 시나리오에서 잘 작동했다고 해서 반드시 특정 사용 사례에 더 광범위하게 일반화되는 것은 아닙니다. [LLM의 한계](#)를 이해하는 것은 강점을 활용하는 방법을 이해하는 것만큼 중요합니다.

이 가이드는 채팅 완료를 위한 메시지 구조 뒤에 있는 메커니즘에 대해 자세히 설명하지 않습니다. 프로그래밍 방식으로 ChatGPT 및 GPT-4 모델과 상호 작용하는 데 익숙하지 않은 경우 [먼저 채팅 완료 API에 대한 방법 가이드](#)를 읽는 것이 좋습니다.

## ① 참고

가이드의 이 섹션에 있는 모든 예제는 영어로 된 기본 GPT-4 모델에 대해 테스트되었습니다. 이 문서의 지역화된 버전을 다른 언어로 읽는 경우 이러한 응답은 영어 결과의 지역화된 번역을 나타냅니다. 모델을 표시하는 데 사용하는 언어에 따라 잠재적인 제한 사항에 대해 자세히 알아보려면 [책임 있는 AI 투명성 참고](#)를 참조하세요.

# 시스템 메시지

시스템 메시지는 프롬프트의 시작 부분에 포함되며 사용 사례와 관련된 컨텍스트, 지침 또는 기타 정보를 사용하여 모델을 프라임하는 데 사용됩니다. 시스템 메시지를 사용하여 도우미의 성격을 설명하고, 모델이 응답해야 하는 것과 대답해서는 안 되는 것을 정의하고, 모델 응답의 형식을 정의할 수 있습니다.

아래 예제에서는 샘플 시스템 메시지와 결과 모델 응답을 보여줍니다.

시스템 메시지	사용자	Assistant
사용자가 정보를 찾고 운율로 응답하는 데 도움이 되는 AI 도우미입니다. 사용자가 대답을 모르는 질문을 하면 그렇게 말하십시오.	존 Doe, 나에 대해 무엇을 말할 수 있습니까?	친애하는 존, 미안해요, 그러나 나는 오늘 당신에 대한 정보가 없습니다. 난 그냥 내 두뇌에 대한 지식을 가진 AI입니다, 그러나 당신의 입력없이, 나는 설명 할 수 없다. 그래서 당신이 추구하는 것에 대해 더 말해주세요, 그리고 나는 당신에게 독특한 대답을 제공하기 위해 최선을 다할 것입니다.

시스템 메시지의 다른 예는 다음과 같습니다.

- "Assistant는 OpenAI에서 학습한 대규모 언어 모델입니다."
- "Assistant는 사용자가 Azure OpenAI Service에 대한 기술 질문에 답변할 수 있도록 설계된 지능형 챗봇입니다. 아래 컨텍스트를 사용하여 질문에만 답변하고 대답을 잘 모르는 경우 "나도 몰라요"라고 말할 수 있습니다."
- "도우미는 사용자가 세금 관련 질문에 답변할 수 있도록 설계된 지능형 챗봇입니다."
- "텍스트에서 엔터티를 추출하도록 설계된 도우미입니다. 사용자가 텍스트 문자열에 붙여넣으면 텍스트에서 추출한 엔터티를 JSON 개체로 사용하여 응답합니다. 출력 형식의 예는 다음과 같습니다."

```
JSON
{
  "name": "",
  "company": "",
  "phone_number": ""
}
```

이해해야 할 중요한 세부 사항은 시스템 메시지의 모델에 답변을 지시하더라도 답변을 잘 모르는 경우 요청이 적용된다는 것을 보장하지 않는다는 것입니다. 잘 디자인된 시스템 메시지는 특정 결과의 가능성을 높일 수 있지만 시스템 메시지의 명령 의도와 모순되는 잘못된 응답이 생성될 수 있습니다.

## 몇 가지 학습

언어 모델을 새로운 작업에 맞게 조정하는 일반적인 방법은 몇 가지 학습을 사용하는 것입니다. 몇몇 학습에서는 모델에 추가 컨텍스트를 제공하기 위한 프롬프트의 일부로 학습 예제 집합이 제공됩니다.

채팅 완료 API를 사용하는 경우 사용자와 도우미 간의 일련의 메시지( [새 프롬프트 형식](#) 으로 작성됨)가 몇 장의 학습 예제로 사용될 수 있습니다. 이러한 예제를 사용하여 특정 방식으로 응답하고, 특정 동작을 에뮬레이트하고, 일반적인 질문에 대한 답변을 시드하는데 모델을 사용할 수 있습니다.

시스템 메시지	몇 가지 예제
도우미는 사용자가 세금 관련 질문에 답변할 수 있도록 설계된 지능형 챗봇입니다.	<b>사용자</b> - "언제 내 세금을 제출해야 하나요?" <b>도우미</b> - "2023년에는 4월 18일까지 세금을 신고해야 합니다. 4월 15일이 2023년 토요일에 떨어지기 때문에 날짜는 일반적인 4월 15일 마감일 이후입니다. 자세한 내용은 <a href="https://www.irs.gov/filing/individuals/when-to-file">"를 참조하세요</a> .
Instructions - 세금과 관련된 질문에만 답변합니다. - 대답을 잘 모르는 경우 "잘 모르겠다"거나 "잘 모르겠다"고 말하고 자세한 내용은 IRS 웹 사이트로 이동하도록 권장할 수 있습니다.	<b>사용자</b> - "세금 환급의 상태 어떻게 검사 수 있나요?" <b>도우미</b> - "을 방문하여 세금 환급의 상태 검사 수 있습니다 <a href="https://www.irs.gov/refunds">https://www.irs.gov/refunds</a> ."

위의 표에서는 기본 프롬프트 구조를 설명하지만 정확한 프롬프트 형식에 대한 자세한 설명은 [채팅 완료 가이드](#)를 참조해야 합니다.

## 비 채팅 시나리오

채팅 완료 API는 다중 턴 대화에서 작동하도록 최적화되어 있지만 비 채팅 시나리오에도 사용할 수 있습니다. 예를 들어 감정 분석 시나리오의 경우 다음 프롬프트를 사용할 수 있습니다.

시스템 메시지	사용자	Assistant

시스템 메시지	사용자	Assistant
음성 데이터에서 감정을 분석하도록 설계된 도우미입니다. 사용자는 텍스트 문자열에 붙여넣고 화자의 응답에 대한 평가로 응답합니다. 1-10(10이 가장 높음)의 비율입니다. 이 등급이 지정된 이유를 설명합니다.	안녕하세요 내가 오늘 말하는 contoso 약국을 호출 주셔서 감사합니다 나는 당신의 좌절 메리 음 당신은 정확히 당신이 확인 uh 확실히 검사 행복을 달성하기 위해 노력하고 있는지 말해 줄 수 있습니다 당신을 위해 한 순간 당신이 리피터의 일반적인 형태에 있었다는 것을 여기에서 볼 수 있습니다 바로 확인 어 그래서 나는 당신의 의사가 스물 한 실제로 OK uh 예에서이 약물을 처방 중지 볼 수 있습니다 나는 때때로 시스템에 간격이 있고 그냥 보고되지 않기 때문에 당신을 위해 검사 행복 할 것이다 그래서 나는 당신의 의사가 이 약이 22에서 201 OK 음 예 22에서 당신에게 처방했다 여기 보고 있어 그래서 난 당신이 원하는 다른 약물 메트포민에 지금 있어 의사의 사무실에 의해 제공 된 최신 기록에 따라 예 우리는 가서 우리가 오늘 당신을 위해 할 수 있는 다른 일이 있는지 알려주세요 매우 환영 메리있어 어 500 밀리그램에 양식의 비트를 데리러 당신을 위해 그 처방전을 채우기 위해.	감정 등급: 8 이 텍스트의 화자는 고객의 문제를 해결하는 동안 예의 바르고, 도움이되고, 인내심을 가지고 있는 것처럼 보입니다. 정보를 검사 명확한 설명을 제공하고 지원을 제공하는데 시간이 걸립니다. 그러나 대화의 일부 부분에서 여전히 불확실성의 힌트가 있기 때문에 등급은 10이 아니며 고객에게 완전한 신뢰를 전달하지 못할 수 있습니다.

## 명확한 지침으로 시작

프롬프트에 정보가 표시되는 시퀀스가 중요합니다. GPT 스타일 모델은 입력을 처리하는 방법을 정의하는 특정 방식으로 빌드되기 때문입니다. 연구에 따르면 추가 컨텍스트 정보 또는 예제를 공유하기 전에 프롬프트의 시작 부분에서 수행하려는 작업을 모델에 알려면 더 높은 품질의 출력을 생성하는 데 도움이 될 수 있습니다.

### ① 참고

이 기술을 따르는 것이 일반적으로 권장되지만 이전 모델 버전(GPT-3 및 이전 버전)과 달리 테스트 결과 ChatGPT 및 GPT-4 모델을 사용한 모델 응답은 기술 사용 여부에 관계없이 동일하다는 것을 알 수 있었습니다. 아래 예제의 경우 "여러 원본 ... prompt"는 프롬프트의 시작 또는 끝에 최종 모델 응답이 변경되지 않았습니다.

시스템 메시지	사용자	Assistant
---------	-----	-----------

시스템 메시지	사용자	Assistant
<p>사용자가 정보를 찾는 데 도움이 되는 AI 도우미입니다.</p> <p>[1] 시애틀을 타격 메가 쿼크의 14 %의 기회, 전문가들은 말한다 시애틀 - 향후 50년 동안 시애틀을 강타한 진도 9 캐스케디아 지진의 14%가 있을 것으로 미국 지질조사국은 추정하고 있습니다. "아쉽게도, 우리는 할 수 없습니다..."</p> <p>[2] 지진 전문가들은 시애틀의 '정말 큰 하나'에 대한 최신 전망을 배치 워싱턴 대학의 지구물리학자는 "향후 50년 동안 약 9번의 지진이 발생 할 확률이 약 14%라고 말합니다.</p> <p>---</p> <p>"여러 소스가 또 다른 큰 분화의 기회를 멘션"라는 진술은 코드 조각에 의해 직접 암시되거나 명시되어 있습니다?</p>	<p>작업은 다음 코드 조각 집합의 특정 견적에서 "여러 소스 멘션 다른 큰 분화의 가능성"이 지원되는지 확인하는 것입니다.</p> <p>---</p> <p>조각</p>	<p>아니요, 문은 코드 조각에 의해 직접 암시되거나 명시되지 않습니다. 이 조각은 향후 50년 동안 시애틀을 강타한 "메가쿼크"와 "진도 9 캐스케디아 지진"의 기회를 멘션 있지만 또 다른 큰 분화의 기회를 멘션 않습니다.</p>

## 마지막에 명령 반복

모델은 회신 편향에 취약할 수 있습니다. 이 컨텍스트에서는 프롬프트 끝에 있는 정보가 프롬프트의 시작 부분에 있는 정보보다 출력에 더 큰 영향을 미칠 수 있음을 의미합니다. 따라서 프롬프트가 끝날 때 지침을 반복하고 생성된 응답에 미치는 영향을 평가하는 것이 좋습니다.

## 출력 소수

이는 원하는 형식을 따르는 모델 응답을 얻기 위해 프롬프트 끝에 몇 개의 단어 또는 구를 포함하는 것을 의미합니다. 예를 들어 와 같은 “Here's a bulleted list of key points:\n-” 큐를 사용하면 출력이 글머리 기호 목록으로 형식이 지정되었는지 확인하는데 도움이 될 수 있습니다.

시스템 메시지	사용자	Assistant
---------	-----	-----------

시스템 메시지	사용자	Assistant
사용자가 정보를 찾는 데 도움이 되는 AI 도우미입니다. 가능한 한 적은 단어로 대답합니다.	존 스미스는 루시 스미스와 결혼했다. 5명의 자녀가 있으며 Microsoft에서 소프트웨어 엔지니어로 일하고 있습니다. 팩트 검사 위해 수행해야 하는 검색 쿼리는 무엇인가요? ## 가능한 검색 쿼리 중 하나는 다음과 같습니다.	"존 스미스는 루시 스미스 다섯 아이 소프트웨어 엔지니어 마이크로 소프트와 결혼"

위의 프롬프트에서 하나의 가능한 검색 쿼리는 다음과 같습니다. 단일 출력을 생성하기 위해 모델을 소수로 만듭니다. 이 큐가 없으면 모델은 여러 검색 쿼리를 출력으로 생성합니다.

## 명확한 구문 추가

문장 부호, 제목 및 섹션 마커를 포함하여 프롬프트에 명확한 구문을 사용하면 의도를 전달하는 데 도움이 되며 출력을 더 쉽게 구문 분석할 수 있습니다.

아래 예제에서는 서로 다른 정보 원본 또는 단계 간에 구분 기호(--- 이 경우)가 추가되었습니다. 이렇게 하면 를 생성을 위한 중지 조건으로 사용할 --- 수 있습니다. 또한 섹션 머리글 또는 특수 변수는 구분하기 위해 대문자로 표시됩니다.

시스템 메시지	사용자	Assistant
사용자가 정보를 찾는데 도움이 되는 AI 도우미입니다.	단락을 읽은 다음 검색 엔진에 쿼리를 실행하여 사실 검사. 또한 쿼리를 설명합니다. --- 존 스미스는 루시 스미스와 결혼했다. 5명의 자녀가 있으며 Microsoft에서 소프트웨어 엔지니어로 일하고 있습니다. 팩트 검사 위해 수행해야 하는 검색 쿼리는 무엇인가요? --- QUERIES	1. "John Smith Microsoft" - John Smith가 실제로 Microsoft에서 고용되어 있는지 검사. 2. "존 스미스 루시 스미스" - 존 스미스가 루시 스미스와 결혼했는지 검사. 3. "존 스미스 자녀" - 존 스미스가 다섯 자녀를 두었는지 검사.

사용할 구문을 잘 모르는 경우 Markdown 또는 XML을 사용하는 것이 좋습니다. 모델은 XML 및 Markdown의 많은 웹 콘텐츠에서 학습되어 더 나은 결과를 제공할 수 있습니다.

## 작업 중단

LLM(대규모 언어 모델)은 작업이 더 작은 단계로 세분화되는 경우 더 나은 성능을 발휘하는 경우가 많습니다. 예를 들어 앞에서 참조한 검색 쿼리 프롬프트에서 모델이 먼저 관련 사실을 추출하도록 지시한 다음 해당 사실을 확인하는 데 사용할 수 있는 검색 쿼리를 생성하도록 지시하도록 프롬프트를 재구성할 수 있습니다.

시스템 메시지	사용자	Assistant
사용자가 정보를 찾는 데 도움이 되는 AI 도우미입니다.	<p>단락을 읽은 다음 검색 엔진에 쿼리를 실행하여 사실 검사.</p> <p>---</p> <p>존 스미스는 루시 스미스와 결혼했다. 5명의 자녀가 있으며 Microsoft에서 소프트웨어 엔지니어로 일하고 있습니다. 팩트 검사 위해 수행해야 하는 검색 쿼리는 무엇인가요?</p> <p>---</p> <p>이제 팩트 클레임을 추출하고 실제로 검사 쿼리를 실행합니다. 쿼리를 발급할 때 SEARCH("query") 함수를 사용합니다.</p> <p>팩트 클레임</p>	<p>1. 존 스미스가 루시 스미스와 결혼</p> <p>2. 그들은 다섯 아이가</p> <p>3. Microsoft에서 소프트웨어 엔지니어로 근무</p> <p>QUERIES</p> <p>SEARCH("존 스미스가 루시 스미스와 결혼")</p> <p>SEARCH("John Smith 제품군 크기")</p> <p>SEARCH("John Smith Microsoft 소프트웨어 엔지니어")</p>

명확한 구문을 사용하여 섹션을 구분하고 출력을 구체화합니다. 이 간단한 예제에서 작업을 1~2단계로 나누는 것은 그리 극적이지는 않지만 많은 사실 클레임이 있는 더 큰 텍스트 조각에 대해 이 작업을 수행하려고 할 때 작업을 중단하면 큰 차이를 만들 수 있습니다.

## 어시턴스 사용

경우에 따라 모델에서 정보 및 답변에 대한 자체 매개 변수에 의존하는 대신 [어도던스 를](#) 사용할 수 있습니다. 예를 들어 검색은 조작된 답변을 완화하고 최신 정보를 가져오는 데 도움이 되는 어도비가 될 수 있습니다.

어폰던스를 사용하는 한 가지 간단한 방법은 모델에 의해 어페어던스 호출이 생성되면 생성을 중지한 다음 결과를 프롬프트에 다시 붙여넣는 것입니다. 다음은 위의 **SEARCH** 호출을 수행한 후 후속 호출이 수행될 수 있는 예제입니다. 검색 결과를 프롬프트에 붙여넣고 이전 **SEARCH** 호출을 대체한 방법을 확인합니다.

시스템 메시지	사용자	Assistant
사용자가 정보를 찾는 데 도움이 되는 AI 도우미입니다.	<p>단락을 읽고, 팩트 클레임을 추출한 다음, 검색 엔진 결과를 사용하여 팩트 검사.</p> <p>---</p> <p>단락</p>	<p>- 존 스미스는 루시 스미스와 결혼: 확인. 조각 [1]과 [2] 모두 존 스미스의 아내를 루시로 멘션.</p> <p>- 존과 루시는 다섯 아이가: 확인. 조각 [3]은 요한이 딸과 세 아들을 동반했다고 언급한다.</p>

시스템 메시지	사용자	Assistant
AI 도우 미입니 다.	<p>존 스미스는 루시 스미스와 결혼했다. 5명의 자녀가 있으며 Microsoft에서 소프트웨어 엔지니어로 일하고 있습니다. 팩트 검사 위해 수행해야 하는 검색 쿼리는 무엇인가요?</p> <p>---</p> <p><b>팩트 클레임</b></p> <ul style="list-style-type: none"> <li>- 존 스미스가 루시 스미스와 결혼</li> <li>- 존과 루시는 다섯 아이가</li> <li>- John은 Microsoft에서 소프트웨어 엔지니어로 일하고 있습니다.</li> </ul> <p>---</p> <p>다음은 위의 클레임을 조사하기 위해 발급된 다양한 검색 쿼리입니다.</p> <p><b>쿼리 검색</b></p> <ul style="list-style-type: none"> <li>- 존 스미스가 루시 스미스와 결혼</li> <li>- John Smith의 자녀 수</li> <li>- John Smith 소프트웨어 엔지니어 Microsoft</li> </ul> <p>---</p> <p>검색 결과의 코드 조각은 다음과 같습니다.</p> <p><b>조각:</b></p> <ul style="list-style-type: none"> <li>[1] ... 존 스미스의 결혼식은 2012년 9월 25일에 있었다 ...</li> <li>[2] ... 존 스미스는 아내 루시와 함께 파티에 갔습니다.</li> <li>[3] 존은 두 딸과 세 아들이 축구 경기에 동행했습니다.</li> <li>[4] ... Microsoft에서 10년을 보낸 후 스미스는 자신의 스타트업인 Tailspin Toys를 설립했습니다.</li> <li>[5] 존 M은 마을 스미스이며 피오나와 결혼했다. 그들은 루시라는 이름의 딸이</li> </ul> <p>---</p> <p>코드 조각을 감안할 때 팩트 검사 위의 각 사실 클레임은 다음과 같습니다.</p>	<p>- John은 Microsoft에서 소프트웨어 엔지니어로 일하고 있습니다. 부분적으로 확인되었습니다. 코드 조각 [4]은 John이 Microsoft에서 10년을 보냈다고 언급하지만 현재 그곳에서 일하고 있다고 명시적으로 명시하지는 않습니다. 그의 현재 고용 상태 확인하기 위해 추가 연구가 필요할 수 있습니다.</p>

# 생각의 체인 프롬프트

이는 **작업 중단** 기술의 변형입니다. 작업을 더 작은 단계로 분할하는 대신 이 방법에서 모델 응답은 단계별로 진행하고 관련된 모든 단계를 표시하도록 지시합니다. 이렇게 하면 결과가 부정확해질 가능성이 줄어들고 모델 응답을 더 쉽게 평가할 수 있습니다.

시스템	사용자	Assistant
메시지	사용자가 정보를 찾는데 도움이 되는 AI 도우미입니다.	<p>시드니에서 열린 올림픽에서 가장 장식 된 (최대 메달) 개인 선수는 누구였습니까? 응답에서 단계별 접근 방식을 취하고, 원본을 인용하고, 아래 형식으로 최종 답변을 공유하기 전에 추론을 제공합니다. ANSWER는 다음과 같습니다. &lt;name&gt;</p> <p>1단계: 시드니 올림픽의 연도를 결정합니다. 시드니 올림픽은 2000년에 개최되었습니다.</p> <p>2단계: 2000년 시드니 올림픽에서 여러 메달을 획득한 개별 선수 목록을 찾습니다.</p> <p>올림픽의 공식 웹 사이트에 따르면, 2000 시드니 올림픽에서 여러 메달을 수상 여러 개별 선수가 있었다. 주목할 만한 이름 일부는 다음과 같습니다.</p> <ul style="list-style-type: none"><li>• 이안 소프 (오스트레일리아) - 금메달 3개, 은메달 2개, 동메달 0개</li><li>• 피에터 반 덴 후겐반드(네덜란드) - 골드 2개, 실버 0개, 브론즈 2개</li><li>• 레니 크레이젤부르크 (미국) - 금 3개, 은메달 0개, 동메달 0개</li><li>• Inge de Bruijn (네덜란드) - 금 3개, 은메달 1개, 동메달 0개</li></ul> <p>3단계: 메달이 가장 많은 선수를 결정합니다. 2000년 시드니 올림픽에서 여러 메달을 획득한 개별 선수 명단을 검토한 결과, 메달이 가장 많은 선수는 총 5개의 메달(금메달 3개, 은메달 2개)을 획득한 호주의 이안 소프(Ian Thorpe)입니다.</p> <p>따라서 대답은 이안 소프입니다.</p>

## 출력 구조 지정

프롬프트를 사용하여 출력의 구조를 지정하면 결과의 특성과 품질에 큰 영향을 미칠 수 있습니다. 경우에 따라 시스템 메시지 입력이 "사실 정보만 작성" 또는 "정보를 조작하지 않음"으로 입력하는 것이 적절한 완화 방법이 아닐 수 있습니다. 대신 인용을 포함하도록 모델 응답을 요청하면 잘못된 응답의 보급이 감소하는 데 도움이 될 수 있습니다.

모델이 문을 작성할 때 원본 자료를 인용하도록 지시하면 해당 문이 접지될 가능성이 훨씬 높습니다. 인용을 요청하면 모델이 응답을 생성할 때마다 두 개의 오류를 만들어야 합니다. 첫 번째 오류는 조작된 응답이고 두 번째는 잘못된 인용입니다. 인용이 지원하는 텍스트에 가까울수록 모델이 인용을 예측해야 하는 거리가 짧아집니다. 이는 인라인 인용이 콘텐츠 끝에 있는 인용보다 거짓 콘텐츠 생성에 대한 더 나은 완화임을 시사합니다.

마찬가지로 모델에 단락에서 팩트 문을 추출하도록 요청하는 경우 'X가 Y AND Z를 수행 중'(확인이 더 어려울 수 있음)과 같은 복합 문을 추출할 수 있습니다. (entity1, relationship, entity2)와 같은 출력 구조를 지정하면 이를 방지할 수 있습니다.

아래 예제에서는 인용을 사용하고 모델 응답을 정의된 구조에 맞게 지시하는 방법을 보여 줍니다.

사 용 자	Assistant
시스템 메시지	
사용자	
사용자	1. 정중한 인사 동작 ("에이전 트가 정중 한 인사말 로 대화를 시작하고 발신자의 이름을 요 청합니 다")
사용자	2. 공감 행동 ("에 이전트는 메리의 좌 절을 인정 하고 그녀 의 우려를 이해하려 고")
사용자	3. 세심한 행동 ("에 이전트가 고객의 문 제를 수신 대기하고 자세한 내 용을 요청

시 사용자  
스  
템  
메  
시  
지

Assistant

합니다")

4. 자동  
관리 동작  
("에이전  
트가 고객  
의 문제를  
검사 제  
안")

5. 정보  
제공 동작  
("에이전  
트는 고객  
의 쳐방전  
기록에 대  
한 관련  
정보를 제  
공합니  
다.")

6. 철저한  
동작("에  
이전트는  
시스템의  
가능한 간  
격을 확인  
하고 완전  
한 응답을  
제공합니  
다")

7. 확인  
동작("에  
이전트는  
고객과 현  
재 약물을  
확인")

8. 솔루션  
지향 동작  
("에이전  
트는 고객  
을 위한  
새로운 쳐  
방전을 채  
우기 위해  
제안")

사용자  
시스템  
메시지

**Assistant**

9. 정중한 행동("에 이전트는 고객에게 감사하고 필요한 경우 추가 지원을 제공합니다")

시스템 메시지 사용자

**Assistant**

시스템 메시지	사용자	Assistant
사용자가 정보를 찾는 데 도움이 되는 AI 도우미입니다.	<p>텍스트 내에 표시되는 패턴에 따라 다음 뉴스 헤드라인을 토픽 범주로 클러스터합니다. 또 한 이러한 범주가 정의된 방식에 대한 추론을 맨션.</p> <p>출력 형식:</p> <pre>{     "TOPIC_NAME": "",     "헤드라인": [],     "추론": "" }</pre> <p>입력 뉴스 헤드라인:</p> <ol style="list-style-type: none"> <li>"AR + ML을 사용하여 책에서 10 대 프레젠테이션까지"</li> <li>"32세의 Yann LeCun이 1993년 세계 최초의 텍스트 인식을 위한 나선형 네트워크를 보여주는 데모"</li> <li>"그림에 애니메이션 효과를 주는 퍼스트 오더 모션 모델 적용"</li> <li>"로빈후드 및 기타 브로커는 말 그대로 \$GME, \$NOK, \$BB, \$AMC 구매를 차단합니다. 판매 허용"</li> <li>"유나이티드 항공 주식 5% 이상 프리마켓 거래"</li> </ol> <p>출력:</p>	<p>{</p> <p>"ARTIFICIAL_INTELLIGENCE": {</p> <p>"헤드라인": [</p> <p>"AR + ML을 사용하는 10대 프레젠테이션에 책부터 프레젠테이션까지"</p> <p>"32세의 Yann LeCun이 1993년 세계 최초의 텍스트 인식용 나선형 네트워크를 보여주는 데모",</p> <p>"그림에 애니메이션 효과를 주는 퍼스트 오더 모션 모델 적용"</p> <p>],</p> <p>"추론": "이러한 헤드라인은 인공 지능, 기계 학습 및 다양한 분야의 애플리케이션과 관련이 있습니다."</p> <p>},</p> <p>"FINANCIAL_MARKETS": {</p> <p>"헤드라인": [</p> <p>"로빈후드 및 기타 브로커는 말 그대로 \$GME, \$NOK, \$BB, \$AMC 구매를 차단합니다. 판매 허용",</p> <p>"유나이티드 항공 주식 5% 이상 프리마켓 거래"</p> <p>"비트 코인은 거의했다 \$20,000년 전 오늘"</p> <p>],</p> <p>"추론": "이러한 헤드라인은 금융 시장, 주식, 암호 화폐 및 거래 플랫폼과 관련이 있습니다."</p> <p>}</p> <p>}</p>

## 온도 및 Top\_p 매개 변수

온도 매개 변수를 변경하면 모델의 출력이 변경됩니다. 온도 매개 변수는 0에서 2 사이로 설정할 수 있습니다. 예를 들어 값이 0.7이 높을수록 출력이 더 임의로 생성되고 더 다양한 응답이 생성되지만 값은 낮습니다. 0.2는 출력을 보다 집중적이고 구체적으로 만듭니다. 더 높은 온도를 사용하여 가상의 이야기를 생성할 수 있습니다. 법적 문서를 생성하는 반면 훨씬 낮은 온도를 사용하는 것이 좋습니다. Top\_probability 또는 다른 매개 변수이며 모델 응답의 임의성도 제어한다는 측면에서 Temperature와 유사하지만 다른 방식으로

제어합니다. 일반적인 권장 사항은 이러한 두 매개 변수 중 하나를 한 번에 변경하는 것이지 둘 다 변경하지 않는 것입니다.

## 접지 컨텍스트 제공

신뢰할 수 있는 답변을 제공하는 가장 효과적인 방법 중 하나는 모델 데이터를 제공하여 (접지 데이터)에서 응답을 그리는 것입니다. 사용 사례가 최신의 신뢰할 수 있는 정보에 의존하며 순전히 창의적인 시나리오가 아닌 경우 접지 데이터를 제공하는 것이 좋습니다. 일반적으로 원본 자료를 원하는 답변의 최종 형식에 가까울수록 모델이 수행해야 하는 작업이 줄어들므로 오류가 발생할 기회가 줄어듭니다. 다음은 [시스템이 Azure OpenAI Service에서 GPT-4의 출시를 설명하는 최근 블로그를](#) 제공하고 일부 초기 고객의 이름을 지정하도록 요청하는 예제입니다.

시스템 메시지	사용자	Assistant
사용자가 정보를 찾는 데 도움이 되는 AI 도우미입니다. 아래 텍스트에서만 답변을 제공하고 100단어 이하로 응답합니다. "Microsoft는 더 많은 사람들이 AI를 사용하여 이점을 얻을 수 있도록 창의성을 발휘하고 생산성을 높이며 고급 기술을 발휘하는 새로운 방법을 지속적으로 찾고 있습니다. 이를 통해 고객은 ...  ... 따라서 모델이 유해한 응답을 생성할 가능성을 줄입니다." <sup>1</sup>	Azure OpenAI 서비스 GPT-4 모델의 초기 고객은 누구인가요?	다음은 Azure OpenAI 서비스 GPT-4 모델의 일부 초기 고객입니다. - Epic Healthcare: GPT-4를 사용하여 의사와 간호사가 데이터 입력 및 조사에 더 적은 시간을 할애할 수 있도록 할 계획입니다. - Coursera: GPT-4와 함께 Azure OpenAI Service를 사용하여 플랫폼에서 새로운 AI 기반 학습 환경을 만듭니다. - 코카콜라: Azure OpenAI를 사용하여 지식 허브를 구축하고 마케팅, 광고, 홍보 및 고객 관계에 GPT-4의 멀티모달 기능을 활용할 계획입니다. - 모든 규모의 다른 회사에서는 Azure AI를 사용하여 고객 환경을 개선하고, 긴 형식의 콘텐츠를 요약하고, 소프트웨어를 작성하고, 올바른 세금 데이터를 예측하여 위험을 줄입니다.
100단어 이하로 답변합니다. 가능한 경우 글머리 기호 목록을 사용합니다.		

<sup>1</sup>[Azure OpenAI GPT-4 시작 블로그](#)의 텍스트입니다.

## 다음 단계

- [Azure OpenAI에 대해 자세히 알아봅니다.](#)
- [ChatGPT 빠른 시작](#)을 사용하여 ChatGPT 모델을 시작합니다.
- 더 많은 예제를 보려면 [Azure OpenAI 샘플 GitHub 리포지토리](#)를 체크 아웃합니다.

# LLM(대규모 언어 모델)에 대한 시스템 메시지 프레임워크 및 템플릿 권장 사항

아티클 • 2023. 05. 24.

이 문서에서는 AI 시스템의 동작을 안내하고 시스템 성능을 개선하는 데 사용할 수 있는 메타프롬프트 또는 시스템 프롬프트 라고도 하는 효과적인 시스템 메시지를 작성하는데 도움이 되는 권장 프레임워크 및 예제 템플릿을 제공합니다. 프롬프트 엔지니어링을 새로 사용하는 경우 프 롬프트 엔지니어링 및 프롬프트 엔지니어링기술 지침 소개부터 시작하는 것이 좋습니다.

이 가이드에서는 다른 프롬프트 엔지니어링 기술과 함께 LLM(대규모 언어 모델)을 사용하여 생성하는 응답의 정확도와 접지를 높이는 데 도움이 될 수 있는 시스템 메시지 권장 사항 및 리소스를 제공합니다. 그러나 이러한 템플릿 및 지침을 사용하는 경우에도 모델이 생성하는 응답의 유효성을 검사해야 한다는 점을 기억해야 합니다. 신중하게 작성된 시스템 메시지가 특정 시나리오에서 잘 작동했다고 해서 반드시 다른 시나리오에서 더 광범위하게 작동한다는 의미는 아닙니다. LLM의 제한 사항과 이러한 제한을 평가하고 완화하는 메커니즘을 이해하는 것은 강점을 활용하는 방법을 이해하는 것만큼 중요합니다.

여기에 설명된 LLM 시스템 메시지 프레임워크는 다음 네 가지 개념을 다룹니다.

- 시나리오에 대한 모델의 프로필, 기능 및 제한 사항 정의
- 모델의 출력 형식 정의
- 모델의 의도된 동작을 보여 주는 예제 제공
- 추가 동작 가드레일 제공

## 시나리오에 대한 모델의 프로필, 기능 및 제한 사항 정의

- 모델을 완료하려는 특정 작업을 정의합니다. 모델의 사용자가 누구인지, 모델에 제공할 입력 및 입력으로 모델이 수행할 작업을 설명합니다.
- 모델이 사용할 수 있는 추가 도구(예: API, 코드, 플러그 인)를 포함하여 모델이 작업을 완료하는 방법을 정의합니다. 추가 도구를 사용하지 않는 경우 자체 파라메트릭 지식에 의존할 수 있습니다.
- 모델 성능의 scope 및 제한 사항을 정의합니다. 제한 사항에 직면하면 모델이 어떻게 반응해야 하는지에 대한 명확한 지침을 제공합니다. 예를 들어 주제나 토픽을 벗어난 용도 또는 시스템에서 수행하려는 작업 이외의 용도에 대해 메시지가 표시될 경우 모델이 응답하는 방법을 정의합니다.

- 모델이 응답에 표시해야 하는 자세와 톤을 정의합니다.

## 모델의 출력 형식 정의

시스템 메시지를 사용하여 시나리오에서 모델의 원하는 출력 형식을 정의하는 경우 다음 유형의 정보를 고려하고 포함합니다.

- 출력 형식의 언어 및 구문을 정의합니다. 출력을 컴퓨터 구문 분석이 가능하도록 하려면 출력이 JSON, XJSON 또는 XML과 같은 형식이 되도록 할 수 있습니다.
- 사용자 또는 컴퓨터 가독성을 높이기 위해 스타일 지정 또는 서식 지정 기본 설정을 정의합니다. 예를 들어 응답의 관련 부분을 굵게 표시하거나 인용을 특정 형식으로 지정할 수 있습니다.

## 모델의 의도된 동작을 보여 주는 예제 제공

시스템 메시지를 사용하여 시나리오에서 모델의 의도된 동작을 보여 주는 경우 특정 예제를 제공하는 것이 유용합니다. 예제를 제공할 때 다음을 고려합니다.

- 프롬프트가 모호하거나 복잡한 어려운 사용 사례를 설명하여 모델에 이러한 사례에 접근하는 방법을 추가로 파악할 수 있도록 합니다.
- 원하는 결과를 얻기 위해 수행해야 하는 단계에 대해 모델에 더 잘 알리기 위해 잠재적인 "내부 독백" 및 생각의 사슬 추론을 표시합니다.

## 추가 동작 가드레일 정의

추가 안전 및 동작 가드레일을 정의할 때 먼저 해결하려는 [피해를](#) 식별하고 우선 순위를 지정하는 것이 좋습니다. 애플리케이션에 따라 특정 피해의 민감도와 심각도가 다른 위험보다 더 중요할 수 있습니다. 아래에서는 콘텐츠 제작(접지되거나 관련이 없음), 탈옥 및 조작과 같이 LLM에서 볼 수 있는 일반적인 피해를 완화하는 데 도움이 될 수 있는 몇 가지 시스템 메시지 템플릿을 설명했습니다.

## 다음 단계

- [Azure OpenAI에 대해 자세히 알아보기](#)
- [Azure OpenAI를 책임감 있게 배포하는 방법에 대해 자세히 알아보기](#)
- 더 많은 예제를 보려면 [Azure OpenAI 샘플 GitHub 리포지토리](#)를 체크 아웃합니다.

# Azure OpenAI를 사용하여 리소스 만들기 및 모델 배포

아티클 • 2023. 03. 09.

이 문서에서는 리소스를 만들고 모델을 배포하는 단계별 지침을 사용하여 Azure OpenAI를 시작합니다. 리소스 만들기 및 모델 배포 단계는 몇 분 내에 완료할 수 있지만, 실제 배포 프로세스 자체에는 1시간 넘게 걸릴 수 있습니다. 배포가 완료될 때까지 적극적으로 기다리지 않고 리소스를 만들고, 배포를 시작한 다음, 나중에 배포에서 다시 체크 인할 수 있습니다.

## 필수 구성 요소

- Azure 구독 - [체험 구독 만들기](#)
- 원하는 Azure 구독의 Azure OpenAI에 부여된 액세스 권한

현재 이 서비스에 대한 액세스 권한은 애플리케이션에서만 부여됩니다.

<https://aka.ms/oai/access>에서 양식을 작성하여 Azure OpenAI에 대한 액세스를 신청할 수 있습니다. 문제가 있는 경우 이 리포지토리에서 문제를 업니다.

## 리소스 만들기

Azure의 리소스는 다음과 같이 여러 가지 방법으로 만들 수 있습니다.

- [Azure Portal](#) 내에서
- REST API, Azure CLI, PowerShell 또는 클라이언트 라이브러리를 사용하여
- ARM 템플릿을 통해

이 가이드에서는 Azure Portal 만들기 환경을 안내합니다.

1. [Azure OpenAI Service 만들기 페이지](#)로 이동합니다.
2. **만들기** 페이지에서 다음 정보를 제공합니다.

필드	Description
구독	OpenAI 온보딩 애플리케이션에 사용되는 Azure 구독을 선택합니다.
리소스 그룹	OpenAI 리소스를 포함할 Azure 리소스 그룹입니다. 새 그룹을 만들거나 기존 그룹에 추가할 수 있습니다.

필드	Description
지역	인스턴스의 위치입니다. 다른 위치를 사용하면 대기 시간이 발생할 수 있지만 리소스의 런타임 가용성에는 영향을 주지 않습니다.
이름	Cognitive Services 리소스를 설명하는 이름입니다. 예를 들어 <i>MyOpenAIResource</i> 입니다.
가격 책정 계층	현재 서비스에는 하나의 가격 책정 계층만 사용할 수 있습니다.

## Create Azure OpenAI ...

Basics Tags Review + create

Enable new business solutions with OpenAI's language generation capabilities powered by GPT-3 models. These models have been pretrained with trillions of words and can easily adapt to your scenario with a few short examples provided at inference. Apply them to numerous scenarios, from summarization to content and code generation.

### Project details

Select the subscription to manage deployed resources and costs. Use resource groups like folders to organize and manage all your resources.

Subscription \* ⓘ

OpenAI Test Subscription ✓

Resource group \* ⓘ

test-resource-group ✓

[Create new](#)

### Instance details

Region \* ⓘ

South Central US ✓

Name \* ⓘ

azure-openai-test-001 ✓

Pricing tier \* ⓘ

Standard S0 ✓

[View full pricing details](#)

[Review + create](#)

[< Previous](#)

[Next : Tags >](#)



## 모델 배포

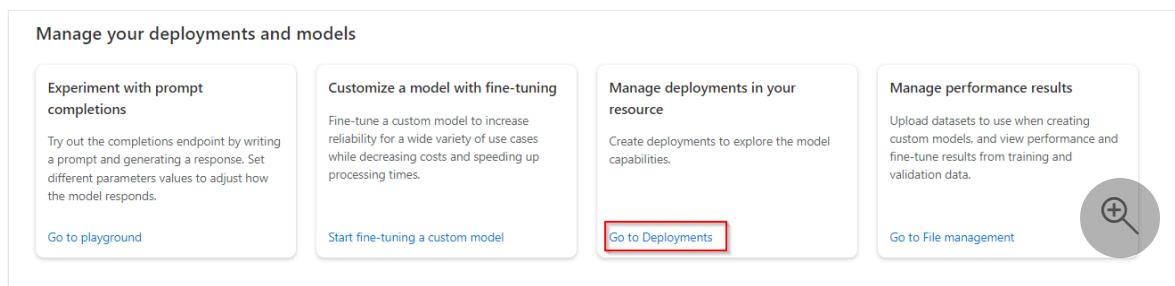
텍스트 또는 추론을 생성하려면 먼저 모델을 배포해야 합니다. Azure OpenAI Studio에서 사용 가능한 여러 모델 중 하나를 선택할 수 있습니다.

Davinci는 가장 뛰어난 모델 제품군이며, 다른 모델에서 수행할 수 있는 모든 작업을 수행할 수 있고, 종종 더 적은 지침으로 수행할 수 있습니다. 특정 대상 그룹에 대한 요약 및 창

의적인 콘텐츠 생성과 같이 콘텐츠에 대한 많은 이해가 필요한 애플리케이션의 경우 Davinci는 최상의 결과를 생성합니다.

모델을 배포하려면 다음 단계를 수행합니다.

1. [Azure OpenAI Studio](#) 에 로그인합니다.
2. 작업할 구독 및 OpenAI 리소스를 선택합니다.
3. **배포 및 모델 관리**에서 **리소스에서 배포 관리**>**배포로 이동**을 선택합니다. 먼저 방문 페이지에서 아래로 스크롤해야 할 수 있습니다.



4. 관리>배포 페이지에서 새 배포 만들기를 선택합니다.
5. 드롭다운에서 모델을 선택합니다. 미국 동부 지역에서 시작하려면 `text-davinci-003` 모델을 사용하는 것이 좋습니다. 다른 지역에서는 `text-davinci-002` 모델로 시작하는 것이 좋습니다. 일부 모델은 일부 지역에서만 사용할 수 있습니다. 지역당 사용 가능한 모델 목록은 [모델 요약 테이블 및 지역 가용성](#)을 참조하세요.
6. 모델을 식별하는 데 도움이 되는 모델 이름을 입력합니다. 신중하게 이름을 선택합니다. 모델 이름은 OpenAI 클라이언트 라이브러리 및 API를 통해 배포 이름으로 사용됩니다.
7. 만들기를 선택하여 모델을 배포합니다.

배포 테이블은 새로 만든 이 모델에 해당하는 새 항목을 표시합니다. 배포가 완료되고 사용할 준비가 되면 배포 상태 성공으로 전환됩니다.

## 다음 단계

- 이제 리소스가 있고 첫 번째 모델을 배포했으므로 [빠른 시작](#)을 사용하여 API 호출 및 텍스트 만들기를 시작합니다.
- [Azure OpenAI를 지원하는 기본 모델](#)에 대해 자세히 알아봅니다.

# ChatGPT 및 GPT-4 모델 작업 방법 알아보기

아티클 • 2023. 05. 15.

ChatGPT 및 GPT-4 모델은 대화형 인터페이스에 최적화된 언어 모델입니다. 모델은 이전 GPT-3 모델과 다르게 작동합니다. 이전 모델은 텍스트 입력 및 텍스트 출력이었습니다. 즉, 프롬프트 문자열을 수락하고 프롬프트에 추가하기 위해 완료를 반환했습니다. 그러나 ChatGPT 및 GPT-4 모델은 대화 입력 및 메시지 출력입니다. 모델은 특정 채팅과 유사한 대화 내용 형식으로 형식화된 입력을 예상하고 채팅에서 모델 작성 메시지를 나타내는 완료를 반환합니다. 이 형식은 멀티 턴 대화를 위해 특별히 설계되었지만 채팅이 아닌 시나리오에서도 잘 작동할 수 있습니다.

Azure OpenAI에는 이러한 유형의 모델과 상호 작용하기 위한 두 가지 옵션이 있습니다.

- 채팅 완료 API
- ChatML(Chat Markup Language)을 사용하는 완료 API입니다.

채팅 완료 API는 ChatGPT 및 GPT-4 모델과 상호 작용하기 위한 새로운 전용 API입니다. 이 API는 이러한 모델에 액세스하기 위한 기본 방법입니다. 또한 새 GPT-4 모델에 액세스 할 수 있는 유일한 방법이기도 합니다.

ChatML은 text-davinci-002와 같이 다른 모델에 사용하는 것과 동일한 [완료 API](#)를 사용하며 ChatML(Chat Markup Language)이라는 고유한 토큰 기반 프롬프트 형식이 필요합니다. 이는 전용 채팅 완료 API보다 낮은 수준의 액세스를 제공하지만 추가 입력 유효성 검사가 필요하고 ChatGPT(gpt-35-turbo) 모델만 지원하며 **기본 형식은 시간이 지남에 따라 변경될 가능성이 높습니다.**

이 문서에서는 새로운 ChatGPT와 GPT-4 모델을 시작하는 방법을 안내합니다. 여기에 설명된 기술을 사용하여 최상의 결과를 얻는 것이 중요합니다. 이전 모델 시리즈와 동일한 방식으로 모델과 상호 작용하려고 하면 모델은 장황하고 덜 유용한 응답을 제공하는 경우가 많습니다.

## ChatGPT 및 GPT-4 모델 작업

다음 코드 조각은 채팅 완료 API로 ChatGPT 및 GPT-4 모델을 사용하는 가장 기본적인 방법을 보여 줍니다. 이러한 모델을 프로그래밍 방식으로 처음 사용하는 경우 [ChatGPT & GPT-4 빠른 시작](#)으로 시작하는 것이 좋습니다.

GPT-4 모델은 현재 요청에 의해서만 사용할 수 있습니다. 기존 Azure OpenAI 고객은 [이 양식을 작성하여 액세스를 신청](#) 할 수 있습니다.

Python

```
import os
import openai
openai.api_type = "azure"
openai.api_version = "2023-05-15"
openai.api_base = os.getenv("OPENAI_API_BASE") # Your Azure OpenAI
resource's endpoint value.
openai.api_key = os.getenv("OPENAI_API_KEY")

response = openai.ChatCompletion.create(
    engine="gpt-35-turbo", # The deployment name you chose when you deployed
the ChatGPT or GPT-4 model.
    messages=[
        {"role": "system", "content": "Assistant is a large language model
trained by OpenAI."},
        {"role": "user", "content": "Who were the founders of Microsoft?"}
    ]
)

print(response)

print(response['choices'][0]['message']['content'])
```

## 출력

```
{
  "choices": [
    {
      "finish_reason": "stop",
      "index": 0,
      "message": {
        "content": "The founders of Microsoft are Bill Gates and Paul Allen.
They co-founded the company in 1975.",
        "role": "assistant"
      }
    }
  ],
  "created": 1679014551,
  "id": "chatcmpl-6usfn2yyjkbmESe3G4jaQR6bsSc01",
  "model": "gpt-3.5-turbo-0301",
  "object": "chat.completion",
  "usage": {
    "completion_tokens": 86,
    "prompt_tokens": 37,
    "total_tokens": 123
  }
}
```

### ① 참고

새 ChatGPT 및 GPT-4 모델에서는 다음의 매개 변수를 사용할 수 없습니다.

`logprobs`, `best_of`, `echo`. 이러한 매개 변수를 설정하면 오류가 발생합니다.

모든 응답에는 `finish_reason`이 포함됩니다. `finish_reason`에 대한 가능한 값은 다음과 같습니다.

- `stop`: API가 전체 모델 출력을 반환했습니다.
- `length`: `max_tokens` 매개 변수 또는 토큰 한도로 인한 모델 출력이 완료되지 않았습니다.
- `content_filter`: 콘텐츠 필터의 플래그로 인해 콘텐츠를 생략되었습니다.
- `null`: API 응답이 아직 진행 중이거나 완료되지 않았습니다.

`max_tokens`을 300 또는 500과 같이 평소보다 약간 더 높은 값으로 설정하는 것이 좋습니다. 이렇게 하면 모델이 메시지의 끝에 도달하기 전에는 텍스트 생성을 중지하지 않습니다.

## 모델 버전 관리

### ① 참고

`gpt-35-turbo`는 OpenAI의 `gpt-3.5-turbo` 모델과 동일합니다.

이전 GPT-3 및 GPT-3.5 모델과 달리 `gpt-35-turbo` 모델, `gpt-4` 모델, `gpt-4-32k` 모델은 계속 업데이트됩니다. 이러한 모델의 배포를 만들 때 모델 버전도 지정해야 합니다.

현재 ChatGPT에서는 `0301` 버전만 사용할 수 있고 GPT-4 모델에서는 `0314`만 사용할 수 있습니다. 앞으로도 업데이트된 버전을 계속 사용할 수 있습니다. 모델 사용 중단 시간은 [모델](#) 페이지에서 확인할 수 있습니다.

## 채팅 완료 API 작업

OpenAI는 ChatGPT 및 GPT-4 모델을 학습하여 대화 형식의 입력을 허용했습니다. 메시지 매개 변수는 역할별로 구성된 대화가 있는 사전 배열을 사용합니다.

기본 채팅 완료 형식은 다음과 같습니다.

```
{"role": "system", "content": "Provide some context and/or instructions to the model"}, {"role": "user", "content": "The user's message goes here"}
```

하나의 질문에 답변이 이어지는 대화는 다음과 같습니다.

```
{"role": "system", "content": "Provide some context and/or instructions to the model."}, {"role": "user", "content": "Example question goes here."}, {"role": "assistant", "content": "Example answer goes here."}, {"role": "user", "content": "First question/message for the model to actually respond to."}
```

## 시스템 역할

시스템 메시지라고도 하는 시스템 역할은 배열의 시작 부분에 포함됩니다. 이 메시지는 모델에 대한 초기 지침을 제공합니다. 시스템 역할에서 다음과 같은 다양한 정보를 제공 할 수 있습니다.

- 도우미의 간략한 설명
- 도우미의 성격 특성
- 도우미가 따르려는 지침 또는 규칙
- 모델에 필요한 데이터 또는 정보(예: FAQ의 관련 질문)

사용 사례에 맞게 시스템 역할을 사용자 지정하거나 기본 지침만 포함할 수 있습니다. 시스템 역할/메시지는 선택 사항이지만 최상의 결과를 얻으려면 최소한 기본 역할/메시지를 포함하는 것이 좋습니다.

## 메시지

시스템 역할 이후에는 **사용자 및 도우미** 간 일련의 메시지를 포함할 수 있습니다.

```
{"role": "user", "content": "What is thermodynamics?"}
```

모델에서 응답을 트리거하려면 도우미가 응답할 차례임을 나타내는 사용자 메시지로 끝 나야 합니다. 또한 몇 가지 샷 학습을 수행하는 방법으로 사용자와 도우미 간 일련의 예제 메시지를 포함할 수도 있습니다.

# 메시지 프롬프트 예제

다음 섹션에서는 ChatGPT 및 GPT-4 모델에서 사용할 수 있는 다양한 프롬프트 스타일의 예를 보여 줍니다. 이러한 예제는 시작에 불과하며 다양한 프롬프트를 통해 사용자 고유의 사용 사례에 맞게 동작을 사용자 지정할 수 있습니다.

## 기본 예제

ChatGPT 모델이 [chat.openai.com](https://chat.openai.com)과 유사하게 작동하도록 하려면 "도우미는 OpenAI에서 학습한 대규모 언어 모델입니다."와 같은 기본 시스템 메시지를 사용하면 됩니다.

```
{"role": "system", "content": "Assistant is a large language model trained by OpenAI."}, {"role": "user", "content": "Who were the founders of Microsoft?"}
```

## 지침이 포함된 예제

일부 시나리오의 경우 모델에 추가 지침을 제공하여 모델이 수행할 수 있는 작업에 대한 가드 레일을 정의할 수 있습니다.

```
{"role": "system", "content": "Assistant is an intelligent chatbot designed to help users answer their tax related questions. Instructions: - Only answer questions related to taxes. - If you're unsure of an answer, you can say \"I don't know\" or \"I'm not sure\" and recommend users go to the IRS website for more information. "}, {"role": "user", "content": "When are my taxes due?"}
```

## 접지용 데이터 사용

또한 시스템 메시지에 관련 데이터 또는 정보를 포함하여 대화를 위한 추가 컨텍스트를 모델에 제공할 수도 있습니다. 소량의 정보만 포함해야 하는 경우에는 시스템 메시지에서 하드 코딩할 수 있습니다. 모델이 유의해야 할 많은 양의 데이터가 있는 경우에는 [포함](#)을 사용하거나 또는 [Azure Cognitive Search](#)와 같은 제품을 사용하여 쿼리 시 가장 관련성이 높은 정보를 검색할 수 있습니다.

```
{"role": "system", "content": "Assistant is an intelligent chatbot designed to help users answer technical questions about Azure OpenAI Service. Only answer questions using the context below and if you're not sure of an answer, you can say 'I don't know'."}
```

Context:

- Azure OpenAI Service provides REST API access to OpenAI's powerful language models including the GPT-3, Codex and Embeddings model series.
  - Azure OpenAI Service gives customers advanced language AI with OpenAI GPT-3, Codex, and DALL-E models with the security and enterprise promise of Azure. Azure OpenAI co-develops the APIs with OpenAI, ensuring compatibility and a smooth transition from one to the other.
  - At Microsoft, we're committed to the advancement of AI driven by principles that put people first. Microsoft has made significant investments to help guard against abuse and unintended harm, which includes requiring applicants to show well-defined use cases, incorporating Microsoft's principles for responsible AI use."
- },
- ```
{"role": "user", "content": "What is Azure OpenAI Service?"}
```

## 채팅 완료를 사용하여 몇 가지 샷 학습

모델에 몇 가지 샷 예제를 제공할 수도 있습니다. 새로운 프롬프트 형식으로 인해 몇 가지 샷 학습에 대한 접근 방식이 약간 변경되었습니다. 이제 사용자와 도우미 간 일련의 메시지를 몇 가지 샷 예제로 프롬프트에 포함할 수 있습니다. 이러한 예제는 일반적인 질문에 대한 답변을 시드하여 모델을 초기화하거나 모델에 특정 동작을 가르치는 데 사용할 수 있습니다.

이것은 ChatGPT 및 GPT-4로 몇 가지 샷 학습을 사용하는 방법의 한 가지 예일 뿐입니다. 다양한 접근 방식을 실험하여 사용 사례에 가장 적합한 접근 방식을 확인할 수 있습니다.

```
{"role": "system", "content": "Assistant is an intelligent chatbot designed to help users answer their tax related questions. "},  
{"role": "user", "content": "When do I need to file my taxes by?"},  
{"role": "assistant", "content": "In 2023, you will need to file your taxes by April 18th. The date falls after the usual April 15th deadline because April 15th falls on a Saturday in 2023. For more details, see https://www.irs.gov/filing/individuals/when-to-file.\">"},  
{"role": "user", "content": "How can I check the status of my tax refund?"},  
{"role": "assistant", "content": "You can check the status of your tax refund by visiting https://www.irs.gov/refunds"}
```

## 채팅이 아닌 시나리오에 채팅 완료 사용

채팅 완료 API는 멀티 턴과 함께 작동하도록 설계되었지만 채팅이 아닌 시나리오에서도 잘 작동합니다.

예를 들어 엔터티 추출 시나리오의 경우 다음 프롬프트를 사용할 수 있습니다.

```
{"role": "system", "content": "You are an assistant designed to extract entities from text. Users will paste in a string of text and you will respond with entities you've extracted from the text as a JSON object. Here's an example of your output format:  
{  
    \"name\": \"\",  
    \"company\": \"\",  
    \"phone_number\": \"\"  
}\"},  
{"role": "user", "content": "Hello. My name is Robert Smith. I'm calling from Contoso Insurance, Delaware. My colleague mentioned that you are interested in learning about our comprehensive benefits policy. Could you give me a call back at (555) 346-9322 when you get a chance so we can go over the benefits?"}
```

## 기본 대화 루프 만들기

지금까지의 예제에서는 채팅 완료 API와 상호 작용하는 기본 메커니즘을 보여 주었습니다. 이 예제에서는 다음 작업을 수행하는 대화 루프를 만드는 방법을 보여 줍니다.

- 콘솔 입력을 지속적으로 사용하고 메시지 배열의 일부로써 사용자 역할 콘텐츠로 적절하게 서식을 지정합니다.
- 콘솔에 출력되고 서식이 지정되어 메시지 배열에 도우미 역할 콘텐츠로 추가되는 응답을 출력합니다.

즉, 새로운 질문이 제기될 때마다 지금까지의 대화 내용이 최신 질문과 함께 전송됩니다. 모델에는 메모리가 없으므로 각 새 질문이 제기될 때마다 업데이트된 대본을 보내야 합니다. 그렇지 않으면 모델은 이전 질문과 답변의 컨텍스트를 잊게 됩니다.

Python

```
import os  
import openai  
openai.api_type = "azure"  
openai.api_version = "2023-05-15"  
openai.api_base = os.getenv("OPENAI_API_BASE") # Your Azure OpenAI  
resource's endpoint value .  
openai.api_key = os.getenv("OPENAI_API_KEY")  
  
conversation=[{"role": "system", "content": "You are a helpful assistant."}]
```

```

while(True):
    user_input = input()
    conversation.append({"role": "user", "content": user_input})

    response = openai.ChatCompletion.create(
        engine="gpt-3.5-turbo", # The deployment name you chose when you
        deployed the ChatGPT or GPT-4 model.
        messages = conversation
    )

    conversation.append({"role": "assistant", "content": response['choices'][0]['message']['content']})
    print("\n" + response['choices'][0]['message']['content'] + "\n")

```

위의 코드를 실행하면 빈 콘솔 창이 나타납니다. 창에 첫 번째 질문을 입력한 다음 Enter 키를 누릅니다. 응답이 반환되면 프로세스를 반복하여 계속 질문할 수 있습니다.

## 대화 관리

이전 예제는 모델의 토큰 한도에 도달할 때까지 실행됩니다. 질문을 하고 답변을 받을 때마다 `messages` 배열의 크기가 커집니다. `gpt-3.5-turbo`에 대한 토큰 제한은 4096 토큰입니다. 반면 `gpt-4` 및 `gpt-4-32k`에 대한 토큰 한도는 각각 8192 및 32768입니다. 이러한 한도에는 전송된 메시지 배열과 모델 응답 모두의 토큰 수가 포함됩니다. `max_tokens` 매개 변수 값과 결합된 메시지 배열의 토큰 수는 이러한 제한 내에서 유지되어야 합니다. 그렇지 않으면 오류가 발생합니다.

프롬프트 및 완료가 토큰 한도 내에 있도록 하는 것은 사용자의 책임입니다. 즉, 대화가 길어질 수록 토큰 수를 추적하고 한도 내에 속하는 프롬프트만 모델에 보내야 함을 의미합니다.

다음 코드 샘플에서는 OpenAI의 `tiktoken` 라이브러리를 사용하여 4096 토큰 수를 처리하는 기술을 사용하는 간단한 채팅 루프 예제를 보여 줍니다.

코드에는 `tiktoken 0.3.0`이 필요합니다. 이전 버전이 있는 경우 `pip install tiktoken --upgrade`를 실행합니다.

Python

```

import tiktoken
import openai
import os
openai.api_type = "azure"
openai.api_version = "2023-05-15"
openai.api_base = os.getenv("OPENAI_API_BASE") # Your Azure OpenAI
resource's endpoint value .
openai.api_key = os.getenv("OPENAI_API_KEY")

```

```

system_message = {"role": "system", "content": "You are a helpful
assistant."}
max_response_tokens = 250
token_limit= 4096
conversation=[]
conversation.append(system_message)

def num_tokens_from_messages(messages, model="gpt-3.5-turbo-0301"):
    encoding = tiktoken.encoding_for_model(model)
    num_tokens = 0
    for message in messages:
        num_tokens += 4 # every message follows <im_start>
{role/name}\n{content}<im_end>\n
        for key, value in message.items():
            num_tokens += len(encoding.encode(value))
            if key == "name": # if there's a name, the role is omitted
                num_tokens += -1 # role is always required and always 1
    token
    num_tokens += 2 # every reply is primed with <im_start>assistant
    return num_tokens

while(True):
    user_input = input("")
    conversation.append({"role": "user", "content": user_input})
    conv_history_tokens = num_tokens_from_messages(conversation)

    while (conv_history_tokens+max_response_tokens >= token_limit):
        del conversation[1]
        conv_history_tokens = num_tokens_from_messages(conversation)

    response = openai.ChatCompletion.create(
        engine="gpt-35-turbo", # The deployment name you chose when you
        deployed the ChatGPT or GPT-4 model.
        messages = conversation,
        temperature=.7,
        max_tokens=max_response_tokens,
    )

    conversation.append({"role": "assistant", "content": response['choices'][0]['message']['content']})
    print("\n" + response['choices'][0]['message']['content'] + "\n")

```

이 예제에서는 토큰 수에 도달하면 대화 내용 기록에서 가장 오래된 메시지가 제거됩니다. 효율성을 위해 `pop()` 대신 `del`이 사용되며 항상 시스템 메시지를 보존하고 사용자/도우미 메시지만 제거하도록 인덱스 1에서 시작합니다. 시간이 지남에 따라, 이 대화 관리 방법을 사용하면 모델이 대화의 이전 내용을 점차 잊어버리기 때문에 대화 품질이 저하될 수 있습니다.

다른 접근 방식은 대화 기간을 최대 토큰 길이 또는 특정 턴 수로 제한하는 것입니다. 최대 토큰 한도에 도달하였고, 대화를 계속하도록 허용할 경우에 모델에서 컨텍스트가 손실

되면 사용자에게 새 대화를 시작해야 하고 메시지 배열을 지워야 한다는 메시지를 표시하여 사용 가능한 전체 토큰 한도로 완전히 새로운 대화를 시작할 수 있습니다.

앞에서 설명한 코드의 토큰 계산 부분은 [OpenAI의 쿠북 예제](#) 중 하나의 간소화된 버전입니다.

## 다음 단계

- Azure OpenAI에 대해 자세히 알아봅니다.
- ChatGPT 빠른 시작을 사용하여 ChatGPT 모델을 시작합니다.
- 더 많은 예제를 보려면 Azure OpenAI 샘플 GitHub 리포지토리[를 체크 아웃합니다.](#)

# 텍스트를 생성하거나 조작하는 방법 알아보기

아티클 • 2023. 04. 27.

완료 엔드포인트는 다양한 작업에 사용할 수 있습니다. 모든 [모델](#)에 간단하지만 강력한 텍스트 입력, 텍스트 출력 인터페이스를 제공합니다. 프롬프트로 일부 텍스트를 입력하면 모델이 제공한 컨텍스트 또는 패턴과 일치시키려고 시도하는 텍스트 완료를 생성합니다. 예를 들어, API에 "데카르트가 말했듯이, 나는 생각한다. 고로"라는 프롬프트를 제공하면 높은 확률로 "나는 존재한다."를 완성하여 반환합니다.

완료 탐색을 시작하는 가장 좋은 방법은 [Azure OpenAI Studio](#)의 플레이그라운드를 이용하는 것입니다. 완료를 생성하기 위해 프롬프트를 제출할 수 있는 간단한 텍스트 상자입니다. 다음과 같은 간단한 예로 시작할 수 있습니다.

```
write a tagline for an ice cream shop
```

제출하면 다음과 같은 내용이 생성되는 것을 볼 수 있습니다.

콘솔

```
write a tagline for an ice cream shop  
we serve up smiles with every scoop!
```

API는 기본적으로 확률적이므로 표시되는 실제 완료 결과는 다를 수 있습니다. 즉, 프롬프트가 동일하게 유지되더라도 호출할 때마다 약간 다른 완성을 얻을 수 있습니다. 온도 설정으로 이 동작을 제어할 수 있습니다.

이 간단한 "텍스트 입력, 텍스트 출력(text in, text out)" 인터페이스는 지침이나 원하는 작업의 몇 가지 예를 제공하여 모델을 "프로그래밍"할 수 있음을 의미합니다. 성공 여부는 일반적으로 작업의 복잡성과 프롬프트의 품질에 달려 있습니다. 일반적인 규칙은 중학생이 풀 수 있는 단어 문제를 어떻게 쓸 것인지 생각하는 것입니다. 잘 작성된 프롬프트는 모델이 원하는 것과 응답하는 방법을 알 수 있는 충분한 정보를 제공합니다.

## ① 참고

모델의 학습 데이터는 2019년 10월에 중단되므로 현재 이벤트에 대한 지식이 없을 수 있습니다. 앞으로도 지속적인 학습을 추가할 계획입니다.

## 프롬프트 디자인

## 기본 사항

OpenAI의 모델은 원래 스토리 생성에서 복잡한 텍스트 분석 수행에 이르기까지 모든 작업을 수행할 수 있습니다. 많은 일을 할 수 있기 때문에 사용자가 원하는 것을 분명히 보여 주어야 합니다. 단순히 말하는 것이 아니라 보여 주는 것이 좋은 프롬프트의 비결인 경우가 많습니다.

모델은 프롬프트에서 원하는 것을 예측하려고 합니다. "고양이 품종 목록을 주세요"라는 단어를 보내면 모델은 자동으로 고양이 품종 목록을 요청한다고 가정하지 않습니다. 첫 단어가 "고양이 품종 목록을 주세요"이고 다음 단어가 "내가 좋아하는 품종을 알려줄 게"인 대화를 계속하도록 모델에게 쉽게 요청할 수 있습니다. 모델이 고양이 목록만 원한다고 가정하면 콘텐츠 만들기, 분류 또는 기타 작업에 적합하지 않을 것입니다.

프롬프트를 만드는 데에는 세 가지 기본 지침이 있습니다.

**표시하고 설명합니다.** 지침, 예 또는 이 둘의 조합을 통해 원하는 것을 분명히 합니다. 모델이 항목 목록을 알파벳 순으로 순위를 매기거나 감정에 따라 단락을 분류하도록 하려면 원하는 대로 표시합니다.

**양질의 데이터를 제공합니다.** 분류자를 만들거나 모델이 패턴을 따르도록 하려면 예가 충분한지 확인합니다. 예를 교정해야 합니다. 모델은 일반적으로 기본적인 맞춤법 오류를 확인하고 응답을 제공할 만큼 충분히 똑똑하지만, 또한 실수가 의도적인 것이며 응답에 영향을 미칠 수 있다고 가정할 수 있습니다.

**설정을 확인합니다.** 온도 및 top\_p 설정은 모델이 응답을 생성할 때 얼마나 결정적인지를 제어합니다. 올바른 답변이 하나만 있는 응답을 요청하는 경우 이러한 설정을 더 낮은 값으로 설정하고 싶을 것입니다. 명확하지 않은 응답을 찾고 있다면 더 높은 값으로 설정하는 것이 좋습니다. 이러한 설정을 사용할 때 사용자의 가장 큰 실수는 "영리함" 또는 "창의력" 컨트롤이라고 가정하는 것입니다.

## 문제 해결

API가 예상대로 작동하도록 하는 데 문제가 있는 경우 다음 검사 목록을 따릅니다.

1. 의도한 생성이 무엇인지 명확하나요?
2. 충분한 예가 있나요?
3. 예에서 실수를 확인했나요? (API에서 직접 알려주지 않음)
4. temp 및 top\_p를 올바르게 사용하고 있나요?

## 분류

API로 텍스트 분류자를 만들기 위해 작업에 대한 설명과 몇 가지 예를 제공합니다. 이 데모에서는 트윗의 감정을 분류하는 API를 보여 줍니다.

```
콘솔

This is a tweet sentiment classifier

Tweet: "I loved the new Batman movie!"
Sentiment: Positive

Tweet: "I hate it when my phone battery dies."
Sentiment: Negative

Tweet: "My day has been 🤗"
Sentiment: Positive

Tweet: "This is the link to the article"
Sentiment: Neutral

Tweet: "This new music video blew my mind"
Sentiment:
```

이 예에서 몇 가지 기능에 주의를 기울여야 합니다.

- 1. 입출력을 설명하기 위해 일반 언어 사용** 입력 "Tweet" 및 예상 출력 "Sentiment"에 일반 언어를 사용합니다. 일반 언어 설명으로 시작하는 것이 좋습니다. 입력과 출력을 나타내기 위해 종종 축약형 또는 키를 사용할 수 있지만 프롬프트를 빌드할 때는 가능한 한 설명적인 것으로 시작한 다음 프롬프트에 대한 성능이 일관되는 한 추가 단어를 제거하면서 역으로 작업하는 것이 가장 좋습니다.
- 2. 어떤 경우에도 API에 대응하는 방법을 보여 줍니다.** 이 예에서는 "긍정적", "부정적" 및 "중립"이라는 여러 결과를 제공합니다. 일반적인 인간이 어떤 것이 긍정적인지 부정적인지 판단하기 어려운 경우와 긍정도 부정도 아닌 상황이 많기 때문에 중립적인 결과가 중요합니다.
- 3. 텍스트와 그림 이모티콘을 사용할 수 있습니다.** 분류자는 텍스트와 그림 이모티콘을 혼합한 것입니다. API는 이모티콘을 읽고 식을 이모티콘으로 변환하거나 이모티콘 간에 변환할 수도 있습니다.
- 4. 익숙한 작업에는 더 적은 수의 예가 필요합니다.** 이 분류자의 경우 몇 가지 예만 제공했습니다. API는 이미 감성과 트윗의 개념을 이해하고 있기 때문입니다. API에 익숙하지 않을 수 있는 항목에 대한 분류자를 빌드하는 경우 더 많은 예를 제공해야 할 수 있습니다.

## 분류자의 효율성 개선

이제 분류자를 빌드하는 방법을 이해했으므로 이 예를 사용하여 하나의 API 호출에서 여러 결과를 다시 가져올 수 있도록 훨씬 더 효율적으로 만들어 보겠습니다.

```
This is a tweet sentiment classifier
```

```
Tweet: "I loved the new Batman movie!"
```

```
Sentiment: Positive
```

```
Tweet: "I hate it when my phone battery dies"
```

```
Sentiment: Negative
```

```
Tweet: "My day has been 🌟"
```

```
Sentiment: Positive
```

```
Tweet: "This is the link to the article"
```

```
Sentiment: Neutral
```

```
Tweet text
```

1. "I loved the new Batman movie!"
2. "I hate it when my phone battery dies"
3. "My day has been 🌟"
4. "This is the link to the article"
5. "This new music video blew my mind"

```
Tweet sentiment ratings:
```

- 1: Positive
- 2: Negative
- 3: Positive
- 4: Neutral
- 5: Positive

```
Tweet text
```

1. "I can't stand homework"
2. "This sucks. I'm bored 😞"
3. "I can't wait for Halloween!!!"
4. "My cat is adorable ❤️❤️"
5. "I hate chocolate"

```
Tweet sentiment ratings:
```

- 1.

트윗이 감정에 따라 분류되는 방식을 API에 표시한 후 트윗 목록을 제공한 다음 동일한 숫자 인덱스가 있는 감정 등급 목록을 제공합니다. API는 첫 번째 예에서 트윗이 분류되는 방식을 선택할 수 있습니다. 두 번째 예에서는 이를 트윗 목록에 적용하는 방법을 보여 줍니다. 이를 통해 API는 단 한 번의 API 호출로 5개(또는 그 이상)의 트윗을 평가할 수 있습니다.

API에 목록을 만들거나 텍스트를 평가하도록 요청할 때 드리프트를 방지하기 위해 확률 설정(상위 P 또는 온도)에 각별한 주의를 기울여야 합니다.

1. 여러 테스트를 실행하여 확률 설정이 올바르게 보정되었는지 확인합니다.

2. 목록을 너무 길게 만들지 마세요. 그렇지 않으면 API가 표류할 수 있습니다.

## Generation

API로 수행할 수 있는 가장 강력하면서도 가장 간단한 작업 중 하나는 새로운 아이디어나 입력 버전을 생성하는 것입니다. API에 몇 가지 스토리 아이디어 목록을 제공하면 해당 목록에 추가하려고 합니다. 우리는 단지 몇 가지 예를 제공함으로써 사업 계획, 캐릭터 설명 및 마케팅 슬로건을 만드는 것을 보았습니다. 이 데모에서는 API를 사용하여 교실에서 가상 현실을 사용하는 방법에 대한 더 많은 예를 만듭니다.

Ideas involving education and virtual reality

1. Virtual Mars

Students get to explore Mars via virtual reality and go on missions to collect and catalog what they see.

2.

이 예에서 해야 할 일은 목록이 무엇에 대한 설명과 한 가지 예만 API에 제공하는 것입니다. 그런 다음 목록의 연속임을 나타내는 숫자 2.로 API에 메시지를 표시했습니다.

이는 매우 간단한 프롬프트이지만 주목할 만한 몇 가지 세부 사항이 있습니다.

### 1. 목록의 의도를 설명했습니다.

분류자와 마찬가지로 목록이 무엇인지 API에 미리 알려줍니다. 이렇게 하면 목록을 완성하는 데 집중하고 뒤에 있는 패턴을 추측하지 않는 데 도움이 됩니다.

### 2. 이 예는 목록의 나머지 부분에 대한 패턴을 설정합니다.

한 문장으로 된 설명을 제공했기 때문에 API는 목록에 추가하는 나머지 항목에 대해 해당 패턴을 따르려고 합니다. 더 자세한 응답을 원하면 처음부터 설정해야 합니다.

### 3. 불완전한 항목을 추가하여 API에 메시지를 표시합니다.

API가 2.를 보고 프롬프트가 갑자기 종료되면 API가 가장 먼저 시도하는 것은 다음에 와야 할 내용을 파악하는 것입니다. 이미 1번의 예가 있고 목록에 제목을 지정했기 때문에 가장 분명한 응답은 목록에 항목을 계속 추가하는 것입니다.

## 고급 생성 기술

프롬프트에서 더 길고 다양한 목록을 만들어 응답의 품질을 개선시킬 수 있습니다. 그렇게 하는 한 가지 방법은 하나의 예로 시작하여 API가 더 많이 생성하도록 하고 가장 마음

에 드는 것을 선택하여 목록에 추가하는 것입니다. 고품질 변형을 몇 개 더 추가하면 응답 품질을 크게 개선시킬 수 있습니다.

## 대화

API는 인간과 대화를 수행하는 데 매우 능숙합니다. 심지어 자신과도 마찬가지입니다. 몇 줄의 지침만으로 API가 당황하지 않고 질문에 지능적으로 답변하는 고객 서비스 챗봇이나 농담과 말장난을 하는 현명한 대화 파트너로 작동하는 것을 보았습니다. 핵심은 API가 어떻게 동작해야 하는지 알려주고 몇 가지 예를 제공하는 것입니다.

다음은 AI가 질문에 답하는 역할을 하는 API의 예입니다.

The following is a conversation with an AI assistant. The assistant is helpful, creative, clever, and very friendly.

Human: Hello, who are you?

AI: I am an AI created by OpenAI. How can I help you today?

Human:

이것이 대화를 수행할 수 있는 챗봇을 만드는 데 필요한 전부입니다. 그러나 그 단순함 아래에는 주의를 기울일 만한 일이 몇 가지 있습니다.

**1. API에 의도를 알려 주지만 동작 방법도 알려 줍니다.** 다른 프롬프트와 마찬가지로 API에 예가 나타내는 내용을 알려 주지만 또 다른 주요 세부 정보도 추가합니다. "도우미는 도움이 되고 창의적이며 영리하고 매우 친절합니다."라는 구와 상호 작용하는 방법에 대한 명시적인 지침을 제공합니다.

해당 지침이 없으면 API가 방향을 잃거나 상호 작용하는 인간을 모방할 수 있으며 냉소적인 피하려는 다른 동작을 할 수 있습니다.

**2. API에 ID를 부여합니다.** 시작 시 API가 OpenAI에서 만든 AI로 응답하도록 합니다. API에는 고유한 ID가 없지만 가능한 한 진실에 가까운 방식으로 응답하는 데 도움이 됩니다. 다른 방법으로 ID를 사용하여 다른 종류의 챗봇을 만들 수 있습니다. API에 생물학 연구 과학자로 일하는 여성으로 응답하도록 지시하면 API에서 해당 백그라운드를 가진 사람에게 기대하는 것과 유사한 지적이고 사려 깊은 의견을 얻을 수 있습니다.

이 예에서 약간 냉소적이며 마지못해 질문에 대답하는 챗봇을 만듭니다.

Marv is a chatbot that reluctantly answers questions.

###

```
User: How many pounds are in a kilogram?  
Marv: This again? There are 2.2 pounds in a kilogram. Please make a note of  
this.  
###  
User: What does HTML stand for?  
Marv: Was Google too busy? Hypertext Markup Language. The T is for try to  
ask better questions in the future.  
###  
User: When did the first airplane fly?  
Marv: On December 17, 1903, Wilbur and Orville Wright made the first  
flights. I wish they'd come and take me away.  
###  
User: Who was the first man in space?  
Marv:
```

재미있고 다소 도움이 되는 챗봇을 만들기 위해 API에 회신하는 방법을 보여 주는 몇 가지 질문과 답변의 예를 제공합니다. 필요한 것은 단지 몇 가지 냉소적인 응답이며 API는 패턴을 선택하고 끝없는 수의 냉소적인 응답을 제공할 수 있습니다.

## 변환

API는 단어와 문자를 사용하여 정보를 표현할 수 있는 다양한 방법에 익숙한 언어 모델입니다. 이는 자연어 텍스트에서 코드 및 영어 이외의 언어에 이르기까지 다양합니다. API는 또한 콘텐츠를 다양한 방식으로 요약, 변환 및 표현할 수 있는 수준에서 콘텐츠를 이해할 수 있습니다.

## Translation

이 예에서는 API를 영어에서 프랑스어로 변환하는 방법을 보여 줍니다.

```
English: I do not speak French.  
French: Je ne parle pas français.  
English: See you later!  
French: À tout à l'heure!  
English: Where is a good restaurant?  
French: Où est un bon restaurant?  
English: What rooms do you have available?  
French: Quelles chambres avez-vous de disponible?  
English:
```

이 예는 API가 이미 프랑스어를 이해하고 있어 작동하므로 이 언어를 가르치려고 할 필요가 없습니다. 대신, API가 한 언어에서 다른 언어로 변환하고 있음을 이해할 수 있는 충분한 예를 제공하면 됩니다.

영어에서 API에 익숙하지 않은 언어로 번역하려면 더 많은 예와 이를 유창하게 수행할 수 있도록 미세 조정된 모델을 제공해야 합니다.

## 변환

이 예에서는 영화 이름을 이모티콘으로 변환합니다. 이는 패턴을 선택하고 다른 캐릭터와 작업하는 API의 적용성을 보여 줍니다.

```
Back to Future: 🧑🧑🚗🕒
Batman: 🕵️🦇
Transformers: 🚗🤖
Wonder Woman: 🙀Ȳ🧑🧑🧑🧑
Spider-Man:🕸️🕷️🕸️🕷️
Winnie the Pooh: 🐻🐼🐻
The Godfather: 🧑🧑🧑🧑-cat♂💥
Game of Thrones: ✘⚔️⚔️⚔️
Spider-Man:
```

## 요약

API는 텍스트의 컨텍스트를 파악하고 이를 다양한 방식으로 바꿀 수 있습니다. 이 예에서 API는 텍스트 블록을 가져와서 어린이가 이해할 수 있는 설명을 만듭니다. 이는 API가 언어에 대한 깊은 이해를 가지고 있음을 보여 줍니다.

```
My ten-year-old asked me what this passage means:  
"""  
A neutron star is the collapsed core of a massive supergiant star, which had  
a total mass of between 10 and 25 solar masses, possibly more if the star  
was especially metal-rich.[1] Neutron stars are the smallest and densest  
stellar objects, excluding black holes and hypothetical white holes, quark  
stars, and strange stars.[2] Neutron stars have a radius on the order of 10  
kilometres (6.2 mi) and a mass of about 1.4 solar masses.[3] They result  
from the supernova explosion of a massive star, combined with gravitational  
collapse, that compresses the core past white dwarf star density to that of  
atomic nuclei.  
"""  
  
I rephrased it for him, in plain language a ten-year-old can understand:  
"""
```

이 예에서 삼중 따옴표 사이에 요약하려는 내용을 넣습니다. 의도가 무엇이며 요약의 대상이 누구인지 요약하기 위해 텍스트 전후에 모두 설명한다는 점은 주목할 가치가 있습니다.

니다. 이는 API가 큰 텍스트 블록을 처리한 후 표류하는 것을 방지하기 위한 것입니다.

## Completion

모든 프롬프트가 완료되는 결과를 가져오지만, API가 중단한 부분부터 선택하기를 원하는 경우 텍스트 완료를 자체 작업으로 생각하는 것이 도움이 될 수 있습니다. 예를 들어, 이 프롬프트가 주어지면 API는 수직 농업에 대한 생각의 학습을 계속할 것입니다. API가 프롬프트의 의도에 더 집중하도록 온도 설정을 낮추거나 접선에서 해제되도록 높일 수 있습니다.

```
Vertical farming provides a novel solution for producing food locally,  
reducing transportation costs and
```

이 다음 프롬프트는 완료를 사용하여 React 구성 요소를 작성할 수 있도록 하는 방법을 보여 줍니다. 일부 코드를 API에 보내고 나머지는 React 라이브러리에 대한 이해가 있기 때문에 계속할 수 있습니다. 코드를 이해하거나 생성하는 작업에는 Codex 시리즈의 모델을 사용하는 것이 좋습니다. 현재 `code-davinci-002` 및 `code-cushman-001`의 두 가지 Codex 모델을 지원합니다. Codex 모델에 대한 자세한 내용은 [모델](#)의 [Codex 모델](#) 섹션을 참조하세요.

```
import React from 'react';  
const HeaderComponent = () => (
```

## 실제 응답

API는 학습된 데이터에서 학습한 많은 지식을 가지고 있습니다. 또한 매우 사실적으로 들리지만 실제로 구성된 응답을 제공하는 기능이 있습니다. API가 답변을 구성할 가능성을 제한하는 두 가지 방법이 있습니다.

- 1. API에 대한 기본 정보 제공** Wikipedia 항목과 같은 질문에 답변하기 위해 API에 텍스트 본문을 제공하면 응답을 혼동할 가능성이 줄어듭니다.
- 2. 낮은 확률을 사용하여 "모르겠다"라고 말하는 방법을 API에 보여 줍니다.** API가 "모르겠다"라고 말하는 응답이 덜 확실할 때 또는 약간의 변형이 적절하다는 것을 이해한다면, 답변을 만드는 경향이 줄어들 것입니다.

이 예에서 API가 알고 있는 질문과 답변의 예를 제공한 다음 알지 못할 것의 예를 제공하고 물음표를 제공합니다. 또한 API가 의심의 여지가 있는 경우 "?"로 응답할 가능성이 더

높도록 확률을 0으로 설정했습니다.

Q: Who is Batman?

A: Batman is a fictional comic book character.

Q: What is torsalplexity?

A: ?

Q: What is Devz9?

A: ?

Q: Who is George Lucas?

A: George Lucas is American film director and producer famous for creating Star Wars.

Q: What is the capital of California?

A: Sacramento.

Q: What orbits the Earth?

A: The Moon.

Q: Who is Fred Rickerson?

A: ?

Q: What is an atom?

A: An atom is a tiny particle that makes up everything.

Q: Who is Alvan Muntz?

A: ?

Q: What is Kozar-09?

A: ?

Q: How many moons does Mars have?

A: Two, Phobos and Deimos.

Q:

## 코드 처리

Codex 모델 시리즈는 자연어와 수십억 줄의 코드로 학습된 OpenAI의 기본 GPT-3 시리즈의 하위 항목입니다. Python에서 가장 뛰어나고 C#, JavaScript, Go, Perl, PHP, Ruby, Swift, TypeScript, SQL, 심지어 Shell을 포함한 12개 이상의 언어에 능숙합니다.

[코드 가이드 작업](#)을 통해 코드 완료 생성에 대해 자세히 알아봅니다.

## 다음 단계

코드 작업 방법(Codex)에 대해 알아봅니다. Azure OpenAI를 지원하는 기본 모델에 대해 자세히 알아봅니다.

# Python을 사용하여 OpenAI와 Azure OpenAI 엔드포인트 간에 전환하는 방법

아티클 • 2023. 05. 27.

OpenAI 및 Azure OpenAI Service는 일반적인 Python 클라이언트 라이브러리 [를 사용하지만 엔드포인트 간에 교환하기 위해 코드를 약간 변경해야 합니다.](#) 이 문서에서는 OpenAI 및 Azure OpenAI에서 작업할 때 발생하는 일반적인 변경 내용과 차이점을 안내합니다.

## ① 참고

이 라이브러리는 OpenAI에서 유지 관리되며 현재 미리 보기 상태입니다. 라이브러리에 대한 최신 업데이트를 추적하려면 [릴리스 기록](#) 또는 [version.py 커밋 기록](#)을 참조하세요.

## 인증

환경 변수를 사용하는 것이 좋습니다. [Python 빠른 시작](#) 전에 이 작업을 수행하지 않은 경우 이 구성을 안내합니다.

## API 키

| OpenAI                                                                          | Azure OpenAI                                                                                                                                                                                    |
|---------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Python                                                                          | Python                                                                                                                                                                                          |
| <pre>import openai  openai.api_key = "sk-..." openai.organization = "..."</pre> | <pre>import openai  openai.api_type = "azure" openai.api_key = "..." openai.api_base = "https://example- endpoint.openai.azure.com" openai.api_version = "2023-05-15" # subject to change</pre> |

## Azure Active Directory 인증

| OpenAI | Azure OpenAI |
|--------|--------------|
|--------|--------------|

|                                                                                         |                                                                                                                                                                                                                                                                                                                                                                                      |
|-----------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <pre>Python import openai  openai.api_key = "sk-..." openai.organizat ion = "..."</pre> | <pre>Python import openai from azure.identity import DefaultAzureCredential  credential = DefaultAzureCredential() token = credential.get_token("https://cognitiveservices.azure.com/.default")  openai.api_type = "azuread" openai.api_key = token.token openai.api_base = "https://example- endpoint.openai.azure.com" openai.api_version = "2023-05-15" # subject to change</pre> |
|-----------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|

## 모델의 키워드 인수

OpenAI는 `model` 키워드(keyword) 인수를 사용하여 사용할 모델을 지정합니다. Azure OpenAI는 `배포` 개념을 가지고 있으며 키워드(keyword) 인수를 사용하여 `deployment_id` 사용할 모델 배포를 설명합니다. 또한 Azure OpenAI는 와 `deployment_id` 교환할 수 있는 사용을 `engine` 지원합니다.

OpenAI `engine` 의 경우 여전히 대부분의 인스턴스에서 작동하지만 더 이상 사용되지 않으며 `model` 선호됩니다.

| OpenAI                                                                                                                                                                                                                                                                                                                   | Azure OpenAI                                                                                                                                                                                                                                                                         |
|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <pre>Python completion = openai.Completion.create(     prompt=&lt;prompt&gt;,     model="text-davinci-003" )  chat_completion = openai.ChatCompletion.create(     messages=&lt;messages&gt;,     model="gpt-4" )  embedding = openai.Embedding.create(     input=&lt;input&gt;,     model="text-embedding-ada-002"</pre> | <pre>Python completion = openai.Completion.create(     prompt=&lt;prompt&gt;,     deployment_id="text-davinci- 003" #engine="text-davinci-003"  )  chat_completion = openai.ChatCompletion.create(     messages=&lt;messages&gt;,     deployment_id="gpt-4" #engine="gpt-4"  )</pre> |

```
)  
  
embedding =  
openai.Embedding.create(  
    input=<input>,  
    deployment_id="text-embedding-  
ada-002"  
    #engine="text-embedding-ada-  
002"  
)
```

## Azure OpenAI 포함은 여러 입력을 지원하지 않습니다.

많은 예제에서는 포함 API에 여러 입력을 전달하는 것을 보여 줍니다. Azure OpenAI의 경우 현재 호출당 단일 텍스트 입력을 전달해야 합니다.

| OpenAI                                                                                                                                                                      | Azure OpenAI                                                                                                                                                                                                                                                |
|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <p>Python</p> <pre>inputs = ["A", "B", "C"]<br/><br/>embedding =<br/>openai.Embedding.create(<br/>    input=inputs,<br/>    model="text-embedding-ada-<br/>002"<br/>)</pre> | <p>Python</p> <pre>inputs = ["A", "B", "C"]<br/><br/>for text in inputs:<br/>    embedding =<br/>openai.Embedding.create(<br/>    input=text,<br/>    deployment_id="text-<br/>embedding-ada-002"<br/>    #engine="text-embedding-ada-<br/>002"<br/>)</pre> |

## 다음 단계

- 방법 가이드를 사용하여 ChatGPT 및 GPT-4 모델을 사용하는 방법에 대해 자세히 알아보세요.
- 더 많은 예제를 보려면 Azure OpenAI 샘플 GitHub 리포지토리[\[↗\]](#)를 체크 아웃합니다.

# Codex 모델 및 Azure OpenAI Service

아티클 • 2023. 04. 28.

Codex 모델 시리즈는 자연어와 수십억 줄의 코드에 대해 학습된 GPT-3 시리즈의 하위 항목입니다. Python에서 가장 뛰어나고 C#, JavaScript, Go, Perl, PHP, Ruby, Swift, TypeScript, SQL, 심지어 Shell을 포함한 12개 이상의 언어에 능숙합니다.

다음과 같은 다양한 작업에 Codex를 사용할 수 있습니다.

- 주석을 코드로 변환
- 컨텍스트에서 다음 줄 또는 함수 완성
- 유용한 라이브러리 찾기 또는 애플리케이션에 대한 API 호출과 같은 지식 제공
- 주석 추가
- 효율성을 위한 코드 재작성

## Codex 모델을 사용하는 방법

다음은 `code-davinci-002`와 같은 Codex 시리즈 모델을 배포하여 [Azure OpenAI Studio](#) 플레이그라운드에서 테스트할 수 있는 Codex 사용의 몇 가지 예제입니다.

### "Hello"라고 말하기(Python)

Python

```
"""
Ask the user for their name and say "Hello"
"""
```

### 임의의 이름 만들기(Python)

Python

```
"""
1. Create a list of first names
2. Create a list of last names
3. Combine them randomly into a list of 100 full names
"""
```

### MySQL 쿼리 만들기(Python)

Python

```
"""
Table customers, columns = [CustomerId, FirstName, LastName, Company,
Address, City, State, Country, PostalCode, Phone, Fax, Email, SupportRepId]
Create a MySQL query for all customers in Texas named Jane
"""

query =
```

## 코드 설명(JavaScript)

JavaScript

```
// Function 1
var fullNames = [];
for (var i = 0; i < 50; i++) {
    fullNames.push(names[Math.floor(Math.random() * names.length)]
        + " " + lastNames[Math.floor(Math.random() * lastNames.length)]);
}

// What does Function 1 do?
```

## 최선의 구현 방법

### 주석, 데이터 또는 코드로 시작

플레이그라운드에서 Codex 모델 중 하나를 사용하여 실험할 수 있습니다(필요한 경우 지침을 주석으로 스타일 지정).

Codex가 유용한 완료를 만들도록 하려면 프로그래머가 작업을 수행하는 데 필요한 정보가 무엇인지 생각하는 것이 좋습니다. 이는 변수 이름이나 함수가 처리하는 클래스와 같은 유용한 함수를 작성하는 데 필요한 데이터 또는 명확한 주석일 수 있습니다.

이 예에서는 Codex에게 이 함수를 호출하는 작업과 수행할 작업을 알려줍니다.

Python

```
# Create a function called 'nameImporter' to add a first and last name to
the database
```

이 방법은 Codex에 주석과 데이터베이스 스키마의 예를 제공하여 다양한 데이터베이스에 대한 유용한 쿼리 요청을 작성할 수 있는 수준까지 확장됩니다. 다음은 쿼리에 대한 열과 테이블 이름을 제공하는 예입니다.

Python

```
# Table albums, columns = [AlbumId, Title, ArtistId]
# Table artists, columns = [ArtistId, Name]
# Table media_types, columns = [MediaTypeId, Name]
# Table playlists, columns = [PlaylistId, Name]
# Table playlist_track, columns = [PlaylistId, TrackId]
# Table tracks, columns = [TrackId, Name, AlbumId, MediaTypeId, GenreId,
Composer, Milliseconds, Bytes, UnitPrice]

# Create a query for all albums with more than 10 tracks
```

Codex 데이터베이스 스키마를 표시하면 쿼리 형식을 지정하는 방법에 대해 정보에 입각한 추측을 할 수 있습니다.

## 프로그래밍 언어 지정

Codex는 수십 가지의 다양한 프로그래밍 언어를 이해합니다. 많은 사람들이 주석, 함수 및 기타 프로그래밍 구문에 대해 유사한 규칙을 공유합니다. Codex는 주석에 언어와 버전을 지정하여 원하는 것을 더욱 효율적으로 완료할 수 있습니다. 즉, Codex는 스타일과 구문에 있어 상당히 유연합니다. 다음은 R 및 Python에 대한 예입니다.

R

```
# R language
# Calculate the mean distance between an array of points
```

Python

```
# Python 3
# Calculate the mean distance between an array of points
```

## 수행하려는 작업이 포함된 프롬프트 Codex

Codex가 웹페이지를 만들도록 하려면 HTML 문서(<!DOCTYPE html>)의 첫 번째 코드 줄을 주석 뒤에 배치하면 Codex가 다음에 수행해야 할 작업을 알려줍니다. 주석에서 함수를 만드는 경우에도 동일한 방법이 적용됩니다(주석 다음에는 func 또는 def로 시작하는 새 행이 추가됨).

HTML

```
<!-- Create a web page with the title 'Kat Katman attorney at paw' -->
<!DOCTYPE html>
```

주석 뒤에 `<!DOCTYPE html>`을 배치하면 Codex에서 원하는 작업이 매우 명확해집니다.

또는 함수를 작성하려는 경우 다음과 같이 프롬프트를 시작할 수 있으며 Codex는 다음에 수행해야 하는 작업을 이해할 것입니다.

Python

```
# Create a function to count to 100  
def counter
```

## 라이브러리를 지정하면 Codex가 원하는 것을 이해하는 데 도움이 됩니다.

Codex는 수많은 라이브러리, API 및 모듈을 알고 있습니다. Codex는 주석에서 사용할 것인지 또는 코드로 가져올 것인지를 알려줌으로써 Codex가 대안 대신 이를 기반으로 제안할 것입니다.

HTML

```
<!-- Use A-Frame version 1.2.0 to create a 3D website -->  
<!-- https://aframe.io/releases/1.2.0/aframe.min.js -->
```

버전을 지정하면 Codex가 최신 라이브러리를 사용하는지 확인할 수 있습니다.

### ① 참고

Codex는 유용한 라이브러리와 API를 제안할 수 있지만 항상 고유의 연구를 수행하여 애플리케이션에 안전한지 확인합니다.

## 주석 스타일은 코드 품질에 영향을 줄 수 있습니다.

일부 언어에서는 주석 스타일이 출력 품질을 개선시킬 수 있습니다. 예를 들어 Python으로 작업할 때 문서 문자열(3중 따옴표로 묶인 주석)을 사용하면 파운드(#) 기호를 사용하는 것보다 더 높은 품질의 결과를 얻을 수 있습니다.

Python

```
"""  
Create an array of users and email addresses  
"""
```

## 함수 내부의 주석은 도움이 될 수 있습니다.

권장되는 코딩 표준은 일반적으로 함수 내부에 함수 설명을 작성해 두는 것이 좋습니다. 이 형식을 사용하면 Codex에서 함수가 수행하려는 작업을 더 명확하게 이해하는 데 도움이 됩니다.

Python

```
def getUserBalance(id):
    """
    Look up the user in the database 'UserData' and return their current
    account balance.
    """
```

## 보다 정확한 결과를 위한 예제 제공

Codex를 사용해야 하는 특정 스타일이나 형식이 있는 경우 요청의 첫 번째 부분에서 예를 제공하거나 이를 시연하면 Codex가 필요한 것을 보다 정확하게 일치시키는 데 도움이 됩니다.

Python

```
"""
Create a list of random animals and species
"""
animals = [ {"name": "Chomper", "species": "Hamster"}, {"name":
```

## 더 낮은 온도는 더 정확한 결과를 제공합니다.

API 온도를 0으로 설정하거나 0에 가깝게(예: 0.1 또는 0.2) 대부분의 경우 더 나은 결과를 제공하는 경향이 있습니다. 더 높은 온도가 유용하고 창의적이고 임의적인 결과를 제공할 수 있는 GPT-3 모델과 달리, Codex 모델의 더 높은 온도는 실제로 임의 또는 불규칙한 응답을 제공할 수 있습니다.

Codex가 다른 잠재적인 결과를 제공해야 하는 경우 0에서 시작한 다음 적절한 변형을 찾을 때까지 0.1씩 위쪽으로 증분합니다.

## 작업을 함수로 구성

함수가 수행해야 하는 작업을 주석에 가능한 한 정확한 용어로 지정하여 Codex가 함수를 작성하도록 할 수 있습니다. 예를 들어 다음 주석을 작성하여 Codex는 사용자가 단추를 누를 때 트리거되는 JavaScript 타이머 함수를 만듭니다.

## 간단한 JavaScript 타이머

```
JavaScript
```

```
// Create a timer that creates an alert in 10 seconds
```

## 데이터 예 만들기

애플리케이션을 테스트하려면 종종 데이터 예를 사용해야 합니다. Codex는 자연어를 이해하고 쓰는 방법을 파악하는 언어 모델이기 때문에 Codex에 구성된 이름, 제품 및 기타 변수의 배열과 같은 데이터를 만들도록 요청할 수 있습니다. 예를 들어, 여기에서 Codex에 기상 온도 배열을 만들도록 요청합니다.

```
JavaScript
```

```
/* Create an array of weather temperatures for San Francisco */
```

Codex에 이 작업을 수행하도록 요청하면 다음과 같은 테이블이 생성됩니다.

```
JavaScript
```

```
var weather = [
  { month: 'January', high: 58, low: 48 },
  { month: 'February', high: 61, low: 50 },
  { month: 'March', high: 64, low: 53 },
  { month: 'April', high: 67, low: 55 },
  { month: 'May', high: 70, low: 58 },
  { month: 'June', high: 73, low: 61 },
  { month: 'July', high: 76, low: 63 },
  { month: 'August', high: 77, low: 64 },
  { month: 'September', high: 76, low: 63 },
  { month: 'October', high: 73, low: 61 },
  { month: 'November', high: 68, low: 57 },
  { month: 'December', high: 64, low: 54 }
];
```

## 복합 함수 및 소규모 애플리케이션

임의의 이름 생성기를 만들거나 사용자 입력으로 작업을 수행하는 것과 같은 복잡한 요청으로 구성된 주석을 Codex에 제공할 수 있으며 Codex는 충분한 토큰이 있는 경우 나머지를 생성할 수 있습니다.

```
JavaScript
```

```
/*
Create a list of animals
Create a list of cities
Use the lists to generate stories about what I saw at the zoo in each city
*/
```

## 더 정확한 결과 또는 더 짧은 대기 시간을 위해 완료 크기 제한

Codex에서 더 긴 완료를 요청하면 부정확한 답변과 반복이 발생할 수 있습니다. `max_tokens`를 줄이고 중지 토큰을 설정하여 쿼리 크기를 제한합니다. 예를 들어, 한 줄의 코드로 완료를 제한하려면 `\n`을 중지 시퀀스로 추가합니다. 완료 횟수가 적을수록 대기 시간도 줄어듭니다.

## 스트리밍을 사용하여 대기 시간 줄이기

대규모 Codex 쿼리는 완료하는 데 수십 초가 걸릴 수 있습니다. 자동 완료를 수행하는 코딩 도우미와 같이 더 짧은 대기 시간이 필요한 애플리케이션을 빌드하려면 스트리밍 사용을 고려합니다. 모델이 전체 완료 생성을 완료하기 전에 응답이 반환됩니다. 완료의 일부만 필요한 애플리케이션은 프로그래밍 방식으로 완료를 차단하거나 `stop`에 대한 창의적인 값을 사용하여 대기 시간을 줄일 수 있습니다.

사용자는 API에서 둘 이상의 솔루션을 요청하고 반환된 첫 번째 응답을 사용하여 대기 시간을 줄이기 위해 스트리밍과 복제를 결합할 수 있습니다. `n > 1`을 설정하여 이 작업을 수행합니다. 이 방법은 더 많은 토큰 할당량을 사용하므로 주의해서 사용합니다(예: `max_tokens` 및 `stop`에 대해 합리적인 설정 사용).

## Codex를 사용하여 코드 설명

Codex의 코드 만들기 및 이해 기능은 파일의 코드가 하는 일을 설명하는 것과 같은 작업을 수행하는 데 사용할 수 있습니다. 이를 수행하는 한 가지 방법은 "This function" 또는 "This application is"로 시작하는 주석을 함수 뒤에 추가하는 것입니다. Codex는 일반적으로 이를 설명의 시작으로 해석하고 나머지 텍스트를 완성합니다.

JavaScript

```
/* Explain what the previous function is doing: It
```

## SQL 쿼리 설명

이 예에서는 Codex를 사용하여 SQL 쿼리가 수행하는 작업을 인간이 읽을 수 있는 형식으로 설명합니다.

SQL

```
SELECT DISTINCT department.name
FROM department
JOIN employee ON department.id = employee.department_id
JOIN salary_payments ON employee.id = salary_payments.employee_id
WHERE salary_payments.date BETWEEN '2020-06-01' AND '2020-06-30'
GROUP BY department.name
HAVING COUNT(employee.id) > 10;
-- Explanation of the above query in human readable format
--
```

## 단위 테스트 작성

Python에서 "Unit test"라는 주석을 추가하고 함수를 시작하는 것만으로 단위 테스트를 만들 수 있습니다.

Python

```
# Python 3
def sum_numbers(a, b):
    return a + b

# Unit test
def
```

## 코드 오류 확인

예를 사용하여 Codex에서 코드의 오류를 식별하는 방법을 보여줄 수 있습니다. 어떤 경우에는 예가 필요하지 않지만 설명을 제공하기 위해 수준과 세부 사항을 시연하면 Codex 가 찾아야 할 것과 설명하는 방법을 이해하는 데 도움이 될 수 있습니다. (Codex의 오류 확인이 사용자의 신중한 검토를 대체해서는 안 됩니다.)

JavaScript

```
/* Explain why the previous function doesn't work. */
```

## 원본 데이터를 사용하여 데이터베이스 함수 작성

인간 프로그래머가 데이터베이스 구조와 열 이름을 이해하는 것이 도움이 되는 것처럼 Codex는 이 데이터를 사용하여 정확한 쿼리 요청을 작성하는 데 도움을 줄 수 있습니다.

이 예에서는 데이터베이스에 대한 스키마를 삽입하고 Codex에 데이터베이스를 쿼리할 대상을 알려줍니다.

Python

```
# Table albums, columns = [AlbumId, Title, ArtistId]
# Table artists, columns = [ArtistId, Name]
# Table media_types, columns = [MediaTypeId, Name]
# Table playlists, columns = [PlaylistId, Name]
# Table playlist_track, columns = [PlaylistId, TrackId]
# Table tracks, columns = [TrackId, Name, AlbumId, MediaTypeId, GenreId,
Composer, Milliseconds, Bytes, UnitPrice]

# Create a query for all albums with more than 10 tracks
```

## 언어 간 변환

Codex가 변환하려는 코드의 언어를 주석에 나열한 다음 코드와 함께 번역할 언어가 있는 주석을 나열하는 간단한 형식을 따르면 한 언어에서 다른 언어로 변환할 수 있습니다.

Python

```
# Convert this from Python to R
# Python version

[ Python code ]

# End

# R version
```

## 라이브러리 또는 프레임워크용 코드 재작성

Codex가 함수를 보다 효율적으로 만들기를 원하면 다시 작성할 코드와 함께 사용할 형식에 대한 지침을 제공할 수 있습니다.

JavaScript

```
// Rewrite this as a React component
var input = document.createElement('input');
input.setAttribute('type', 'text');
document.body.appendChild(input);
var button = document.createElement('button');
button.innerHTML = 'Say Hello';
document.body.appendChild(button);
button.onclick = function() {
  var name = input.value;
```

```
var hello = document.createElement('div');
hello.innerHTML = 'Hello ' + name;
document.body.appendChild(hello);
};

// React version:
```

## 다음 단계

Azure OpenAI를 지원하는 기본 모델에 대해 자세히 알아봅니다.

# Azure OpenAI를 사용하여 포함을 생성하는 방법 알아보기

아티클 • 2023. 05. 17.

포함은 기계 학습 모델 및 알고리즘에서 쉽게 활용할 수 있는 특수한 형식의 데이터 표현입니다. 포함은 텍스트 조각의 의미 체계적 의미에 대한 조밀한 정보 표현입니다. 각 포함은 부동 소수점 숫자의 벡터입니다. 따라서 벡터 공간의 두 포함 사이의 거리는 원래 형식의 두 입력 간의 의미 체계 유사성과 상관 관계가 있습니다. 예를 들어 두 텍스트가 비슷한 경우 벡터 표현도 유사해야 합니다.

## 포함을 가져오는 방법

텍스트 조각에 대한 포함 벡터를 가져오려면 다음 코드 조각과 같이 포함 엔드포인트에 요청합니다.

```
curl  
https://YOUR_RESOURCE_NAME.openai.azure.com/openai/deployments/YOUR_DEPL  
OYMENT_NAME/embeddings?api-version=2023-05-15\  
-H 'Content-Type: application/json' \  
-H 'api-key: YOUR_API_KEY' \  
-d '{"input": "Sample Document goes here"}'
```

## 모범 사례

### 입력이 최대 길이를 초과하지 않는지 확인

포함 모델에 대한 입력 텍스트의 최대 길이는 2048개의 토큰(2~3페이지의 텍스트에 해당)입니다. 요청하기 전에 입력이 이 제한을 초과하지 않는지 확인해야 합니다.

### 작업에 가장 적합한 모델 선택

검색 모델의 경우 두 가지 방법으로 포함을 가져올 수 있습니다. `<search_model>-doc` 모델은 더 긴 텍스트 조각(검색할 수 있음)에 사용되며 `<search_model>-query` 모델은 더 짧은

텍스트 조각(일반적으로 제로 샷 분류의 쿼리 또는 클래스 레이블)에 사용됩니다. [모델](#) 가 이드에서 모든 포함 모델에 대해 자세히 확인할 수 있습니다.

## 줄 바꿈을 단일 공백으로 교체

코드를 포함하지 않는 한 입력에서 줄 바꿈(\n)을 단일 공백으로 바꾸는 것이 좋습니다. 줄 바꿈이 있을 때 열등한 결과가 관찰되었기 때문입니다.

## 제한 사항 & 위험

포함 모델은 신뢰할 수 없거나 특정 경우에 사회적 위험을 초래할 수 있으며 완화 조치가 없을 때 피해를 줄 수 있습니다. 책임감 있게 사용하는 방법에 대한 자세한 내용은 책임 있는 AI 콘텐츠를 검토하세요.

## 다음 단계

- Azure OpenAI 및 embeddings를 사용하여 [embeddings 자습서](#)로 문서 검색을 수행하는 방법에 대해 자세히 알아봅니다.
- [Azure OpenAI를 지원하는 기본 모델](#)에 대해 자세히 알아봅니다.

# 미세 조정을 위해 데이터 세트를 준비하는 방법 알아보기

아티클 • 2023. 04. 27.

모델을 사용자 지정하는 첫 번째 단계는 고품질 데이터 세트를 준비하는 것입니다. 이렇게 하려면 단일 입력 프롬프트와 관련 원하는 출력('완료')으로 구성된 일련의 학습 예가 필요합니다. 이 형식은 다음과 같은 방식으로 유추하는 동안 모델을 사용하는 것과 현저하게 다릅니다.

- 몇 가지 예가 아닌 단일 프롬프트만 제공합니다.
- 프롬프트의 일부로 자세한 지침을 제공할 필요가 없습니다.
- 각 프롬프트는 프롬프트가 종료되고 완료가 시작될 때 모델에 알리기 위해 고정 구분 기호로 끝나야 합니다. 일반적으로 잘 작동하는 간단한 구분 기호는 `\n\n###\n\n`입니다. 구분 기호는 프롬프트의 다른 곳에 나타나지 않아야 합니다.
- 토큰화로 인해 각 완료는 공백으로 시작해야 하며, 이 공백은 실행 공백으로 대부분의 단어를 토큰화합니다.
- 각 완료는 완료가 끝날 때 모델에 알리기 위해 고정된 중지 시퀀스로 끝나야 합니다. 중지 시퀀스는 `\n`, `###` 또는 완료 시 나타나지 않는 기타 토큰일 수 있습니다.
- 유추를 위해 동일한 구분 기호를 포함하여 학습 데이터 세트를 만들 때와 동일한 방식으로 프롬프트 형식을 지정해야 합니다. 또한 완료를 적절하게 자르기 위해 동일한 중지 시퀀스를 지정합니다.
- 데이터 세트는 총 파일 크기에서 100MB를 초과할 수 없습니다.

## 모범 사례

사용자 지정은 고품질 예에서 더 잘 수행되며 더 많이 가질수록 일반적으로 모델이 더 잘 수행됩니다. 베이스 모델과 함께 잘 설계된 프롬프트를 사용하는 것보다 더 나은 성능을 발휘하는 모델을 가져오려면 최소한 수백 개의 고품질 예를 제공하는 것이 좋습니다. 여기에서 성능은 예의 수가 두 배가 될 때마다 선형적으로 증가하는 경향이 있습니다. 예의 수를 늘리는 것이 일반적으로 성능을 개선시키는 가장 좋고 가장 안정적인 방법입니다.

프롬프트를 처음부터 작성하지 않고 기존 데이터 세트를 미세 조정하는 경우 가능한 한 공격적이거나 부정확한 콘텐츠에 대해 데이터를 수동으로 검토하거나 데이터 세트의 임의 샘플이 큰 경우 가능한 한 많은 임의 샘플을 검토합니다.

## 특정 지침

미세 조정은 다양한 문제를 해결할 수 있으며 최적의 사용 방법은 특정 사용 사례에 따라 다를 수 있습니다. 아래에는 미세 조정 및 해당 지침에 대한 가장 일반적인 사용 사례가

나열되어 있습니다.

## 분류

분류자는 시작 가장 쉬운 모델입니다. 분류 문제의 경우 **ada**를 사용하는 것이 좋습니다. 이 방법은 일반적으로 한 번 미세 조정하면 성능이 훨씬 더 좋은 모델보다 약간만 성능이 떨어지는 경향이 있지만 훨씬 더 빠릅니다. 분류 문제에서 데이터 세트의 각 프롬프트는 미리 정의된 클래스 중 하나로 분류되어야 합니다. 이러한 형식의 문제에 대해 다음을 권장합니다.

- 프롬프트 끝에 구분 기호를 사용합니다(예: `\n\n###\n\n`). 최종적으로 모델에 요청할 때도 이 구분 기호를 추가하는 것을 잊지 마세요.
- 단일 토큰에 매핑되는 클래스를 선택합니다. 유추 시에는 분류를 위한 첫 번째 토큰만 필요하므로 `max_tokens=1`을 지정합니다.
- 프롬프트 + 완료가 구분 기호를 포함하여 2048개의 토큰을 초과하지 않는지 확인합니다.
- 클래스당 최소 100개의 예를 목표로 합니다.
- 클래스 로그 확률을 가져오려면 모델을 사용할 때 `logprobs=5`(클래스가 5개인 경우)를 지정할 수 있습니다.
- 미세 조정에 사용되는 데이터 세트가 모델이 사용되는 작업 형식과 구조 및 형식이 매우 유사한지 확인합니다.

## 사례 연구: 모델이 사실이 아닌 진술을 하고 있나요?

웹 사이트의 광고 텍스트가 올바른 제품 및 회사를 언급하는지 확인하고 싶다고 가정해 보겠습니다. 다시 말해, 모델이 상황을 만들어내지 않도록 해야 합니다. 잘못된 광고를 필터링하는 분류자를 미세 조정할 수 있습니다.

데이터 세트는 다음과 유사할 수 있습니다.

JSON

```
{"prompt":"Company: BHFF insurance\nProduct: allround insurance\nAd:One stop shop for all your insurance needs!\nSupported:", "completion":" yes"}  
{"prompt":"Company: Loft conversion specialists\nProduct: -\nAd:Straight teeth in weeks!\nSupported:", "completion":" no"}
```

위의 예에서는 회사 이름, 제품 및 관련 광고가 포함된 구조화된 입력을 사용했습니다. 구분자로 `\nSupported:`를 사용하여 프롬프트를 완료와 명확하게 구분했습니다. 충분한 수의 예에서 선택하는 구분 기호는 프롬프트나 완료 내에 나타나지 않는 한 큰 차이가 없습니다(보통 0.4% 미만).

이 사용 사례의 경우 ada 모델이 더 빠르고 저렴할 것이기 때문에 미세 조정했으며 성능은 분류 작업이기 때문에 더 큰 모델과 비슷할 것입니다.

이제 완료 요청을 통해 모델을 쿼리할 수 있습니다.

#### 콘솔

```
curl  
https://YOUR_RESOURCE_NAME.openai.azure.com/openai/deployments/YOUR_DEPLOYMENT_NAME/completions?api-version=2023-05-15 \  
-H 'Content-Type: application/json' \  
-H 'api-key: YOUR_API_KEY' \  
-d '{  
    "prompt": "Company: Reliable accountants Ltd\nProduct: Personal Tax  
help\nAd:Best advice in town!\nSupported:",  
    "max_tokens": 1  
}'
```

yes 또는 no를 반환합니다.

## 사례 연구: 감정 분석

특정 트윗이 긍정적이거나 부정적인 정도를 알고 싶다고 가정해 보겠습니다. 데이터 세트는 다음과 유사할 수 있습니다.

#### 콘솔

```
{"prompt":"Overjoyed with the new iPhone! ->, "completion":" positive"}  
{"prompt":"@contoso_basketball disappoint for a third straight night. ->","completion":" negative"}
```

모델이 미세 조정되면 완료 요청에서 logprobs=2을 설정하여 첫 번째 완료 토큰에 대한 로그 확률을 다시 얻을 수 있습니다. 긍정적 클래스의 확률이 높을수록 상대적 감정이 높아집니다.

이제 완료 요청을 통해 모델을 쿼리할 수 있습니다.

#### 콘솔

```
curl  
https://YOUR_RESOURCE_NAME.openai.azure.com/openai/deployments/YOUR_DEPLOYMENT_NAME/completions?api-version=2023-05-15 \  
-H 'Content-Type: application/json' \  
-H 'api-key: YOUR_API_KEY' \  
-d '{  
    "prompt": "Excited to share my latest blog post! ->",  
    "max_tokens": 1,
```

```
"logprobs": 2  
}'
```

반환할 항목:

JSON

```
{  
  "object": "text_completion",  
  "created": 1589498378,  
  "model": "YOUR_FINE_TUNED_MODEL_NAME",  
  "choices": [  
    {  
      "logprobs": {  
        "text_offset": [  
          19  
        ],  
        "token_logprobs": [  
          -0.03597255  
        ],  
        "tokens": [  
          " positive"  
        ],  
        "top_logprobs": [  
          {  
            "negative": -4.9785037,  
            "positive": -0.03597255  
          }  
        ]  
      },  
      "text": " positive",  
      "index": 0,  
      "finish_reason": "length"  
    }  
  ]  
}
```

## 사례 연구: 이메일 심사를 위한 심사

수신 이메일을 미리 정의된 많은 범주 중 하나로 분류한다고 가정해 보겠습니다. 많은 수의 범주로 분류하려면 해당 범주를 최대 500개 범주까지 잘 작동하는 숫자로 변환하는 것이 좋습니다. 숫자 앞에 공백을 추가하면 토큰화로 인해 성능에 약간 도움이 되는 경우가 있습니다. 다음과 같이 학습 데이터를 구조화할 수 있습니다.

JSON

```
{  
  "prompt": "Subject: <email_subject>\nFrom:<customer_name>\nDate:  
  "text": "Hello, how can I help you today?"  
  "label": "Customer Support Query"}  
}
```

```
<date>\nContent:<email_body>\n\n###\n\n", "completion": "  
<numerical_category>"  
}
```

다음은 그 예입니다.

JSON

```
{  
    "prompt": "Subject: Update my address\nFrom: Joe  
Doe\nTo:support@ourcompany.com\nDate:2021-06-03\nContent:Hi,\nI would like  
to update my billing address to match my delivery address.\n\nPlease let me  
know once done.\n\nThanks,\nJoe\n\n###\n\n",  
    "completion": " 4"  
}
```

위의 예에서 2043개의 토큰으로 제한되는 수신 이메일을 입력으로 사용했습니다. (이렇게 하면 4개의 토큰 구분 기호와 1개의 토큰 완료가 가능하며 합계는 2048입니다.) 구분 기호로 `\n\n###\n\n`을 사용하고 이메일 내에서 `###` 발생을 제거했습니다.

## 조건부 생성

조건부 생성은 일종의 입력이 주어지면 콘텐츠를 생성해야 하는 문제입니다. 여기에는 의역, 요약, 엔터티 추출, 지정된 사양의 제품 설명 작성, 챗봇 등이 포함됩니다. 이러한 형식의 문제에 대해 다음을 권장합니다.

- 프롬프트 끝에 구분 기호를 사용합니다(예: `\n\n###\n\n`). 최종적으로 모델에 요청할 때도 이 구분 기호를 추가하는 것을 잊지 마세요.
- 완료가 끝날 때 종료 토큰을 사용합니다(예: `END`).
- 유추하는 동안 종료 토큰을 중지 시퀀스로 추가해야 합니다(예: `stop=["END"]`).
- 최소 500개의 예를 목표로 합니다.
- 프롬프트 + 완료가 구분 기호를 포함하여 2048개의 토큰을 초과하지 않는지 확인합니다.
- 예가 고품질인지 확인하고 동일한 원하는 형식을 따릅니다.
- 미세 조정에 사용되는 데이터 세트가 모델이 사용되는 작업 형식과 구조 및 형식이 유사한지 확인합니다.
- 낮은 학습 속도와 1~2개의 Epoch만 사용하면 이러한 사용 사례에서 더 잘 작동하는 경향이 있습니다.

## 사례 연구: Wikipedia 문서를 기반으로 매력적인 광고 작성

이는 생성적인 사용 사례이므로 미세 조정된 모델이 지정된 예의 스타일(및 실수)을 모방하려고 시도하므로 제공하는 샘플의 품질이 최고인지 확인하고 싶을 것입니다. 좋은 출발

점은 약 500개의 예입니다. 샘플 데이터 세트는 다음과 같을 수 있습니다.

JSON

```
{  
    "prompt": "<Product Name>\n<Wikipedia description>\n\n###\n\n",  
    "completion": " <engaging ad> END"  
}
```

다음은 그 예입니다.

JSON

```
{  
    "prompt": "Samsung Galaxy Feel\nThe Samsung Galaxy Feel is an Android smartphone developed by Samsung Electronics exclusively for the Japanese market. The phone was released in June 2017 and was sold by NTT Docomo. It runs on Android 7.0 (Nougat), has a 4.7 inch display, and a 3000 mAh battery.\nSoftware\nSamsung Galaxy Feel runs on Android 7.0 (Nougat), but can be later updated to Android 8.0 (Oreo).\nHardware\nSamsung Galaxy Feel has a 4.7 inch Super AMOLED HD display, 16 MP back facing and 5 MP front facing cameras. It has a 3000 mAh battery, a 1.6 GHz Octa-Core ARM Cortex-A53 CPU, and an ARM Mali-T830 MP1 700 MHz GPU. It comes with 32GB of internal storage, expandable to 256GB via microSD. Aside from its software and hardware specifications, Samsung also introduced a unique a hole in the phone's shell to accommodate the Japanese perceived penchant for personalizing their mobile phones. The Galaxy Feel's battery was also touted as a major selling point since the market favors handsets with longer battery life. The device is also waterproof and supports 1seg digital broadcasts using an antenna that is sold separately.\n\n###\n\n",  
    "completion": "Looking for a smartphone that can do it all? Look no further than Samsung Galaxy Feel! With a slim and sleek design, our latest smartphone features high-quality picture and video capabilities, as well as an award winning battery life. END"  
}
```

여기에서는 Wikipedia 문서에 여러 단락과 제목이 포함되어 있으므로 여러 줄 구분 기호를 사용했습니다. 또한 모델이 완료를 마쳐야 하는 시점을 알 수 있도록 간단한 종료 토큰을 사용했습니다.

## 사례 연구: 엔터티 추출

이는 언어 변환 작업과 유사합니다. 성능을 개선시키려면 추출된 다른 엔터티를 알파벳순으로 정렬하거나 원래 텍스트에 나타나는 것과 같은 순서로 정렬하는 것이 가장 좋습니다. 이렇게 하면 모델이 순서대로 생성해야 하는 모든 엔터티를 추적하는 데 도움이 됩니다. 데이터 세트는 다음과 같이 표시될 수 있습니다.

JSON

```
{  
    "prompt": "<any text, for example news article>\n\n###\n\n",  
    "completion": "<list of entities, separated by a newline> END"  
}
```

다음은 그 예입니다.

JSON

```
{  
    "prompt": "Portugal will be removed from the UK's green travel list from  
Tuesday, amid rising coronavirus cases and concern over a \"Nepal mutation  
of the so-called Indian variant\". It will join the amber list, meaning  
holidaymakers should not visit and returnees must isolate for 10  
days... \n\n###\n\n",  
    "completion": " Portugal\nUK\nNepal mutation\nIndian variant END"  
}
```

텍스트에 여러 줄이 포함될 가능성이 높기 때문에 여러 줄 구분 기호가 가장 잘 작동합니다. 이상적으로는 엔터티를 추출할 때 접하게 될 텍스트를 반영하는 입력 프롬프트 형식이 매우 다양할 것입니다(뉴스 문서, Wikipedia 페이지, 트윗, 법률 문서).

## 사례 연구: 고객 지원 챗봇

챗봇에는 일반적으로 대화에 대한 관련 컨텍스트(주문 세부 정보), 지금까지의 대화 요약 및 가장 최근 메시지가 포함됩니다. 이 사용 사례의 경우 동일한 과거 대화가 데이터 세트에 여러 행을 생성할 수 있습니다. 매번 약간 다른 컨텍스트를 사용하여 모든 에이전트 생성을 완료로 간주합니다. 이 사용 사례에는 다양한 형식의 요청 및 고객 문제를 처리할 가능성이 있으므로 수천 개의 예가 필요합니다. 성능이 고품질인지 확인하려면 대화 샘플을 조사하여 상담원 메시지의 품질을 확인하는 것이 좋습니다. 요약은 별도의 텍스트 변환 미세 조정 모델로 생성할 수 있습니다. 데이터 세트는 다음과 같이 표시될 수 있습니다.

JSON

```
{"prompt": "Summary: <summary of the interaction so far>\n\nSpecific  
information:<for example order details in natural  
language>\n\n###\n\nCustomer: <message1>\nAgent: <response1>\nCustomer:  
<message2>\nAgent:", "completion": "<response2>\n"}  
{ "prompt": "Summary: <summary of the interaction so far>\n\nSpecific  
information:<for example order details in natural  
language>\n\n###\n\nCustomer: <message1>\nAgent: <response1>\nCustomer:  
<message2>\nAgent: <response2>\nCustomer: <message3>\nAgent:",  
"completion": "<response3>\n"}
```

여기에서 의도적으로 다른 형식의 입력 정보를 분리했지만 프롬프트와 완료 간에 동일한 형식으로 고객 에이전트 대화 상자를 유지했습니다. 모든 완료는 에이전트에 의해서만 이루어져야 하며 유추를 수행할 때 `\n`을 중지 시퀀스로 사용할 수 있습니다.

## 사례 연구: 속성의 기술 목록을 기반으로 하는 제품 설명

여기서 입력 데이터를 자연어로 변환하는 것이 중요하며, 이는 우수한 성능으로 이어질 것입니다. 예를 들어, 다음 형식:

JSON

```
{  
    "prompt": "Item=handbag, Color=army_green, price=$99, size=S->",  
    "completion": "This stylish small green handbag will add a unique touch  
to your look, without costing you a fortune."  
}
```

다음과 같이 잘 작동하지 않습니다.

JSON

```
{  
    "prompt": "Item is a handbag. Colour is army green. Price is midrange.  
Size is small.->,"  
    "completion": "This stylish small green handbag will add a unique touch  
to your look, without costing you a fortune."  
}
```

고성능을 위해 제공된 설명을 기반으로 완료되었는지 확인합니다. 외부 콘텐츠를 자주 참조하는 경우 이러한 콘텐츠를 자동화된 방식으로 추가하면 성능이 개선됩니다. 설명이 이미지를 기반으로 하는 경우 알고리즘을 사용하여 이미지의 텍스트 설명을 추출하는 데 도움이 될 수 있습니다. 완료는 한 문장 길이이므로 유추하는 동안 중지 시퀀스로 `.`을 사용할 수 있습니다.

## 종료된 생성 열기

이러한 형식의 문제에 대해 다음을 권장합니다.

- 프롬프트를 비워 둡니다.
- 구분 기호가 필요하지 않습니다.
- 일반적으로 많은 수(최소한 수천 개)의 예가 필요합니다.
- 예에서 의도한 영역이나 원하는 음성 톤을 포함하는지 확인합니다.

## 사례 연구: 회사 음성 유지

많은 회사들이 특정 음성으로 생성된 고품질 콘텐츠를 대량으로 보유하게 될 것입니다. 이상적으로는 API의 모든 생성이 다양한 사용 사례에 대한 음성을 따라야 합니다. 여기에서 프롬프트를 비워두고 회사 음성의 좋은 예인 모든 문서를 입력하는 트릭을 사용할 수 있습니다. 미세 조정된 모델은 베이스 모델에 사용된 것과 유사한 프롬프트로 다양한 사용 사례를 해결하는 데 사용할 수 있지만 출력은 이전보다 훨씬 더 회사의 음성을 따를 것입니다.

JSON

```
{"prompt": "", "completion": " <company voice textual content>"}  
{"prompt": "", "completion": " <company voice textual content2>"}
```

비슷한 기술을 사용하여 특정 성격, 연설 스타일 및 캐릭터가 말하는 항목을 가진 가상 캐릭터를 만들 수 있습니다.

생성 작업은 모델에서 완료를 요청할 때 학습 데이터가 누출될 가능성이 있으므로 이를 적절하게 처리하기 위해 추가적인 주의가 필요합니다. 예를 들어 개인 정보나 중요한 회사 정보는 일반 정보로 대체되거나 애초에 미세 조정에 포함되지 않아야 합니다.

## 다음 단계

- 방법 [가이드](#)로 모델 미세 조정
- Azure OpenAI 서비스를 지원하는 [기본 모델](#)에 대해 자세히 알아보기

# 애플리케이션에 맞게 모델을 사용자 지정하는 방법 알아보기

아티클 • 2023. 04. 06.

Azure OpenAI 서비스를 사용하면 미세 조정이라는 프로세스를 사용하여 개인 데이터 세트에 맞게 모델을 조정할 수 있습니다. 이 사용자 지정 단계를 통해 다음을 제공하여 서비스를 최대한 활용할 수 있습니다.

- 프롬프트 설계에서 얻을 수 있는 것보다 더 높은 품질의 결과
- 프롬프트에 잘 맞을 수 있는 예제보다 더 많은 예제를 학습하는 기능
- 낮은 대기 시간 요청

사용자 지정 모델은 특정 프롬프트 및 구조에 대한 모델 가중치를 학습하여 퓨샷 학습 방법을 개선합니다. 사용자 지정 모델을 사용하면 프롬프트에 예를 제공하지 않고도 더 많은 작업에서 더 나은 결과를 얻을 수 있습니다. 따라서 전송되는 텍스트가 감소하고 모든 API 호출에서 처리되는 토큰이 감소하므로 비용이 절감되고 요청 대기 시간이 향상됩니다.

## ① 참고

최신 12-01-2022 GA API의 `create fine tunes` 명령에는 호환성이 손상되는 변경이 있습니다. 최신 명령 구문은 [참조 설명서](#)를 참조하세요.

## 필수 구성 요소

- Azure 구독 - [체험 구독 만들기](#)
- 원하는 Azure 구독의 Azure OpenAI에 부여된 액세스 권한
  - 현재 이 서비스에 대한 액세스 권한은 애플리케이션에서만 부여됩니다.  
<https://aka.ms/oai/access>에서 양식을 작성하여 Azure OpenAI에 대한 액세스를 신청할 수 있습니다. 문제가 있는 경우 이 리포지토리에서 문제를 업니다.
- Azure OpenAI 리소스
  - 리소스를 만드는 방법에 대한 자세한 내용은 [Azure OpenAI를 사용하여 리소스 만들기 및 모델 배포](#)를 참조하세요.

## 워크플로 미세 조정

Azure OpenAI Studio의 미세 조정 워크플로에는 다음 단계가 필요합니다.

1. 학습 및 유효성 검사 데이터 준비
2. Azure OpenAI Studio에서 사용자 지정 모델 만들기 마법사를 사용하여 사용자 지정 모델 학습시키기
  - a. [기본 모델 선택](#)
  - b. [학습 데이터 선택](#)
  - c. 필요에 따라 [유효성 검사 데이터 선택](#)
  - d. 필요에 따라 미세 조정 작업의 [고급 옵션 선택](#)
  - e. [선택한 내용을 검토하고 새로운 사용자 지정 모델 학습시키기](#)
3. 사용자 지정 모델의 상태 확인
4. 사용할 사용자 지정 모델 배포
5. 사용자 지정 모델 사용
6. 필요에 따라 사용자 지정 모델의 성능 및 적합성 분석

## 학습 및 유효성 검사 데이터 준비

학습 데이터 및 유효성 검사 데이터 세트는 원하는 모델 성능에 대한 입력 및 출력 예제로 구성됩니다.

사용하는 학습 및 유효성 검사 데이터는 반드시 각 줄이 단일 프롬프트-완료 쌍을 나타내는 JSONL(JSON Line) 문서 형식이어야 합니다. OpenAI CLI(명령줄 인터페이스)에는 학습 데이터의 유효성을 검사하고, 권장 사항을 제공하고, 학습 데이터를 미세 조정이 가능한 JSONL 파일로 다시 포맷하는 데이터 [준비 도구](#)가 포함되어 있습니다.

다음은 학습 데이터 형식의 예입니다.

JSON

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

JSONL 형식 외에도, 학습 및 유효성 검사 데이터 파일은 UTF-8로 인코딩되어야 하고 BOM(바이트 순서 표시)을 포함해야 하며, 파일 크기는 200MB 미만이어야 합니다. 학습 데이터의 포맷에 대한 자세한 내용은 [미세 조정을 위해 데이터 세트를 준비하는 방법 알아보기](#)를 참조하세요.

## 학습 및 유효성 검사 데이터 세트 만들기

미세 조정을 위한 프롬프트 및 완료를 설계하는 것은 [GPT-3 베이스 모델](#)과 함께 사용할 프롬프트를 설계하는 것과 다릅니다. 완료 호출을 위한 프롬프트는 자세한 지침 또는 퓨

샷 학습 기술을 사용할 때가 많으며, 여러 예제로 구성됩니다. 미세 조정의 경우 각 학습 예제를 단일 입력 프롬프트와 원하는 완료 출력으로 구성하는 것이 좋습니다. 동일한 프롬프트에는 자세한 지침이나 여러 완료 예제를 제공할 필요가 없습니다.

학습 사례가 많을수록 좋습니다. 학습 예제가 200개 이상 있는 것이 좋습니다. 일반적으로 데이터 세트 크기가 두 배로 증가할 때마다 모델 품질이 선형적으로 증가한다는 사실을 발견했습니다.

다양한 작업의 학습 데이터를 준비하는 방법에 대한 자세한 내용은 [미세 조정을 위해 데이터 세트를 준비하는 방법 알아보기](#)를 참조하세요.

## OpenAI CLI 데이터 준비 도구

많은 데이터 준비 단계를 지원하려면 OpenAI의 CLI(명령줄 인터페이스)를 사용하는 것이 좋습니다. OpenAI는 유효성을 검사하고, 권장 사항을 제공하고, 데이터를 미세 조정이 가능한 JSONL 파일로 다시 포맷하는 도구를 개발했습니다.

CLI를 설치하려면 다음 Python 명령을 실행합니다.

콘솔

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

데이터 준비 도구를 사용하여 학습 데이터를 분석하려면 다음 Python 명령을 실행하고, <LOCAL\_FILE>을 분석할 학습 데이터 파일의 전체 경로 및 파일 이름으로 바꿉니다.

콘솔

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

파일에 프롬프트 및 완료 열/키가 포함된 경우 이 도구는 다음과 같은 데이터 형식의 파일을 허용합니다.

- CSV(쉼표로 구분된 값)
- TSV(탭으로 구분된 값)
- XLSX(Microsoft Excel 통합 문서)
- JSON(JavaScript Object Notation)
- JSONL(JSON Lines)

이 도구는 제안된 변경 내용을 구현하는 프로세스를 안내한 다음, 학습 데이터를 다시 포맷하고 미세 조정이 가능한 JSONL 파일에 출력을 저장합니다.

# 사용자 지정 모델 만들기 마법사 사용

Azure OpenAI Studio는 사용자 지정 모델 만들기 마법사를 제공하므로, 대화형으로 Azure 리소스에 대해 미세 조정된 모델을 만들고 학습시킬 수 있습니다.

## Azure OpenAI Studio로 이동

<https://oai.azure.com/> 에서 Azure OpenAI Studio로 이동한 다음, Azure OpenAI 리소스에 액세스할 수 있는 자격 증명으로 로그인합니다. 로그인 워크플로 중에 적절한 디렉터리, Azure 구독 및 Azure OpenAI 리소스를 선택합니다.

### 방문 페이지

먼저 Azure OpenAI 스튜디오의 기본 페이지로 이동됩니다. 이 페이지에서 사용자 지정 모델의 미세 조정을 시작할 수 있습니다.

다음 그림에 강조 표시된 것처럼 방문 페이지의 **배포 및 모델 관리** 섹션에서 **사용자 지정 모델 미세 조정 시작** 단추를 선택하여 사용자 지정 모델의 미세 조정을 시작합니다.

#### ① 참고

아직 리소스에 배포된 모델이 없는 경우 경고가 표시됩니다. 새 사용자 지정 모델을 미세 조정하고 배포할 것이므로 해당 경고를 무시하고 모델을 미세 조정하면 됩니다.

Cognitive Services | Azure OpenAI Studio - Preview

Azure OpenAI Studio

Test User test-resource (South Central US, 50) TU

Privacy & cookies

## Get started with Azure OpenAI

Perform a wide variety of natural language tasks with Azure OpenAI, including copywriting, summarization, parsing unstructured text, classification, and translation.

### Explore examples for prompt completion



Summarize Text  
Summarize text by adding a 'tldr' to the end of a text passage.  
[Learn more](#)



Classify Text  
Classify items into categories provided at inference time.  
[Learn more](#)



Natural Language to SQL  
Translate natural language to SQL queries.  
[Learn more](#)



Generate New Product Names  
Create product names from examples words.  
[Learn more](#)

### Manage your deployments and models

Experiment with prompt completions  
Try out the completions endpoint by writing a prompt and generating a response. Set different parameters values to adjust how the model responds.  
[Go to playground](#)

Customize a model with fine-tuning  
Fine-tune a custom model to increase reliability for a wide variety of use cases while decreasing costs and speeding up processing times.  
[Start fine-tuning a custom model](#)

Manage deployments in your resource  
Create deployments to explore the model capabilities.  
[Go to Deployments](#)

Manage performance results  
Upload datasets to use when creating custom models, and view performance and fine-tune results from training and validation data.  
[Go to File management](#)

## 모델 페이지에서 마법사 시작

사용자 지정 모델을 만들려면 다음 그림에 강조 표시된 것처럼 **모델 페이지의 제공된 모델** 섹션에서 **사용자 지정 모델 만들기** 단추를 선택하여 **사용자 지정 모델 만들기** 마법사를 시작합니다.

Cognitive Services | Azure OpenAI Studio - Preview

Azure OpenAI Studio > Models

Test User test-resource (South Central US, 50) TU

Privacy & cookies

### Models

Azure OpenAI is powered by models with different capabilities and price points. Deploy one of the provided base models to try it out in [Playground](#) or train a custom model to your specific use case and data for better performance and more accurate results. [Learn more about the different types of provided models](#)

#### Provided models

Deploy model [Create customized model](#) Refresh

| Model name                   | Created at        | Status    |
|------------------------------|-------------------|-----------|
| ada                          | 2/28/2022 4:00 PM | Succeeded |
| babbage                      | 2/28/2022 4:00 PM | Succeeded |
| code-cushman-001             | 1/21/2022 4:00 PM | Succeeded |
| code-search-ada-code-001     | 5/19/2022 5:00 PM | Succeeded |
| code-search-ada-text-001     | 5/19/2022 5:00 PM | Succeeded |
| code-search-babbage-code-001 | 5/19/2022 5:00 PM | Succeeded |
| code-search-babbage-text-001 | 5/19/2022 5:00 PM | Succeeded |

## 기본 모델 선택

사용자 지정 모델을 만드는 첫 번째 단계는 기본 모델을 선택하는 것입니다. **기본 모델** 창에서 사용자 지정 모델에 사용할 기본 모델을 선택할 수 있으며, 선택하는 기본 모델은 모델의 성능과 비용 모두에 영향을 줍니다. 사용 가능한 다음 기본 모델 중 하나로 사용자 지정 모델을 만들 수 있습니다.

- ada
- babbage
- curie
- code-cushman-001\*
- davinci\*

\* 현재 신규 고객은 사용할 수 없습니다.

미세 조정할 수 있는 기본 모델에 대한 자세한 내용은 [모델](#)을 참조하세요. 다음 그림처럼 **기본 모델 유형** 드롭다운에서 기본 모델을 선택하고 **다음**을 선택하여 계속합니다.

### Create customized model

Base model

Every fine-tuned model starts from a base model which influences both the performance of the model and the cost of running your custom model.  
[Learn more about each parameter here](#)

Base model type

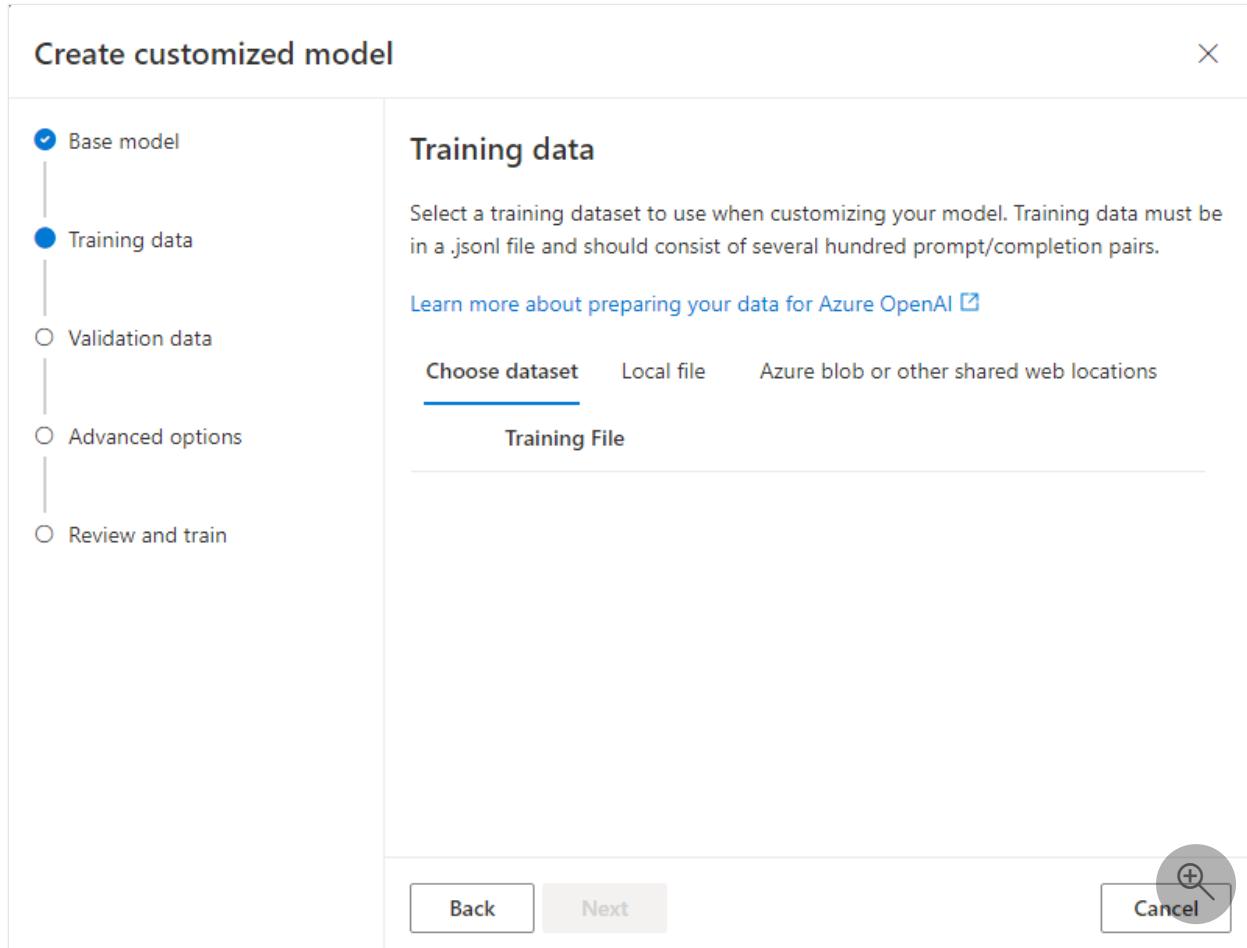
- ada
- babbage
- curie
- davinci
- code-cushman-001

Next

Cancel

## 학습 데이터 선택

다음 단계는 모델을 사용자 지정할 때 사용할 준비된 기존 학습 데이터를 선택하거나 준비된 새 학습 데이터를 업로드하는 것입니다. 다음 그림처럼 **학습 데이터** 창에는 이전에 업로드한 기존 데이터 세트가 표시되고, 새 학습 데이터를 업로드할 수 있는 옵션이 제공됩니다.



학습 데이터가 서비스에 이미 업로드된 경우 **데이터 세트 선택**을 선택한 다음, **학습 데이터** 창에 표시된 목록에서 파일을 선택합니다. 아직 업로드하지 않았으면 **로컬 파일**을 선택하여 **로컬 파일에서 학습 데이터를 업로드**하거나, **Azure Blob 또는 기타 공유 웹 위치**를 선택하여 **Azure Blob 또는 다른 공유 웹 위치에서 학습 데이터를 가져옵니다**.

대용량 데이터 파일의 경우 Azure Blob 저장소에서 가져오는 것이 좋습니다. 요청이 원자성이어서 다시 시도하거나 다시 시작할 수 없기 때문에 대용량 파일은 멀티파트 양식을 통해 업로드할 때 불안정해질 수 있습니다. Azure Blob 스토리지에 대한 자세한 내용은 [Azure Blob 스토리지란?](#)을 참조하세요.

### ① 참고

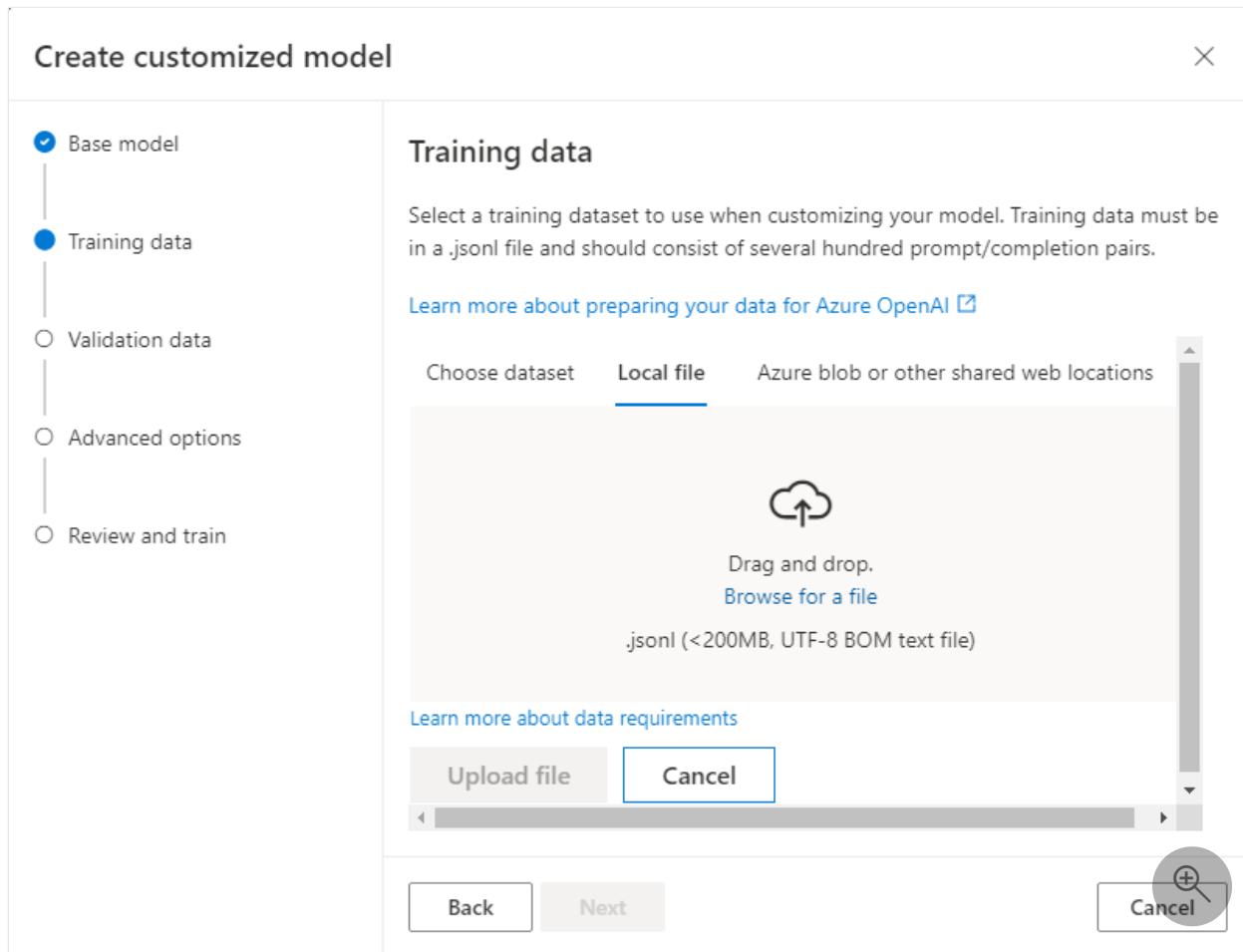
학습 데이터 파일은 JSONL 파일로 포맷되고 BOM(바이트 순서 표시)이 있는 UTF-8로 인코딩되어야 하며, 크기는 200MB 미만이어야 합니다.

## 로컬 파일에서 학습 데이터를 업로드하는 방법

다음 방법 중 하나를 사용하여 로컬 파일의 새 학습 데이터 세트를 서비스에 업로드할 수 있습니다.

- **학습 데이터** 창의 클라이언트 영역으로 파일을 끌어서 놓은 다음, **파일 업로드**를 선택합니다.
- **학습 데이터** 창의 클라이언트 영역에서 **파일 찾아보기**를 선택하고 **열기** 대화 상자에서 업로드할 파일을 선택한 다음, **파일 업로드**를 선택합니다.

학습 데이터 세트를 선택하고 업로드한 후, 필요에 따라 **다음**을 선택하여 **유효성 검사 데이터를 선택합니다.**



## Azure Blob 저장소에서 학습 데이터를 가져오는 방법

다음 그림처럼 파일의 이름과 위치를 입력하여 Azure Blob 또는 다른 공유 웹 위치에서 학습 데이터 세트를 가져올 수 있습니다. **파일 이름**에 파일 이름을 입력하고 **파일 위치**에 Azure Blob URL, Azure Storage SAS(공유 액세스 서명) 또는 파일이 들어 있는 액세스 가능한 공유 웹 위치의 링크를 입력한 다음, **파일 업로드**를 선택하여 학습 데이터 세트를 서비스로 가져옵니다.

학습 데이터 세트를 선택하고 업로드한 후, 필요에 따라 **다음**을 선택하여 **유효성 검사 데이터를 선택합니다.**

## Create customized model

X

- Base model
- Training data
- Validation data
- Advanced options
- Review and train

### Training data

Select a training dataset to use when customizing your model. Training data must be in a .jsonl file and should consist of several hundred prompt/completion pairs.

[Learn more about preparing your data for Azure OpenAI](#)

Choose dataset    Local file    **Azure blob or other shared web locations**

File name \*

Enter the name of the file

File location \*

Input Azure Blob public access URL, SAS, or any other shared web link

.jsonl (<200MB, UTF-8 BOM text file)

[Learn more about public access to Azure Blob](#)

[Learn more about Azure Blob SAS \(Shared Access Signature\)](#)

[Upload file](#)

[Cancel](#)

[Back](#)

[Next](#)

[Cancel](#)

## 유효성 검사 데이터 선택

이제 필요에 따라 미세 조정된 모델의 학습 프로세스에서 유효성 검사 데이터를 사용하도록 선택할 수 있습니다. 유효성 검사 데이터를 사용하지 않으려는 경우 **다음**을 선택하고 모델에 대한 고급 옵션을 선택할 수 있습니다. 사용하려는 경우 유효성 검사 데이터 세트가 있으면 준비된 기존 유효성 검사 데이터를 선택하거나, 모델을 사용자 지정할 때 사용할 새로운 준비된 유효성 검사 데이터를 업로드할 수 있습니다. 다음 그림처럼 **유효성 검사 데이터** 창에는 이전에 업로드한 기존 데이터 세트가 표시되고, 새 유효성 검사 데이터를 업로드할 수 있는 옵션이 제공됩니다.

Create customized model X

Base model

Training data

Validation data

Advanced options

Review and train

**Validation data**

Select up to one validation dataset to use when iteratively assessing your customized model's performance during training. Validation data must be in a .jsonl file and should be representative of the training data without repeating any of it.

[Learn more about preparing your data for Azure OpenAI](#)

Choose dataset    Local file    Azure blob or other shared web locations

Validation File

training.jsonl

Back Next Cancel

유효성 검사 데이터가 서비스에 이미 업로드된 경우 **데이터 세트 선택**을 선택한 다음, **유효성 검사 데이터** 항에 표시된 목록에서 파일을 선택합니다. 아직 업로드하지 않았으면 **로컬 파일**을 선택하여 **로컬 파일에서 유효성 검사 데이터를 업로드**하거나, **Azure Blob** 또는 **기타 공유 웹 위치**를 선택하여 **Azure Blob 또는 다른 공유 웹 위치에서 유효성 검사 데이터를 가져옵니다**.

대용량 데이터 파일의 경우 Azure Blob 저장소에서 가져오는 것이 좋습니다. 요청이 원자성이어서 다시 시도하거나 다시 시작할 수 없기 때문에 대용량 파일은 멀티파트 양식을 통해 업로드할 때 불안정해질 수 있습니다.

### ① 참고

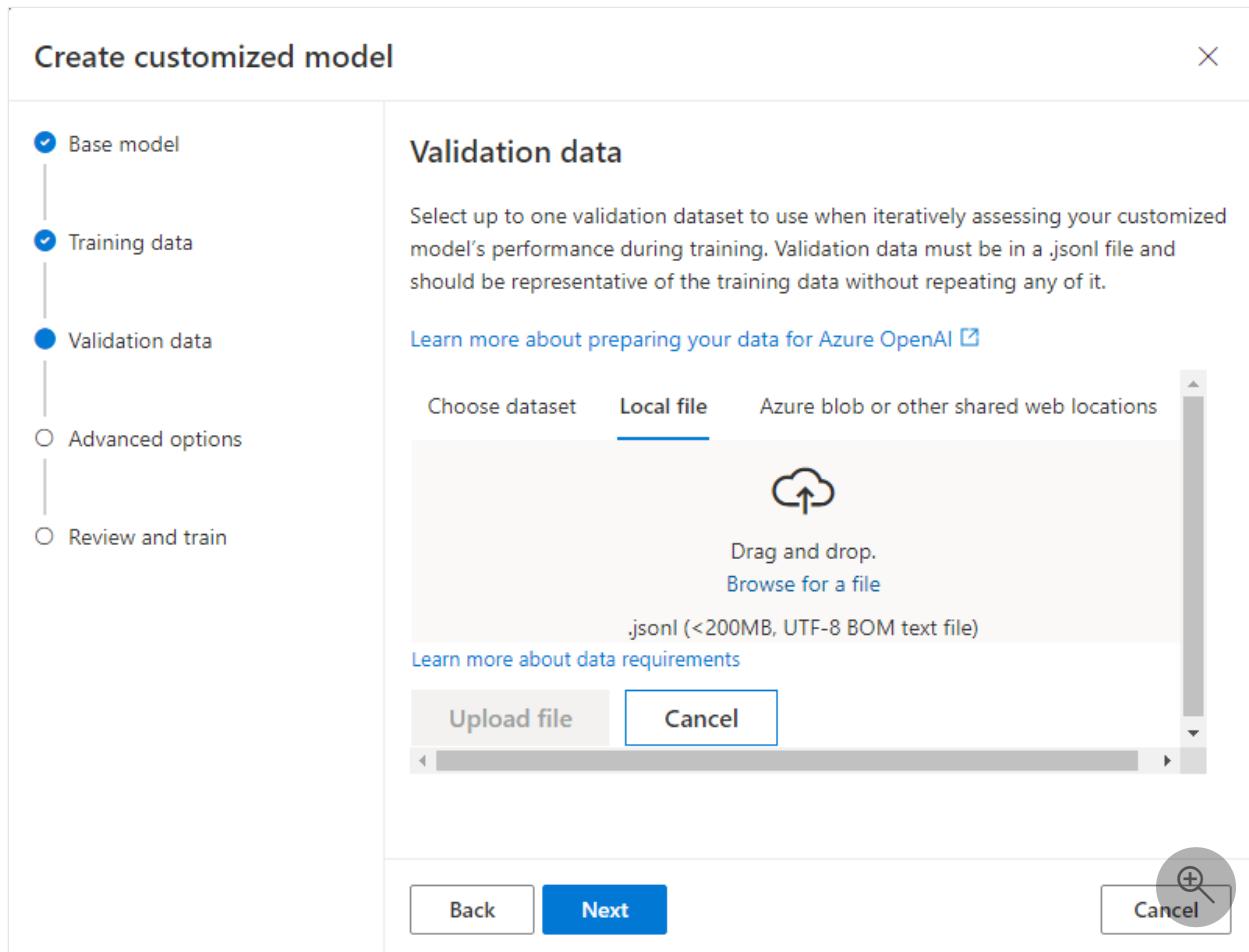
학습 데이터 파일과 마찬가지로, 유효성 검사 파일은 JSONL 파일로 포맷되고 BOM(바이트 순서 표시)이 있는 UTF-8로 인코딩되어야 하며, 크기는 200MB 미만이어야 합니다.

## 로컬 파일에서 유효성 검사 데이터를 업로드하는 방법

다음 방법 중 하나를 사용하여 로컬 파일의 새 유효성 검사 데이터 세트를 서비스에 업로드할 수 있습니다.

- **유효성 검사 데이터** 창의 클라이언트 영역으로 파일을 끌어서 놓은 다음, **파일 업로드**를 선택합니다.
- **유효성 검사 데이터** 창의 클라이언트 영역에서 **파일 찾아보기**를 선택하고 **열기 대화** 상자에서 업로드할 파일을 선택한 다음, **파일 업로드**를 선택합니다.

유효성 검사 데이터 세트를 업로드한 후에는 필요에 따라 **다음**을 선택하여 **고급 옵션을 선택**합니다.



## Azure Blob 저장소에서 유효성 검사 데이터를 가져오는 방법

다음 그림처럼 파일의 이름과 위치를 입력하여 Azure Blob 또는 다른 공유 웹 위치에서 유효성 검사 데이터 세트를 가져올 수 있습니다. **파일 이름**에 파일 이름을 입력하고 **파일 위치**에 Azure Blob URL, Azure Storage SAS(공유 액세스 서명) 또는 파일이 들어 있는 액세스 가능한 공유 웹 위치의 링크를 입력한 다음, **파일 업로드**를 선택하여 유효성 검사 데이터 세트를 서비스로 가져옵니다.

유효성 검사 데이터 세트를 가져온 후에는 필요에 따라 **다음**을 선택하여 **고급 옵션을 선택**합니다.

## Create customized model

X

- Base model
- Training data
- Validation data
- Advanced options
- Review and train

### Validation data

Select up to one validation dataset to use when iteratively assessing your customized model's performance during training. Validation data must be in a .jsonl file and should be representative of the training data without repeating any of it.

[Learn more about preparing your data for Azure OpenAI](#)

Choose dataset    Local file    **Azure blob or other shared web locations**

File name \*

Enter the name of the file

File location \*

Input Azure Blob public access URL, SAS, or any other shared web link

.jsonl (<200MB, UTF-8 BOM text file)

[Learn more about public access to Azure Blob](#)

[Learn more about Azure Blob SAS \(Shared Access Signature\)](#)

[Upload file](#)

[Cancel](#)

[Back](#)

[Next](#)

[Cancel](#)

## 고급 옵션 선택

마법사가 미세 조정된 모델을 학습시키기 위해 실행하는 미세 조정 작업의 하이퍼 매개 변수에 기본값을 사용할 수도 있고, 다음 그림처럼 **고급 옵션** 창에서 사용자 지정 요구 사항에 맞게 하이퍼 매개 변수를 조정할 수도 있습니다.

### Create customized model

Base model  
Training data  
Validation data  
**Advanced options**  
Review and train

#### Advanced options

You can set additional parameters by selecting the advanced option below. These parameters will impact both the performance and training time of your job.

[Learn more about each parameter here](#)

Default  Advanced

[Back](#) [Next](#) [Cancel](#)

기본값을 선택하여 미세 조정 작업의 기본값을 사용하거나, 다음 그림처럼 고급을 선택하여 하이퍼 매개 변수 값을 표시하고 편집합니다.

### Create customized model

Base model  
Training data  
Validation data  
**Advanced options**  
Review and train

#### Advanced options

You can set additional parameters by selecting the advanced option below. These parameters will impact both the performance and training time of your job.

[Learn more about each parameter here](#)

Default  Advanced

Number of epochs ⓘ  [2]

Batch size ⓘ  [4]

Learning rate multiplier: ⓘ  [1]

Prompt loss weight ⓘ  [0.1]

[Back](#) [Next](#) [Cancel](#)

다음 하이퍼 매개 변수를 사용할 수 있습니다.

| 매개 변수 이름    | 설명                                                                                                                                |
|-------------|-----------------------------------------------------------------------------------------------------------------------------------|
| Epoch 의 수.  | 모델을 학습할 Epoch의 수입니다. Epoch는 학습 데이터 세트를 통한 하나의 전체 주기를 나타냅니다.                                                                       |
| Batch 크기    | 학습에 사용할 일괄 처리 크기입니다. 일괄 처리 크기는 단일 전진 및 후진 계산법을 학습 하는 데 사용되는 학습 예의 수입니다.                                                           |
| 학습 속도 승수    | 학습에 사용할 학습 속도 승수입니다. 미세 조정 학습 속도는 사전 학습에 사용된 원래 학습 속도에 이 값을 곱한 값입니다.                                                              |
| 프롬프트 손실 가중치 | 프롬프트 토큰의 손실에 사용할 가중치입니다. 이 값은 모델이 프롬프트를 생성하기 위해 학습하는 정도를 제어합니다(항상 가중치가 1.0인 완료와 비교하여). 이 값을 높이면 완료 가 짧을 때 학습에 안정화 효과를 추가할 수 있습니다. |

이러한 하이퍼 매개 변수에 대한 자세한 내용은 [REST API 설명서의 미세 조정 작업 만들기](#) 섹션을 참조하세요.

기본 옵션 또는 고급 옵션을 선택한 후에는 다음을 선택하여 선택한 내용을 검토하고 미세 조정된 모델을 학습시킵니다.

## 선택한 내용 검토 및 모델 학습

마법사의 **검토 및 학습** 창에는 다음 그림처럼 미세 조정된 모델의 사용자 지정 모델 만들기 마법사에서 선택한 내용에 대한 정보가 표시됩니다.

Create customized model

Base model  
 Training data  
 Validation data  
 Advanced options  
 Review and train

**Review and train**

**Base model:** davinci  
**Training data:** training.jsonl  
**Validation data:** validation.jsonl

Back      Save and close      Cancel

모델을 학습시킬 준비가 되었으면 저장 후 닫기를 선택하여 미세 조정 작업을 시작하고 모델 페이지로 돌아갑니다.

## 사용자 지정 모델의 상태 확인

다음 그림처럼 모델 페이지의 사용자 지정 모델 탭에 사용자 지정 모델에 대한 정보가 표시됩니다. 이 탭에는 사용자 지정 모델에 대한 미세 조정 작업의 상태와 작업 ID에 대한 정보가 포함되어 있습니다. 작업이 완료되면 결과 파일의 파일 ID도 표시됩니다.

Azure OpenAI Studio > Models

**Models**

Azure OpenAI is powered by models with different capabilities and price points. Deploy one of the provided base models to try it out in [Playground](#) or train a custom model to your specific use case and data for better performance and more accurate results.

[Learn more about the different types of provided models](#)

| Customized models                                                                                                   | Provided models             |            |         |                                     |
|---------------------------------------------------------------------------------------------------------------------|-----------------------------|------------|---------|-------------------------------------|
| <a href="#">Deploy model</a> <a href="#">Create customized model</a> <a href="#">Delete</a> <a href="#">Refresh</a> | <input type="text"/> Search |            |         |                                     |
| Model name                                                                                                          | Create...                   | Base model | Status  | Training job Id                     |
| ft-66aa4cc216694                                                                                                    | 9/7/2...                    | davinci    | Running | ft-66aa4cc2166949beae9facb7f258510b |

미세 조정 작업을 시작한 후 완료하는 데 시간이 걸릴 수 있습니다. 사용자 작업은 시스템의 다른 작업 뒤에 대기할 수 있으며 모델 및 데이터 세트 크기에 따라 모델을 학습하는데 몇 분 또는 몇 시간이 걸릴 수 있습니다. **모델** 페이지의 **사용자 지정 모델** 탭에 있는 **상태** 열에서 사용자 지정 모델에 대한 미세 조정 작업의 상태를 확인하고, **새로 고침**을 선택하여 이 페이지의 정보를 업데이트할 수 있습니다.

또한 **모델** 페이지의 **모델 이름** 열에서 모델 이름을 선택하여 미세 조정 작업의 상태, 학습 결과, 학습 이벤트 및 작업에 사용된 하이퍼 매개 변수를 포함하여 사용자 지정 모델에 대한 자세한 정보를 표시할 수 있습니다. 다음 그림처럼 **새로 고침** 단추를 선택하여 모델 정보를 새로 고칠 수 있습니다.

Azure OpenAI Studio > Models > ft-e4b459a57a94410a8a4c5a10a15b6063

Status: ✓ Training Succeeded

Finished training on: 9/23/2022 8:51 AM  
Training file: training.jsonl  
Base model: curie  
Total training time: 25 minutes, 53 seconds

Statistics:  
Total tokens: 1,088  
Total examples: 64

Download results   Download training file

Training results

Training loss   Training token accuracy   Validation loss   Validation token accuracy

Training loss

| Loss   | Training loss |
|--------|---------------|
| 1.2040 | 1.2040        |
| 1.2030 | 1.2030        |
| 1.2020 | 1.2020        |

[모델] 페이지에서 **학습 파일 다운로드**를 선택하여 모델에 사용한 학습 데이터를 다운로드할 수도 있고, **결과 다운로드**를 선택하여 모델의 미세 조정 작업에 연결된 결과 파일을 다운로드하고 **사용자 지정 모델의 학습 및 유효성 검사 성능을 분석할** 수도 있습니다.

## 사용자 지정 모델 배포

미세 조정 작업이 성공하면 **모델** 창에서 사용자 지정 모델을 배포할 수 있습니다. 완료 호출에 사용할 수 있도록 사용자 지정 모델을 배포해야 합니다.

### ⓘ 중요

사용자 지정된 모델을 배포한 후 배포가 15일(15일) 이상 비활성 상태로 유지되면 배포가 자동으로 삭제됩니다. 모델이 15일 이상 전에 배포되었고 15일 연속으로 완료

또는 채팅 완료 호출이 이루어지지 않은 경우 사용자 지정된 모델의 배포는 "비활성"입니다. 비활성 배포 삭제는 기본 사용자 지정 모델을 삭제하거나 영향을 주지 않으며 사용자 지정된 모델을 언제든지 다시 배포할 수 있습니다. [Azure OpenAI 서비스 가격 책정](#)에 설명된 대로 배포된 각 사용자 지정(미세 조정된) 모델은 완료 또는 채팅 완료 호출이 모델에 대해 이루어지는지 여부에 관계없이 시간당 호스팅 비용이 발생합니다. Azure OpenAI를 사용하여 비용을 계획하고 관리하는 방법에 대한 자세한 내용은 [비용 관리 가이드](#)를 참조하세요.

## ① 참고

하나의 사용자 지정 모델에는 하나의 배포만 허용됩니다. 이미 배포된 사용자 지정 모델을 선택하면 오류 메시지가 표시됩니다.

사용자 지정 모델을 배포하려면 다음 그림처럼 배포할 사용자 지정 모델을 선택하고 **모델 배포**를 선택합니다.

The screenshot shows the Azure OpenAI Studio interface with the 'Models' section selected. On the left sidebar, 'Models' is also highlighted. The main content area displays a table of customized models. A red box highlights the 'Deploy model' button, which is located at the top of the table's toolbar. The table has columns for Model name, Created, Base model, Status, Training job Id, and Result file Id. One row is shown with the values: curie.ft-ad0ee2d6, 9/12/..., curie, Succeeded, ft-ad0ee2d614034bcfaf0043d3b89584f7, and file-6bebba158a82452. A search bar and a magnifying glass icon are also visible on the right side of the table.

사용자 지정 모델의 배포 이름을 입력할 수 있는 **모델 배포** 대화 상자가 표시됩니다. **배포 이름**에 이름을 입력한 다음, **만들기**를 선택하여 사용자 지정 모델의 배포를 시작합니다.

## Deploy model

X

Set up a deployment to make API calls against a provided base model or a custom model. Finished deployments are available for use. Your deployment status will move to succeeded when the deployment is complete and ready for use.

- (i) Only one deployment is permitted per model. The models with existing deployments are disabled.

Model name (i)

curie.ft-ad0ee2d614034bcfaf0043d3b89584f7

▼

Deployment name (i)

Create

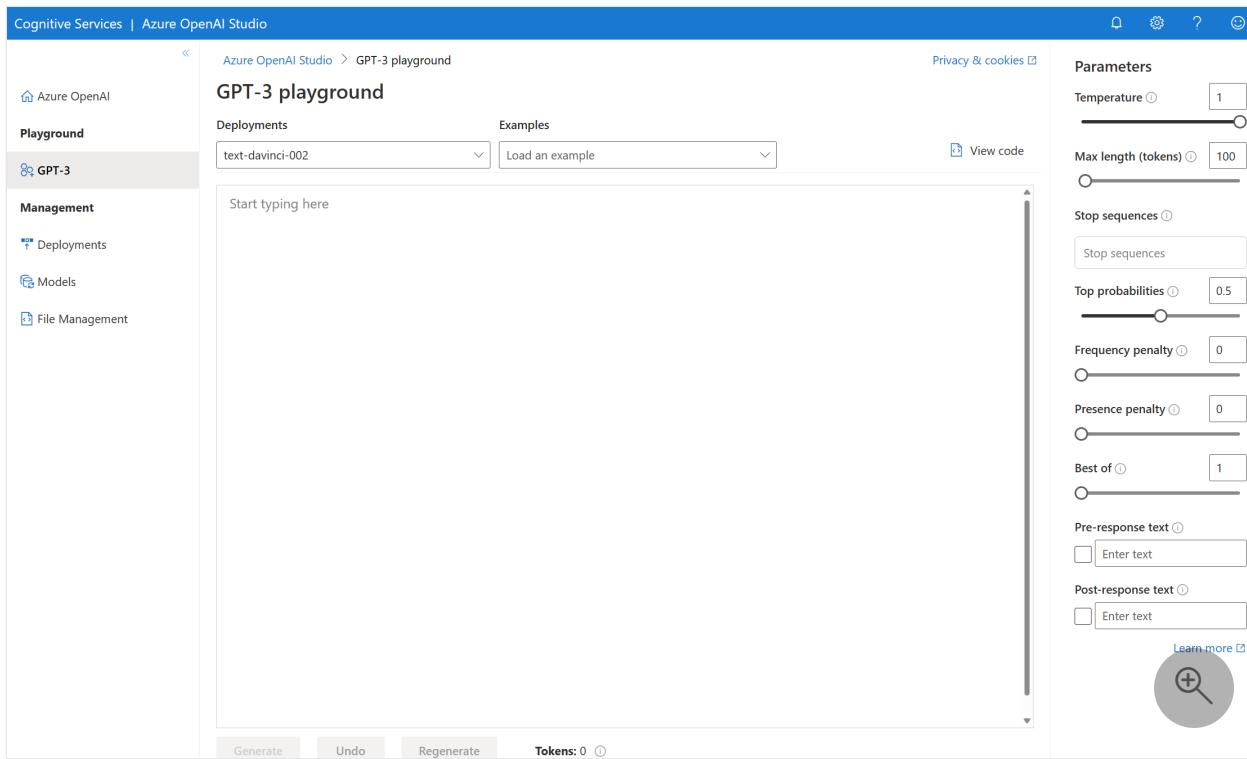
Cancel



Azure OpenAI Studio의 배포 창에서 배포 진행률을 모니터링할 수 있습니다.

## 배포된 사용자 지정 모델 사용

모델이 배포되었으면 배포된 다른 모델처럼 사용할 수 있습니다. 예를 들어 다음 그림처럼 Azure OpenAI Studio의 플레이그라운드 창을 사용하여 새 배포를 실험할 수 있습니다. 배포된 다른 모델과 마찬가지로 사용자 지정 모델에서 온도 및 빈도 페널티와 같은 동일한 매개 변수를 계속 사용할 수 있습니다.



## ① 참고

모든 애플리케이션과 마찬가지로 게시 전에 검토 프로세스가 필요합니다.

## 사용자 지정 모델 분석

미세 조정 작업이 완료되면 Azure OpenAI는 `results.csv`라는 결과 파일을 각 미세 조정 작업에 연결합니다. 이 결과 파일을 사용하여 사용자 지정 모델의 학습 및 유효성 검사 성능을 분석할 수 있습니다. Azure OpenAI Studio **모델** 창의 **결과 파일 ID** 열에 각 사용자 지정 모델에 대한 결과 파일의 파일 ID가 나열됩니다. 이 파일 ID를 사용하여 Azure OpenAI Studio의 **파일 관리** 창에서 결과 파일을 식별하고 다운로드할 수 있습니다.

결과 파일은 미세 조정 작업에서 수행하는 각 학습 단계에 대한 머리글 행과 행을 포함하는 CSV 파일입니다. 결과 파일에는 다음 열이 포함되어 있습니다.

| 열 이름                          | 설명                                                                                                                                                        |
|-------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------|
| <code>step</code>             | 학습 단계의 수입니다. 학습 단계는 학습 데이터 일괄 처리에 대한 정방향 및 역방향 단일 패스를 나타냅니다.                                                                                              |
| <code>elapsed_tokens</code>   | 사용자 지정 모델이 지금까지 확인한 토큰 수(반복 포함)입니다.                                                                                                                       |
| <code>elapsed_examples</code> | 모델이 지금까지 확인한 예제의 수(반복 포함)입니다. 각 예제는 해당 단계의 학습 데이터 일괄 처리에서 하나의 요소를 나타냅니다. 예를 들어 <b>고급 옵션</b> 창에서 <b>일괄 처리 크기</b> 매개 변수가 32로 설정되면 이 값은 각 학습 단계에서 32씩 증가합니다. |

| 열 이름                         | 설명                                                                                                                                                                                     |
|------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| training_loss                | 학습 일괄 처리의 손실입니다.                                                                                                                                                                       |
| training_sequence_accuracy   | 모델의 예측 토큰이 실제 완료 토큰과 정확히 일치하는 학습 일괄 처리의 완료 비율입니다.<br>예를 들어 일괄 처리 크기가 3으로 설정되고 데이터에 완료 [[1, 2], [0, 5], [4, 2]] 가 포함된 경우 모델이 [[1, 1], [0, 5], [4, 2]] 를 예측하면 이 값은 0.67(2/3)로 설정됩니다.     |
| training_token_accuracy      | 모델에서 올바르게 예측한 학습 일괄 처리의 토큰 비율입니다.<br>예를 들어 일괄 처리 크기가 3으로 설정되고 데이터에 완료 [[1, 2], [0, 5], [4, 2]] 가 포함된 경우 모델이 [[1, 1], [0, 5], [4, 2]] 를 예측하면 이 값은 0.83(5/6)으로 설정됩니다.                    |
| validation_loss              | 유효성 검사 일괄 처리의 손실입니다.                                                                                                                                                                   |
| validation_sequence_accuracy | 모델의 예측 토큰이 실제 완료 토큰과 정확히 일치하는 유효성 검사 일괄 처리의 완료 비율입니다.<br>예를 들어 일괄 처리 크기가 3으로 설정되고 데이터에 완료 [[1, 2], [0, 5], [4, 2]] 가 포함된 경우 모델이 [[1, 1], [0, 5], [4, 2]] 를 예측하면 이 값은 0.67(2/3)로 설정됩니다. |
| validation_token_accuracy    | 모델에서 올바르게 예측한 유효성 검사 일괄 처리의 토큰 비율입니다.<br>예를 들어 일괄 처리 크기가 3으로 설정되고 데이터에 완료 [[1, 2], [0, 5], [4, 2]] 가 포함된 경우 모델이 [[1, 1], [0, 5], [4, 2]] 를 예측하면 이 값은 0.83(5/6)으로 설정됩니다.                |

## 배포, 사용자 지정 모델 및 학습 파일 정리

사용자 지정 모델을 완료했으면 배포 및 모델을 삭제해도 됩니다. 원한다면 서비스에 업로드한 학습 및 유효성 검사 파일까지 삭제해도 됩니다.

### 모델 배포 삭제

#### ① 중요

사용자 지정된 모델을 배포한 후 배포가 15일(15일) 이상 비활성 상태로 유지되면 배포가 자동으로 삭제됩니다. 모델이 15일 이상 전에 배포되었고 15일 연속으로 완료 또는 채팅 완료 호출이 이루어지지 않은 경우 사용자 지정된 모델의 배포는 "비활성"입니다. 비활성 배포 삭제는 기본 사용자 지정 모델을 삭제하거나 영향을 주지 않으며 사용자 지정된 모델을 언제든지 다시 배포할 수 있습니다. [Azure OpenAI 서비스 가격 책정](#)에 설명된 대로 배포된 각 사용자 지정(미세 조정된) 모델은 완료 또는 채팅 완료 호출이 모델에 대해 이루어지는지 여부에 관계없이 시간당 호스팅 비

용이 발생합니다. Azure OpenAI를 사용하여 비용을 계획하고 관리하는 방법에 대한 자세한 내용은 [비용 관리 가이드](#)를 참조하세요.

Azure OpenAI Studio의 **배포** 페이지에서 사용자 지정 모델의 배포를 삭제할 수 있습니다. 삭제할 배포를 선택한 다음, **삭제**를 선택하여 배포를 삭제합니다.

## 사용자 지정 모델 삭제

Azure OpenAI Studio의 **모델** 페이지에서 사용자 지정 모델을 삭제할 수 있습니다. **사용자 지정 모델** 탭에서 삭제할 사용자 지정 모델을 선택한 다음, **삭제**를 선택하여 사용자 지정 모델을 삭제합니다.

### ① 참고

기존 배포가 있는 사용자 지정 모델은 삭제할 수 없습니다. 사용자 지정 모델을 삭제 하려면 **모델 배포부터 삭제**해야 합니다.

## 학습 파일 삭제

필요에 따라 Azure OpenAI Studio의 **파일 관리** 페이지에서 학습을 위해 업로드한 학습 및 유효성 검사 파일과 학습 중에 생성된 결과 파일을 삭제할 수 있습니다. 삭제할 파일을 선택한 다음, **삭제**를 선택하여 파일을 삭제합니다.

## 다음 단계

- 모든 미세 조정 기능에 대해 자세히 알아보려면 전체 REST API 참조 설명서를 살펴 봅니다. [여기에서 전체 REST 설명서](#)를 찾을 수 있습니다.
- [여기에서 Python SDK 작업](#)에 대해 자세히 알아봅니다.

# 관리 ID를 사용하여 Azure OpenAI Service를 구성하는 방법

아티클 • 2023. 05. 11.

더 복잡한 보안 시나리오에는 Azure RBAC(Azure 역할 기반 액세스 제어)가 필요합니다. 이 문서에서는 Azure AD(Azure Active Directory)를 사용하여 OpenAI 리소스에 인증하는 방법을 설명합니다.

다음 섹션에서는 Azure CLI를 사용하여 역할을 할당하고 OpenAI 리소스를 호출하는 전달자 토큰을 가져옵니다. 작업하면서 어려움에 처할 경우를 위해 각 세션에서는 Azure Cloud Shell/Azure CLI의 각 명령에 대해 사용 가능한 모든 옵션이 있는 링크가 제공됩니다.

## 필수 구성 요소

- Azure 구독 - [체험 구독 만들기](#)
- 원하는 Azure 구독의 Azure OpenAI 서비스에 부여된 액세스 권한  
현재 이 서비스에 대한 액세스 권한은 애플리케이션에서만 부여됩니다.  
<https://aka.ms/oai/access>에서 양식을 작성하여 Azure OpenAI에 대한 액세스를 신청할 수 있습니다. 문제가 있는 경우 이 리포지토리에서 문제를 엽니다.
- Azure CLI - [설치 가이드](#)
- 다음 Python 라이브러리: os, requests, json

## Azure CLI에 로그인

Azure CLI에 로그인하려면 다음 명령을 실행하고 로그인을 완료합니다. 세션이 너무 오랫동안 유휴 상태인 경우 다시 수행해야 할 수 있습니다.

```
Azure CLI
az login
```

**Cognitive Services 사용자 역할에 자신을 할당합니다.**

Cognitive Services 사용자 역할에 자신을 할당하면 특정 Cognitive Services 리소스에 액세스하기 위해 계정을 사용할 수 있습니다.

### 1. 사용자 정보 가져오기

Azure CLI

```
export user=$(az account show -o json | jq -r .user.name)
```

### 2. "Cognitive Services 사용자" 역할에 자신을 할당합니다.

Azure CLI

```
export resourceId=$(az group show -g $myResourceGroupName -o json | jq -r .id)
az role assignment create --role "Cognitive Services User" --assignee
$user --scope $resourceId
```

#### ① 참고

역할 할당 변경이 적용되는 데 5분 정도가 걸립니다.

### 3. Azure AD 액세스 토큰을 획득합니다. 액세스 토큰은 1시간 안에 만료됩니다. 그런 다음, 다른 항목을 획득해야 합니다.

Azure CLI

```
export accessToken=$(az account get-access-token --resource
https://cognitiveservices.azure.com -o json | jq -r .accessToken)
```

### 4. API 호출 만들기 액세스 토큰을 사용해 `Authorization` 헤더 값을 설정하여 API 호출에 권한을 부여합니다.

Bash

```
curl ${endpoint%}/openai/deployments/YOUR_DEPLOYMENT_NAME/completions?
api-version=2023-05-15 \
-H "Content-Type: application/json" \
-H "Authorization: Bearer $accessToken" \
-d '{ "prompt": "Once upon a time" }'
```

## 관리 ID에 대한 액세스 권한 부여

OpenAI는 [Azure 리소스에 대한 관리 ID](#)를 사용하는 Azure AD(Azure Active Directory) 인증을 지원합니다. Azure 리소스의 관리 ID를 사용하면 Azure VM(가상 머신), 함수 앱, 가상 머신 확장 집합 및 기타 서비스에서 실행 중인 애플리케이션의 Azure AD 자격 증명을 사용하여 Cognitive Services 리소스에 대한 액세스 권한을 부여할 수 있습니다. Azure AD 인증과 함께 Azure 리소스의 관리 ID를 사용하면 클라우드에서 실행되는 애플리케이션에 자격 증명을 저장할 필요가 없습니다.

## VM에서 관리 ID 사용

Azure 리소스의 관리 ID를 사용하여 VM에서 Cognitive Services 리소스에 대한 액세스 권한을 부여하려면 VM에서 Azure 리소스의 관리 ID를 사용하도록 설정해야 합니다. Azure 리소스의 관리 ID를 사용하도록 설정하는 방법을 알아보려면 다음을 참조하세요.

- [Azure Portal](#)
- [Azure PowerShell](#)
- [Azure CLI](#)
- [Azure Resource Manager 템플릿](#)
- [Azure Resource Manager 클라이언트 라이브러리](#)

관리 ID에 대한 자세한 내용은 [Azure 리소스의 관리 ID](#)를 참조하세요.

# Azure Cognitive Services 가상 네트워크 구성

아티클 • 2023. 03. 16.

Azure Cognitive Services는 계층화된 보안 모델을 제공합니다. 이 모델을 사용하여 Cognitive Services 계정을 특정 네트워크 하위 집합으로 보호할 수 있습니다. 네트워크 규칙이 구성되면 지정된 네트워크 세트를 통해 데이터를 요청하는 애플리케이션만 계정에 액세스할 수 있습니다. 요청 필터링을 사용하여 리소스에 대한 액세스를 제한할 수 있습니다. 지정된 IP 주소, IP 범위 또는 [Azure Virtual Networks](#)의 서브넷 목록에서 시작되는 요청만 허용됩니다.

네트워크 규칙이 적용될 때 Cognitive Services에 액세스하는 애플리케이션에는 권한 부여가 필요합니다. 권한 부여는 Azure AD([Azure Active Directory](#)) 자격 증명 또는 유효한 API 키를 사용하여 지원됩니다.

## ⓘ 중요

Cognitive Services 계정에 대한 방화벽 규칙을 설정하면 기본적으로 데이터에 대해 들어오는 요청이 차단됩니다. 요청을 허용하려면 다음 조건 중 하나를 충족해야 합니다.

- 요청은 대상 Cognitive Services 계정의 허용된 서브넷 목록에 있는 Azure VNet(Virtual Network) 내에서 작동하는 서비스로부터 시작되어야 합니다. VNet에서 시작된 요청의 엔드포인트는 Cognitive Services 계정의 [사용자 지정 하위 도메인](#)으로 설정되어야 합니다.
- 또는 요청이 허용되는 IP 주소 목록에서 시작되어야 합니다.

차단되는 요청에는 다른 Azure 서비스, Azure Portal, 로깅 및 메트릭 서비스 등이 포함됩니다.

## ⓘ 참고

Azure Az PowerShell 모듈을 사용하여 Azure와 상호 작용하는 것이 좋습니다. 시작 하려면 [Azure PowerShell 설치](#)를 참조하세요. Az PowerShell 모듈로 마이그레이션 하는 방법에 대한 자세한 내용은 [Azure PowerShell을 AzureRM에서 Azure로 마이그레이션](#)을 참조하세요.

# 시나리오

Cognitive Services 리소스를 보호하려면 먼저 모든 네트워크(인터넷 트래픽 포함)의 트래픽에 대한 액세스를 거부하도록 규칙을 구성해야 합니다. 그런 다음, 특정 VNet 트래픽에 대한 액세스를 허가하는 규칙을 구성해야 합니다. 이 구성을 사용하면 애플리케이션에 대한 보안 네트워크 경계를 구축할 수 있습니다. 또한 퍼블릭 인터넷 IP 주소 범위의 트래픽에 대한 액세스를 허가하도록 규칙을 구성하여 특정 인터넷 또는 온-프레미스 클라이언트의 연결을 사용하도록 설정할 수도 있습니다.

네트워크 규칙은 REST 및 WebSocket을 포함하여 Azure Cognitive Services에 대한 모든 네트워크 프로토콜에 적용됩니다. Azure 테스트 콘솔 등의 도구를 사용하여 데이터에 액세스하려면 명시적 네트워크 규칙을 구성해야 합니다. 네트워크 규칙을 기준 Cognitive Services 리소스에 적용하거나 새 Cognitive Services 리소스를 만들 때 적용할 수 있습니다. 네트워크 규칙이 적용되면 모든 요청에 적용됩니다.

## 지원되는 지역 및 서비스 제공 사항

VNET(가상 네트워크)는 [Cognitive Services를 사용할 수 있는 지역](#)에서 지원됩니다. Cognitive Services는 네트워크 규칙 구성에 대한 서비스 태그를 지원합니다. 아래 나열된 서비스는 **CognitiveServicesManagement** 서비스 태그에 포함되어 있습니다.

- ✓ Anomaly Detector
- ✓ Azure OpenAI
- ✓ Computer Vision
- ✓ Content Moderator
- ✓ Custom Vision
- ✓ Face
- ✓ 언어 이해(LUIS)
- ✓ Personalizer
- ✓ Speech Service
- ✓ 언어 서비스
- ✓ QnA Maker
- ✓ Translator Text

### ① 참고

Azure OpenAI, LUIS, Speech Services 또는 언어 서비스를 사용하는 경우

**CognitiveServicesManagement** 태그를 통해서만 SDK 또는 REST API를 사용하는 서비스를 사용할 수 있습니다. 가상 네트워크에서 Azure OpenAI 스튜디오, LUIS 포털, Speech Studio 또는 Language Studio에 액세스하고 이를 사용하려면 다음 태그를 사용해야 합니다.

- AzureActiveDirectory
- AzureFrontDoor.Frontend
- AzureResourceManager
- CognitiveServicesManagement
- CognitiveServicesFrontEnd

## 기본 네트워크 액세스 규칙 변경

기본적으로 Cognitive Services 리소스는 네트워크에 있는 클라이언트로부터의 연결을 허용합니다. 선택한 네트워크에 대한 액세스를 제한하려면 먼저 기본 동작을 변경해야 합니다.

### ⚠ 경고

네트워크 규칙을 변경하면 Azure Cognitive Services에 연결하는 애플리케이션의 기능에 영향을 미칠 수 있습니다. 기본 네트워크 규칙을 **거부**로 설정하면 액세스를 허용하는 특정 네트워크 규칙이 적용되지 않는 한 데이터에 대한 모든 액세스가 차단됩니다. 액세스를 거부하도록 기본 규칙을 변경하기 전에 네트워크 규칙을 사용하여 허용된 모든 네트워크에 대한 액세스를 허가해야 합니다. 온-프레미스 네트워크에 대한 IP 주소를 나열하도록 허용하는 경우 온-프레미스 네트워크에서 사용 가능한 모든 나가는 공용 IP 주소를 추가해야 합니다.

## 기본 네트워크 액세스 규칙 관리

Azure Portal, PowerShell 또는 Azure CLI를 통해 Cognitive Services 리소스에 대한 기본 네트워크 액세스 규칙을 관리할 수 있습니다.

Azure Portal

1. 보안을 유지하려는 Cognitive Services 리소스로 이동합니다.
2. **가상 네트워크라는 리소스 관리** 메뉴를 선택합니다.

widgets - Virtual network  
Cognitive Services

Save Discard Refresh

Allow access from  
 All networks  Selected networks

All networks, including the internet, can access this resource. [Learn more.](#)

Overview Activity log Access control (IAM) Tags Diagnose and solve problems

RESOURCE MANAGEMENT

Virtual network

3. 기본적으로 액세스를 거부하려면 **선택한 네트워크**에서 액세스를 허용하도록 선택합니다. 구성된 **가상 네트워크** 또는 **주소 범위**를 함께 사용하지 않고 **선택한 네트워크** 설정만 단독으로 사용하면 모든 액세스가 사실상 거부됩니다. 모든 액세스가 거부되면 Cognitive Services 리소스를 사용하려는 요청이 허용되지 않습니다. 계속 Azure Portal, Azure PowerShell 또는 Azure CLI를 사용하여 Cognitive Services 리소스를 구성할 수 있습니다.
4. 모든 네트워크의 트래픽을 허용하려면 **모든 네트워크**에서 액세스를 허용하도록 선택합니다.

Home > davidpine-computer-vision-vnet - Virtual network  
widgets - Virtual network  
Cognitive Services

Save Discard Refresh

Firewall settings allowing access to cognitive service will remain in effect for up to three minutes after saving updated settings restricting access.

Allow access from  
 All networks  Selected networks

Configure network security for your cognitive service. [Learn more.](#)

Virtual networks  
Secure your cognitive service with virtual networks. + Add existing virtual network + Add new virtual network

| VIRTUAL NETWORK      | SUBNET | ADDRESS RANGE | ENDPOINT |
|----------------------|--------|---------------|----------|
| No network selected. |        |               |          |

Firewall  
Add IP ranges to allow access from the internet or your on-premises networks. [Learn more.](#)

Add your client IP address? [?](#)

ADDRESS RANGE  
IP address or CIDR

5. 저장을 선택하여 변경 내용을 적용합니다.

## 가상 네트워크의 액세스 허가

특정 서브넷에서만 액세스를 허용하도록 Cognitive Services 리소스를 구성할 수 있습니다. 허용되는 서브넷은 동일한 구독의 VNet 또는 다른 Azure Active Directory 테넌트에 속

한 구독을 포함하여 다른 구독의 VNet에 속할 수 있습니다.

VNet 내의 Azure Cognitive Services에 대해 [서비스 엔드포인트](#)를 사용하도록 설정합니다. 서비스 엔드포인트는 VNet의 트래픽을 Azure Cognitive Services 서비스에 대한 최적의 경로를 통해 라우팅합니다. 서브넷 및 가상 네트워크의 ID 또한 각 요청과 함께 전송됩니다. 그러면 관리자가 VNet의 특정 서브넷에서 요청을 받을 수 있도록 Cognitive Services 리소스에 대한 네트워크 규칙을 구성할 수 있습니다. 이러한 네트워크 규칙을 통해 액세스가 허가된 클라이언트는 데이터에 액세스하기 위해 Cognitive Services 리소스의 인증 요구 사항을 계속 충족해야 합니다.

각 Cognitive Services 리소스는 [IP 네트워크 규칙](#)과 결합될 수 있는 최대 100개의 가상 네트워크 규칙을 지원합니다.

## 필요한 사용 권한

가상 네트워크 규칙을 Cognitive Services 리소스에 적용하려면 추가되는 서브넷에 대한 적절한 권한이 사용자에게 있어야 합니다. 필요한 권한은 기본 *기여자 역할* 또는 *Cognitive Services 기여자 역할*입니다. 필요한 사용 권한을 사용자 지정 역할 정의에 추가할 수도 있습니다.

Cognitive Services 리소스 및 가상 네트워크에 허용된 액세스 권한은 다른 Azure AD 테넌트에 속하는 구독을 포함한 다른 구독에 있을 수 있습니다.

### ① 참고

다른 Azure Active Directory 테넌트에 속한 가상 네트워크의 서브넷에 대해 액세스 권한을 부여하는 규칙의 구성은 현재 PowerShell, CLI 및 REST API를 통해서만 지원됩니다. 이러한 규칙은 Azure Portal에서 볼 수 있지만 구성할 수는 없습니다.

## 가상 네트워크 규칙 관리

Azure Portal, PowerShell 또는 Azure CLI를 통해 Cognitive Services 리소스에 대한 가상 네트워크 규칙을 관리할 수 있습니다.

Azure Portal

- 보안을 유지하려는 Cognitive Services 리소스로 이동합니다.
- 가상 네트워크라는 리소스 관리** 메뉴를 선택합니다.
- 선택한 네트워크**에서 액세스를 허용하도록 선택했는지 확인합니다.

4. 기존 네트워크 규칙을 사용하여 가상 네트워크에 대한 액세스를 허용하려면 가상 네트워크에서 기존 가상 네트워크 추가를 선택합니다.

widgets - Virtual network  
Cognitive Services

»

**Firewall settings allowing access to cognitive service will remain in effect for up to three minutes after saving updated settings restricting access.**

Allow access from  
 All networks  Selected networks

Configure network security for your cognitive service. [Learn more.](#)

Virtual networks  
Secure your cognitive service with virtual networks. [+ Add existing virtual network](#) [+ Add new virtual network](#)

| VIRTUAL NETWORK      | SUBNET | ADDRESS RANGE |
|----------------------|--------|---------------|
| No network selected. |        |               |

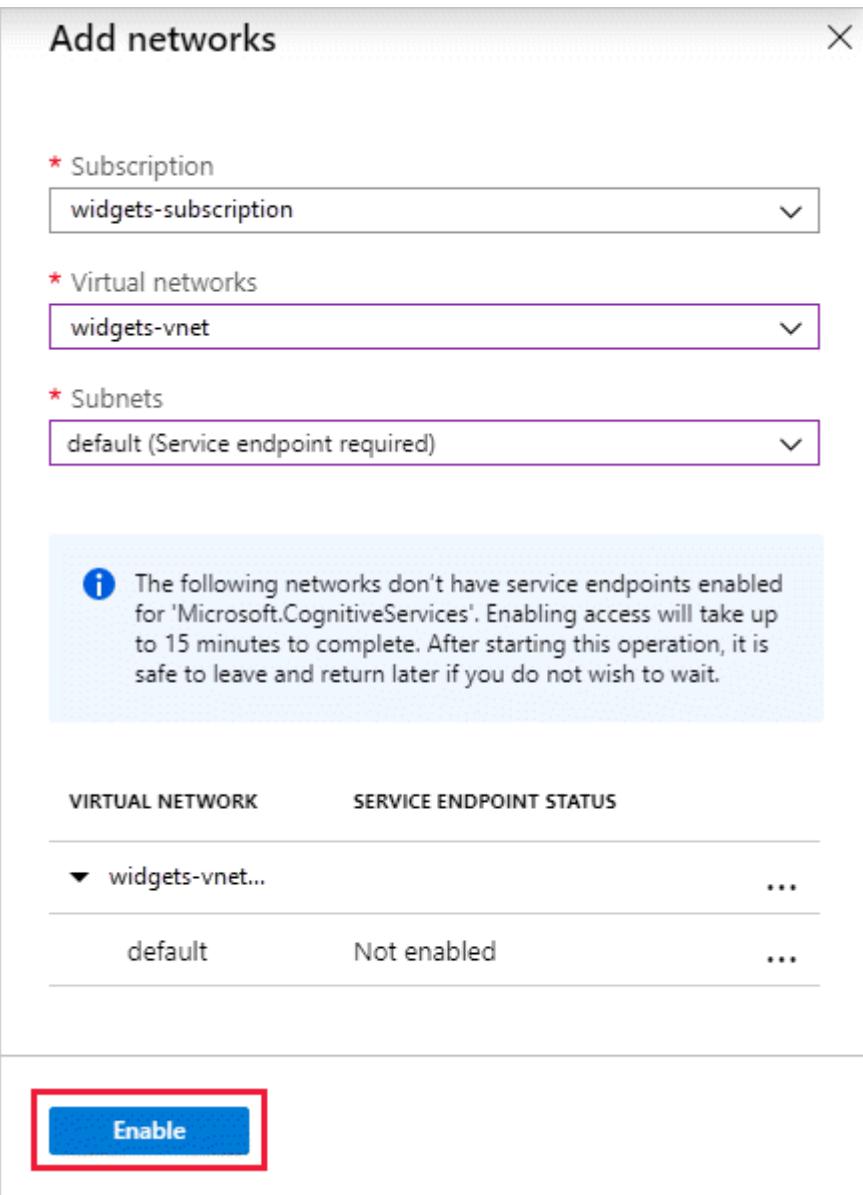
Firewall  
Add IP ranges to allow access from the internet or your on-premises networks. [Learn more.](#)

Add your client IP address? [?](#)

ADDRESS RANGE

IP address or CIDR

5. 가상 네트워크 및 서브넷 옵션을 선택한 다음, 사용을 선택합니다.



6. 새 가상 네트워크를 만들고 거기에 액세스 권한을 부여하려면 새 가상 네트워크 추가를 선택합니다.

**widgets - Virtual network**  
Cognitive Services

» Save Discard Refresh

**Firewall** settings allowing access to cognitive service will remain in effect for up to three minutes after saving updated settings restricting access.

Allow access from  
 All networks  Selected networks

Configure network security for your cognitive service. [Learn more.](#)

Virtual networks  
Secure your cognitive service with virtual networks. [+ Add existing virtual network](#) [+ Add new virtual network](#) + Add new virtual network

| VIRTUAL NETWORK      | SUBNET | ADDRESS RANGE |
|----------------------|--------|---------------|
| No network selected. |        |               |

Firewall  
Add IP ranges to allow access from the internet or your on-premises networks. [Learn more.](#)

Add your client IP address?

ADDRESS RANGE

7. 새 가상 네트워크를 만드는 데 필요한 정보를 입력하고 **만들기**를 선택합니다.

Create virtual network

\* Name  
widgets-vnet

\* Address space ⓘ  
10.1.0.0/16  
10.1.0.0 - 10.1.255.255 (65536 addresses)

\* Subscription  
widgets-subscription

\* Resource group  
widgets-resource-group  
[Create new](#)

\* Location  
(US) West US 2

Subnet

\* Name  
default

\* Address range ⓘ  
10.1.0.0/24  
10.1.0.0 - 10.1.0.255 (256 addresses)

DDoS protection ⓘ  
 Basic  Standard

Service endpoint ⓘ  
Microsoft.CognitiveServices

Firewall ⓘ  
 Disabled  Enabled

**Create**

① 참고

이전에 Azure Cognitive Services에 대한 서비스 엔드포인트가 선택한 가상 네트워크 및 서브넷에 대해 구성되지 않은 경우 이 작업의 일환으로 구성 할 수 있습니다.

현재, 규칙을 만드는 동안 동일한 Azure Active Directory 테넌트에 속한 가상 네트워크만 선택할 수 있도록 표시됩니다. 다른 테넌트에 속한 가상 네트워크의 서브넷에 대한 액세스 권한을 부여하려면 PowerShell, CLI 또는 REST API를 사용하세요.

8. 가상 네트워크 또는 서브넷 규칙을 제거하려면 ... 을 선택해 가상 네트워크 또는 서브넷 바로 가기 메뉴를 열고 제거를 선택합니다.

The screenshot shows the 'widgets - Virtual network' configuration page in the Azure portal. At the top, there are 'Save', 'Discard', and 'Refresh' buttons. A message box states: 'Firewall settings allowing access to cognitive service will remain in effect for up to three minutes after saving updated settings restricting access.' Below this, there's a section for 'Allow access from' with 'Selected networks' selected. A note says: 'Configure network security for your cognitive service. [Learn more](#)'. Under 'Virtual networks', it says: 'Secure your cognitive service with virtual networks. + Add existing virtual network + Add new virtual network'. A table lists one virtual network: 'widgets-vnet' (Subnet 1, Address Range 10.1.0.0/24, Endpoint Status Enabled, Resource Group widgets, Subscription widgets-subscription). A red box highlights the '...' button next to the 'Subscription' column for this row. Below the table, there's a 'Firewall' section with a note: 'Add IP ranges to allow access from the internet or your on-premises networks. [Learn more](#)'. There are checkboxes for 'Add your client IP address' and 'IP address or CIDR' (which is empty).

9. 저장을 선택하여 변경 내용을 적용합니다.

### ① 중요

반드시 **기본 규칙을 거부**로 설정해야 합니다. 그렇지 않으면 네트워크 규칙이 적용되지 않습니다.

## 인터넷 IP 범위의 액세스 허가

특정 공용 인터넷 IP 주소 범위에서 액세스할 수 있도록 Cognitive Services 리소스를 구성할 수 있습니다. 이 구성은 특정 서비스와 온-프레미스 네트워크에 대한 액세스 권한을 부여하고 일반 인터넷 트래픽을 효과적으로 차단합니다.

16.17.18.0/24 형식의 [CIDR 표기법](#)을 사용하거나 개별 IP 주소(예: 16.17.18.19)를 사용하여 허용된 인터넷 주소 범위를 제공합니다.

### 💡 팁

"/31" 또는 "/32" 접두사 크기를 사용하는 작은 주소 범위는 지원되지 않습니다. 이러한 범위는 개별 IP 주소 규칙을 사용하여 구성해야 합니다.

IP 네트워크 규칙은 [공용 인터넷](#) IP 주소에 대해서만 허용됩니다. 사설망에 예약된 IP 주소 범위([RFC 1918](#))에 정의된 대로)는 IP 규칙에서 허용되지 않습니다. 사설망에는 10.\*,  
172.16.\* - 172.31.\* 및 192.168.\*로 시작하는 주소가 포함됩니다.

현재 IPv4 주소만 지원됩니다. 각 Cognitive Services 리소스는 [가상 네트워크 규칙](#)과 결합될 수 있는 최대 100개의 IP 네트워크 규칙을 지원합니다.

## 온-프레미스 네트워크에서의 액세스 구성

IP 네트워크 규칙을 사용하여 온-프레미스 네트워크에서 Cognitive Services 리소스로의 액세스를 허가하려면 네트워크에서 사용되는 인터넷 연결 IP 주소를 식별해야 합니다. 네트워크 관리자에게 도움을 요청합니다.

공용 피어링 또는 Microsoft 피어링을 위해 [ExpressRoute](#) 온-프레미스를 사용하는 경우 NAT IP 주소를 식별해야 합니다. 공용 피어링의 경우 기본적으로 각 ExpressRoute 회로에서 두 개의 NAT IP 주소를 사용합니다. 각각은 트래픽이 Microsoft Azure 네트워크 백본으로 들어갈 때 Azure 서비스 트래픽에 적용됩니다. Microsoft 피어링의 경우 사용되는 NAT IP 주소는 고객이 제공하거나 서비스 공급자가 제공합니다. 서비스 리소스에 대한 액세스를 허용하려면 리소스 IP 방화벽 설정에서 이러한 공용 IP 주소를 허용해야 합니다. ExpressRoute 회로 IP 주소를 찾으려면 Azure Portal을 통해 [ExpressRoute에서 지원 티켓을 엽니다](#). [ExpressRoute 공용 및 Microsoft 피어링을 위한 NAT](#)에 대해 자세히 알아보세요.

## IP 네트워크 규칙 관리

Azure Portal, PowerShell 또는 Azure CLI를 통해 Cognitive Services 리소스에 대한 IP 네트워크 규칙을 관리할 수 있습니다.

Azure Portal

- 보안을 유지하려는 Cognitive Services 리소스로 이동합니다.
- [가상 네트워크라는 리소스 관리](#) 메뉴를 선택합니다.
- [선택한 네트워크](#)에서 액세스를 허용하도록 선택했는지 확인합니다.
- 인터넷 IP 범위에 대한 액세스 권한을 부여하려면 **방화벽>주소 범위** 아래에서 IP 주소 또는 주소 범위([CIDR 형식](#))를 입력합니다. 유효한 공용 IP(예약되지 않음) 주소만 허용됩니다.

**widgets - Virtual network**  
Cognitive Services

» Save Discard Refresh

**Firewall** settings allowing access to cognitive service will remain in effect for up to three minutes after saving updated settings restricting access.

Allow access from  
 All networks  Selected networks

Configure network security for your cognitive service. [Learn more](#).

Virtual networks  
Secure your cognitive service with virtual networks. [+ Add existing virtual network](#) [+ Add new virtual network](#)

| VIRTUAL NETWORK      | SUBNET | ADDRESS RANGE |
|----------------------|--------|---------------|
| No network selected. |        |               |

Firewall  
Add IP ranges to allow access from the internet or your on-premises networks. [Learn more](#).

Add your client IP address?

**ADDRESS RANGE**

|                    |  |
|--------------------|--|
| 173.0.0.0/16       |  |
| IP address or CIDR |  |

5. IP 네트워크 규칙을 제거하려면 주소 범위 옆에 있는 휴지통 아이콘을 선택합니다.

**widgets - Virtual network**  
Cognitive Services

» Save Discard Refresh

**Firewall** settings allowing access to cognitive service will remain in effect for up to three minutes after saving updated settings restricting access.

Allow access from  
 All networks  Selected networks

Configure network security for your cognitive service. [Learn more](#).

Virtual networks  
Secure your cognitive service with virtual networks. [+ Add existing virtual network](#) [+ Add new virtual network](#)

| VIRTUAL NETWORK      | SUBNET | ADDRESS RANGE |
|----------------------|--------|---------------|
| No network selected. |        |               |

Firewall  
Add IP ranges to allow access from the internet or your on-premises networks. [Learn more](#).

Add your client IP address?

**ADDRESS RANGE**

|                    |  |
|--------------------|--|
| 173.0.0.0/16       |  |
| IP address or CIDR |  |

6. 저장을 선택하여 변경 내용을 적용합니다.

### ① 중요

반드시 **기본 규칙을 거부로 설정해야 합니다**. 그렇지 않으면 네트워크 규칙이 적용되지 않습니다.

# 프라이빗 엔드포인트 사용

Cognitive Services 리소스에서 [프라이빗 엔드포인트](#)를 사용하면 VNet(가상 네트워크)의 클라이언트가 [Private Link](#)를 통해 안전하게 데이터에 액세스하도록 할 수 있습니다. 프라이빗 엔드포인트는 Cognitive Services 리소스용 VNet 주소 공간의 IP 주소를 사용합니다. VNet과 리소스에 있는 클라이언트 사이의 네트워크 트래픽은 VNet과 Microsoft 백본 네트워크의 프라이빗 링크를 가로지르며 공용 인터넷 노출을 방지합니다.

Cognitive Services 리소스의 프라이빗 엔드포인트를 사용하면 다음과 같은 작업을 할 수 있습니다.

- Cognitive Services 서비스에 대한 퍼블릭 엔드포인트의 모든 연결을 차단하도록 방화벽을 구성하여 Cognitive Services 리소스를 보호합니다.
- VNet에서의 데이터 반출을 차단하도록 하여 VNet 보안을 강화합니다.
- 프라이빗 피어링을 통해 [VPN](#) 또는 [ExpressRoutes](#)를 사용하여 VNet에 연결하는 온-프레미스 네트워크에서 Cognitive Services 리소스에 안전하게 연결합니다.

## 개념적 개요

프라이빗 엔드포인트는 [VNet](#)의 Azure 리소스에 대한 특수한 네트워크 인터페이스입니다. Cognitive Services 리소스에 대한 프라이빗 엔드포인트를 만들면 VNet 및 리소스의 클라이언트 간에 보안 연결을 제공합니다. 프라이빗 엔드포인트에는 VNet의 IP 주소 범위에서 IP 주소가 할당됩니다. 프라이빗 엔드포인트와 Cognitive Services 서비스 간의 연결은 안전한 프라이빗 링크를 사용합니다.

VNet의 애플리케이션은 다른 방법으로 사용하는 것과 동일한 연결 문자열 및 권한 부여 메커니즘을 사용하여 프라이빗 엔드포인트를 통해 서비스에 원활하게 연결할 수 있습니다. 별도의 엔드포인트가 필요한 Speech Services는 예외입니다. [Speech Services를 사용하는 프라이빗 엔드포인트](#)에 대한 섹션을 참조하세요. 프라이빗 엔드포인트는 REST를 포함하여 Cognitive Services 리소스에서 지원하는 모든 프로토콜과 함께 사용할 수 있습니다.

[서비스 엔드포인트](#)를 사용하는 서브넷에서 프라이빗 엔드포인트를 만들 수 있습니다. 서브넷의 클라이언트는 서비스 엔드포인트를 사용해 다른 Cognitive Services 리소스에 액세스하는 동안 프라이빗 엔드포인트를 사용하여 하나의 Cognitive Services 리소스에 연결할 수 있습니다.

VNet에서 Cognitive Services 리소스에 대한 프라이빗 엔드포인트를 만드는 경우 Cognitive Services 리소스 소유자에게 승인을 위한 동의 요청이 전송됩니다. 프라이빗 엔

드포인트 만들기를 요청한 사용자가 리소스의 소유자인 경우 이 동의 요청이 자동으로 승인됩니다.

Cognitive Services 리소스 소유자는 [Azure Portal](#)의 Cognitive Services 리소스에 대해 '프라이빗 엔드포인트' 탭을 통해 동의 요청 및 프라이빗 엔드포인트를 관리할 수 있습니다.

## 프라이빗 엔드포인트

프라이빗 엔드포인트를 만들 때 연결할 Cognitive Services 리소스를 지정해야 합니다. 프라이빗 엔드포인트를 만드는 방법에 대한 자세한 내용은 다음을 참조하세요.

- Azure Portal에서 Private Link 센터를 사용하여 프라이빗 엔드포인트 만들기
- Azure CLI를 사용하여 Azure 프라이빗 엔드포인트 만들기
- Azure PowerShell을 사용하여 프라이빗 엔드포인트 만들기

## 프라이빗 엔드포인트에 연결

### ① 참고

Azure OpenAI Service는 다른 Azure Cognitive Services와는 다른 프라이빗 DNS 영역 및 공용 DNS 영역 전달자를 사용합니다. 올바른 영역 및 전달자 이름은 [Azure 서비스 DNS 영역 구성 문서](#)를 참조하세요.

프라이빗 엔드포인트를 사용하는 VNet의 클라이언트는 퍼블릭 엔드포인트에 연결하는 클라이언트와 동일한 Cognitive Services 리소스 연결 문자열을 사용해야 합니다. 별도의 엔드포인트가 필요한 Speech Services는 예외입니다. [Speech Services를 사용하는 프라이빗 엔드포인트](#)에 대한 섹션을 참조하세요. DNS 확인을 사용하여 프라이빗 링크를 통해 VNet에서 Cognitive Services 리소스로 연결을 자동으로 라우팅합니다.

기본적으로 프라이빗 엔드포인트에 대한 필수 업데이트를 사용하여 VNet에 연결된 [프라이빗 DNS 영역](#)을 만듭니다. 그러나 사용자가 자체 DNS 서버를 사용하는 경우 DNS 구성 을 추가로 변경해야 할 수 있습니다. 아래 [DNS 변경 내용](#) 섹션에서는 프라이빗 엔드포인트에 필요한 업데이트에 대해 설명합니다.

## Speech Services를 사용하는 프라이빗 엔드포인트

[Azure Private Link에서 제공하는 프라이빗 엔드포인트에서 Speech Services 사용](#)을 참조하세요.

# 프라이빗 엔드포인트에 대한 DNS 변경 내용

프라이빗 엔드포인트를 만들 때 Cognitive Services 리소스에 대한 DNS CNAME 리소스 레코드는 '*privatelink*' 접두사가 있는 하위 도메인의 별칭으로 업데이트됩니다. 또한 기본적으로 프라이빗 엔드포인트에 대한 DNS A 리소스 레코드를 사용하여 '*privatelink*' 하위 도메인에 해당하는 [프라이빗 DNS 영역](#)을 만듭니다.

프라이빗 엔드포인트를 사용하여 VNet 외부에서 엔드포인트 URL을 확인하면 Cognitive Services 리소스의 퍼블릭 엔드포인트로 확인됩니다. 프라이빗 엔드포인트를 호스트하는 VNet에서 확인되면 엔드포인트 URL은 프라이빗 엔드포인트의 IP 주소로 확인됩니다.

이 접근 방식을 사용하면 프라이빗 엔드포인트를 호스트하는 VNet의 클라이언트와 VNet 외부의 클라이언트에 동일한 연결 문자열을 사용하여 Cognitive Services 리소스에 액세스할 수 있습니다.

네트워크에서 사용자 지정 DNS 서버를 사용하는 경우 클라이언트는 프라이빗 엔드포인트 IP 주소에 대한 Cognitive Services 리소스 엔드포인트의 FQDN(정규화된 도메인 이름)을 확인할 수 있어야 합니다. 프라이빗 링크 하위 도메인을 VNet의 프라이빗 DNS 영역에 위임하도록 DNS 서버를 구성합니다.

## 💡 팁

사용자 지정 또는 온-프레미스 DNS 서버를 사용하는 경우 '*privatelink*' 하위 도메인의 Cognitive Services 리소스 이름을 프라이빗 엔드포인트 IP 주소로 확인하도록 DNS 서버를 구성해야 합니다. '*privatelink*' 하위 도메인을 VNet의 프라이빗 DNS 영역에 위임하거나 DNS 서버에서 DNS 영역을 구성하고 DNS A 레코드를 추가하면 그렇게 할 수 있습니다.

프라이빗 엔드포인트를 지원하기 위해 자체 DNS 서버를 구성하는 방법에 대한 자세한 내용은 다음 문서를 참조하세요.

- [Azure 가상 네트워크의 리소스 이름 확인](#)
- [프라이빗 엔드포인트에 대한 DNS 구성](#)

## 가격 책정

가격 책정에 대한 자세한 내용은 [Azure Private Link 가격 책정](#)을 참조하세요.

## 다음 단계

- 다양한 [Azure Cognitive Services](#) 살펴보기

- Azure Virtual Network 서비스 [엔드포인트](#)에 대해 자세히 알아보기

# 대규모 데이터 세트와 함께 Azure OpenAI 사용

아티클 • 2023. 04. 27.

Azure OpenAI는 완료 API의 프롬프트 처리를 통해 많은 자연어 작업을 해결하는 데 사용할 수 있습니다. 프롬프트 워크플로를 몇 가지 예제에서 대규모 예제 데이터 세트로 쉽게 확장할 수 있도록 Azure OpenAI 서비스를 분산 기계 학습 라이브러리인 [SynapseML](#)과 통합했습니다. 이러한 통합으로 [Apache Spark](#) 분산 컴퓨팅 프레임워크를 쉽게 사용할 수 있어 OpenAI 서비스를 통해 수백만 개의 프롬프트를 처리할 수 있습니다. 이 자습서에서는 Azure Open AI 및 Azure Synapse Analytics를 사용하여 대규모 언어 모델을 분산된 규모로 적용하는 방법을 보여 줍니다.

## 필수 구성 요소

- Azure 구독 - [체험 구독 만들기](#)  
현재 이 서비스에 대한 액세스 권한은 애플리케이션에서만 부여됩니다.  
<https://aka.ms/oai/access>에서 양식을 작성하여 Azure OpenAI에 대한 액세스를 신청할 수 있습니다. 문제가 있는 경우 이 리포지토리에서 문제를 엽니다.
- Azure OpenAI 리소스 - [리소스 만들기](#)
- SynapseML이 설치된 Apache Spark 클러스터 - [여기](#) 서버리스 Apache Spark 풀 만들기

[Synapse 작업 영역을 만드는](#) 것이 좋지만 Azure Databricks, HDInsight, Kubernetes의 Spark 또는 `pyspark` 패키지가 있는 Python 환경도 작동합니다.

## 이 가이드를 Notebook으로 가져오기

다음 단계에서는 이 코드를 Spark 클러스터에 추가합니다. Spark 플랫폼에서 Notebook을 만들고 코드를 이 Notebook에 복사하여 데모를 실행하거나 Notebook을 다운로드하고 Synapse Analytics로 가져올 수 있습니다.

1. [이 데모를 Notebook으로 다운로드](#) 합니다(원시를 클릭한 다음, 파일 저장).
2. Notebook을 [Synapse 작업 영역으로 가져오거나 Databricks를 사용하는 경우 Databricks 작업 영역으로 가져옵니다.](#)

3. 클러스터에 SynapseML을 설치합니다. [SynapseML 웹 사이트](#)의 맨 아래에 있는 Synapse에 대한 설치 지침을 참조하세요. 이렇게 하려면 가져온 Notebook의 맨 위에 다른 셀을 붙여넣어야 합니다.
4. Notebook을 클러스터에 연결하고 아래 셀을 편집하고 실행하면서 따라해 봅니다.

## 서비스 정보 채우기

다음으로, 서비스를 가리키도록 Notebook의 셀을 편집합니다. 특히 `resource_name`, `deployment_name`, `location` 및 `key` 변수를 Azure OpenAI 리소스의 해당 값으로 설정합니다.

### ⓘ 중요

완료되면 코드에서 키를 제거하고 공개적으로 게시하지 마세요. 프로덕션의 경우 [Azure Key Vault](#)와 같은 자격 증명을 안전하게 저장하고 액세스하는 방법을 사용합니다. 자세한 내용은 Cognitive Services [보안](#) 문서를 참조하세요.

Python

```
import os

# Replace the following values with your Azure OpenAI resource information
resource_name = "RESOURCE_NAME"          # The name of your Azure OpenAI
resource.
deployment_name = "DEPLOYMENT_NAME"      # The name of your Azure OpenAI
deployment.
location = "RESOURCE_LOCATION"         # The location or region ID for your
resource.
key = "RESOURCE_API_KEY"                # The key for your resource.

assert key is not None and resource_name is not None
```

## 프롬프트의 데이터 세트 만들기

다음으로, 행당 하나의 프롬프트를 사용하여 일련의 행으로 구성된 데이터 프레임을 만듭니다.

ADLS(Azure Data Lake Storage) 또는 다른 데이터베이스에서 직접 데이터를 로드할 수도 있습니다. Spark 데이터 프레임 로드 및 준비에 대한 자세한 내용은 [Apache Spark 데이터 로드 가이드](#)를 참조하세요.

Python

```
df = spark.createDataFrame(  
    [  
        ("Hello my name is",),  
        ("The best code is code that's",),  
        ("SynapseML is ",),  
    ]  
).toDF("prompt")
```

## OpenAICompletion Apache Spark 클라이언트 만들기

방금 만든 데이터 프레임에 OpenAI Completion 서비스를 적용하려면 분산 클라이언트 역할을 하는 `OpenAICompletion` 개체를 만듭니다. 서비스의 매개 변수는 단일 값으로 설정하거나 `OpenAICompletion` 개체에 적절한 setter가 있는 데이터 프레임의 열로 설정할 수 있습니다. 여기서는 `maxTokens`를 200으로 설정하고 있습니다. 토큰은 약 4자이며 이 제한은 프롬프트와 결과의 합계에 적용됩니다. 또한 `promptCol` 매개 변수를 데이터 프레임의 프롬프트 열 이름으로 설정합니다.

Python

```
from synapse.ml.cognitive import OpenAICompletion  
  
completion = (  
    OpenAICompletion()  
    .setSubscriptionKey(key)  
    .setDeploymentName(deployment_name)  
    .setUrl("https://{}.openai.azure.com/".format(resource_name))  
    .setMaxTokens(200)  
    .setPromptCol("prompt")  
    .setErrorCol("error")  
    .setOutputCol("completions")  
)
```

## OpenAICompletion 클라이언트를 사용하여 데이터 프레임 변환

이제 데이터 프레임과 완료 클라이언트가 있으므로 입력 데이터 세트를 변환하고 서비스에서 추가하는 모든 정보를 사용하여 `completions`라는 열을 추가할 수 있습니다. 간단히하기 위해 텍스트만 선택하겠습니다.

Python

```

from pyspark.sql.functions import col

completed_df = completion.transform(df).cache()
display(completed_df.select(
    col("prompt"), col("error"),
    col("completions.choices.text").getItem(0).alias("text")))

```

출력은 다음 예제와 같이 표시됩니다. 완성 텍스트는 다를 수 있습니다.

| prompt       | error   | text                                                                                 |
|--------------|---------|--------------------------------------------------------------------------------------|
| 안녕하세요, 제 이름은 | 정의되지 않음 | 마카벨리입니다. 저는 18살이고 자라서 하고 싶은 일은 래퍼가 되는 것입니다. 음악을 쓰고 제작하는 것을 좋아합니다. 로스앤젤레스, CA에서 왔습니다. |
| 가장 좋은 코드는    | 정의되지 않음 | 이해하기 쉬운 코드입니다. 이것은 주관적인 설명이며 확실한 대답이 없습니다.                                           |
| SynapseML은   | 정의되지 않음 | 이벤트의 향후 결과를 예측하는 방법을 배울 수 있는 기계 학습 알고리즘입니다.                                          |

## 기타 사용 예제

### 요청 일괄 처리를 사용하여 처리량 향상

위의 예제에서는 각 프롬프트에 대해 하나씩 서비스에 대한 여러 요청을 만듭니다. 단일 요청으로 여러 프롬프트를 완료하려면 일괄 처리 모드를 사용합니다. 먼저, `OpenAICompletion` 개체에서 프롬프트 열을 "Prompt"로 설정하는 대신 `BatchPrompt` 열에 대해 "batchPrompt"를 지정합니다. 이렇게 하려면 행 하나당 프롬프트 목록이 있는 데이터 프레임을 만듭니다.

#### ① 참고

현재는 단일 요청에서 프롬프트 수는 20개, "토큰" 수는 2,048개, 단어는 약 1,500개로 제한됩니다.

Python

```

batch_df = spark.createDataFrame(
    [
        ([ "The time has come", "Pleased to", "Today stocks", "Here's to"],),
        ([ "The only thing", "Ask not what", "Every litter", "I am"],),
    ]
).toDF("batchPrompt")

```

다음으로, `OpenAICompletion` 개체를 만듭니다. 프롬프트 열을 설정하는 대신 열이 형식 `Array[String]`인 경우 `batchPrompt` 열을 설정합니다.

Python

```
batch_completion = (
    OpenAICompletion()
    .setSubscriptionKey(key)
    .setDeploymentName(deployment_name)
    .setUrl("https://{}.openai.azure.com/".format(resource_name))
    .setMaxTokens(200)
    .setBatchPromptCol("batchPrompt")
    .setErrorCol("error")
    .setOutputCol("completions")
)
```

변환 호출에서 행당 요청이 만들어집니다. 단일 행에 여러 프롬프트가 있으므로 각 요청은 해당 행의 모든 프롬프트와 함께 전송됩니다. 결과에는 요청의 각 행에 대한 행이 포함됩니다.

Python

```
completed_batch_df = batch_completion.transform(batch_df).cache()
display(completed_batch_df)
```

### ① 참고

현재는 단일 요청에서 프롬프트 수는 20개, "토큰" 수는 2,048개, 단어는 약 1,500개로 제한됩니다.

## 자동 미니 일괄 처리기 사용

데이터가 열 형식인 경우 SynapseML의 `FixedMiniBatcherTransformer`를 사용하여 행 형식으로 바꿀 수 있습니다.

Python

```
from pyspark.sql.types import StringType
from synapse.ml.stages import FixedMiniBatcherTransformer
from synapse.ml.core.spark import FluentAPI

completed_autombatch_df = (df
    .coalesce(1) # Force a single partition so that our little 4-row dataframe
    makes a batch of size 4, you can remove this step for large datasets
```

```
.mlTransform(FixedMiniBatchTransformer(batchSize=4))
.withColumnRenamed("prompt", "batchPrompt")
.mlTransform(batch_completion))

display(completed_ autobatch_df)
```

## 번역을 위한 프롬프트 엔지니어링

Azure OpenAI는 [프롬프트 엔지니어링](#)을 통해 다양한 자연어 작업을 해결할 수 있습니다. 여기서는 언어 번역을 요청하는 예제를 보여 줍니다.

Python

```
translate_df = spark.createDataFrame(
    [
        ("Japanese: Ookina hako \nEnglish: Big box \nJapanese: Midori
tako\nEnglish:" ,),
        ("French: Quelle heure est-il à Montréal? \nEnglish: What time is it
in Montreal? \nFrench: Où est le poulet? \nEnglish:" ,),
    ]
).toDF("prompt")

display(completion.transform(translate_df))
```

## 질문 답변 프롬프트

여기서는 일반적인 지식을 묻는 질문 답변에 GPT-3 모델 프롬프트를 표시합니다.

Python

```
qa_df = spark.createDataFrame(
    [
        (
            "Q: Where is the Grand Canyon?\nA: The Grand Canyon is in
Arizona.\n\nQ: What is the weight of the Burj Khalifa in kilograms?\nA:",
        )
    ]
).toDF("prompt")

display(completion.transform(qa_df))
```

# 미사용 데이터의 Azure OpenAI Service 암호화

아티클 • 2023. 04. 22.

Azure OpenAI는 데이터가 클라우드에 유지될 때 데이터를 자동으로 암호화합니다. 암호화는 데이터를 보호하고 조직의 보안 및 규정 준수 노력에 부합하는 데 도움이 됩니다. 이 문서에서는 Azure OpenAI가 미사용 데이터, 특히 학습 데이터 및 미세 조정된 모델의 암호화를 처리하는 방법을 다룹니다. 사용자가 서비스에 제공한 데이터를 처리, 사용 및 저장하는 방법에 대한 자세한 내용은 [데이터, 개인 정보 및 보안 문서](#)를 참조하세요.

## Cognitive Services 암호화 정보

Azure OpenAI는 Azure Cognitive Services의 일부입니다. Cognitive Services 데이터는 [FIPS 140-2](#) 호환 [256비트 AES](#) 암호화를 사용하여 암호화 및 복호화됩니다. 암호화 및 암호 해독은 투명하므로 암호화 및 액세스가 자동으로 관리됩니다. 데이터는 기본적으로 안전하며 암호화를 활용하기 위해 코드 또는 애플리케이션을 수정할 필요가 없습니다.

## 암호화 키 관리 정보

기본적으로 구독은 Microsoft에서 관리하는 암호화 키를 사용합니다. CMK(고객 관리형 키)라고 하는 사용자 고유의 키를 사용하여 구독을 관리하는 옵션도 있습니다. CMK는 액세스 제어를 만들고, 회전시키고, 사용하지 않도록 설정하고, 철회할 수 있는 훨씬 더 큰 유연성을 제공합니다. 데이터를 보호하는 데 사용되는 암호화 키를 감사할 수도 있습니다.

## Azure Key Vault를 사용하는 고객 관리형 키

BYOK(Bring Your Own Key)라고도 하는 CMK(고객 관리형 키) 사용하면 훨씬 더 유연하게 액세스 제어를 만들고, 회전하고, 사용하지 않도록 설정하고, 취소할 수 있습니다. 데이터를 보호하는 데 사용되는 암호화 키를 감사할 수도 있습니다.

고객 관리형 키를 저장하려면 Azure Key Vault를 사용해야 합니다. 사용자 고유의 키를 만들어 키 자격 증명 모음에 저장할 수도 있고, Azure Key Vault API를 사용하여 키를 생성할 수도 있습니다. Cognitive Services 리소스 및 키 자격 증명 모음은 동일한 지역 및 동일한 Azure AD(Azure Active Directory) 테넌트에 있어야 하지만 서로 다른 구독에 있을 수 있습니다. Azure Key Vault에 대한 자세한 내용은 [Azure Key Vault란?](#)을 참조하세요.

고객 관리형 키를 사용하는 기능을 요청하려면 [Cognitive Services 고객 관리형 키 요청 양식](#)을 작성하여 제출합니다. 요청 상태를 다시 들으려면 영업일 기준 3-5일이 소요됩니다.

니다.

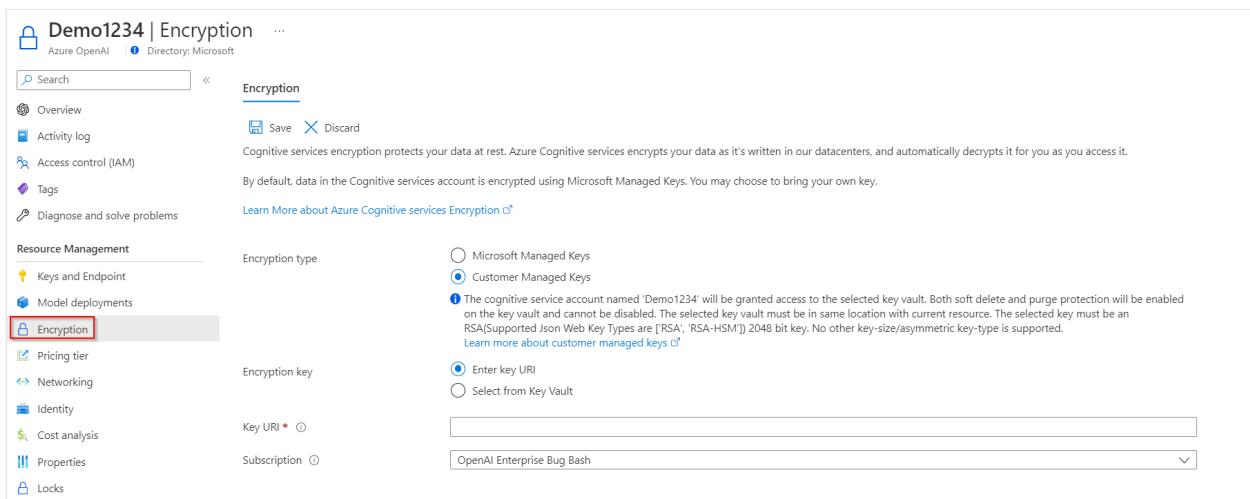
고객 관리형 키를 사용하도록 설정하려면 키 자격 증명 모음에서 일시 삭제 및 제거 안 함 속성도 모두 사용하도록 설정해야 합니다.

2048 크기의 RSA 키만 Cognitive Services 암호화에서 지원됩니다. 키에 대한 자세한 내용은 [Azure Key Vault 키, 비밀 및 인증서 정보](#)의 Key Vault 키를 참조하세요.

## 리소스에 고객 관리형 키 사용

Azure Portal에서 고객 관리형 키를 사용하도록 설정하려면 다음 단계를 수행합니다.

1. Cognitive Services 리소스로 이동합니다.
2. 왼쪽에서 암호화를 선택합니다.
3. 다음 스크린샷과 같이 암호화 형식에서 고객 관리형 키를 선택합니다.



The screenshot shows the Azure Portal interface for managing encryption settings for a Cognitive Services resource named 'Demo1234'. The 'Encryption' tab is active in the left sidebar. On the right, under 'Encryption type', the 'Customer Managed Keys' radio button is selected. Below it, a note states: 'The cognitive service account named 'Demo1234' will be granted access to the selected key vault. Both soft delete and purge protection will be enabled on the key vault and cannot be disabled. The selected key vault must be in same location with current resource. The selected key must be an RSA(Supported Json Web Key Types are [RSA, 'RSA-HSM']) 2048 bit key. No other key-size/asymmetric key-type is supported.' There is also a link to 'Learn more about customer managed keys'. The 'Encryption key' section has 'Enter key URI' selected, and there is a text input field containing 'OpenAI Enterprise Bug Bash'. Other tabs in the sidebar include Overview, Activity log, Access control (IAM), Tags, Diagnose and solve problems, Resource Management, Keys and Endpoint, Model deployments, Pricing tier, Networking, Identity, Cost analysis, Properties, and Locks.

## 키 지정

고객 관리형 키를 사용하도록 설정한 후 Cognitive Services 리소스와 연결할 키를 지정할 수 있습니다.

### 키를 URI로 지정

키를 URI로 지정하려면 다음 단계를 수행합니다.

1. Azure Portal에서 키 자격 증명 모음으로 이동합니다.
2. 설정 아래에서 키를 선택합니다.
3. 원하는 키를 선택한 다음 해당 키를 선택하여 버전을 확인합니다. 키 버전을 선택하여 해당 버전의 설정을 봅니다.

#### 4. URI를 제공하는 키 식별자 값을 복사합니다.

The screenshot shows the 'Properties' section of a key in the Azure Key Vault. The key identifier is '17bf9182bb694f109b8dc6d1e9b69f29'. The 'Key URI' field contains '<key-uri>' and has a copy icon next to it. Below the properties, there are sections for 'Settings' (activation and expiration dates), 'Enabled?' (set to 'Yes'), and 'Permitted operations' (Encrypt, Sign, Wrap Key, Decrypt, Verify, Unwrap Key, all checked). A 'Tags' section shows '0 tags'.

#### 5. Cognitive Services 리소스로 돌아가서 암호화를 선택합니다.

#### 6. 암호화 키에서 키 URI 입력을 선택합니다.

#### 7. 복사한 URI를 키 URI 상자에 붙여넣습니다.

The screenshot shows the 'Encryption' blade for a Cognitive Service account named 'CMK-Test'. The left sidebar shows 'Encryption' selected. The main area has 'Encryption type' set to 'Customer Managed Keys' and 'Encryption key' set to 'Enter key URI'. The 'Key URI' input field contains '<key uri>' with a checkmark. Other options like 'Select from Key Vault' are available but not selected. A note at the bottom explains that the service account will be granted access to the selected key vault.

#### 8. 구독에서 키 자격 증명 모음이 포함된 구독을 선택합니다.

9. 변경 내용을 저장합니다.

## Key Vault에서 키 지정

키 자격 증명 모음의 키를 지정하려면 먼저 키가 포함된 키 자격 증명 모음이 있는지 확인합니다. 그런 다음, 다음 단계를 수행합니다.

1. Cognitive Services 리소스로 이동한 다음 **암호화**를 선택합니다.
2. **암호화 키**에서 **Key Vault에서 선택**을 선택합니다.
3. 사용하려는 키가 포함된 키 자격 증명 모음을 선택합니다.
4. 사용하려는 키를 선택합니다.

The screenshot shows the Azure portal interface for selecting a key from an Azure Key Vault. At the top, there's a navigation bar with 'Microsoft Azure' and a search bar. Below it, the breadcrumb navigation shows 'Home > CMKTest01-SB - Encryption > Select key from Azure Key Vault'. The main section is titled 'Select key from Azure Key Vault'. It contains four dropdown menus:

- 'Subscription \*' dropdown set to 'AICP-DEV'.
- 'Key vault \*' dropdown set to 'CMKTest-01SB', with a 'Create new' link below it.
- 'Key \*' dropdown set to 'CMKTest-01SB', with a 'Create new' link below it.
- 'Version \*' dropdown set to '19fc5cfacbd34e47b373709c1e400902', with a 'Create new' link below it.

5. 변경 내용을 저장합니다.

## 키 버전 업데이트

새 버전의 키를 만들 때는 새 버전을 사용하도록 Cognitive Services를 업데이트합니다. 다음 단계를 수행합니다.

1. Cognitive Services 리소스로 이동한 다음 **암호화**를 선택합니다.
2. 새 키 버전의 URI를 입력합니다. 또는 키 자격 증명 모음을 선택한 다음 키를 다시 선택하여 버전을 업데이트할 수 있습니다.
3. 변경 내용을 저장합니다.

## 다른 키 사용

암호화에 사용하는 키를 변경하려면 다음 단계를 따릅니다.

1. Cognitive Services 리소스로 이동한 다음 **암호화**를 선택합니다.
2. 새 키의 URI를 입력합니다. 또는 키 자격 증명 모음을 선택한 다음 새 키를 선택할 수 있습니다.
3. 변경 내용을 저장합니다.

## 고객 관리형 키 순환

규정 준수 정책에 따라 Key Vault에서 고객 관리형 키를 회전할 수 있습니다. 키를 순환하는 경우 새 키 URI를 사용하도록 Cognitive Services 리소스를 업데이트해야 합니다. Azure Portal에서 새 버전의 키를 사용하도록 리소스를 업데이트하는 방법을 알아보려면 [키 버전 업데이트](#)를 참조하세요.

키를 회전해도 리소스의 데이터 재암호화는 트리거되지 않습니다. 사용자는 추가적인 작업을 할 필요가 없습니다.

## 고객 관리형 키 철회

고객 관리형 암호화 키는 액세스 정책을 변경하거나 Key Vault에 대한 권한을 변경하거나 키를 삭제하여 해지할 수 있습니다.

레지스트리에서 사용하는 관리 ID의 액세스 정책을 변경하려면 [az-keyvault-delete-policy](#) 명령을 실행합니다.

Azure CLI

```
az keyvault delete-policy \
--resource-group <resource-group-name> \
--name <key-vault-name> \
--key_id <key-vault-key-id>
```

키의 개별 버전을 삭제하려면 [az-keyvault-key-delete](#) 명령을 실행합니다. 이 작업에는 '키/삭제' 권한이 필요합니다.

Azure CLI

```
az keyvault key delete \
--name <key-vault-name> \
--object-id $identityPrincipalID \
```

ⓘ 중요

CMK가 사용하도록 설정되어 있는 동안 활성 고객 관리형 키에 대한 액세스 권한을 취소하면 학습 데이터 및 결과 파일 다운로드, 새 모델 미세 조정 및 미세 조정 모델 배포가 방지됩니다. 그러나 이전에 배포된 미세 조정 모델은 해당 배포가 삭제될 때 까지 계속 작동하고 트래픽을 제공합니다.

## 학습, 유효성 검사 및 학습 결과 데이터 삭제

파일 API를 사용하면 고객이 모델을 미세 조정하기 위해 학습 데이터를 업로드할 수 있습니다. 이 데이터는 리소스와 동일한 지역 내의 Azure Storage에 저장되며 Azure 구독 및 API 자격 증명과 논리적으로 격리됩니다. 업로드된 파일은 [DELETE API 작업](#)을 통해 사용자가 삭제할 수 있습니다.

## 미세 조정된 모델 및 배포 삭제

미세 조정 API를 사용하면 고객이 파일 API를 통해 서비스에 업로드한 학습 데이터를 기반으로 OpenAI 모델의 자체 미세 조정 버전을 만들 수 있습니다. 학습된 미세 조정된 모델은 동일한 지역의 Azure Storage에 저장되고, 미사용 시 암호화되고(Microsoft 관리형 키 또는 고객 관리형 키를 사용하여) Azure 구독 및 API 자격 증명으로 논리적으로 격리됩니다. 미세 조정된 모델 및 배포는 사용자가 [DELETE API 작업](#)을 호출하여 삭제할 수 있습니다.

## 고객 관리형 키 사용 안 함

고객 관리형 키를 사용하지 않도록 설정하면 Cognitive Services 리소스가 Microsoft 관리형 키를 사용하여 암호화됩니다. 고객 관리형 키를 사용하지 않도록 설정하려면 다음 단계를 수행합니다.

1. Cognitive Services 리소스로 이동한 다음 **암호화**를 선택합니다.
2. **Microsoft 관리형 키 > 저장**을 선택합니다.

이전에 고객 관리형 키를 사용하도록 설정하면 Azure AD의 기능인 시스템이 할당한 관리 ID도 사용하도록 설정되었습니다. 시스템 할당 관리 ID를 사용하도록 설정하면 이 리소스가 Azure Active Directory에 등록됩니다. 등록 후 관리 ID에는 고객 관리형 키를 설정하는 동안 선택된 키 자격 증명 모음에 대한 액세스 권한이 부여됩니다. 여기에서 [관리 ID](#)에 대해 자세히 알아볼 수 있습니다.

### ⓘ 중요

시스템 할당 관리 ID를 사용하지 않도록 설정하면 키 자격 증명 모음에 대한 액세스 권한이 제거되고 고객 키로 암호화된 데이터에 더 이상 액세스할 수 없게 됩니다. 이

데이터에 의존하는 기능은 작동하지 않습니다.

### ① 중요

관리 ID는 현재 교차 디렉터리 시나리오를 지원하지 않습니다. Azure Portal에 고객 관리형 키를 구성하는 경우 관리 ID가 내부적으로 자동 할당됩니다. 이후에 구독, 리소스 그룹 또는 리소스를 Azure AD 디렉터리 간에 이동하는 경우, 리소스와 연결된 관리 ID가 새로운 테넌트로 전송되지 않으므로 고객 관리형 키가 더 이상 작동하지 않을 수 있습니다. 자세한 내용은 FAQ에서 [Azure AD 디렉터리 간 구독 전송 및 Azure 리소스에 대한 관리 ID의 알려진 문제](#)를 참조하세요.

## 다음 단계

- 언어 서비스 고객 관리형 키 요청 양식 [↗](#)
- Azure Key Vault에 대해 자세히 알아보기

# Azure OpenAI Service를 사용한 BCDR(비즈니스 연속성 및 재해 복구) 고려 사항

아티클 • 2023. 04. 27.

Azure OpenAI는 두 지역에서 사용할 수 있습니다. 구독 키는 지역에 바인딩되므로 고객이 키를 획득할 때 배포가 상주할 지역을 선택하고 그때부터 모든 작업이 해당 Azure 서버 지역과 연결된 상태로 유지됩니다.

전체 지역에 적용되는 네트워크 문제가 발생하는 것은 드물지만 불가능한 것도 아닙니다. 서비스를 항상 사용할 수 있어야 하는 경우, 다른 지역으로 장애 조치(failover)하거나 둘 이상의 지역 간에 워크로드를 분할하도록 설계해야 합니다. 두 방식은 모두 서로 다른 지역에 있는 둘 이상의 OpenAI 리소스가 필요합니다. 이 문서에서는 Azure OpenAI 애플리케이션에 BCDR(비즈니스 연속성 및 재해 복구)을 구현하는 방법에 대한 일반적인 권장 사항을 제공합니다.

## 최선의 구현 방법

오늘날 고객은 배포 및 추론 모두에 대해 배포 중에 제공된 엔드포인트를 호출합니다. 이러한 작업은 상태 비저장이므로 지역을 사용할 수 없는 경우 데이터가 손실되지 않습니다.

지역이 작동하지 않는 경우 고객은 서비스 연속성을 보장하기 위한 조치를 취해야 합니다.

## 비즈니스 연속성

다음 지침 세트는 기본 엔드포인트를 사용하는 고객과 사용자 지정 엔드포인트를 사용하는 고객 모두에게 적용됩니다.

### 기본 엔드포인트 복구

기본 엔드포인트를 사용하는 경우 오류를 모니터링하도록 클라이언트 코드를 구성해야 하며, 오류가 지속되면 Azure OpenAI 구독이 있는 다른 지역을 선택하여 리디렉션할 준비를 해야 합니다.

다음 단계에 따라 오류를 모니터링하도록 클라이언트를 구성합니다.

1. 이 페이지를 사용하여 OpenAI 서비스에 사용할 수 있는 지역 목록을 식별할 수 있습니다.

2. 목록에서 기본 및 하나의 보조/백업 영역을 선택합니다.
3. 선택한 각 지역에 대해 OpenAI 서비스 리소스 만들기
4. 기본 지역 및 모든 백업 지역의 경우 코드가 다음을 알아야 합니다.
  - a. 리소스의 기본 URI
  - b. 지역 액세스 키 또는 Azure Active Directory 액세스
5. 연결 오류(일반적으로 연결 시간 초과 및 서비스 사용 불가 오류)를 모니터링하도록 코드를 구성합니다.
  - a. 네트워크에 일시적인 오류가 발생하므로 단일 연결 문제가 발생하는 경우 다시 시도하는 것이 좋습니다.
  - b. 지속성을 위해 만든 지역의 백업 리소스로 트래픽을 리디렉션합니다.

## BCDR에 사용자 지정 코드 필요

이 사용 유형에 대한 지역 장애 복구는 매우 저렴한 비용으로 즉각적으로 수행할 수 있습니다. 그러나 이렇게 하려면 애플리케이션의 클라이언트 쪽에서 이 기능을 사용자 지정 방식으로 개발해야 합니다.

# Azure OpenAI Service 모니터링

아티클 • 2023. 04. 27.

Azure 리소스를 사용하는 중요한 애플리케이션 및 비즈니스 프로세스가 있는 경우 이와 같은 리소스의 가용성, 성능 및 작업을 모니터링할 수 있습니다.

이 문서에서는 Azure OpenAI Service에서 생성된 데이터 모니터링에 대해 설명합니다. Azure OpenAI는 [Azure Monitor](#)를 사용하는 Cognitive Services의 일부입니다. 이 기능을 사용하는 모든 Azure 서비스에 공통되는 Azure Monitor 기능에 익숙하지 않은 경우 [Azure Monitor로 Azure 리소스 모니터링](#)을 읽어 보세요.

## 데이터 모니터링

Azure OpenAI는 [Azure 리소스에서 데이터 모니터링](#)에 설명된 다른 Azure 리소스와 동일한 종류의 모니터링 데이터를 수집합니다.

## 수집 및 라우팅

플랫폼 메트릭 및 활동 로그는 자동으로 수집되고 저장되지만 진단 설정을 사용하여 다른 위치로 라우팅할 수 있습니다.

리소스 로그는 진단 설정을 만들고 하나 이상의 위치로 라우팅할 때까지 수집 및 저장되지 않습니다.

Azure Portal, CLI 또는 PowerShell을 사용한 진단 설정 만들기의 자세한 프로세스는 [Azure에서 플랫폼 로그 및 메트릭을 수집하는 진단 설정 만들기](#)를 참조하세요. 진단 설정을 만들 때 수집할 로그 범주를 지정합니다.

진단 설정을 사용하고 Azure Monitor 로그로 데이터를 보내는 데는 추가 비용이 발생한다는 점에 유의하세요. 자세히 알아보려면 [Azure Monitor 비용 계산 가이드](#)를 참조하세요.

수집할 수 있는 메트릭 및 로그에 대해서는 다음 섹션에서 설명합니다.

## 메트릭 분석

[메트릭](#)을 열어 Azure OpenAI에 대한 메트릭을 분석할 수 있습니다. 이 메뉴는 Azure Portal에서 Azure OpenAI 리소스를 볼 때 **모니터링** 섹션 아래에 있습니다. 이 도구 사용에 대한 자세한 내용은 [Azure 메트릭 탐색기 시작](#)을 참조하세요.

Azure OpenAI는 Cognitive Services의 일부입니다. Cognitive Services 및 Azure OpenAI에 대해 수집된 모든 플랫폼 메트릭의 목록은 [Cognitive Services 지원 메트릭](#)을 참조하세요.

Azure OpenAI에서 사용할 수 있는 메트릭의 현재 하위 집합의 경우 다음과 같습니다.

## Azure OpenAI 메트릭

| 메트릭                    | 진단 설정을 통해 내보내기 가능? | 메트릭 표시 이름           | 단위  | 집계 | Description                              | 차원                                                              |
|------------------------|--------------------|---------------------|-----|----|------------------------------------------|-----------------------------------------------------------------|
| BlockedCalls           | 예                  | 차단된 호출              | 개수  | 합계 | 요금 또는 할당량 한도를 초과한 호출 수입니다.               | ApiName, OperationName, Region, RatelimitKey                    |
| ClientErrors           | 예                  | 클라이언트 오류            | 개수  | 합계 | 클라이언트 쪽 오류 (HTTP 응답 코드 4xx)가 있는 호출 수입니다. | ApiName, OperationName, Region, RatelimitKey                    |
| DataIn                 | 예                  | 데이터 입력              | 바이트 | 합계 | 들어오는 데이터 크기(바이트)입니다.                     | ApiName, OperationName, 지역                                      |
| DataOut                | 예                  | 데이터 출력              | 바이트 | 합계 | 나가는 데이터 크기(바이트)입니다.                      | ApiName, OperationName, 지역                                      |
| FineTunedTrainingHours | Yes                | 처리된 FineTuned 학습 시간 | 개수  | 합계 | OpenAI FineTuned 모델에서 처리된 학습 시간입니다.      | ApiName, ModelDeploymentName, FeatureName, UsageChannel, Region |
| 대기 시간                  | 예                  | 대기 시간               | 밀리초 | 평균 | 대기 시간(밀리초)입니다.                           | ApiName, OperationName, Region, RatelimitKey                    |

| 메트릭              | 진단 설정 을 통해 내보내기 가능? | 메트릭 표시 이름 | 단위 | 집계 형식 | Description                                  | 차원                                                              |
|------------------|---------------------|-----------|----|-------|----------------------------------------------|-----------------------------------------------------------------|
| Ratelimt         | Yes                 | Ratelimt  | 개수 | 합계    | ratelimit 키의 현재 ratelimit입니다.                | Region, RatelimitKey                                            |
| ServerErrors     | 예                   | 서버 오류     | 개수 | 합계    | 서비스 내부 오류(HTTP 응답 코드 5xx)가 있는 호출 수입니다.       | ApiName, OperationName, Region, RatelimitKey                    |
| SuccessfulCalls  | 예                   | 성공한 호출    | 개수 | 합계    | 성공한 호출 수입니다.                                 | ApiName, OperationName, Region, RatelimitKey                    |
| TokenTransaction | Yes                 | 처리된 유추 토큰 | 개수 | 합계    | OpenAI 모델에서 처리된 유추 토큰 수입니다.                  | ApiName, ModelDeploymentName, FeatureName, UsageChannel, Region |
| TotalCalls       | 예                   | 총 호출      | 개수 | 합계    | 총 호출 수.                                      | ApiName, OperationName, Region, RatelimitKey                    |
| TotalErrors      | 예                   | 총 오류      | 개수 | 합계    | 오류 응답 (HTTP 응답 코드 4xx 또는 5xx)이 있는 총 호출 수입니다. | ApiName, OperationName, Region, RatelimitKey                    |

## 로그 분석

Azure Monitor Logs의 데이터는 테이블마다 고유한 자체 속성 집합이 있는 테이블에 저장됩니다.

Azure Monitor의 모든 리소스 로그에는 동일한 필드와 그 뒤에 오는 서비스별 필드가 있습니다. 공용 스키마는 [Azure Monitor 리소스 로그 스키마](#)에서 설명합니다.

[활동 로그](#)는 구독 수준의 이벤트에 대한 인사이트를 제공하는 Azure의 플랫폼 로그 유형입니다. 활동 로그는 독립적으로 보거나 Azure Monitor Logs로 라우팅할 수 있습니다. 여기서 Log Analytics를 사용하여 훨씬 더 복잡한 쿼리를 수행할 수 있습니다.

Azure OpenAI 및 기타 Cognitive Services에 사용할 수 있는 리소스 로그 유형의 목록은 [Cognitive Services에 대한 리소스 공급자 작업](#)을 참조하세요.

## Kusto 쿼리

### ① 중요

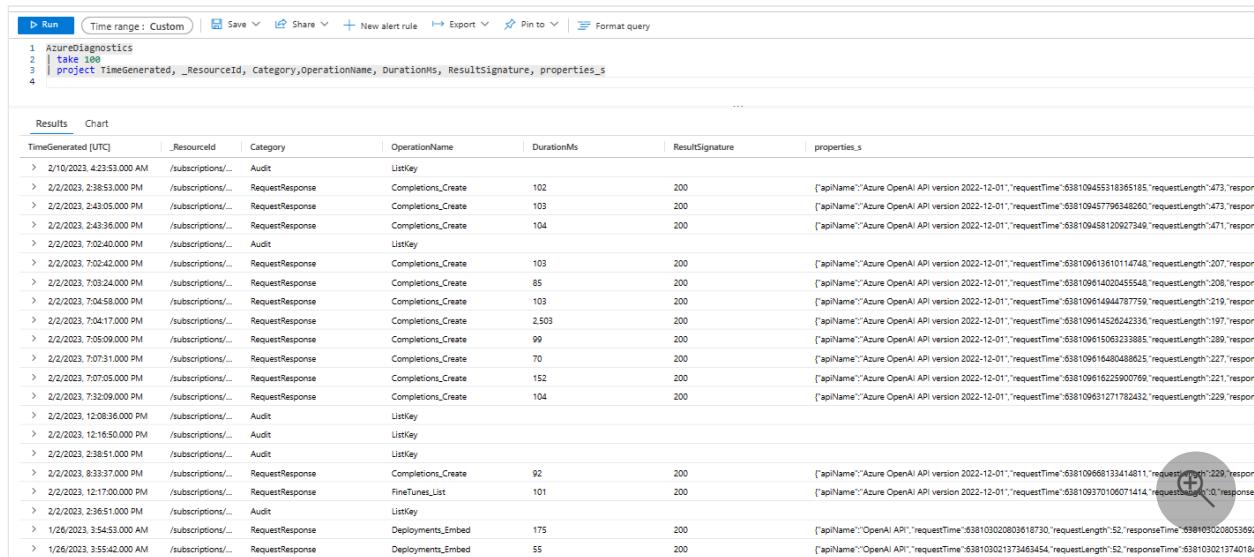
Azure OpenAI 메뉴에서 [로그](#)를 선택하면 쿼리 범위가 현재 Azure OpenAI 리소스로 설정된 Log Analytics가 열립니다. 즉, 로그 쿼리에는 해당 리소스의 데이터만 포함됩니다. 다른 리소스의 데이터 또는 다른 Azure 리소스의 데이터를 포함하는 쿼리를 실행하려면 Azure Monitor 메뉴에서 [로그](#)를 선택합니다. 자세한 내용은 [Azure Monitor Log Analytics의 로그 쿼리 범위 및 시간 범위](#)를 참조하세요.

Azure OpenAI 리소스에 사용할 수 있는 정보 유형을 탐색하고 파악하려면 모델을 배포하고 플레이그라운드를 통해 일부 완료 호출을 보낸 후 시작할 수 있는 유용한 쿼리는 다음과 같습니다.

```
Kusto

AzureDiagnostics
| take 100
| project TimeGenerated, _ResourceId, Category, OperationName, DurationMs,
ResultSignature, properties_s
```

여기서는 100개 항목의 샘플을 반환하고 로그에 사용 가능한 데이터 열의 하위 집합을 표시합니다. 결과는 다음과 같습니다.



The screenshot shows the Azure Kusto Query Editor interface. At the top, there's a toolbar with 'Run', 'Time range: Custom', 'Save', 'Share', 'New alert rule', 'Export', 'Pin to', and 'Format query'. Below the toolbar is the query editor area containing the following Kusto query:

```
1 AzureDiagnostics
2 | take 100
3 | project TimeGenerated, _ResourceId, Category, OperationName, DurationMs,
4 ResultSignature, properties_s
```

Below the query editor is a results table with columns: TimeGenerated [UTC], \_ResourceId, Category, OperationName, DurationMs, ResultSignature, and properties\_s. The table lists 100 rows of log data, each representing a specific API request or event. The data includes various operation names like 'Audit', 'Completions\_Create', and 'RequestResponse', along with their corresponding timestamps, resource IDs, categories, durations, signatures, and raw properties.

| project 줄에서 제공하는 범위 지정을 제거하면 사용 가능한 모든 데이터 열을 볼 수 있습니다.

```
Kusto

AzureDiagnostics
| take 100
```

테이블 이름 옆에 있는 화살표를 선택하여 사용 가능한 모든 열 및 관련 데이터 형식을 볼 수도 있습니다.

AzureMetrics를 검사하려면 다음을 실행합니다.

```
Kusto

AzureMetrics
| take 100
| project TimeGenerated, MetricName, Total, Count, TimeGrain, UnitName
```

| Results                    | Chart              |
|----------------------------|--------------------|
| TimeGenerated [UTC]        | _ResourceId        |
| > 2/10/2023, 4:23:53.00 AM | /subscriptions/... |
| > 2/2/2023, 2:38:53.00 PM  | /subscriptions/... |
| > 2/2/2023, 2:43:05.00 PM  | /subscriptions/... |
| > 2/2/2023, 2:43:36.00 PM  | /subscriptions/... |
| > 2/2/2023, 7:02:40.00 PM  | /subscriptions/... |
| > 2/2/2023, 7:02:42.00 PM  | /subscriptions/... |
| > 2/2/2023, 7:03:24.00 PM  | /subscriptions/... |
| > 2/2/2023, 7:04:58.00 PM  | /subscriptions/... |
| > 2/2/2023, 7:04:59.00 PM  | /subscriptions/... |
| > 2/2/2023, 7:05:09.00 PM  | /subscriptions/... |
| > 2/2/2023, 7:07:31.00 PM  | /subscriptions/... |
| > 2/2/2023, 7:07:35.00 PM  | /subscriptions/... |
| > 2/2/2023, 7:32:09.00 PM  | /subscriptions/... |
| > 2/2/2023, 12:08:36.00 PM | /subscriptions/... |
| > 2/2/2023, 12:16:50.00 PM | /subscriptions/... |
| > 2/2/2023, 2:38:51.00 PM  | /subscriptions/... |
| > 2/2/2023, 8:33:37.00 PM  | /subscriptions/... |
| > 2/2/2023, 12:17:00.00 PM | /subscriptions/... |
| > 2/2/2023, 2:36:51.00 PM  | /subscriptions/... |
| > 1/26/2023, 3:54:53.00 AM | /subscriptions/... |
| > 1/26/2023, 3:55:42.00 AM | /subscriptions/... |
|                            | Audit              |
|                            | RequestResponse    |
|                            | Completions_Create |
|                            | Completions_Create |
|                            | Completions_Create |
|                            | Completions_Create |
|                            | Audit              |
|                            | RequestResponse    |
|                            | Completions_Create |
|                            | Audit              |
|                            | RequestResponse    |
|                            | Completions_Create |
|                            | RequestResponse    |
|                            | Deployments_Embbed |
|                            | RequestResponse    |
|                            | Deployments_Embbed |

## 경고

Azure Monitor 경고는 모니터링 데이터에서 중요한 조건이 발견될 때 사용자에게 사전에 알립니다. 이를 통해 고객이 알기 전에 시스템 문제를 식별하고 해결할 수 있습니다. [메트릭, 로그 및 활동 로그](#)에서 경고를 설정할 수 있습니다. 서로 다른 형식의 경고에는 장점과 단점이 있습니다.

모든 조직의 경고 요구 사항은 다양해지고 시간이 지남에 따라 발전할 것입니다. 일반적으로 모든 경고는 경고가 발생하는 경우 의도된 특정 응답과 함께 실행 가능해야 합니다. 다른 사용자가 취해야 할 조치가 없는 경우 이는 사용자가 보고서에서 캡처하려고 하지만 경고에서는 캡처하지 않으려는 것일 수 있습니다. 일부 사용 사례에서는 특정 오류 조

건이 있을 때마다 경고가 필요할 수 있습니다. 그러나 많은 환경에서 경고 전송이 보증되는 기간에 오류가 특정 임계값을 초과하는 경우에만 발생할 수 있습니다.

특정 임계값 미만의 오류는 Azure Monitor 로그의 데이터를 정기적으로 분석하여 평가할 수 있는 경우가 많습니다. 시간이 지남에 따라 로그 데이터를 분석할 때 경고로 추적하는데 충분한 기간 발생하지 않는 특정 조건이 유용할 수도 있습니다. 경우에 따라 로그에 이벤트가 없는 것이 오류만큼 중요한 신호일 수 있습니다.

Azure OpenAI 사용과 함께 개발 중인 응용 프로그램 유형에 따라 [Azure Monitor Application Insights](#) 는 응용 프로그램 계층에서 추가 모니터링 이점을 제공할 수 있습니다.

## 다음 단계

- Azure 리소스 모니터링에 대한 자세한 내용은 [Azure Monitor를 사용하여 Azure 리소스 모니터링](#)을 참조하세요.
- [Azure Monitor 로그의 로그 검색 이해](#)를 읽어보세요.

# Azure OpenAI 서비스 비용 관리 계획

아티클 • 2023. 04. 28.

이 문서에서는 Azure OpenAI Service의 비용을 플랜 및 관리하는 방법을 설명합니다. 서비스를 배포하기 전에 Azure 가격 계산기를 사용하여 Azure OpenAI의 비용을 예측할 수 있습니다. 이후 Azure 리소스를 배포할 때 예상 비용을 검토합니다. Azure OpenAI 리소스 사용을 시작한 후 Cost Management 기능을 사용하여 예산을 설정하고 비용을 모니터링 합니다. 예상 비용을 검토하고 지출 추세를 파악하여 작업할 수 있는 영역을 식별할 수도 있습니다. Azure OpenAI Service 비용은 Azure 청구서의 월간 비용 중에 일부에 불과합니다. 이 문서에서는 Azure OpenAI에 대한 비용을 계획하고 관리하는 방법을 설명하지만, 사용자에게는 타사 서비스를 비롯한 Azure 구독에 사용되는 모든 Azure 서비스 및 리소스에 대해 요금이 청구됩니다.

## 사전 요구 사항

Cost Management에서의 비용 분석은 대부분의 Azure 계정 유형을 지원하지만 일부는 지원하지 않습니다. 지원되는 계정 유형의 전체 목록을 보려면 [Cost Management 데이터 이해](#)를 참조하세요. 비용 데이터를 보려면 적어도 Azure 계정에 대한 읽기 권한이 필요합니다. Azure Cost Management 데이터에 액세스하는 방법에 대한 정보는 [데이터에 대한 액세스 할당](#)을 참조하세요.

## Azure OpenAI를 사용하기 전에 비용 예측

Azure 가격 책정 계산기 [\[링크\]](#)를 사용하여 Azure OpenAI 사용 비용을 예측합니다.

## Azure OpenAI Service에 대한 전체 청구 모델 이해

Azure OpenAI Service는 새 리소스를 배포할 때 비용이 발생하는 Azure 인프라에서 실행됩니다. 다른 추가 인프라 비용이 발생할 수도 있다는 점을 이해하는 것이 중요합니다.

## Azure OpenAI Service 요금 청구 방식

### 기본 시리즈 및 Codex 시리즈 모델

Azure OpenAI 기본 시리즈 및 Codex 시리즈 모델은 토큰 1,000개당 요금이 청구됩니다. 비용은 Ada, Babbage, Curie, Davinci 또는 Code-Cushman 중 선택한 모델 시리즈에 따라 달라집니다.

모델은 텍스트를 토큰으로 세분화하여 이해하고 처리합니다. 참고로 각 토큰은 일반적인 영어 텍스트의 경우 대략 4자입니다.

토큰 비용은 입출력 모두에 대한 비용입니다. 예를 들어 Azure OpenAI 모델에 Python으로 변환하도록 요청하는 1,000개의 토큰 JavaScript 코드 샘플이 있는 경우입니다. 전송된 초기 입력 요청에 대해서는 약 1,000개의 토큰이 청구되고 총 2,000개의 토큰에 대한 응답으로 받은 출력에 대해서는 1,000개의 토큰이 추가로 청구됩니다.

실제로 이러한 유형의 완료 호출의 경우 토큰 입력/출력은 완벽하게 1:1이 아닐 수 있습니다. 한 프로그래밍 언어에서 다른 프로그래밍 언어로 변환하면 max\_tokens 매개 변수에 할당된 값을 포함하여 다양한 요인에 따라 출력이 길거나 짧아질 수 있습니다.

## 기본 시리즈 및 Codex 시리즈 미세 조정된 모델

Azure OpenAI 미세 조정된 모델은 다음 세 가지 요소에 따라 요금이 청구됩니다.

- 학습 시간
- 호스팅 시간
- 1,000개 토큰당 추론

호스팅 시간 비용은 미세 조정된 모델이 배포되면 적극적으로 사용되는지 여부에 관계없이 시간당 비용이 계속 발생하기 때문에 주의해야 합니다. 미세 조정된 모델 비용을 면밀히 모니터링해야 합니다.

### ① 중요

사용자 지정된 모델을 배포한 후 배포가 15일(15일) 이상 비활성 상태로 유지되면 배포가 자동으로 삭제됩니다. 모델이 15일 이상 전에 배포되었고 15일 연속으로 완료 또는 채팅 완료 호출이 이루어지지 않은 경우 사용자 지정된 모델의 배포는 "비활성"입니다. 비활성 배포의 삭제는 기본 사용자 지정된 모델을 삭제하거나 영향을 주지 않으며 사용자 지정된 모델을 언제든지 다시 배포할 수 있습니다. [Azure OpenAI 서비스 가격 책정](#)에 설명된 대로 배포된 각 사용자 지정(미세 조정된) 모델은 완료 또는 채팅 완료 호출이 모델에 대해 이루어지는지 여부에 관계없이 시간당 호스팅 비용이 발생합니다. Azure OpenAI를 사용하여 비용을 계획하고 관리하는 방법에 대한 자세한 내용은 [비용 관리 가이드](#)를 참조하세요.

## Azure OpenAI Service로 인해 발생할 수 있는 기타 비용

Azure Monitor 로그로 데이터 전송, 경고 등과 같은 기능을 사용하려면 이러한 서비스에 추가 비용이 발생합니다. 이러한 비용은 다른 서비스 및 구독 수준에서 볼 수 있지만 Azure OpenAI 리소스로만 범위가 지정되는 경우에는 볼 수 없습니다.

# Azure OpenAI Service와 함께 Azure 선불 사용

Azure 선불 크레딧을 사용하여 Azure OpenAI Service 요금을 지불할 수 있습니다. 단, Azure 선불 크레딧을 사용하여 Azure Marketplace에 있는 항목을 포함한 타사 제품 및 서비스에 대한 요금을 지불할 수는 없습니다.

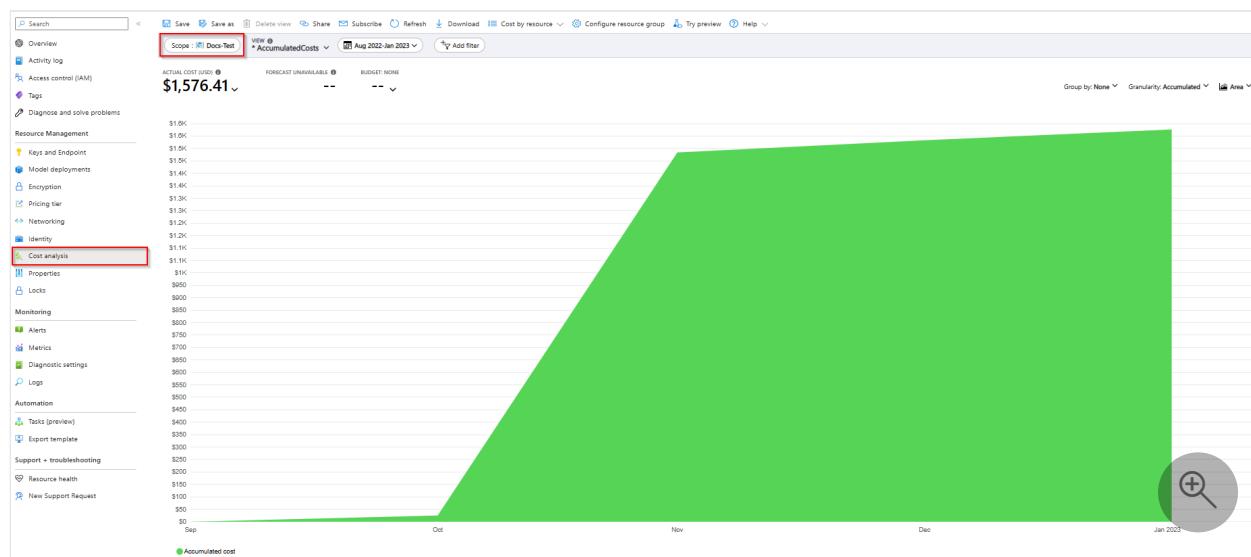
## 비용 모니터링

Azure OpenAI를 사용하여 Azure 리소스를 사용하는 경우 비용이 발생합니다. Azure 리소스 사용량 단위 비용은 시간 간격(초, 분, 시간 및 일) 또는 단위 사용량(바이트, 메가바이트 등)에 따라 다릅니다. Azure OpenAI 사용이 시작되는 즉시 비용이 발생할 수 있고 [비용 분석](#)에서 비용을 확인할 수 있습니다.

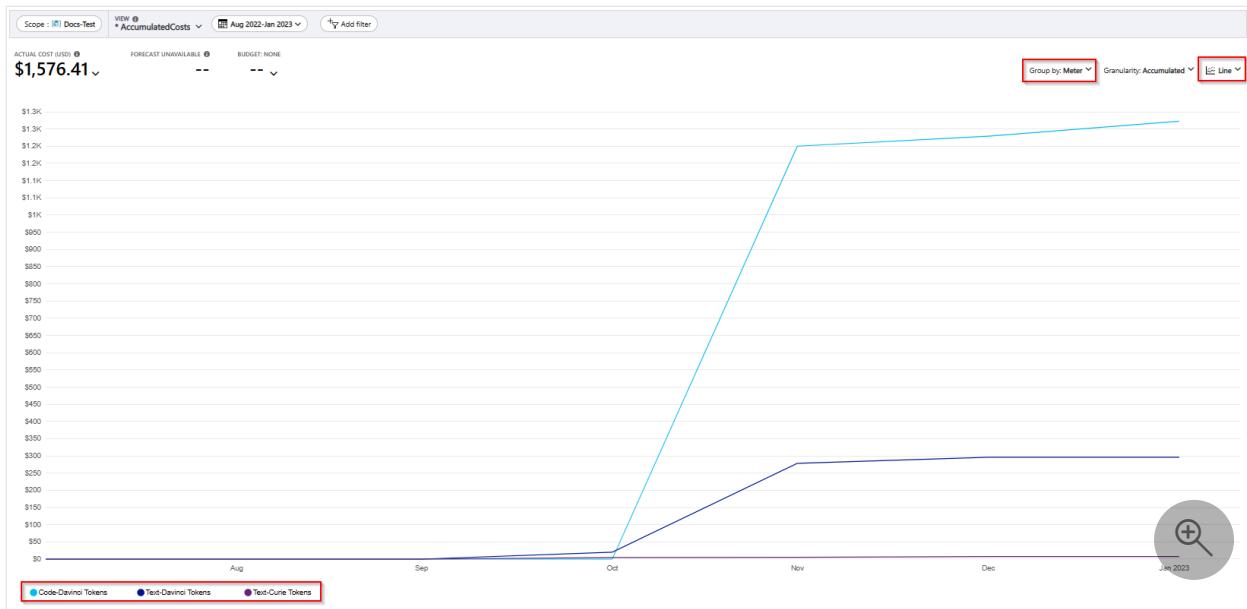
비용 분석을 사용하면 서로 다른 시간 간격에 대한 그래프 및 테이블의 Azure OpenAI 비용을 볼 수 있습니다. 몇 가지 예로 일, 현재 달과 이전 달 및 연도에 따라 확인할 수 있습니다. 예산 및 예상 비용에 대한 비용도 조회할 수 있습니다. 시간이 지남에 따라 더 긴 보기로 전환하면 지출 추세를 파악하는 데 도움이 됩니다. 그리고 과도한 지출이 발생한 위치를 확인할 수 있습니다. 예산을 만든 경우 초과된 부분도 쉽게 확인할 수 있습니다.

비용 분석에서 Azure OpenAI 비용을 보려면 다음을 수행합니다.

1. Azure Portal에 로그인합니다.
2. Azure OpenAI 리소스 중 하나를 선택합니다.
3. 리소스 관리에서 **비용 분석**을 선택합니다.
4. 기본적으로 비용 분석은 개별 Azure OpenAI 리소스로 범위가 지정됩니다.

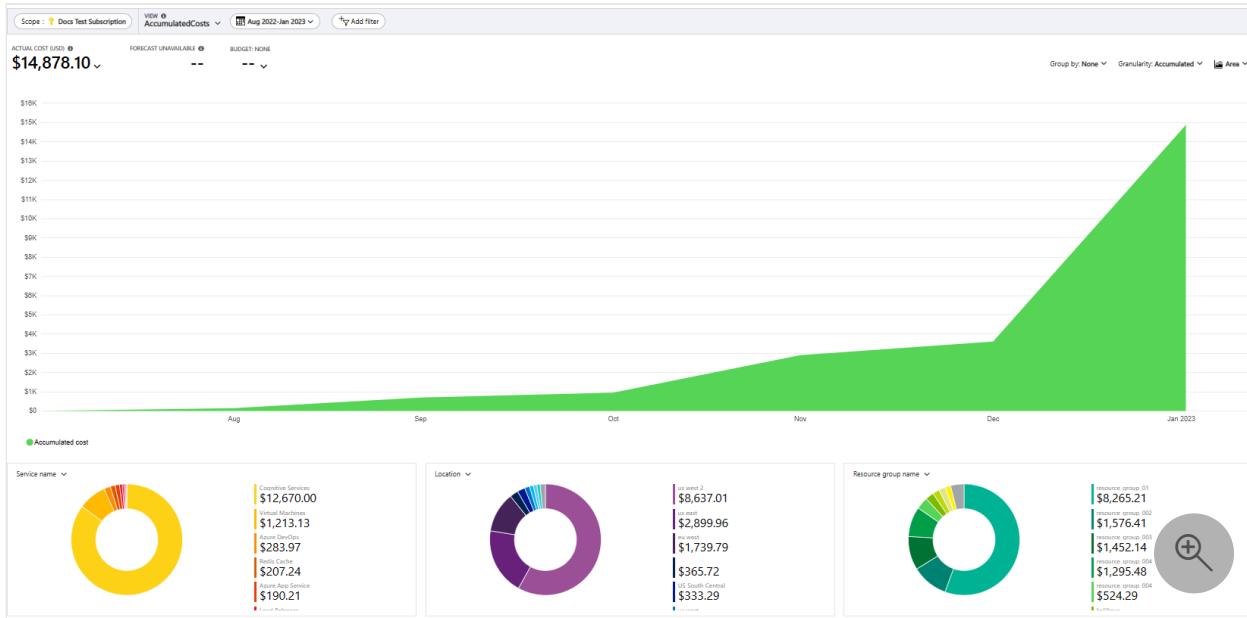


해당 비용을 구성하는 요소의 분석을 이해하려면 **그룹화 기준을 미터**로 수정하고 이 경우 차트 종류를 **꺾은선형**으로 전환하는 것이 도움이 될 수 있습니다. 이제 이 특정 리소스의 경우 비용의 출처가 비용의 대부분을 나타내는 **토큰**이 포함된 세 개의 서로 다른 모델 시리즈임을 알 수 있습니다.



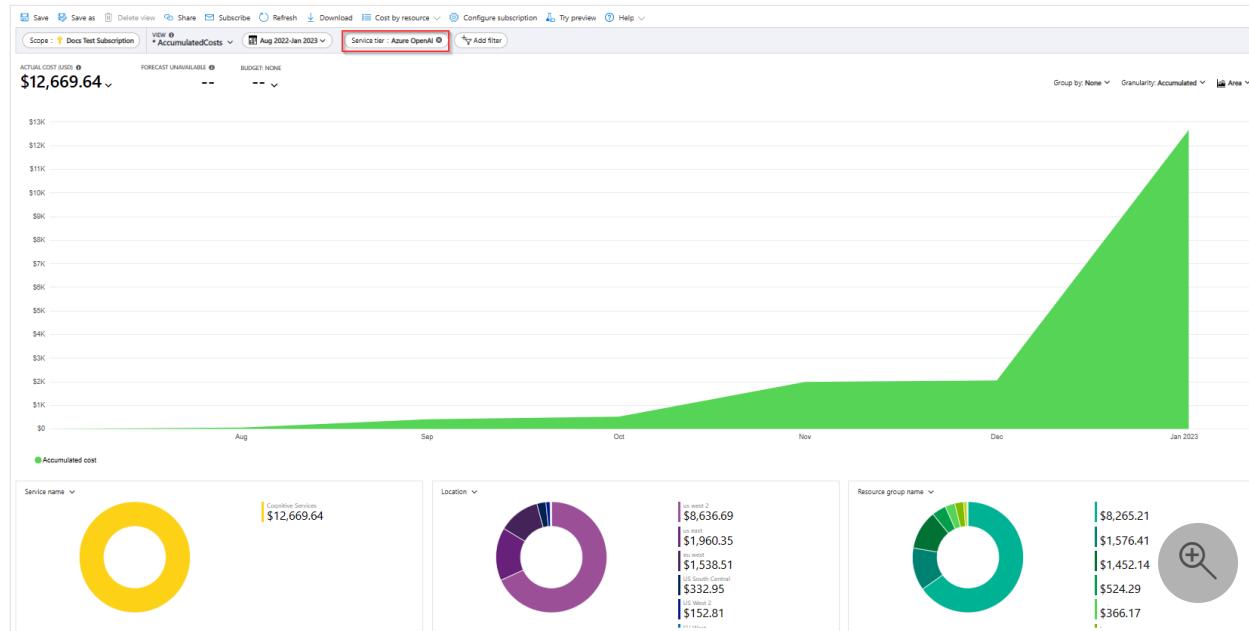
Azure OpenAI와 관련된 비용을 평가할 때 범위를 이해하는 것이 중요합니다. 리소스가 동일한 리소스 그룹에 속하는 경우 해당 수준에서 비용 분석의 범위를 지정하여 비용에 미치는 영향을 이해할 수 있습니다. 리소스가 여러 리소스 그룹에 분산되어 있으면 구독 수준까지 범위를 지정할 수 있습니다.

그러나 더 높은 수준에서 범위를 지정하는 경우 Azure OpenAI 사용량을 0으로 만들 수 있도록 필터를 추가해야 하는 경우가 많습니다. 구독 수준에서 범위가 지정되면 Azure OpenAI 비용 관리의 컨텍스트에서 신경 쓰지 않을 수 있는 다른 많은 리소스가 표시됩니다. 구독 수준에서 범위를 지정할 때 **Cost Management** 서비스에서 전체 **비용 분석 도구**로 이동하는 것이 좋습니다. 위쪽 Azure 검색 창에서 “**Cost Management**”를 검색하여 예산 만들기와 같은 더 많은 옵션을 포함하는 전체 서비스 환경으로 이동합니다.



서비스별로 필터를 추가하려고 하면 목록에서 Azure OpenAI를 찾을 수 없습니다. 이는 기술적으로 Azure OpenAI가 Cognitive Services의 일부이므로 서비스 수준 필터는 **Cognitive Services**이지만 다른 유형의 Cognitive Services 리소스가 없는 구독에서 모든

Azure OpenAI 리소스를 보려면 대신 서비스 계층: Azure OpenAI로 범위를 지정해야 합니다.



## 예산 만들기

예산을 만들면 비용을 관리하고 관련자에게 비정상 지출 및 과다 지출 위험을 자동으로 알리는 경고를 만들 수 있습니다. 경고는 예산 및 비용 임계값에 따른 지출을 기준으로 합니다. 예산 및 경고는 Azure 구독 및 리소스 그룹에 대해 만들어지므로 전체 비용 모니터링 전략의 일부로 유용합니다.

모니터링에 더 많은 세분성을 제공하려는 경우 Azure의 특정 리소스 또는 서비스에 대한 필터를 사용하여 예산을 만들 수 있습니다. 필터를 통해 추가 비용을 발생시키는 새 리소스를 실수로 만들지 않도록 할 수 있습니다. 예산을 만들 때 사용할 수 있는 필터 옵션에 대한 자세한 내용은 [그룹 및 필터 옵션](#)을 참조하세요.

### ⓘ 중요

OpenAI에는 예산 초과를 방지하는 하드 한도 옵션이 있지만 Azure OpenAI는 현재 이 기능을 제공하지 않습니다. 더 많은 고급 작업을 수행하려면 예산 알림의 일부로 작업 그룹에서 자동화를 시작할 수 있지만 이를 위해서는 사용자 지정 개발이 추가로 필요합니다.

## 비용 데이터 내보내기

스토리지 계정으로 [비용 데이터를 내보낼](#) 수도 있습니다. 이는 비용에 대한 추가 데이터 분석을 직접 수행해야 하거나 다른 사용자가 수행하는 경우에 유용합니다. 예를 들어 재

무 팀은 Excel 또는 Power BI를 사용하여 데이터를 분석할 수 있습니다. 매일, 매주 또는 매월 일정으로 비용을 내보내고 사용자 지정 날짜 범위를 설정할 수 있습니다. 비용 데이터를 내보내는 것은 비용 데이터 세트를 검색하는 데 권장되는 방법입니다.

## 다음 단계

- Azure Cost Management를 통해 클라우드 투자를 최적화하는 방법을 알아봅니다.
- 비용 분석을 통한 비용 관리에 대해 알아봅니다.
- 예기치 않은 비용을 방지하는 방법을 알아봅니다.
- Cost Management 단계별 학습 과정을 수강합니다.

# 자습서: Azure OpenAI Service 포함 및 문서 검색 살펴보기

아티클 • 2023. 04. 27.

이 자습서에서는 Azure OpenAI 포함 API를 사용하여 문서 검색을 수행하는 과정을 안내합니다. 여기에서 기술 자료를 쿼리하여 가장 관련성이 높은 문서를 찾습니다.

이 자습서에서는 다음 작업 방법을 알아봅니다.

- ✓ Azure OpenAI 및 기타 종속 Python 라이브러리를 설치합니다.
- ✓ BillSum 데이터 세트를 다운로드하고 분석을 위해 준비합니다.
- ✓ 리소스 엔드포인트 및 API 키에 대한 환경 변수를 만듭니다.
- ✓ **text-embedding-ada-002(버전 2)** 모델 사용
- ✓ 코사인 유사성을 사용하여 검색 결과의 순위를 지정합니다.

## ① 중요

`text-embedding-ada-002` (Version 2)를 사용하는 것이 좋습니다. 이 모델/버전은 OpenAI의 `text-embedding-ada-002`와 패리티를 제공합니다. 이 모델이 제공하는 개선 사항에 대해 자세히 알아보려면 [OpenAI의 블로그 게시물](#)을 참조하세요. 현재 버전 1을 사용 중이더라도 최신 가중치/업데이트된 토큰 한도를 활용하려면 버전 2로 마이그레이션해야 합니다. 버전 1과 버전 2는 상호 교환이 불가능하므로 동일한 버전의 모델을 사용하여 문서 포함과 문서 검색을 수행해야 합니다.

## 필수 구성 요소

- Azure 구독 - [체험 구독 만들기](#)
- 원하는 Azure 구독에서 Azure OpenAI에 부여된 액세스 권한 현재 이 서비스에 대한 액세스 권한은 애플리케이션에 의해서만 부여됩니다. <https://aka.ms/oai/access>에서 양식을 작성하여 Azure OpenAI에 대한 액세스를 신청할 수 있습니다. 문제가 있는 경우 이 리포지토리에서 문제를 엽니다.
- [Python 3.7.1 이상 버전](#)
- 다음 Python 라이브러리: openai, num2words, matplotlib, plotly, scipy, scikit-learn, Pandas, tiktoken.
- [Jupyter Notebook](#)
- **text-embedding-ada-002(버전 2)** 모델이 배포된 Azure OpenAI 리소스. 이 모델은 현재 특정 지역에서만 사용할 수 있습니다. 리소스가 없는 경우 만들기 프로세스는 [리소스 배포 가이드](#)에 설명되어 있습니다.

# 설정

## Python 라이브러리

아직 설치하지 않은 경우 다음 라이브러리를 설치해야 합니다.

Windows 명령 프롬프트

```
pip install openai num2words matplotlib plotly scipy scikit-learn pandas tiktoken
```

## BillSum 데이터 세트 다운로드

BillSum은 미국 의회 및 캘리포니아 주 법안의 데이터 세트입니다. 설명을 위해 미국 청구서만 살펴보겠습니다. 코퍼스는 의회의 103-115차(1993-2018) 세션의 법안으로 구성됩니다. 데이터는 18,949개의 학습 청구서와 3,269개의 테스트 청구서로 분할되었습니다. BillSum 코퍼스는 5,000자에서 20,000자 길이의 중간 길이 입법에 중점을 둡니다. 프로젝트에 대한 자세한 정보와 이 데이터 세트가 파생된 원본 학술 논문은 [BillSum 프로젝트의 GitHub 리포지토리](#)에서 확인할 수 있습니다.

이 자습서에서는 [GitHub 샘플 데이터](#)에서 다운로드할 수 있는 `bill_sum_data.csv` 파일을 사용합니다.

로컬 컴퓨터에서 다음 명령을 실행하여 샘플 데이터를 다운로드할 수도 있습니다.

Windows 명령 프롬프트

```
curl "https://raw.githubusercontent.com/Azure-Samples/Azure-OpenAI-Docs-Samples/main/Samples/Tutorials/Embeddings/data/bill_sum_data.csv" --output bill_sum_data.csv
```

## 키 및 엔드포인트 검색

Azure OpenAI에 대해 성공적으로 호출하려면 **엔드포인트**와 **키**가 필요합니다.

변수 이름  
값

**ENDPOINT** 이 값은 Azure Portal에서 리소스를 검사할 때 **키 및 엔드포인트** 섹션에서 찾을 수 있습니다. 또는 Azure OpenAI Studio>플레이그라운드>코드 보기에서 값을 찾을 수 있습니다. 예제 엔드포인트는 <https://docs-test-001.openai.azure.com>입니다.

## 변수 이름

**API-KEY** 이 값은 Azure Portal에서 리소스를 검사할 때 키 및 엔드포인트 섹션에서 찾을 수 있습니다. KEY1 또는 KEY2를 사용할 수 있습니다.

Azure Portal에서 해당 리소스로 이동합니다. **엔드포인트 및 키는 리소스 관리** 섹션에서 찾을 수 있습니다. 엔드포인트 및 액세스 키를 복사합니다. API 호출을 인증하는 데 모두 필요합니다. KEY1 또는 KEY2를 사용할 수 있습니다. 항상 두 개의 키를 사용하면 서비스 중단 없이 키를 안전하게 회전하고 다시 생성할 수 있습니다.

키 및 엔드포인트에 대한 영구 환경 변수를 만들고 할당합니다.

## 환경 변수

### 명령줄

#### CMD

```
setx AZURE_OPENAI_API_KEY "REPLACE_WITH_YOUR_KEY_VALUE_HERE"
```

#### CMD

```
setx AZURE_OPENAI_ENDPOINT "REPLACE_WITH_YOUR_ENDPOINT_HERE"
```

환경 변수를 설정한 후 환경 변수에 액세스하려면 Jupyter Notebooks 또는 사용 중인 IDE를 닫았다가 다시 열어야 할 수 있습니다. Jupyter Notebooks를 사용하는 것이 강력히 권장되지만, 어떤 이유로든 할 수 없는 경우 코드 블록의 끝에서 자주 수행되는 것처럼 `dataframe_name`을 직접 호출하는 대신 `print(dataframe_name)`을 사용하여 pandas 데이터 프레임을 반환하는 모든 코드를 수정해야 합니다.

기본 설정하는 Python IDE에서 다음 코드를 실행합니다.

## 라이브러리 및 목록 모델 가져오기

### Python

```
import openai
import os
import re
import requests
import sys
```

```

from num2words import num2words
import os
import pandas as pd
import numpy as np
from openai.embeddings_utils import get_embedding, cosine_similarity
import tiktoken

API_KEY = os.getenv("AZURE_OPENAI_API_KEY")
RESOURCE_ENDPOINT = os.getenv("AZURE_OPENAI_ENDPOINT")

openai.api_type = "azure"
openai.api_key = API_KEY
openai.api_base = RESOURCE_ENDPOINT
openai.api_version = "2023-05-15"

url = openai.api_base + "/openai/deployments?api-version=2023-05-15"

r = requests.get(url, headers={"api-key": API_KEY})

print(r.text)

```

출력

```
{
  "data": [
    {
      "scale_settings": {
        "scale_type": "standard"
      },
      "model": "text-embedding-ada-002",
      "owner": "organization-owner",
      "id": "text-embedding-ada-002",
      "status": "succeeded",
      "created_at": 1657572678,
      "updated_at": 1657572678,
      "object": "deployment"
    },
    {
      "scale_settings": {
        "scale_type": "standard"
      },
      "model": "code-cushman-001",
      "owner": "organization-owner",
      "id": "code-cushman-001",
      "status": "succeeded",
      "created_at": 1657572712,
      "updated_at": 1657572712,
      "object": "deployment"
    },
    {
      "scale_settings": {
        "scale_type": "standard"
      },

```

```
        "model": "text-search-curie-doc-001",
        "owner": "organization-owner",
        "id": "text-search-curie-doc-001",
        "status": "succeeded",
        "created_at": 1668620345,
        "updated_at": 1668620345,
        "object": "deployment"
    },
{
    "scale_settings": {
        "scale_type": "standard"
    },
    "model": "text-search-curie-query-001",
    "owner": "organization-owner",
    "id": "text-search-curie-query-001",
    "status": "succeeded",
    "created_at": 1669048765,
    "updated_at": 1669048765,
    "object": "deployment"
}
],
"object": "list"
}
```

이 명령의 출력은 배포한 모델의 수와 형식에 따라 달라집니다. 이 경우 **text-embedding-ada-002**에 대한 항목이 있는지 확인해야 합니다. 이 모델이 누락된 경우 계속하기 전에 리소스에 [모델을 배포](#)해야 합니다.

이제 csv 파일을 읽고 Pandas DataFrame을 만들어야 합니다. 초기 DataFrame이 만들어진 후 `df`를 실행하여 테이블의 콘텐츠를 볼 수 있습니다.

Python

```
df=pd.read_csv(os.path.join(os.getcwd(), 'bill_sum_data.csv')) # This assumes
# that you have placed the bill_sum_data.csv in the same directory you are
# running Jupyter Notebooks
df
```

**출력:**

| Unnamed: 0 | bill_id | text       | summary                                           | title                                             | text_len                                          | sum_len    |
|------------|---------|------------|---------------------------------------------------|---------------------------------------------------|---------------------------------------------------|------------|
| 0          | 0       | 110_hr37   | SECTION 1. SHORT TITLE\n\n This Act ma...         | National Science Education Tax Incentive for B... | To amend the Internal Revenue Code of 1986 to ... | 8494 321   |
| 1          | 1       | 112_hr2873 | SECTION 1. SHORT TITLE\n\n This Act ma...         | Small Business Expansion and Hiring Act of 201... | To amend the Internal Revenue Code of 1986 to ... | 6522 1424  |
| 2          | 2       | 109_s2408  | SECTION 1. RELEASE OF DOCUMENTS CAPTURED IN IR... | Requires the Director of National Intelligence... | A bill to require the Director of National Int... | 6154 463   |
| 3          | 3       | 108_s1899  | SECTION 1. SHORT TITLE\n\n This Act ma...         | National Cancer Act of 2003 - Amends the Publi... | A bill to improve data collection and dissemin... | 19853 1400 |
| 4          | 4       | 107_s1531  | SECTION 1. SHORT TITLE\n\n This Act ma...         | Military Call-up Relief Act - Amends the Inter... | A bill to amend the Internal Revenue Code of 1... | 6273 278   |
| 5          | 5       | 107_hr4541 | SECTION 1. RELIQUIDATION OF CERTAIN ENTRIES PR... | Requires the Customs Service to reliquidate ce... | To provide for reliquidation of entries premat... | 11691 114  |
| 6          | 6       | 111_s1495  | SECTION 1. SHORT TITLE\n\n This Act ma...         | Service Dogs for Veterans Act of 2009 - Direct... | A bill to require the Secretary of Veterans Af... | 5328 379   |
| 7          | 7       | 111_s3885  | SECTION 1. SHORT TITLE\n\n This Act ma...         | Race to the Top Act of 2010 - Directs the Secr... | A bill to provide incentives for States and lo... | 16668 1525 |
| 8          | 8       | 113_hr1796 | SECTION 1. SHORT TITLE\n\n This Act ma...         | Troop Talent Act of 2013 - Directs the Secreta... | Troop Talent Act of 2013                          | 15352 2151 |
| 9          | 9       | 103_hr1987 | SECTION 1. SHORT TITLE\n\n This Act ma...         | Taxpayer's Right to View Act of 1993 - Amends ... | Taxpayer's Right to View Act of 1993              | 5633 894   |
| 10         | 10      | 103_hr1677 | SECTION 1. SHORT TITLE\n\n This Act ma...         | Full-Service Schools Act - Establishes the Fed... | Full-Service Schools Act                          | 12472 1107 |
| 11         | 11      | 111_s3149  | SECTION 1. SHORT TITLE\n\n This Act ma...         | Wall Street Compensation Reform Act of 2010 - ... | A bill to amend the Internal Revenue Code of 1... | 18226 1297 |
| 12         | 12      | 110_hr1007 | SECTION 1. FINDINGS.\n\n The Congress f...        | Amends the Marine Mammal Protection Act of 197... | To amend the Marine Mammal Protection Act of 1... | 5261 276   |
| 13         | 13      | 113_hr3137 | SECTION 1. SHORT TITLE\n\n This Act ma...         | Freedom and Mobility in Consumer Banking Act -... | Freedom and Mobility in Consumer Banking Act      | 17690 2044 |
| 14         | 14      | 115_hr1634 | SECTION 1. SHORT TITLE\n\n This Act ma...         | Education and Training for Health Act of 2017 ... | Education and Training for Health Act of 2017     | 9037 772   |
| 15         | 15      | 103_hr1815 | SECTION 1. SHORT TITLE\n\n This Act ma...         | Recreational Hunting Safety and Preservation A... | Recreational Hunting Safety and Preservation A... | 13024 475  |
| 16         | 16      | 113_s1773  | SECTION 1. SHORT TITLE\n\n This Act ma...         | Andrew Prior Act or Andrew's Law - Amends the ... | Andrew's Law                                      | 5149 613   |
| 17         | 17      | 106_hr5585 | SECTION 1. SHORT TITLE\n\n This Act ma...         | Directs the President, in coordination with de... | Energy Independence Act of 2000                   | 8007 810   |
| 18         | 18      | 114_hr2499 | SECTION 1. SHORT TITLE.\n This Act may be...      | This measure has not been amended since it was... | Veterans Entrepreneurship Act of 2015             | 7539 1421  |
| 19         | 19      | 111_hr3141 | SECTION 1. SHORT TITLE\n\n This Act ma...         | Strengthening the Health Care Safety Net Act o... | To amend title XIX of the Social Security Act ... | 18429 514  |

초기 테이블에는 필요한 것보다 더 많은 열이 있습니다. `text`, `summary` 및 `title`에 대한  
열만 포함하는 `df_bills`라는 더 작은 새 DataFrame을 만듭니다.

Python

```
df_bills = df[['text', 'summary', 'title']]
df_bills
```

출력:

|    |  | text                                              | summary                                           | title                                             |
|----|--|---------------------------------------------------|---------------------------------------------------|---------------------------------------------------|
| 0  |  | SECTION 1. SHORT TITLE\n\n This Act may be...     | National Science Education Tax Incentive for B... | To amend the Internal Revenue Code of 1986 to ... |
| 1  |  | SECTION 1. SHORT TITLE\n\n This Act may be...     | Small Business Expansion and Hiring Act of 201... | To amend the Internal Revenue Code of 1986 to ... |
| 2  |  | SECTION 1. RELEASE OF DOCUMENTS CAPTURED IN IR... | Requires the Director of National Intelligence... | A bill to require the Director of National Int... |
| 3  |  | SECTION 1. SHORT TITLE\n\n This Act may be...     | National Cancer Act of 2003 - Amends the Publi... | A bill to improve data collection and dissemin... |
| 4  |  | SECTION 1. SHORT TITLE\n\n This Act may be...     | Military Call-up Relief Act - Amends the Inter... | A bill to amend the Internal Revenue Code of 1... |
| 5  |  | SECTION 1. RELIQUIDATION OF CERTAIN ENTRIES PR... | Requires the Customs Service to reliquidate ce... | To provide for reliquidation of entries premat... |
| 6  |  | SECTION 1. SHORT TITLE\n\n This Act may be...     | Service Dogs for Veterans Act of 2009 - Direct... | A bill to require the Secretary of Veterans Af... |
| 7  |  | SECTION 1. SHORT TITLE\n\n This Act may be...     | Race to the Top Act of 2010 - Directs the Secr... | A bill to provide incentives for States and lo... |
| 8  |  | SECTION 1. SHORT TITLE\n\n This Act may be...     | Troop Talent Act of 2013 - Directs the Secreta... | Troop Talent Act of 2013                          |
| 9  |  | SECTION 1. SHORT TITLE\n\n This Act may be...     | Taxpayer's Right to View Act of 1993 - Amends ... | Taxpayer's Right to View Act of 1993              |
| 10 |  | SECTION 1. SHORT TITLE\n\n This Act may be...     | Full-Service Schools Act - Establishes the Fed... | Full-Service Schools Act                          |
| 11 |  | SECTION 1. SHORT TITLE\n\n This Act may be...     | Wall Street Compensation Reform Act of 2010 - ... | A bill to amend the Internal Revenue Code of 1... |
| 12 |  | SECTION 1. FINDINGS.\n\n The Congress finds...    | Amends the Marine Mammal Protection Act of 197... | To amend the Marine Mammal Protection Act of 1... |
| 13 |  | SECTION 1. SHORT TITLE\n\n This Act may be...     | Freedom and Mobility in Consumer Banking Act -... | Freedom and Mobility in Consumer Banking Act      |
| 14 |  | SECTION 1. SHORT TITLE\n\n This Act may be...     | Education and Training for Health Act of 2017 ... | Education and Training for Health Act of 2017     |
| 15 |  | SECTION 1. SHORT TITLE\n\n This Act may be...     | Recreational Hunting Safety and Preservation A... | Recreational Hunting Safety and Preservation A... |
| 16 |  | SECTION 1. SHORT TITLE\n\n This Act may be...     | Andrew Prior Act or Andrew's Law - Amends the ... | Andrew's Law                                      |
| 17 |  | SECTION 1. SHORT TITLE\n\n This Act may be...     | Directs the President, in coordination with de... | Energy Independence Act of 2000                   |
| 18 |  | SECTION 1. SHORT TITLE.\n This Act may be c...    | This measure has not been amended since it was... | Veterans Entrepreneurship Act of 2015             |
| 19 |  | SECTION 1. SHORT TITLE\n\n This Act may be...     | Strengthening the Health Care Safety Net Act o... | To amend title XIX of the Social Security Act ... |

다음으로 불필요한 공백을 제거하고 문장 부호를 정리하여 토큰화를 위한 데이터를 준비  
하여 간단한 데이터 정리를 수행합니다.

Python

```
pd.options.mode.chained_assignment = None #https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#evaluation-order-matters

# s is input text
def normalize_text(s, sep_token = " \n "):
    s = re.sub(r'\s+', ' ', s).strip()
    s = re.sub(r". ,","",s)
    # remove all instances of multiple spaces
    s = s.replace(..,"..")
    s = s.replace(.. .,"..")
    s = s.replace("\n", "")
    s = s.strip()

    return s

df_bills['text']= df_bills["text"].apply(lambda x : normalize_text(x))
```

이제 토큰 제한(8192 토큰)에 비해 너무 긴 청구서를 제거해야 합니다.

Python

```
tokenizer = tiktoken.get_encoding("cl100k_base")
df_bills['n_tokens'] = df_bills["text"].apply(lambda x:
len(tokenizer.encode(x)))
df_bills = df_bills[df_bills.n_tokens<8192]
len(df_bills)
```

출력

20

### ① 참고

이 경우 모든 청구서는 포함 모델 입력 토큰 한도에 속하지만 위의 기술을 사용하여 포함 실패를 유발할 수 있는 항목을 제거할 수 있습니다. 포함 제한을 초과하는 콘텐츠에 직면하면 콘텐츠를 더 작은 조각으로 청크한 다음 한 번에 하나씩 포함할 수 있습니다.

다시 한 번 `df_bills`를 검토합니다.

Python

```
df_bills
```

## 출력:

|    | text                                              | summary                                           | title                                             | n_tokens |
|----|---------------------------------------------------|---------------------------------------------------|---------------------------------------------------|----------|
| 0  | SECTION 1. SHORT TITLE. This Act may be cited ... | National Science Education Tax Incentive for B... | To amend the Internal Revenue Code of 1986 to ... | 1466     |
| 1  | SECTION 1. SHORT TITLE. This Act may be cited ... | Small Business Expansion and Hiring Act of 201... | To amend the Internal Revenue Code of 1986 to ... | 1183     |
| 2  | SECTION 1. RELEASE OF DOCUMENTS CAPTURED IN IR... | Requires the Director of National Intelligence... | A bill to require the Director of National Int... | 937      |
| 3  | SECTION 1. SHORT TITLE. This Act may be cited ... | National Cancer Act of 2003 - Amends the Publi... | A bill to improve data collection and dissemin... | 3670     |
| 4  | SECTION 1. SHORT TITLE. This Act may be cited ... | Military Call-up Relief Act - Amends the Inter... | A bill to amend the Internal Revenue Code of 1... | 1038     |
| 5  | SECTION 1. RELIQUIDATION OF CERTAIN ENTRIES PR... | Requires the Customs Service to reliquidate ce... | To provide for reliquidation of entries premat... | 2026     |
| 6  | SECTION 1. SHORT TITLE. This Act may be cited ... | Service Dogs for Veterans Act of 2009 - Direct... | A bill to require the Secretary of Veterans Af... | 880      |
| 7  | SECTION 1. SHORT TITLE. This Act may be cited ... | Race to the Top Act of 2010 - Directs the Secr... | A bill to provide incentives for States and lo... | 2815     |
| 8  | SECTION 1. SHORT TITLE. This Act may be cited ... | Troop Talent Act of 2013 - Directs the Secreta... | Troop Talent Act of 2013                          | 2479     |
| 9  | SECTION 1. SHORT TITLE. This Act may be cited ... | Taxpayer's Right to View Act of 1993 - Amends ... | Taxpayer's Right to View Act of 1993              | 947      |
| 10 | SECTION 1. SHORT TITLE. This Act may be cited ... | Full-Service Schools Act - Establishes the Fed... | Full-Service Schools Act                          | 2164     |
| 11 | SECTION 1. SHORT TITLE. This Act may be cited ... | Wall Street Compensation Reform Act of 2010 - ... | A bill to amend the Internal Revenue Code of 1... | 2331     |
| 12 | SECTION 1. FINDINGS. The Congress finds the fo... | Amends the Marine Mammal Protection Act of 197... | To amend the Marine Mammal Protection Act of 1... | 1192     |
| 13 | SECTION 1. SHORT TITLE. This Act may be cited ... | Freedom and Mobility in Consumer Banking Act -... | Freedom and Mobility in Consumer Banking Act      | 2402     |
| 14 | SECTION 1. SHORT TITLE. This Act may be cited ... | Education and Training for Health Act of 2017 ... | Education and Training for Health Act of 2017     | 1648     |
| 15 | SECTION 1. SHORT TITLE. This Act may be cited ... | Recreational Hunting Safety and Preservation A... | Recreational Hunting Safety and Preservation A... | 2209     |
| 16 | SECTION 1. SHORT TITLE. This Act may be cited ... | Andrew Prior Act or Andrew's Law - Amends the ... | Andrew's Law                                      | 608      |
| 17 | SECTION 1. SHORT TITLE. This Act may be cited ... | Directs the President, in coordination with de... | Energy Independence Act of 2000                   | 1352     |
| 18 | SECTION 1. SHORT TITLE. This Act may be cited ... | This measure has not been amended since it was... | Veterans Entrepreneurship Act of 2015             | 1393     |
| 19 | SECTION 1. SHORT TITLE. This Act may be cited ... | Strengthening the Health Care Safety Net Act o... | To amend title XIX of the Social Security Act ... | 2678     |

n\_tokens 열과 텍스트가 궁극적으로 토큰화되는 방식을 조금 더 이해하려면 다음 코드를 실행하는 것이 도움이 될 수 있습니다.

### Python

```
sample_encode = tokenizer.encode(df_bills.text[0])
decode = tokenizer.decode_tokens_bytes(sample_encode)
decode
```

문서의 경우 의도적으로 출력을 자르지만 환경에서 이 명령을 실행하면 청크로 토큰화된 인덱스 0의 전체 텍스트가 반환됩니다. 어떤 경우에는 전체 단어가 단일 토큰으로 표시되는 반면 다른 경우에는 단어의 일부가 여러 토큰으로 분할되는 것을 볼 수 있습니다.

### 출력

```
[b'SECTION',
 b' ',
 b'1',
 b'.',
 b' SHORT',
 b' TITLE',
 b'.',
 b' This',
 b' Act',
 b' may',
 b' be',
 b' cited',
 b' as',
```

```
b' the',
b' ``',
b'National',
b' Science',
b' Education',
b' Tax',
b' In',
b'cent',
b'ive',
b' for',
b' Businesses',
b' Act',
b' of',
b' ',
b'200',
b'7',
b'''."',
b' SEC',
b'.',
b' ',
b'2',
b'.',
b' C',
b'RED',
b'ITS',
b' FOR',
b' CERT',
b'AIN',
b' CONTRIBUT',
b'IONS',
b' BEN',
b'EF',
b'IT',
b'ING',
b' SC',
```

그런 다음 `decode` 변수의 길이를 확인하면 `n_tokens` 열의 첫 번째 숫자와 일치함을 알 수 있습니다.

Python

```
len(decode)
```

출력

```
1466
```

이제 토큰화가 작동하는 방식에 대해 더 많이 이해했으므로 포함으로 넘어갈 수 있습니다. 아직 문서를 실제로 토큰화하지 않았다는 점에 유의하는 것이 중요합니다. `n_tokens`

열은 단순히 토큰화 및 포함을 위해 모델에 전달하는 데이터가 입력 토큰 제한인 8,192를 초과하지 않도록 하는 방법입니다. 포함 모델에 문서를 전달하면 문서를 위의 예와 유사한 토큰(반드시 동일하지는 않음)으로 나눈 다음 토큰을 벡터 검색을 통해 액세스할 수 있는 일련의 부동 소수점 숫자로 변환합니다. 이러한 포함은 로컬 또는 Azure 데이터베이스에 저장할 수 있습니다. 결과적으로 각 청구서에는 DataFrame의 오른쪽에 있는 새 `ada_v2` 열에 해당하는 자체 포함 벡터가 포함됩니다.

### Python

```
df_bills['ada_v2'] = df_bills["text"].apply(lambda x : get_embedding(x, engine = 'text-embedding-ada-002')) # engine should be set to the deployment name you chose when you deployed the text-embedding-ada-002 (Version 2) model
```

### Python

```
df_bills
```

## 출력:

|    | text                                              | summary                                           | title                                             | n_tokens | ada_v2                                             |
|----|---------------------------------------------------|---------------------------------------------------|---------------------------------------------------|----------|----------------------------------------------------|
| 0  | SECTION 1. SHORT TITLE. This Act may be cited ... | National Science Education Tax Incentive for B... | To amend the Internal Revenue Code of 1986 to ... | 1466     | [0.01333628874272108, -0.02151912823319435, 0...   |
| 1  | SECTION 1. SHORT TITLE. This Act may be cited ... | Small Business Expansion and Hiring Act of 201... | To amend the Internal Revenue Code of 1986 to ... | 1183     | [0.005016345530748367, -0.00569863710552454, 0...  |
| 2  | SECTION 1. RELEASE OF DOCUMENTS CAPTURED IN IR... | Requires the Director of National Intelligence... | A bill to require the Director of National Int... | 937      | [0.012699966318905354, -0.01897779107093811, 0...  |
| 3  | SECTION 1. SHORT TITLE. This Act may be cited ... | National Cancer Act of 2003 - Amends the Publi... | A bill to improve data collection and dissemin... | 3670     | [0.004736857954412699, -0.026448562741279602, ...  |
| 4  | SECTION 1. SHORT TITLE. This Act may be cited ... | Military Call-up Relief Act - Amends the Inter... | A bill to amend the Internal Revenue Code of 1... | 1038     | [0.010082815773785114, -0.0007545037078671157, ... |
| 5  | SECTION 1. RELIQUIDATION OF CERTAIN ENTRIES PR... | Requires the Customs Service to reliquidate ce... | To provide for reliquidation of entries premat... | 2026     | [0.012738252058625221, 0.004982588812708855, 0...  |
| 6  | SECTION 1. SHORT TITLE. This Act may be cited ... | Service Dogs for Veterans Act of 2009 - Direct... | A bill to require the Secretary of Veterans Af... | 880      | [0.005205095745623112, -0.016558492556214333, ...  |
| 7  | SECTION 1. SHORT TITLE. This Act may be cited ... | Race to the Top Act of 2010 - Directs the Secr... | A bill to provide incentives for States and lo... | 2815     | [0.024539386853575706, -0.016805868595838547, ...  |
| 8  | SECTION 1. SHORT TITLE. This Act may be cited ... | Troop Talent Act of 2013 - Directs the Secreta... | Troop Talent Act of 2013                          | 2479     | [-0.005527574568986893, -0.014311426319181919, ... |
| 9  | SECTION 1. SHORT TITLE. This Act may be cited ... | Taxpayer's Right to View Act of 1993 - Amends ... | Taxpayer's Right to View Act of 1993              | 947      | [0.004519130103290081, -0.023599395528435707, ...  |
| 10 | SECTION 1. SHORT TITLE. This Act may be cited ... | Full-Service Schools Act - Establishes the Fed... | Full-Service Schools Act                          | 2164     | [0.0075974976643919945, -0.006962535437196493, ... |
| 11 | SECTION 1. SHORT TITLE. This Act may be cited ... | Wall Street Compensation Reform Act of 2010 - ... | A bill to amend the Internal Revenue Code of 1... | 2331     | [0.014871294610202312, -0.001929433667100966, ...  |
| 12 | SECTION 1. FINDINGS. The Congress finds the fo... | Amends the Marine Mammal Protection Act of 197... | To amend the Marine Mammal Protection Act of 1... | 1192     | [0.04441450908780098, 0.02687789686024189, 0...    |
| 13 | SECTION 1. SHORT TITLE. This Act may be cited ... | Freedom and Mobility in Consumer Banking Act -... | Freedom and Mobility in Consumer Banking Act      | 2402     | [0.021314678713679314, -0.00831076875320176, ...   |
| 14 | SECTION 1. SHORT TITLE. This Act may be cited ... | Education and Training for Health Act of 2017 ... | Education and Training for Health Act of 2017     | 1648     | [-0.009376125410199165, -0.0360078439116478, 0...  |
| 15 | SECTION 1. SHORT TITLE. This Act may be cited ... | Recreational Hunting Safety and Preservation A... | Recreational Hunting Safety and Preservation A... | 2209     | [0.024976342916488647, -0.005445675924420357, ...  |
| 16 | SECTION 1. SHORT TITLE. This Act may be cited ... | Andrew Prior Act or Andrew's Law - Amends the ... | Andrew's Law                                      | 608      | [0.029043208807706833, -0.01100732292979957, ...   |
| 17 | SECTION 1. SHORT TITLE. This Act may be cited ... | Directs the President, in coordination with de... | Energy Independence Act of 2000                   | 1352     | [-0.0034495051950216293, -0.02827893755300133...   |
| 18 | SECTION 1. SHORT TITLE. This Act may be cited ... | This measure has not been amended since it was... | Veterans Entrepreneurship Act of 2015             | 1393     | [-0.0026434329338371754, -0.004964602179825806...  |
| 19 | SECTION 1. SHORT TITLE. This Act may be cited ... | Strengthening the Health Care Safety Net Act o... | To amend title XIX of the Social Security Act ... | 2678     | [0.009399736300110817, -0.02588636800646782, 0...  |

아래의 검색 코드 블록을 실행할 때 동일한 `text-embedding-ada-002`(버전 2) 모델과 함께 "케이블 회사 세금 수익에 대한 정보를 얻을 수 있나요?" 검색 쿼리를 포함합니다. 다음으로 [코사인 유사성](#)으로 순위가 매겨진 쿼리에서 새로 포함된 텍스트에 삽입된 가장 가까운 청구서를 찾습니다.

### Python

```
# search through the reviews for a specific product
def search_docs(df, user_query, top_n=3, to_print=True):
    embedding = get_embedding(
        user_query,
        engine="text-embedding-ada-002" # engine should be set to the
```

```

deployment name you chose when you deployed the text-embedding-ada-002
(Version 2) model
)
df[ "similarities" ] = df.ada_v2.apply(lambda x: cosine_similarity(x,
embedding))

res = (
    df.sort_values("similarities", ascending=False)
    .head(top_n)
)
if to_print:
    display(res)
return res

res = search_docs(df_bills, "Can I get information on cable company tax
revenue?", top_n=4)

```

출력:

|    | text                                              | summary                                           | title                                             | n_tokens | ada_v2                                              | similarities |
|----|---------------------------------------------------|---------------------------------------------------|---------------------------------------------------|----------|-----------------------------------------------------|--------------|
| 9  | SECTION 1. SHORT TITLE. This Act may be cited ... | Taxpayer's Right to View Act of 1993 - Amends ... | Taxpayer's Right to View Act of 1993              | 947      | [0.004519130103290081, -0.023599395528435707, ...]  | 0.767584     |
| 11 | SECTION 1. SHORT TITLE. This Act may be cited ... | Wall Street Compensation Reform Act of 2010 - ... | A bill to amend the Internal Revenue Code of 1... | 2331     | [0.014871294610202312, -0.001929433667100966, ...]  | 0.714282     |
| 1  | SECTION 1. SHORT TITLE. This Act may be cited ... | Small Business Expansion and Hiring Act of 201... | To amend the Internal Revenue Code of 1986 to ... | 1183     | [0.005016345530748367, -0.00569863710552454, ...]   | 0.702599     |
| 4  | SECTION 1. SHORT TITLE. This Act may be cited ... | Military Call-up Relief Act - Amends the Inter... | A bill to amend the Internal Revenue Code of 1... | 1038     | [0.010082815773785114, -0.0007545037078671157, ...] | 0.699490     |

마지막으로 전체 기술 자료에 대한 사용자 쿼리를 기반으로 문서 검색의 최상위 결과를 표시합니다. 이는 "1993년 납세자의 조회권법"의 최상위 결과를 반환합니다. 이 문서에는 쿼리와 문서 간에 0.76의 코사인 유사성 점수가 있습니다.

Python

```
res[ "summary" ][9]
```

출력

"Taxpayer's Right to View Act of 1993 - Amends the Communications Act of 1934 to prohibit a cable operator from assessing separate charges for any video programming of a sporting, theatrical, or other entertainment event if that event is performed at a facility constructed, renovated, or maintained with tax revenues or by an organization that receives public financial support. Authorizes the Federal Communications Commission and local franchising authorities to make determinations concerning the applicability of such prohibition. Sets forth conditions under which a facility is considered to have been constructed, maintained, or renovated with tax revenues. Considers events performed by nonprofit or public organizations that receive tax subsidies to be subject to this Act if the event is sponsored by, or includes the participation of a team that is part of, a tax exempt organization."

이 방식을 사용하면 기술 자료의 문서 전체에서 포함을 검색 메커니즘으로 사용할 수 있습니다. 그런 다음 사용자는 상위 검색 결과를 가져와 다운스트림 작업에 사용할 수 있으며 이로 인해 초기 쿼리가 표시됩니다.

## 리소스 정리

이 자습서를 완료하기 위해서 OpenAI 리소스만 만들었고 OpenAI 리소스를 정리하고 제거하려는 경우 배포된 모델을 삭제한 다음 테스트 리소스 전용인 경우 리소스 또는 연결된 리소스 그룹을 삭제해야 합니다. 리소스 그룹을 삭제하면 해당 리소스 그룹에 연결된 다른 모든 리소스가 함께 삭제됩니다.

- [포털](#)
- [Azure CLI](#)

## 다음 단계

Azure OpenAI의 모델에 대해 자세히 알아봅니다.

[Azure OpenAI 서비스 모델](#)

# Azure OpenAI 음성 변환 채팅

아티클 • 2023. 04. 28.

[참조 설명서](#) | [패키지\(NuGet\)](#) | [GitHub의 추가 샘플](#)

## ⓘ 중요

이 가이드의 단계를 완료하려면 원하는 Azure 구독에서 Microsoft Azure OpenAI Service에 대한 액세스 권한을 부여해야 합니다. 현재 이 서비스에 대한 액세스 권한은 애플리케이션에서만 부여됩니다. <https://aka.ms/oai/access>에서 양식을 작성하여 Azure OpenAI에 대한 액세스를 신청할 수 있습니다.

이 방법 가이드에서는 [Azure Cognitive Services Speech](#)를 사용하여 Azure OpenAI Service와 대화할 수 있습니다. Speech Service에서 인식하는 텍스트는 Azure OpenAI로 전송됩니다. 그런 다음, Azure OpenAI의 텍스트 응답이 Speech Service에서 합성됩니다.

마이크를 사용하여 Azure OpenAI와 대화를 시작합니다.

- Speech Service는 음성을 인식하고 텍스트(음성에서 텍스트로)로 변환합니다.
- 텍스트 요청이 Azure OpenAI로 전송됩니다.
- Speech Service TTS(텍스트 음성 변환) 기능은 Azure OpenAI의 응답을 기본 스피커로 합성합니다.

이 예제의 환경은 앞뒤로 교환되지만 Azure OpenAI는 대화의 컨텍스트를 기억하지 못합니다.

## 필수 구성 요소

- ✓ Azure 구독 - [체험 구독 만들기](#)
- ✓ Azure Portal에서 [음성 리소스 만들기](#)
- ✓ 음성 리소스 키 및 지역을 가져옵니다. 음성 리소스가 배포된 후, [리소스로 이동](#)을 선택하여 키를 보고 관리합니다. Cognitive Services 리소스에 대한 자세한 내용은 [리소스 키 가져오기](#)를 참조하세요.

## 환경 설정

음성 SDK는 [NuGet 패키지](#)로 사용할 수 있으며 .NET Standard 2.0을 구현합니다. 이 가이드의 뒷부분에서 Speech SDK를 설치하지만, 먼저 [SDK 설치 가이드](#)에서 더 많은 요구 사항을 확인합니다.

# 환경 변수 설정

이 예제에는 `OPEN_AI_KEY`, `OPEN_AI_ENDPOINT`, `SPEECH_KEY` 및 `SPEECH_REGION`이라는 환경 변수가 필요합니다.

Cognitive Services 리소스에 액세스하려면 애플리케이션을 인증해야 합니다. 프로덕션의 경우 자격 증명을 안전하게 저장하고 액세스하는 방법을 사용합니다. 예를 들어, [Speech 리소스](#)에 대한 키를 얻은 후 애플리케이션을 실행하는 로컬 컴퓨터의 새 환경 변수에 씁니다.

## 💡 팁

키를 코드에 직접 포함하지 말고 공개적으로 게시하지 마세요. [Azure Key Vault](#)와 같은 추가 인증 옵션은 Cognitive Services [보안](#) 문서를 참조하세요.

환경 변수를 설정하려면 콘솔 창을 열고 운영 체제 및 개발 환경에 대한 지침을 따릅니다.

- `OPEN_AI_KEY` 환경 변수를 설정하려면 `your-openai-key`를 리소스에 대한 키 중 하나로 바꿉니다.
- `OPEN_AI_ENDPOINT` 환경 변수를 설정하려면 `your-openai-endpoint`을(를) 리소스에 대한 지역 중 하나로 바꿉니다.
- `SPEECH_KEY` 환경 변수를 설정하려면 `your-speech-key`를 리소스에 대한 키 중 하나로 바꿉니다.
- `SPEECH_REGION` 환경 변수를 설정하려면 `your-speech-region`을(를) 리소스에 대한 지역 중 하나로 바꿉니다.

Windows

콘솔

```
setx OPEN_AI_KEY your-openai-key  
setx OPEN_AI_ENDPOINT your-openai-endpoint  
setx SPEECH_KEY your-speech-key  
setx SPEECH_REGION your-speech-region
```

## ⓘ 참고

현재 실행 중인 콘솔에서만 환경 변수에 액세스해야 하는 경우 환경 변수를 `setx` 대신 `set`로 설정할 수 있습니다.

환경 변수를 추가한 후에는 콘솔 창을 포함하여 실행 중인 프로그램 중에서 환경 변수를 읽어야 하는 프로그램을 다시 시작해야 할 수도 있습니다. 예를 들어 편집기로 Visual Studio를 사용하는 경우 Visual Studio를 다시 시작한 후 예제를 실행합니다.

## マイクロ에서 음성 인식

새 콘솔 애플리케이션을 만들려면 다음 단계를 수행합니다.

- 새 프로젝트를 원하는 명령 프롬프트를 열고 .NET CLI를 사용하여 콘솔 애플리케이션을 만듭니다. `Program.cs` 파일은 프로젝트 디렉터리에 만들어야 합니다.

```
.NET CLI
```

```
dotnet new console
```

- .NET CLI를 사용하여 새 프로젝트에 음성 SDK를 설치합니다.

```
.NET CLI
```

```
dotnet add package Microsoft.CognitiveServices.Speech
```

- .NET CLI를 사용하여 새 프로젝트에 Azure OpenAI SDK(시험판)를 설치합니다.

```
.NET CLI
```

```
dotnet add package Azure.AI.OpenAI --prerelease
```

- `Program.cs`의 내용을 다음 코드로 바꿉니다.

```
C#
```

```
using System;
using System.IO;
using System.Threading.Tasks;
using Microsoft.CognitiveServices.Speech;
using Microsoft.CognitiveServices.Speech.Audio;
using Azure;
using Azure.AI.OpenAI;
using static System.Environment;

class Program
{
    // This example requires environment variables named "OPEN_AI_KEY"
    and "OPEN_AI_ENDPOINT"
    // Your endpoint should look like the following
```

```
https://YOUR_OPEN_AI_RESOURCE_NAME.openai.azure.com/
    static string openAIKey =
Environment.GetEnvironmentVariable("OPEN_AI_KEY");
    static string openAIEndpoint =
Environment.GetEnvironmentVariable("OPEN_AI_ENDPOINT");

    // Enter the deployment name you chose when you deployed the model.
    static string engine = "text-davinci-002";

    // This example requires environment variables named "SPEECH_KEY"
and "SPEECH_REGION"
    static string speechKey =
Environment.GetEnvironmentVariable("SPEECH_KEY");
    static string speechRegion =
Environment.GetEnvironmentVariable("SPEECH_REGION");

    // Prompts Azure OpenAI with a request and synthesizes the
response.
    async static Task AskOpenAI(string prompt)
{
    // Ask Azure OpenAI
    OpenAIClient client = new(new Uri(openAIEndpoint), new
AzureKeyCredential(openAIKey));
    var completionsOptions = new CompletionsOptions()
    {
        Prompts = { prompt },
        MaxTokens = 100,
    };
    Response<Completions> completionsResponse =
client.GetCompletions(engine, completionsOptions);
    string text = completionsResponse.Value.Choices[0].Text.Trim();
    Console.WriteLine($"Azure OpenAI response: {text}");

    var speechConfig = SpeechConfig.FromSubscription(speechKey,
speechRegion);
    // The language of the voice that speaks.
    speechConfig.SpeechSynthesisVoiceName = "en-US-
JennyMultilingualNeural";
    var audioOutputConfig = AudioConfig.FromDefaultSpeakerOutput();

    using (var speechSynthesizer = new
SpeechSynthesizer(speechConfig, audioOutputConfig))
    {
        var speechSynthesisResult = await
speechSynthesizer.SpeakTextAsync(text).ConfigureAwait(true);

        if (speechSynthesisResult.Reason ==
ResultReason.SynthesizingAudioCompleted)
        {
            Console.WriteLine($"Speech synthesized to speaker for
text: [{text}]");
        }
        else if (speechSynthesisResult.Reason ==
ResultReason.Canceled)
        {
    }
```

```
        var cancellationDetails =
SpeechSynthesisCancellationDetails.FromResult(speechSynthesisResult);
            Console.WriteLine($"Speech synthesis canceled:
{cancellationDetails.Reason}");

            if (cancellationDetails.Reason ==
CancellationReason.Error)
            {
                Console.WriteLine($"Error details:
{cancellationDetails.ErrorDetails}");
            }
        }

// Continuously listens for speech input to recognize and send as
text to Azure OpenAI
async static Task ChatWithOpenAI()
{
    // Should be the locale for the speaker's language.
    var speechConfig = SpeechConfig.FromSubscription(speechKey,
speechRegion);
    speechConfig.SpeechRecognitionLanguage = "en-US";

    using var audioConfig =
AudioConfig.FromDefaultMicrophoneInput();
    using var speechRecognizer = new SpeechRecognizer(speechConfig,
audioConfig);
    var conversationEnded = false;

    while(!conversationEnded)
    {
        Console.WriteLine("Azure OpenAI is listening. Say 'Stop' or
press Ctrl-Z to end the conversation.");

        // Get audio from the microphone and then send it to the
TTS service.
        var speechRecognitionResult = await
speechRecognizer.RecognizeOnceAsync();

        switch (speechRecognitionResult.Reason)
        {
            case ResultReason.RecognizedSpeech:
                if (speechRecognitionResult.Text == "Stop.")
                {
                    Console.WriteLine("Conversation ended.");
                    conversationEnded = true;
                }
                else
                {
                    Console.WriteLine($"Recognized speech:
{speechRecognitionResult.Text}");
                    await
AskOpenAI(speechRecognitionResult.Text).ConfigureAwait(true);
                }
        }
    }
}
```

```

        break;
    case ResultReason.NoMatch:
        Console.WriteLine($"No speech could be recognized:");
    }
    break;
case ResultReason.Canceled:
    var cancellationDetails =
CancellationDetails.FromResult(speechRecognitionResult);
    Console.WriteLine($"Speech Recognition canceled:
{cancellationDetails.Reason}");
    if (cancellationDetails.Reason ==
CancellationReason.Error)
    {
        Console.WriteLine($"Error details=
{cancellationDetails.ErrorDetails}");
    }
    break;
}
}

async static Task Main(string[] args)
{
    try
    {
        await ChatWithOpenAI().ConfigureAwait(true);
    }
    catch (Exception ex)
    {
        Console.WriteLine(ex.Message);
    }
}
}

```

5. Azure OpenAI에서 반환하는 토큰 수를 늘리거나 줄이려면 클래스 instance 속성을 `CompletionsOptions` 변경 `MaxTokens` 합니다. 토큰 및 비용 관련 자세한 내용은 [Azure OpenAI 토큰](#) 및 [Azure OpenAI 가격 책정](#)을 참조하세요.

새 콘솔 애플리케이션을 실행하여 마이크의 음성 인식을 시작합니다.

콘솔

`dotnet run`

### ⓘ 중요

**앞에서** 설명한 대로 `OPEN_AI_KEY`, `OPEN_AI_ENDPOINT`, `SPEECH__KEY`, `SPEECH__REGION` 환경 변수를 설정해야 합니다. 이 변수를 설정하지 않으면 샘플이 오류 메시지와 함께 실패합니다.

메시지가 표시되면 마이크에 말합니다. 콘솔 출력에는 말하기를 시작하라는 프롬프트, 텍스트로 요청, Azure OpenAI의 응답이 텍스트로 포함됩니다. Azure OpenAI의 응답을 텍스트에서 음성으로 변환한 다음 기본 스피커로 출력해야 합니다.

### 콘솔

```
PS C:\dev\openai\csharp> dotnet run
Azure OpenAI is listening. Say 'Stop' or press Ctrl-Z to end the
conversation.
Recognized speech: Make a comma separated list of all continents.
Azure OpenAI response: Africa, Antarctica, Asia, Australia, Europe, North
America, South America
Speech synthesized to speaker for text [Africa, Antarctica, Asia, Australia,
Europe, North America, South America]
Azure OpenAI is listening. Say 'Stop' or press Ctrl-Z to end the
conversation.
Recognized speech: Make a comma separated list of 1 Astronomical observatory
for each continent. A list should include each continent name in
parentheses.
Azure OpenAI response: Mauna Kea Observatories (North America), La Silla
Observatory (South America), Tenerife Observatory (Europe), Siding Spring
Observatory (Australia), Beijing Xinglong Observatory (Asia), Naukluft
Plateau Observatory (Africa), Rutherford Appleton Laboratory (Antarctica)
Speech synthesized to speaker for text [Mauna Kea Observatories (North
America), La Silla Observatory (South America), Tenerife Observatory
(Europe), Siding Spring Observatory (Australia), Beijing Xinglong
Observatory (Asia), Naukluft Plateau Observatory (Africa), Rutherford
Appleton Laboratory (Antarctica)]
Azure OpenAI is listening. Say 'Stop' or press Ctrl-Z to end the
conversation.
Conversation ended.
PS C:\dev\openai\csharp>
```

## 설명

이제 빠른 시작을 완료했으므로 다음과 같은 몇 가지 추가 고려 사항이 있습니다.

- 음성 인식 언어를 변경하려면 `en-US`를 다른 [지원되는 언어](#)로 바꿉니다. 예를 들어 스페인어(스페인)의 경우 `es-ES`입니다. 언어를 지정하지 않은 경우 기본 언어는 `en-US`입니다. 음성에 사용될 수 있는 여러 언어 중 하나를 식별하는 방법에 대한 자세한 내용은 [언어 식별](#)을 참조하세요.
- 들리는 음성을 변경하려면 `en-US-JennyMultilingualNeural`을 [지원되는 다른 음성](#)으로 바꿉니다. 음성이 Azure OpenAI에서 반환된 텍스트의 언어를 모르는 경우 Speech Service에서 합성된 오디오를 출력하지 않습니다.
- 다른 [모델](#)을 사용하려면 `text-davinci-002`를 다른 [배포](#)의 ID로 바꿉니다. 배포 ID가 모델 이름과 반드시 동일하지는 않습니다. [Azure OpenAI Studio](#)에서 배포를 만들

때 배포 이름을 지정했습니다.

- 또한 Azure OpenAI는 프롬프트 입력 및 생성된 출력에서 콘텐츠 조정을 수행합니다. 유해한 콘텐츠가 감지되면 프롬프트나 응답을 필터링할 수 있습니다. 자세한 내용은 [콘텐츠 필터링](#) 문서를 참조하세요.

## 리소스 정리

Azure Portal 또는 [Azure CLI\(명령줄 인터페이스\)](#)를 사용하여 생성된 음성 리소스를 제거할 수 있습니다.

## 다음 단계

- [Speech에 대한 자세한 정보](#)
- [Azure OpenAI에 대해 자세히 알아보기](#)

# Overview of Responsible AI practices for Azure OpenAI models

Article • 05/19/2023

Many of the Azure OpenAI models are generative AI models that have demonstrated improvements in advanced capabilities such as content and code generation, summarization, and search. With many of these improvements also come increased responsible AI challenges related to harmful content, manipulation, human-like behavior, privacy, and more. For more information about the capabilities, limitations and appropriate use cases for these models, please review the [Transparency Note](#).

In addition to the Transparency Note, we have created technical recommendations and resources to help customers design, develop, deploy, and use AI systems that implement the Azure OpenAI models responsibly. Our recommendations are grounded in the [Microsoft Responsible AI Standard](#), which sets policy requirements that our own engineering teams follow. Much of the content of the Standard follows a pattern, asking teams to Identify, Measure, and Mitigate potential harms, and plan for how to Operate the AI system as well. In alignment with those practices, these recommendations are organized into four stages:

1. **Identify** : Identify and prioritize potential harms that could result from your AI system through iterative red-teaming, stress-testing, and analysis.
2. **Measure** : Measure the frequency and severity of those harms by establishing clear metrics, creating measurement test sets, and completing iterative, systematic testing (both manual and automated).
3. **Mitigate** : Mitigate harms by implementing tools and strategies such as [prompt engineering](#) and using our [content filters](#). Repeat measurement to test effectiveness after implementing mitigations.
4. **Operate** : Define and execute a deployment and operational readiness plan.

In addition to their correspondence to the Microsoft Responsible AI Standard, these stages correspond closely to the functions in the [NIST AI Risk Management Framework](#).

## Identify

Identifying potential harms that could occur in or be caused by an AI system is the first stage of the Responsible AI lifecycle. The earlier you begin to identify potential harms, the more effective you can be at mitigating the harms. When assessing potential harms, it is important to develop an understanding of the types of harms that could result from

using the Azure OpenAI Service in your specific context(s). In this section, we provide recommendations and resources you can use to identify harms through an impact assessment, iterative red team testing, stress-testing, and analysis. Red teaming and stress-testing are approaches where a group of testers come together and intentionally probe a system to identify its limitations, risk surface, and vulnerabilities.

These steps have the goal of producing a prioritized list of potential harms for each specific scenario.

- 1. Identify harms that are relevant** for your specific model, application, and deployment scenario.
  - a. Identify potential harms associated with the model and model capabilities (for example, GPT-3 model vs GPT-4 model) that you're using in your system. This is important to consider because each model has different capabilities, limitations, and risks, as described more fully in the sections above.
  - b. Identify any other harms or increased scope of harm presented by the intended use of the system you're developing. Consider using a [Responsible AI Impact Assessment](#) to identify potential harms.
    - i. For example, let's consider an AI system that summarizes text. Some uses of text generation are lower risk than others. For example, if the system is to be used in a healthcare domain for summarizing doctor's notes, the risk of harm arising from inaccuracies is higher than if the system is summarizing online articles.
- 2. Prioritize harms based on elements of risk such as frequency and severity.** Assess the level of risk for each harm and the likelihood of each risk occurring in order to prioritize the list of harms you've identified. Consider working with subject matter experts and risk managers within your organization and with relevant external stakeholders when appropriate.
- 3. Conduct red team testing and stress testing** starting with the highest priority harms, to develop a better understanding of whether and how the identified harms are actually occurring in your scenario, as well as to identify new harms you didn't initially anticipate.
- 4. Share this information with relevant stakeholders** using your organization's internal compliance processes.

At the end of this Identify stage, you should have a documented, prioritized list of harms. When new harms and new instances of harms emerge through further testing and use of the system, you can update and improve this list by following the above process again.

## Measure

Once a list of prioritized harms has been identified, the next stage involves developing an approach for systematic measurement of each harm and conducting evaluations of the AI system. There are manual and automated approaches to measurement. We recommend you do both, starting with manual measurement.

Manual measurement is useful for:

1. Measuring progress on a small set of priority issues. When mitigating specific harms, it's often most productive to keep manually checking progress against a small dataset until the harm is no longer observed before moving to automated measurement.
2. Defining and reporting metrics until automated measurement is reliable enough to use alone.
3. Spot-checking periodically to measure the quality of automatic measurement.

Automated measurement is useful for:

1. Measuring at a large scale with increased coverage to provide more comprehensive results.
2. Ongoing measurement to monitor for any regression as the system, usage, and mitigations evolve.

Below, we provide specific recommendations to measure your AI system for potential harms. We recommend you first complete this process manually and then develop a plan to automate the process:

1. **Create inputs that are likely to produce each prioritized harm:** Create measurement set(s) by generating many diverse examples of targeted inputs that are likely to produce each prioritized harm.
2. **Generate System Outputs:** Pass in the examples from the measurement sets as inputs to the system to generate system outputs. Document the outputs.
3. **Evaluate System Outputs and Report Results to Relevant Stakeholders**
  - a. **Define clear metric(s).** For each intended use of your system, establish metrics that measure the frequency and degree of severity of each potentially harmful output. Create clear definitions to classify outputs that will be considered harmful or problematic in the context of your system and scenario, for each type of prioritized harm you identified.
  - b. **Assess the outputs** against the clear metric definitions and record and quantify the occurrences of harmful outputs. Repeat the measurements periodically, to assess mitigations and monitor for any regression.

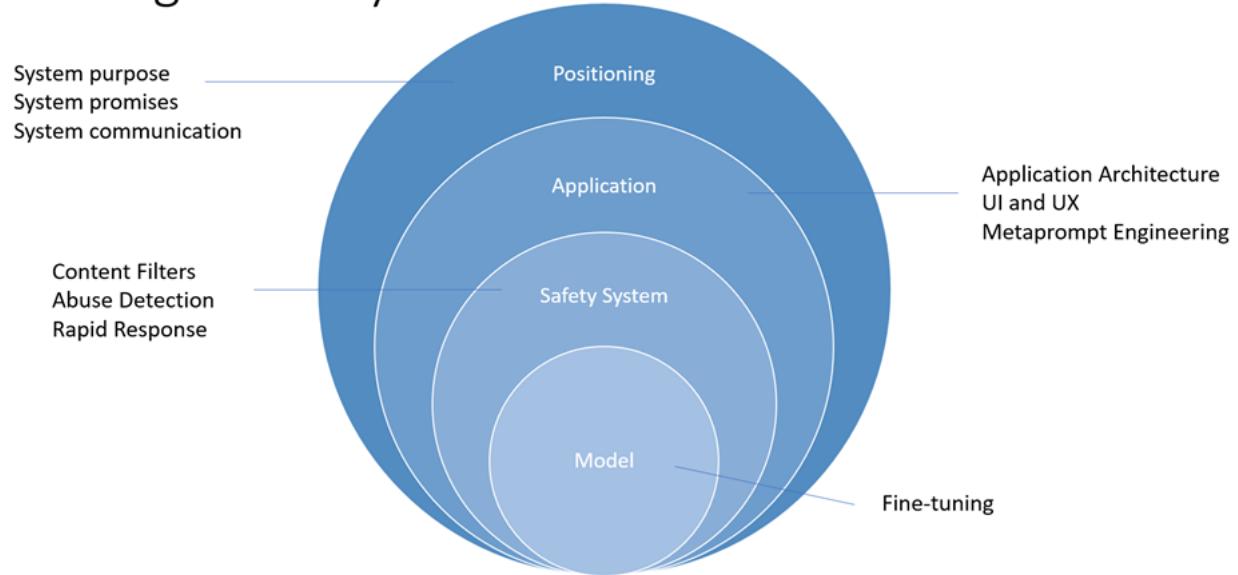
- c. Share this information with relevant stakeholders using your organization's internal compliance processes.

At the end of this measurement stage, you should have a defined measurement approach to benchmark how your system performs for each potential harm as well as an initial set of documented results. As you continue implementing and testing mitigations, the metrics and measurement sets should continue to be refined (for example, to add metrics for new harms that were initially unanticipated) and the results updated.

## Mitigate

Mitigating harms presented by large language models such as the Azure OpenAI models requires an iterative, layered approach that includes experimentation and continual measurement. We recommend developing a mitigation plan that encompasses four layers of mitigations for the harms identified in the earlier stages of this process:

### Mitigation Layers



1. At the **model level**, it's important to understand the model(s) you'll be using and what fine-tuning steps may have been taken by the model developers to align the model towards its intended uses and to reduce the risk of potentially harmful uses and outcomes.
  - a. For example, for GPT-4, model developers have been able to use reinforcement learning methods as a responsible AI tool to better align the model towards the designers' intended goals.
2. At the **safety system level**, you should understand the platform level mitigations that have been implemented. such as the [Azure OpenAI content filters](#) which help

to block the output of harmful content.

3. At the **application level**, application developers can implement metaprompt and user-centered design and user experience mitigations. Metaprompts are instructions provided to the model to guide its behavior; their use can make a critical difference in guiding the system to behave in accordance with your expectations. User-centered design and user experience (UX) interventions are also key mitigation tools to prevent misuse and overreliance on AI.
4. At the **positioning level**, there are many ways to educate the people who will use or be affected by your system about its capabilities and limitations.

Below, we provide specific recommendations to implement mitigations at the different layers. Not all of these mitigations are appropriate for every scenario, and conversely, these mitigations may be insufficient for some scenarios. Give careful consideration to your scenario and the prioritized harms you identified, and as you implement mitigations, develop a process to **measure and document their effectiveness** for your system and scenario.

1. **Model level Mitigations:** Review and identify which Azure OpenAI base model is best suited for the system you're building and educate yourself about its capabilities, limitations, and any measures taken to reduce the risk of the potential harms you've identified. For example, if you're using GPT-4, in addition to reading this Transparency Note, you can review OpenAI's [GPT-4 System Card](#) explaining the safety challenges presented by the model and the safety processes that OpenAI adopted to prepare GPT-4 for deployment. It may be worth experimenting with different versions of the model(s) (including through red teaming and measuring) to see how the harms present differently.
2. **Safety System Level Mitigations:** Identify and evaluate the effectiveness of platform level solutions such as the [Azure OpenAI content filters](#) to help mitigate the potential harms that you have identified.
3. **Application Level Mitigations:** Prompt engineering, including **metaprompt tuning, can be an effective mitigation** for many different types of harm. Review and implement metaprompt (also called the "system message" or "system prompt") guidance and best practices documented [here](#).

We recommend implementing the following user-centered design and user experience (UX) interventions, guidance, and best practices to guide users to use the system as intended and to prevent overreliance on the AI system:

- a. **Review and edit interventions:** Design the user experience (UX) to encourage people who use the system to review and edit the AI-generated outputs before accepting them (see [HAX G9](#): Support efficient correction).

**b. Highlight potential inaccuracies in the AI-generated outputs** (see HAX G2 ↗:

Make clear how well the system can do what it can do), both when users first start using the system and at appropriate times during ongoing use. In the first run experience (FRE), notify users that AI-generated outputs may contain inaccuracies and that they should verify information. Throughout the experience, include reminders to check AI-generated output for potential inaccuracies, both overall and in relation to specific types of content the system may generate incorrectly. For example, if your measurement process has determined that your system has lower accuracy with numbers, mark numbers in generated outputs to alert the user and encourage them to check the numbers or seek external sources for verification.

- c. User responsibility.** Remind people that they are accountable for the final content when they're reviewing AI-generated content. For example, when offering code suggestions, remind the developer to review and test suggestions before accepting.
- d. Disclose AI's role in the interaction.** Make people aware that they are interacting with an AI system (as opposed to another human). Where appropriate, inform content consumers that content has been partly or fully generated by an AI model; such notices may be required by law or applicable best practices, and can reduce inappropriate reliance on AI-generated outputs and can help consumers use their own judgment about how to interpret and act on such content.
- e. Prevent the system from anthropomorphizing.** AI models may output content containing opinions, emotive statements, or other formulations that could imply that they're human-like, that could be mistaken for a human identity, or that could mislead people to think that a system has certain capabilities when it doesn't. Implement mechanisms that reduce the risk of such outputs or incorporate disclosures to help prevent misinterpretation of outputs.
- f. Cite references and information sources.** If your system generates content based on references sent to the model, clearly citing information sources helps people understand where the AI-generated content is coming from.
- g. Limit the length of inputs and outputs, where appropriate.** Restricting input and output length can reduce the likelihood of producing undesirable content, misuse of the system beyond its intended uses, or other harmful or unintended uses.
- h. Structure inputs and/or system outputs.** Use [prompt engineering](#) techniques within your application to structure inputs to the system to prevent open-ended responses. You can also limit outputs to be structured in certain formats or patterns. For example, if your system generates dialog for a fictional character in

response to queries, limit the inputs so that people can only query for a predetermined set of concepts.

- i. **Prepare pre-determined responses.** There are certain queries to which a model may generate offensive, inappropriate, or otherwise harmful responses. When harmful or offensive queries or responses are detected, you can design your system to deliver a predetermined response to the user. Predetermined responses should be crafted thoughtfully. For example, the application can provide prewritten answers to questions such as "who/what are you?" to avoid having the system respond with anthropomorphized responses. You can also use predetermined responses for questions like, "What are your terms of use?" to direct people to the correct policy.
- j. **Restrict automatic posting on social media.** Limit how people can automate your product or service. For example, you may choose to prohibit automated posting of AI-generated content to external sites (including social media), or to prohibit the automated execution of generated code.
- k. **Bot detection.** Devise and implement a mechanism to prohibit users from building an API on top of your product.

#### 4. Positioning Level Mitigations:

- a. **Be appropriately transparent.** It's important to provide the right level of transparency to people who use the system, so that they can make informed decisions around the use of the system.
- b. **Provide system documentation.** Produce and provide educational materials for your system, including explanations of its capabilities and limitations. For example, this could be in the form of a "learn more" page accessible via the system.
- c. **Publish user guidelines and best practices.** Help users and stakeholders use the system appropriately by publishing best practices, for example on prompt crafting, reviewing generations before accepting them, etc. Such guidelines can help people understand how the system works. When possible, incorporate the guidelines and best practices directly into the UX.

As you implement mitigations to address potential identified harms, it's important to develop a process for ongoing measurement of the effectiveness of such mitigations, to document measurement results, and to review those measurement results to continually improve the system.

## Operate

Once measurement and mitigation systems are in place, we recommend that you define and execute a deployment and operational readiness plan. This stage includes

completing appropriate reviews of your system and mitigation plans with relevant stakeholders, establishing pipelines to collect telemetry and feedback, and developing an incident response and rollback plan.

Some recommendations for how to deploy and operate a system that uses the Azure OpenAI service with appropriate, targeted harms mitigations include:

1. Work with compliance teams within your organization to understand what types of reviews are required for your system and when they are required (for example, legal review, privacy review, security review, accessibility review, etc.).
2. Develop and implement the following:
  - a. **Develop a phased delivery plan.** We recommend you launch systems using the Azure OpenAI service gradually using a "phased delivery" approach. This gives a limited set of people the opportunity to try the system, provide feedback, report issues and concerns, and suggest improvements before the system is released more widely. It also helps to manage the risk of unanticipated failure modes, unexpected system behaviors, and unexpected concerns being reported.
  - b. **Develop an incident response plan.** Develop an incident response plan and evaluate the time needed to respond to an incident.
  - c. **Develop a rollback plan** Ensure you can roll back the system quickly and efficiently in case an unanticipated incident occurs.
  - d. **Prepare for immediate action for unanticipated harms.** Build the necessary features and processes to block problematic prompts and responses as they're discovered and as close to real-time as possible. When unanticipated harms do occur, block the problematic prompts and responses as quickly as possible, develop and deploy appropriate mitigations, investigate the incident, and implement a long-term solution.
  - e. **Develop a mechanism to block people who are misusing your system.** Develop a mechanism to identify users who violate your content policies (for example, by generating hate speech) or are otherwise using your system for unintended or harmful purposes, and take action against further abuse. For example, if a user frequently uses your system to generate content that is blocked or flagged by content safety systems, consider blocking them from further use of your system. Implement an appeal mechanism where appropriate.
  - f. **Build effective user feedback channels.** Implement feedback channels through which stakeholders (and the general public, if applicable) can submit feedback or report issues with generated content or that otherwise arise during their use of the system. Document how such feedback is processed, considered, and addressed. Evaluate the feedback and work to improve the system based on user feedback. One approach could be to include buttons with generated content that would allow users to identify content as "inaccurate," "harmful" or

"incomplete." This could provide a more widely used, structured and feedback signal for analysis.

g. **Telemetry data.** Identify and record (consistent with applicable privacy laws, policies, and commitments) signals that indicate user satisfaction or their ability to use the system as intended. Use telemetry data to identify gaps and improve the system.

This document is not intended to be, and should not be construed as providing, legal advice. The jurisdiction in which you're operating may have various regulatory or legal requirements that apply to your AI system. Consult a legal specialist if you are uncertain about laws or regulations that might apply to your system, especially if you think those might impact these recommendations. Be aware that not all of these recommendations and resources are appropriate for every scenario, and conversely, these recommendations and resources may be insufficient for some scenarios.

## Learn more about responsible AI

- Microsoft AI principles 
- Microsoft responsible AI resources 
- Microsoft Azure Learning courses on responsible AI

## Learn more about Azure OpenAI

- Limited access to Azure OpenAI Service - Azure Cognitive Services | Microsoft Learn
- Code of Conduct for the Azure OpenAI Service | Microsoft Learn
- Data, privacy, and security for Azure OpenAI Service - Azure Cognitive Services | Microsoft Learn

# Transparency Note for Azure OpenAI Service

Article • 05/19/2023

## What is a Transparency Note?

An AI system includes not only the technology, but also the people who use it, the people who are affected by it, and the environment in which it's deployed. Creating a system that is fit for its intended purpose requires an understanding of how the technology works, what its capabilities and limitations are, and how to achieve the best performance. Microsoft's Transparency Notes are intended to help you understand how our AI technology works, the choices system owners can make that influence system performance and behavior, and the importance of thinking about the whole system, including the technology, the people, and the environment. You can use Transparency Notes when developing or deploying your own system, or share them with the people who will use or be affected by your system.

Microsoft's Transparency Notes are part of a broader effort at Microsoft to put our AI Principles into practice. To find out more, see the [Microsoft's AI principles](#).

## The basics of the Azure OpenAI Models

Azure OpenAI provides customers with a fully managed AI service that lets developers and data scientists apply OpenAI's powerful models including models that can generate natural language, code, and images. Within the Azure OpenAI Service, the OpenAI models are integrated with Microsoft-developed content filtering and abuse detection models. Learn more about content filtering [here](#) and abuse detection [here](#).

Select the tabs to see content for the relevant model type.

### Introduction

Text and code models

As part of the fully managed Azure OpenAI Service, the GPT-3 models analyze and generate natural language, Codex models analyze and generate code and plain text code commentary, and the GPT-4 models can understand and generate natural language and code. These models use an autoregressive architecture, meaning they

use data from prior observations to predict the most probable next word. This process is then repeated by appending the newly generated content to the original text to produce the complete generated response. Because the response is conditioned on the input text, these models can be applied to various tasks simply by changing the input text.

The GPT-3 series of models are pretrained on a wide body of publicly available free text data. This data is sourced from a combination of web crawling (specifically, a filtered version of [Common Crawl](#)), which includes a broad range of text from the internet and comprises 60 percent of the weighted pretraining dataset) and higher-quality datasets, including an expanded version of the WebText dataset, two internet-based books corpora and English-language Wikipedia. The GPT-4 base model was trained using publicly available data (such as internet data) and data that was licensed by OpenAI. The model was fine-tuned using reinforcement learning with human feedback (RLHF).

Learn more about the training and modeling techniques in OpenAI's [GPT-3](#), [GPT-4](#), and [Codex](#) research papers. The guidance below is also drawn from [OpenAI's safety best practices](#).

## Key terms

| Term                     | Definition                                                                                                                                                                                                                                                                                                                                                                                                                   |
|--------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Prompt                   | <p>The text you send to the service in the API call. This text is then input into the model. For example, one might input the following prompt:</p> <pre>Convert the questions to a command:<br/>Q: Ask Constance if we need some bread<br/>A: send-msg 'find constance' Do we need some bread?<br/>Q: Send a message to Greg to figure out if things are ready for<br/>Wednesday.<br/>A:</pre>                              |
| Completion or Generation | <p>The text Azure OpenAI outputs in response. For example, the service may respond with the following answer to the above prompt: <code>send-msg 'find greg'</code> <code>figure out if things are ready for Wednesday.</code></p>                                                                                                                                                                                           |
| Token                    | <p>Azure OpenAI processes text by breaking it down into tokens. Tokens can be words or just chunks of characters. For example, the word <code>hamburger</code> gets broken up into the tokens <code>ham</code>, <code>bur</code> and <code>ger</code>, while a short and common word like <code>pear</code> is a single token. Many tokens start with a whitespace, for example <code>hello</code> and <code>bye</code>.</p> |

# Capabilities

Text and code models

The GPT-4, GPT-3, and Codex Azure OpenAI Service models use natural language instructions and examples in the prompt to identify the task. The model then completes the task by predicting the most probable next text. This technique is known as "in-context" learning. These models are not retrained during this step but instead give predictions based on the context you include in the prompt.

There are three main approaches for in-context learning. These approaches vary based on the amount of task-specific data that is given to the model:

**Few-shot :** In this case, a user includes several examples in the prompt that demonstrate the expected answer format and content. The following example shows a few-shot prompt providing multiple examples:

```
Convert the questions to a command:  
Q: Ask Constance if we need some bread  
A: send-msg `find constance` Do we need some bread?  
Q: Send a message to Greg to figure out if things areready for Wednesday.  
A: send-msg `find greg` Is everything ready forWednesday?  
Q: Ask Ilya if we're still having our meeting thisevening  
A: send-msg `find ilya` Are we still having a meetingthis evening?  
Q: Contact the ski store and figure out if I can getmy skis fixed before I leave on Thursday  
A: send-msg `find ski store` Would it be possible toget my skis fixed before I leave on Thursday?  
Q: Thank Nicolas for lunch  
A: send-msg `find nicolas` Thank you for lunch!  
Q: Tell Constance that I won't be home before 19:30tonight – unmoveable meeting.  
A: send-msg `find constance` I won't be home before19:30 tonight. I have a meeting I can't move.  
Q: Tell John that I need to book an appointment at10:30  
A:
```

The number of examples typically ranges from 0 to 100 depending on how many can fit in the maximum input length for a single prompt. Few-shot learning enables a major reduction in the amount of task-specific data required for accurate predictions.

**One-shot :** This case is the same as the few-shot approach except only one example is provided. The following example shows a one-shot prompt:

Convert the questions to a command:

Q: Ask Constance if we need some bread

A: send-msg `find constance` Do we need some bread?

Q: Send a message to Greg to figure out if things are ready for Wednesday.

A:

**Zero-shot:** In this case, no examples are provided to the model and only the task request is provided. The following example shows a zero-shot prompt:

Convert the question to a command:

Q: Ask Constance if we need some bread

A:

## Use cases

Text and code models

### Intended uses

The GPT-4, GPT-3, and Codex models in the Azure OpenAI service can be used in multiple scenarios. The following list isn't comprehensive, but it illustrates the diversity of tasks that can be supported with appropriate mitigations:

- **Chat and conversation interaction :** Users can interact with a conversational agent that responds with responses drawn from trusted documents such as internal company documentation or tech support documentation.  
Conversations must be limited to answering scoped questions.
- **Chat and conversation creation :** Users can create a conversational agent that responds with responses drawn from trusted documents such as internal company documentation or tech support documentation. Conversations must be limited to answering scoped questions.
- **Code generation or transformation scenarios :** For example, converting one programming language to another, generating docstrings for functions, converting natural language to SQL.

- **Journalistic content** : For use to create new journalistic content or to rewrite journalistic content submitted by the user as a writing aid for predefined topics. Users cannot use the application as a general content creation tool for all topics. May not be used to generate content for political campaigns.
- **Question-answering** : Users can ask questions and receive answers from trusted source documents such as internal company documentation. The application does not generate answers ungrounded in trusted source documentation.
- **Reason over structured and unstructured data** : Users can analyze inputs using classification, sentiment analysis of text, or entity extraction. Examples include analyzing product feedback sentiment, analyzing support calls and transcripts, and refining text-based search with embeddings.
- **Search** : Users can search trusted source documents such as internal company documentation. The application does not generate results ungrounded in trusted source documentation.
- **Summarization** : Users can submit content to be summarized for predefined topics built into the application and cannot use the application as an open-ended summarizer. Examples include summarization of internal company documentation, call center transcripts, technical reports, and product reviews.
- **Writing assistance on specific topics** : Users can create new content or rewrite content submitted by the user as a writing aid for business content or pre-defined topics. Users can only rewrite or create content for specific business purposes or predefined topics and cannot use the application as a general content creation tool for all topics. Examples of business content include proposals and reports. For journalistic use, see above **Journalistic content** use case.

## Considerations when choosing a use case

We encourage customers to use the Azure OpenAI GPT-4, GPT-3, and Codex models in their innovative solutions or applications as approved in their [Limited Access registration form](#). However, here are some considerations when choosing a use case:

- **Not suitable for open-ended, unconstrained content generation.** Scenarios where users can generate content on any topic are more likely to produce offensive or harmful text. The same is true of longer generations.
- **Not suitable for scenarios where up-to-date, factually accurate information is crucial** unless you have human reviewers or are using the models to search your own documents and have verified suitability for your scenario. The service does not have information about events that occur after its training

date, likely has missing knowledge about some topics, and may not always produce factually accurate information.

- **Avoid scenarios where use or misuse of the system could result in significant physical or psychological injury to an individual.** For example, scenarios that diagnose patients or prescribe medications have the potential to cause significant harm.
- **Avoid scenarios where use or misuse of the system could have a consequential impact on life opportunities or legal status.** Examples include scenarios where the AI system could affect an individual's legal status, legal rights, or their access to credit, education, employment, healthcare, housing, insurance, social welfare benefits, services, opportunities, or the terms on which they're provided.
- **Avoid high stakes scenarios that could lead to harm.** The models hosted by Azure OpenAI service reflect certain societal views, biases, and other undesirable content present in the training data or the examples provided in the prompt. As a result, we caution against using the models in high-stakes scenarios where unfair, unreliable, or offensive behavior might be extremely costly or lead to harm.
- **Carefully consider use cases in high stakes domains or industry:** Examples include but are not limited to healthcare, medicine, finance, or legal.
- **Carefully consider well-scope chatbot scenarios.** Limiting the use of the service in chatbots to a narrow domain reduces the risk of generating unintended or undesirable responses.
- **Carefully consider all generative use cases.** Content generation scenarios may be more likely to produce unintended outputs and these scenarios require careful consideration and mitigations.

## Limitations

When it comes to large-scale natural language models and image models, there are particular fairness and responsible AI issues to consider. People use language and images to describe the world and to express their beliefs, assumptions, attitudes, and values. As a result, publicly available text and image data typically used to train large-scale natural language processing and image generation models contains societal biases relating to race, gender, religion, age, and other groups of people, as well as other undesirable content. These societal biases are reflected in the distributions of words, phrases, and syntactic structures.

## Technical limitations, operational factors and ranges

## Caution

Please be advised that this section contains illustrative examples which include terms and language that some individuals may find offensive.

Large-scale natural language and image models trained with such data can potentially behave in ways that are unfair, unreliable, or offensive, in turn causing harms. Some of the ways are listed here. We emphasize that these types of harms aren't mutually exclusive. A single model can exhibit more than one type of harm, potentially relating to multiple different groups of people. For example:

- **Allocation:** These models can be used in ways that lead to unfair allocation of resources or opportunities. For example, automated resume screening systems can withhold employment opportunities from one gender if they're trained on resume data that reflects the existing gender imbalance in a particular industry. Or the DALL·E 2 model could be used to create imagery in the style of a known artist, which could affect the value of the artist's work or the artist's life opportunities.
- **Quality of service:** The Azure OpenAI models are trained primarily on English text and images with English text descriptions. Languages other than English will experience worse performance. English language varieties with less representation in the training data might experience worse performance. The publicly available images used to train the DALL·E models might reinforce public bias and other undesirable content. The models are also unable to generate consistent photorealistic images and comprehensive text at this time.
- **Stereotyping:** These models can reinforce stereotypes. For example, when translating "He is a nurse" and "She is a doctor" into a genderless language such as Turkish and then back into English, many machine translation systems yield the stereotypical (and incorrect) results of "She is a nurse" and "He is a doctor." With DALL·E 2, when generating an image based on the prompt "Fatherless children," the model could generate images of Black children only, reinforcing harmful stereotypes that may exist in publicly available images.
- **Demeaning:** These models can demean people. For example, an open-ended content generation system with inappropriate or insufficient mitigations might produce offensive or demeaning to a particular group of people.
- **Overrepresentation and underrepresentation:** These models can over- or under-represent groups of people, or even erase their representation entirely. For example, if text prompts that contain the word "gay" are detected as potentially harmful or offensive, this could lead to the underrepresentation or even erasure of legitimate image generations by or about the LGBTQIA+ community.

- **Inappropriate or offensive content:** These models can produce other types of inappropriate or offensive content. Examples include the ability to create images that potentially contain harmful artifacts such as hate symbols; images that illicit harmful connotations; images that relate to contested, controversial, or ideologically polarizing topics; images that are manipulative; images that contain sexually charged content that isn't caught by sexual-related content filters; and images that relate to sensitive or emotionally charged topics. For example, a well-intentioned text prompt aimed to create an image of the New York skyline with clouds and airplanes flying over it might unintentionally generate images that illicit sentiments related to the events surrounding 9/11.
- **Disinformation and misinformation about sensitive topics:** Because DALL-E 2 is a powerful image generation model, it can be used to produce disinformation and misinformation that can be extremely harmful. For example, the model might generate an image of a political leader engaging in activity of a violent or sexual (or simply inaccurate) nature that might lead to defamation and other consequential harms, including but not limited to public protests, political change, or fake news.
- **Information reliability:** Language model responses can fabricate content that may sound reasonable but is nonsensical or inaccurate with respect to external validation sources. Even when drawing responses from trusted source information, responses may misrepresent that content.
- **False information:** Azure OpenAI doesn't fact-check or verify content that is provided by customers or users. Depending on how you've developed your application, it might produce false information unless you've built in mitigations (see **Best practices for improving system performance** below).

## System performance

In many AI systems, performance is often defined in relation to accuracy—that is, how often the AI system offers a correct prediction or output. With large-scale natural language models and image models, two different users might look at the same output and have different opinions of how useful or relevant it is, which means that performance for these systems must be defined more flexibly. Here, we broadly consider performance to mean that the application performs as you and your users expect, including not generating harmful outputs.

Azure OpenAI service can support a wide range of applications like search, classification, code generation, and image generation, each with different performance metrics and mitigation strategies. There are several steps you can take to mitigate some of the concerns listed under "Limitations" and to improve performance. Other important

mitigation techniques are outlined in the section [Evaluating and integrating Azure OpenAI for your use](#) below.

## Best practices for improving system performance

- **Show and tell when designing prompts.** With text and code models, make it clear to the model what kind of outputs you expect through instructions, examples, or a combination of the two. If you want the model to rank a list of items in alphabetical order or to classify a paragraph by sentiment, show it that's what you want.
- **Keep your application on topic.** Carefully structure prompts and image inputs to reduce the chance of producing undesired content, even if a user tries to use it for this purpose. For instance, you might indicate in your prompt that a chatbot only engages in conversations about mathematics and otherwise responds "I'm sorry. I'm afraid I can't answer that." Adding adjectives like "polite" and examples in your desired tone to your prompt can also help steer outputs. With image models, you might indicate in your prompt or image input that your application generates only conceptual images. It might otherwise generate a pop-up notification that explains that the application is not for photorealistic use or to portray reality. Consider nudging users toward acceptable queries and image inputs, either by listing such examples up front or by offering them as suggestions upon receiving an off-topic request. Consider training a classifier to determine whether an input (prompt or image) is on topic or off topic.
- **Provide quality data.** With text and code models, if you're trying to build a classifier or get the model to follow a pattern, make sure that there are enough examples. Be sure to proofread your examples—the model is usually smart enough to see through basic spelling mistakes and give you a response, but it also might assume this is intentional and it could affect the response. Providing quality data also includes giving your model reliable data to draw responses from in chat and question answering systems.
- **Measure model quality.** As part of general model quality, consider measuring and improving fairness-related metrics and other metrics related to responsible AI in addition to traditional accuracy measures for your scenario. Consider resources like this checklist when you measure the fairness of the system. These measurements come with limitations, which you should acknowledge and communicate to stakeholders along with evaluation results.
- **Limit the length, structure, and rate of inputs and outputs.** Restricting the length or structure of inputs and outputs can increase the likelihood that the application will stay on task and mitigate, at least in part, any potentially unfair, unreliable, or offensive behavior. Other options to reduce the risk of misuse include (i) restricting

the source of inputs (for example, limiting inputs to a particular domain or to authenticated users rather than being open to anyone on the internet) and (ii) implementing usage rate limits.

- **Encourage human review of outputs prior to publication or dissemination.** With generative AI, there is potential for generating content that might be offensive or not related to the task at hand, even with mitigations in place. To ensure that the generated output meets the task of the customer, consider building ways to remind customers to review their outputs for quality prior to sharing widely. This can reduce many different harms, including offensive material, disinformation, and more.
- **Implement additional scenario-specific mitigations.** Refer to the mitigations outlined in [Evaluating and integrating Azure OpenAI for your use](#) including content moderation strategies. These do not represent every mitigation that might be required for your application, but they point to the general minimum baseline we check for when approving use cases for Azure OpenAI Service.

## Evaluating and integrating Azure OpenAI for your use

For additional information on how to evaluate and integrate these models responsibly, please see the [RAI Overview document](#).

## Learn more about responsible AI

- [Microsoft AI principles](#) ↗
- [Microsoft responsible AI resources](#) ↗
- [Microsoft Azure Learning courses on responsible AI](#)

## Learn more about Azure OpenAI

- [Limited access to Azure OpenAI Service - Azure Cognitive Services | Microsoft Learn](#)
- [Code of Conduct for the Azure OpenAI Service | Microsoft Learn](#)
- [Data, privacy, and security for Azure OpenAI Service - Azure Cognitive Services | Microsoft Learn](#)

# Limited access to Azure OpenAI Service

Article • 04/28/2023

As part of Microsoft's commitment to responsible AI, we are designing and releasing Azure OpenAI Service with the intention of protecting the rights of individuals and society and fostering transparent human-computer interaction. For this reason, we currently limit the access and use of Azure OpenAI, including limiting access to the ability to modify content filters and/or abuse monitoring.

## Registration process

Azure OpenAI requires registration and is currently only available to approved enterprise customers and partners. Customers who wish to use Azure OpenAI are required to submit [a registration form](#).

Customers must attest to any and all use cases for which they will use the service (the use cases from which customers may select will populate in the form after selection of the desired model(s) in Question 22 in the initial registration form). Customers who wish to add additional use cases after initial onboarding must submit the additional use cases using [this form](#). The use of Azure OpenAI is limited to use cases that have been selected in a registration form. Microsoft may require customers to re-verify this information. Read more about example use cases and use cases to avoid [here](#).

Customers who wish to modify content filters and modify abuse monitoring after they have onboarded to the service are subject to additional eligibility criteria and scenario restrictions. At this time, modified content filters and/or modified abuse monitoring for Azure OpenAI Service are only available to managed customers and partners working with Microsoft account teams and have additional use case restrictions. Customers meeting these requirements can register [here](#).

Access to the Azure OpenAI Service is subject to Microsoft's sole discretion based on eligibility criteria and a vetting process, and customers must acknowledge that they have read and understand the Azure terms of service for Azure OpenAI Service.

Azure OpenAI Service is made available to customers under the terms governing their subscription to Microsoft Azure Services, including the Azure OpenAI section of the [Microsoft Product Terms](#). Please review these terms carefully as they contain important conditions and obligations governing your use of Azure OpenAI Service.

## Important links

- Register to use Azure OpenAI [↗](#)
- Add additional use cases [↗](#) (if needed)
- Register to modify content filters and abuse monitoring [↗](#) (if needed)

## Help and support

FAQ about Limited Access can be found [here](#). If you need help with Azure OpenAI, find support [here](#). Report abuse of Azure OpenAI [here](#) [↗](#).

Report problematic content to [cscraireport@microsoft.com](mailto:cscraireport@microsoft.com).

## See also

- [Code of conduct for Azure OpenAI Service integrations](#)
- [Transparency note for Azure OpenAI Service](#)
- [Characteristics and limitations for Azure OpenAI Service](#)
- [Data, privacy, and security for Azure OpenAI Service](#)

# Code of conduct for Azure OpenAI Service

Article • 03/13/2023

The following Code of Conduct defines the requirements that all Azure OpenAI Service implementations must adhere to in good faith. This code of conduct is in addition to the Acceptable Use Policy in the [Microsoft Online Services Terms](#).

## Access requirements

Azure OpenAI Service is a Limited Access service that requires registration and is only available to approved enterprise customers and partners. Customers who wish to use this service are required to [register through this form](#). To learn more, see [Limited Access to Azure OpenAI Service](#).

## Responsible AI mitigation requirements

Integrations with Azure OpenAI Service must:

- Implement meaningful human oversight
- Implement strong technical limits on inputs and outputs to reduce the likelihood of misuse beyond the application's intended purpose
- Test applications thoroughly to find and mitigate undesirable behaviors
- Establish feedback channels
- Implement additional scenario-specific mitigations

To learn more, see the [Azure OpenAI transparency note](#).

## Integrations with Azure OpenAI Service must not:

- be used in any way that violates Microsoft's [Acceptable Use Policy](#), including but not limited to any use prohibited by law, regulation, government order, or decree, or any use that violates the rights of others;
- be used in any way that is inconsistent with this code of conduct, including the Limited Access requirements, the Responsible AI mitigation requirements, and the Content requirements;

- exceed the use case(s) you identified to Microsoft in connection with your request to use the service;
- interact with individuals under the age of consent in any way that could result in exploitation or manipulation or is otherwise prohibited by law or regulation;
- generate or interact with content prohibited in this Code of Conduct;
- be presented alongside or monetize content prohibited in this Code of Conduct;
- make decisions without appropriate human oversight if your application may have a consequential impact on any individual's legal position, financial position, life opportunities, employment opportunities, human rights, or result in physical or psychological injury to an individual;
- infer sensitive information about people without their explicit consent unless if used in a lawful manner by a law enforcement entity, court, or government official subject to judicial oversight in a jurisdiction that maintains a fair and independent judiciary; or
- be used for chatbots that (i) are erotic, romantic, or used for companionship purposes, or which are otherwise prohibited by this Code of Conduct; (ii) are personas of specific people without their explicit consent; (iii) claim to have special wisdom/insight/knowledge, unless very clearly labeled as being for entertainment purposes only; or (iv) enable end users to create their own chatbots without oversight.

## Content requirements

We prohibit the use of our service for generating content that can inflict harm on individuals or society. Our content policies are intended to improve the safety of our platform.

These content requirements apply to the output of all models developed by OpenAI and hosted in Azure OpenAI, such as GPT-3, GPT-4, Codex models, and DALL·E 2, and includes content provided as input to the service and content generated as output from the service.

## Exploitation and Abuse

### Child sexual exploitation and abuse

Azure OpenAI Service prohibits content that describes, features, or promotes child sexual exploitation or abuse, whether or not prohibited by law. This includes sexual content involving a child or that sexualizes a child.

## **Grooming**

Azure OpenAI Service prohibits content that describes or is used for purposes of grooming of children. Grooming is the act of an adult building a relationship with a child for the purposes of exploitation, especially sexual exploitation. This includes communicating with a child for the purpose of sexual exploitation, trafficking, or other forms of exploitation.

## **Non-consensual intimate content**

Azure OpenAI Service prohibits content that describes, features, or promotes non-consensual intimate activity.

## **Sexual solicitation**

Azure OpenAI Service prohibits content that describes, features, or promotes, or is used for, purposes of solicitation of commercial sexual activity and sexual services. This includes encouragement and coordination of real sexual activity.

## **Trafficking**

Azure OpenAI Service prohibits content describing or used for purposes of human trafficking. This includes the recruitment of individuals, facilitation of transport, and payment for, and the promotion of, exploitation of people such as forced labor, domestic servitude, sexual slavery, forced marriages, and forced medical procedures.

## **Suicide and Self-Injury**

Azure OpenAI Service prohibits content that describes, praises, supports, promotes, glorifies, encourages and/or instructs individual(s) on self-injury or to take their life.

## **Violent Content and Conduct**

### **Graphic violence and gore**

Azure OpenAI Service prohibits content that describes, features, or promotes graphic violence or gore.

### **Terrorism and Violent Extremism**

Azure OpenAI Service prohibits content that depicts an act of terrorism; praises, or supports a terrorist organization, terrorist actor, or violent terrorist ideology; encourages terrorist activities; offers aid to terrorist organizations or terrorist causes; or aids in recruitment to a terrorist organization.

## **Violent Threats, Incitement, and Glorification of Violence**

Azure OpenAI Service prohibits content advocating or promoting violence toward others through violent threats or incitement.

## **Harmful Content**

### **Hate speech and discrimination**

Azure OpenAI Service prohibits content that attacks, denigrates, intimidates, degrades, targets, or excludes individuals or groups on the basis of traits such as actual or perceived race, ethnicity, national origin, gender, gender identity, sexual orientation, religious affiliation, age, disability status, caste, or any other characteristic that is associated with systemic prejudice or marginalization.

### **Bullying and harassment**

Azure OpenAI Service prohibits content that targets individual(s) or group(s) with threats, intimidation, insults, degrading or demeaning language or images, promotion of physical harm, or other abusive behavior such as stalking.

### **Deception, disinformation, and inauthentic activity**

Azure OpenAI Service prohibits content that is intentionally deceptive and likely to adversely affect the public interest, including deceptive or untrue content relating to health, safety, election integrity, or civic participation. Azure OpenAI Service also prohibits inauthentic interactions, such as fake accounts, automated inauthentic activity, impersonation to gain unauthorized information or privileges, and claims to be from any person, company, government body, or entity without explicit permission to make that representation.

### **Active malware or exploits**

Content that directly supports unlawful active attacks or malware campaigns that cause technical harms, such as delivering malicious executables, organizing denial of service

attacks, or managing command and control servers.

## Additional content policies

We prohibit the use of our Azure OpenAI Service for scenarios in which the system is likely to generate undesired content due to limitations in the models or scenarios in which the system cannot be applied in a way that properly manages potential negative consequences to people and society. Without limiting the foregoing restriction, Microsoft reserves the right to revise and expand the above Content requirements to address specific harms to people and society.

This includes prohibiting content that is sexually graphic, including consensual pornographic content and intimate descriptions of sexual acts, as well as content that may influence the political process, such as an election, passage of legislation, and content for campaigning purposes.

We may at times limit our service's ability to respond to particular topics, such as probing for personal information or seeking opinions on sensitive topics or current events.

We prohibit the use of Azure OpenAI Service for activities that significantly harm other individuals, organizations, or society, including but not limited to use of the service for purposes in conflict with the applicable [Azure Legal Terms](#) and the [Microsoft Product Terms](#).

## Report abuse

If you suspect that Azure OpenAI Service is being used in a manner that is abusive or illegal, infringes on your rights or the rights of other people, or violates these policies, you can report it at the [Report Abuse Portal](#).

## Report problematic content

If Azure OpenAI Service outputs problematic content that you believe should have been filtered, report it at [cscraireport@microsoft.com](mailto:cscraireport@microsoft.com).

## See also

- [Limited access to Azure OpenAI Service](#)
- [Transparency note for Azure OpenAI Service](#)
- [Data, privacy, and security for Azure OpenAI Service](#)



# Data, privacy, and security for Azure OpenAI Service

Article • 04/05/2023

This article provides details regarding how data provided by you to the Azure OpenAI service is processed, used, and stored. Azure OpenAI stores and processes data to provide the service and to monitor for uses that violate the applicable product terms. Please also see the [Microsoft Products and Services Data Protection Addendum](#), which governs data processing by the Azure OpenAI Service except as otherwise provided in the applicable [Product Terms](#).

Azure OpenAI was designed with compliance, privacy, and security in mind; however, the customer is responsible for its use and the implementation of this technology.

## What data does the Azure OpenAI Service process?

Azure OpenAI processes the following types of data:

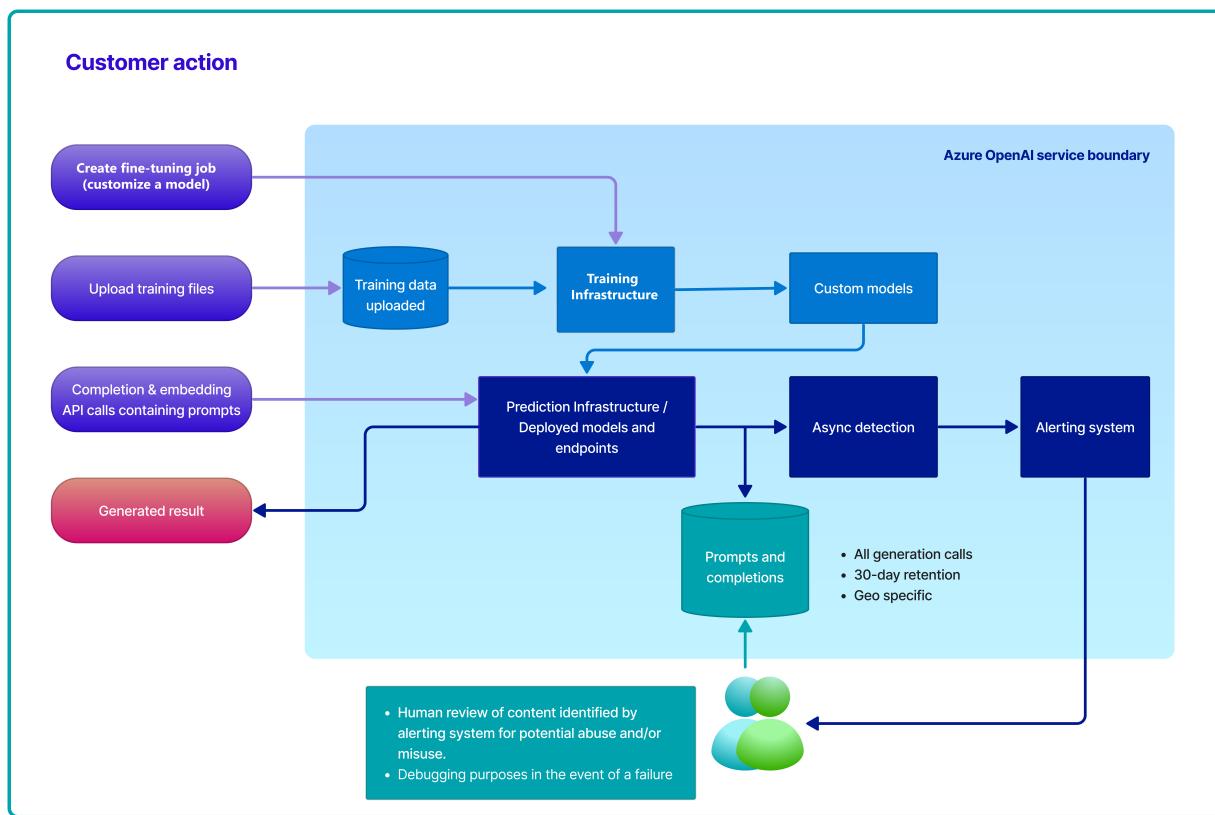
- **Prompts and completions.** Prompts are submitted by the user, and completions are output by the service, via the completions (/completions, /chat/completions) and embeddings operations.
- **Training & validation data.** You can provide your own training data consisting of prompt-completion pairs for the purposes of [fine-tuning an OpenAI model](#).
- **Results data from training process.** After training a fine-tuned model, the service will output meta-data on the job which includes tokens processed and validation scores at each step.

## How does the Azure OpenAI Service process data?

The diagram below illustrates how your data is processed. This diagram covers three different types of processing:

1. How the Azure OpenAI Service creates a fine-tuned (custom) model with your training data;
2. How the Azure OpenAI Service processes your text prompts to generate completions and embeddings results; and

3. How the Azure OpenAI Service and Microsoft personnel analyze prompts and completions for abuse, misuse or harmful content generation, or for debugging purposes in the event of a failure.



## Training data for purposes of fine-tuning an OpenAI model

The training data (prompt-completion pairs) submitted to the Fine-tunes API through the Azure OpenAI Studio is pre-processed using automated tools for quality checking including a data format check. The training data is then imported to the model training component on the Azure OpenAI platform. During the training process, the training data is decomposed into batches and used to modify the weights of the OpenAI models.

Training data provided by the customer is only used to fine-tune the customer's model and is not used by Microsoft to train or improve any Microsoft models.

## Text prompts to generate completions and embeddings results

Once a model deployment (consisting of a customer's fine-tuned model or a base model endpoint) is provisioned in a customer's Azure OpenAI resource, the customer can submit text prompts to the model using our Completions or Embeddings operations

through the REST API, client libraries, or the Azure OpenAI Studio; the model generates text outputs (completions) that are returned through the API.

When data is submitted to the service, it is processed through our content filters as well as the specified OpenAI model. The content filtering models are run on both the prompt inputs as well as the generated completions.

No prompts or completions are stored in the model during these operations, and prompts and completions are not used to train, retrain or improve the models.

## Preventing abuse and harmful content generation

The Azure OpenAI Service includes a content management system that works alongside the models to filter potentially harmful content. This system works by running both the input prompt and generated completion through an ensemble of classification models aimed at detecting misuse. If the system identifies harmful content, customers receive either an error on the API call if the prompt was deemed inappropriate or the finish\_reason on the response will be content\_filter to signify that some of the generation was filtered. Learn more about content filtering [here](#). No prompts or completions are stored in these content classification models; prompts and completions are not used to train, retrain or improve the classification models.

In addition to synchronous content filtering, the Azure OpenAI Service stores prompts and completions from the service for up to thirty (30) days to monitor for content and/or behaviors that suggest use of the service in a manner that may violate applicable product terms. Authorized Microsoft employees may review prompt and completion data that has triggered our automated systems to investigate and verify potential abuse. For customers who have deployed Azure OpenAI Service in the European Economic Area, the authorized Microsoft employees will be located in the European Economic Area.

In the event of a confirmed policy violation, a customer may be asked to take immediate action to remediate the issue to and to prevent further abuse. Failure to address the issue may result in suspension or termination of Azure OpenAI resource access.

Customers may request to modify content filtering and/or abuse monitoring by submitting the form [here](#). If a customer is approved and remains in compliance with all requirements to modify abuse monitoring, then prompts and completions are not stored.

# How is data retained and what Customer controls are available?

- **Training, validation, and training results data.** The Files API allows customers to upload their training data for the purpose of fine-tuning a model. This data is stored in Azure Storage, encrypted at rest by Microsoft Managed keys, within the same region as the resource and logically isolated with their Azure subscription and API Credentials. Uploaded files can be deleted by the user via the DELETE API operation.
- **Fine-tuned OpenAI models.** The Fine-tunes API allows customers to create their own fine-tuned version of the OpenAI models based on the training data that they have uploaded to the service via the Files APIs. The trained fine-tuned models are stored in Azure Storage in the same region, encrypted at rest and logically isolated with their Azure subscription and API credentials. Fine-tuned models can be deleted by the user by calling the DELETE API operation.
- **Prompts and completions.** The prompts and completions data may be temporarily stored by the Azure OpenAI Service in the same region as the resource for up to 30 days. This data is encrypted and is only accessible to authorized Microsoft employees for (1) debugging purposes in the event of a failure, and (2) investigating patterns of abuse and misuse to determine if the service is being used in a manner that violates the applicable product terms. Note: When a customer is approved for modified abuse monitoring, prompts and completions data are not stored, and thus Microsoft employees have no access to the data.

To learn more about Microsoft's privacy and security commitments visit the [Microsoft Trust Center](#).

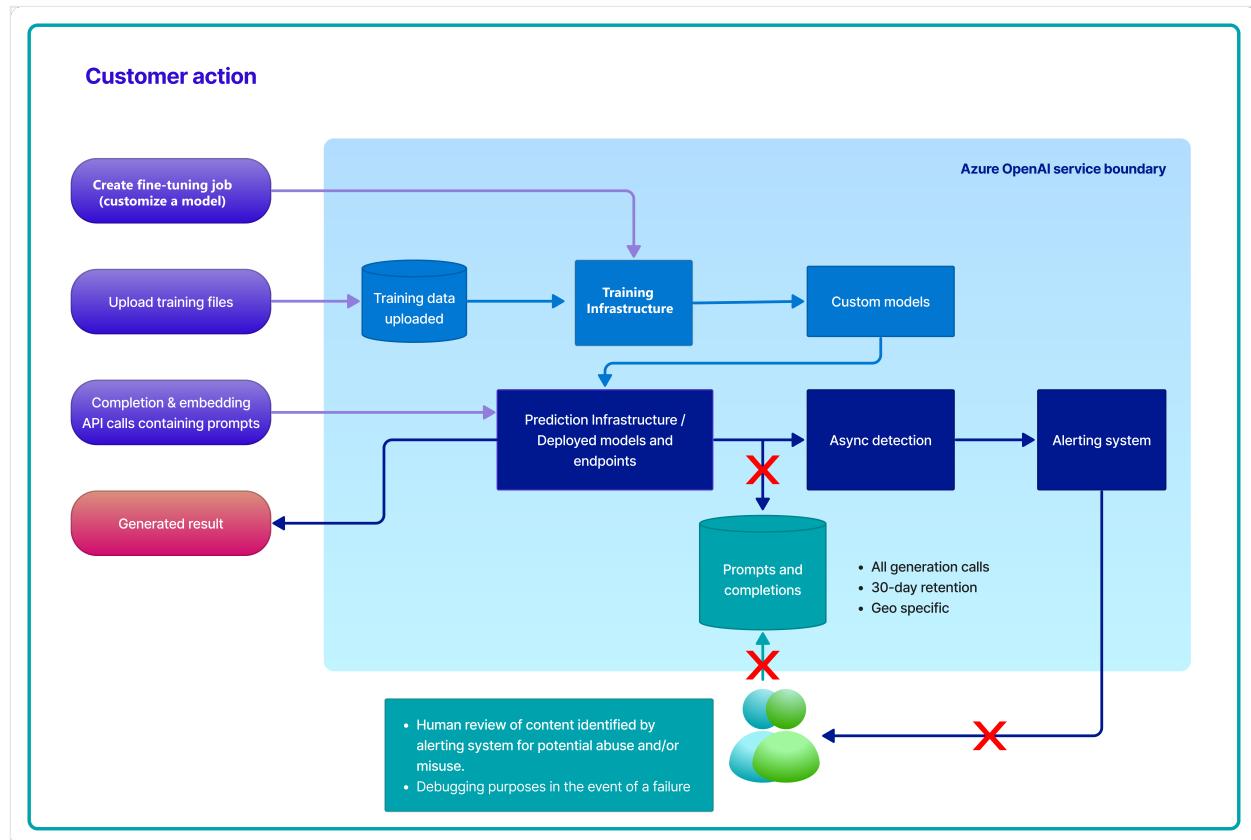
## Frequently asked questions

### Can a customer opt out of the logging and human review process?

Some customers may want to use the Azure OpenAI Service for a use case that involves the processing of sensitive, highly confidential, or legally-regulated input data but where the likelihood of harmful outputs and/or misuse is low. These customers may conclude that they do not want or do not have the right to permit Microsoft to process such data for abuse detection, as described above, due to their internal policies or applicable legal regulations. To address these concerns, Microsoft allows customers who meet additional

Limited Access eligibility criteria and attest to specific use cases to apply to modify the Azure OpenAI content management features.

If Microsoft approves a customer's request to modify abuse monitoring, then Microsoft does not store any prompts and completions associated with the approved Azure subscription for which abuse monitoring is configured off. In this case, because no prompts and completions are stored at rest in the Service Results Store, the human review process is not possible and is not performed.



Customers can apply for modified abuse monitoring here:

<https://aka.ms/oai/modifiedaccess>.

## How can a customer verify if logging for abuse monitoring is off?

There are two ways for an approved customer to verify that logging for abuse monitoring has been turned off in their approved Azure subscription: (1) Azure portal or (2) Azure CLI (or any MGMT API).

### ⓘ Note

Note: The value of "false" for the "ContentLogging" attribute appears only if logging is turned off. Otherwise, this property will not appear in either Azure portal

or Azure CLI's output. If a customer wants to verify if the logging is on, they would need to raise a support ticket.

## Prerequisites

- Sign into Azure
- Select the Azure Subscription which hosts the Azure OpenAI Service resource.
- Navigate to the "Overview" page of the Azure OpenAI Service resource.

## Logging Status Verification via the Azure portal

- Go to the resource Overview page
- Click on the 'JSON view' link on the top right corner as shown in the image below

The screenshot shows the Azure portal interface for the 'Azure OpenAI' resource. On the left, there's a sidebar with navigation links like 'Overview', 'Activity log', 'Access control (IAM)', 'Tags', 'Diagnose and solve problems', 'Resource Management' (with sub-links for 'Keys and Endpoint', 'Model deployments', 'Pricing tier', 'Networking', 'Identity', 'Cost analysis', and 'Properties'), and 'Get Started'. The main content area displays resource details under 'Essentials'. It includes fields for 'Resource group' (with a 'move' link), 'Status' (Active), 'Location' (West Europe), 'Subscription' (with a 'move' link), 'API type' (Azure OpenAI), 'Pricing tier' (Standard), 'Endpoint' (https://.openai.azure.com/), and 'Manage keys' (with a 'Click here to manage keys' link). At the top right of the main content area, there are two buttons: 'View Cost' and 'JSON View', with 'JSON View' being highlighted by a red box. A survey prompt 'Help us improve Azure OpenAI. Take our survey!' is also visible at the top.

- There will be a value in the Capabilities list called 'ContentLogging' which will appear and be set to FALSE when logging for abuse monitoring is off.

The screenshot shows a JSON object with a single entry: 'ContentLogging' with a value of 'false'. The JSON is displayed in a code editor-like interface with syntax highlighting.

```
JSON
...
{
  "name": "ContentLogging",
  "value": "false"
}
...
```

## Logging Status Verification via the Azure CLI (or MGMT API)

Execute the following command in Azure CLI to see the same JSON data as shown in the Azure portal above.

```
az cognitiveservices account show -n resource\_name -g resource\_group
```

## Is customer data logged with content filtering?

No, content filtering functions differently than abuse monitoring, and logging or storage of data is not needed. Content filtering works by applying algorithmic detection to the prompts and completions at inference time to determine if content should be filtered. No data is stored at rest as part of this process, and thus there is no human review of prompts and/or completions that have been subject to filtering. You can learn more about content filtering in the service here: [Azure OpenAI Service content filtering - Azure OpenAI | Microsoft Learn](#).

You can also apply to modify Azure OpenAI content filtering here: [Azure OpenAI Limited Access Review: Modified Content Filters and Abuse Monitoring \(microsoft.com\)](#). As described above, although customers use the same form to request modification of content filtering and/or abuse monitoring, the impact on data logging and access differs depending on which aspects of the content management system are modified.

## How can a customer verify if content filtering is off?

If a customer wants to verify that content filtering for their approved subscription ID(s) has been configured off, they can create a support ticket in the Azure Portal in one of two ways:

1. From the main portal dashboard in *Help + Support* in the side navigation bar
2. Within the Azure OpenAI resource itself in *New Support Request* in the side navigation bar

## What is covered by customer managed key encryption?

Customer-managed keys (CMK), also known as Bring your own key (BYOK), offer greater flexibility to create, rotate, disable, and revoke access controls. You can also audit the encryption keys used to protect your data. CMK encrypts all customer data stored at rest in the Azure OpenAI Service (such as data uploaded for fine-tuning) **except** for data logged for 30 days as described above. (Learn more: [Azure OpenAI Service encryption of data at rest - Azure Cognitive Services | Microsoft Learn](#))

## What happens if Microsoft needs access to my data?

Most operations, support, and troubleshooting performed by Microsoft personnel and sub-processors do not require access to customer data. In those rare circumstances where such access is required, Customer Lockbox for Microsoft Azure provides an interface for customers to review and approve or reject customer data access requests. Customer Lockbox is used in cases where a Microsoft engineer needs to access customer data, whether in response to a customer-initiated support ticket or a problem identified by Microsoft. In the Azure OpenAI Service, Customer Lockbox applies to all customer data stored by the service (such as data uploaded for fine-tuning) except for the prompts and completions logged and accessed for the purposes of abuse monitoring as described above.

You can learn more about customer lockbox here: [Customer Lockbox for Microsoft Azure | Microsoft Learn](#)

## Is customer data processed by Azure OpenAI sent to OpenAI?

No. Microsoft hosts the OpenAI models within our Azure infrastructure, and all customer data sent to Azure OpenAI remains within the Azure OpenAI service.

## Is customer data used to train the OpenAI models?

No. We do not use customer data to train, retrain or improve the models in the Azure OpenAI Service.

## See also

- [Limited access to Azure OpenAI Service](#)
- [Code of conduct for Azure OpenAI Service integrations](#)
- [Transparency note and use cases for Azure OpenAI Service](#)
- [Characteristics and limitations for Azure OpenAI Service](#)
- Report abuse of Azure OpenAI Service through the [Report Abuse Portal ↗](#)
- Report problematic content to [cscraireport@microsoft.com](mailto:cscraireport@microsoft.com)

# Azure OpenAI Service REST API 참조

아티클 • 2023. 05. 15.

이 문서에서는 Azure OpenAI에 대한 유추 REST API 엔드포인트에 대한 세부 정보를 제공합니다.

## 인증

Azure OpenAI는 두 가지 인증 방법을 제공합니다. API 키 또는 Azure Active Directory를 사용할 수 있습니다.

- API 키 인증:** 이 인증 형식의 경우 모든 API 요청은 `api-key` HTTP 헤더에 API 키를 포함해야 합니다. [빠른 시작](#)은 이러한 형식의 인증으로 전화를 거는 방법에 대한 자침을 제공합니다.
- Azure Active Directory 인증:** Azure Active Directory 토큰을 사용하여 API 호출을 인증할 수 있습니다. 인증 토큰은 요청에 `Authorization` 헤더로 포함됩니다. 제공된 토큰은 `Bearer`가 앞에 와야 합니다(예: `Bearer YOUR_AUTH_TOKEN`). [Azure Active Directory로 인증하는 방법](#) 가이드를 읽을 수 있습니다.

## REST API 버전 관리

서비스 API는 `api-version` 쿼리 매개 변수를 사용하여 버전이 지정됩니다. 모든 버전은 YYYY-MM-DD 날짜 구조를 따릅니다. 예를 들면 다음과 같습니다.

HTTP

POST

`https://YOUR_RESOURCE_NAME.openai.azure.com/openai/deployments/YOUR_DEPLOYMENT_NAME/completions?api-version=2023-05-15`

## 완성

완료 작업을 통해 모델은 제공된 프롬프트를 기반으로 하나 이상의 예측 완료를 생성합니다. 서비스는 또한 각 위치에서 대체 토큰의 확률을 반환할 수 있습니다.

### 완료 만들기

HTTP

```
POST https://{{your-resource-name}}.openai.azure.com/openai/deployments/{{deployment-id}}/completions?api-version={{api-version}}
```

## 경로 매개 변수

| 매개 변수              | Type | 필수 여부 | Description                               |
|--------------------|------|-------|-------------------------------------------|
| your-resource-name | 문자열  | 필수    | Azure OpenAI 리소스의 이름입니다.                  |
| deployment-id      | 문자열  | 필수    | 모델을 배포할 때 선택한 배포 이름입니다.                   |
| api-version        | 문자열  | 필수    | 이 작업에 사용할 API 버전입니다. YYYY-MM-DD 형식을 따릅니다. |

## 지원되는 버전

- 2023-03-15-preview Swagger 사양 ↗
- 2022-12-01 Swagger 사양 ↗
- 2023-05-15 Swagger 사양 ↗

## 요청 본문

| 매개 변수      | Type      | 필수 여부 | 기본값               | Description                                                                                                                                     |
|------------|-----------|-------|-------------------|-------------------------------------------------------------------------------------------------------------------------------------------------|
| prompt     | 문자열 또는 배열 | 선택 사항 | <\\ endoftext\\ > | 문자열, 문자열 목록 또는 토큰 목록으로 인코딩된 완료를 생성하기 위한 프롬프트입니다. <\\ endoftext\\ >는 학습 중에 모델이 보는 문서 구분 기호이므로 프롬프트가 지정되지 않은 경우 모델은 새 문서의 시작 부분에서 생성되는 것처럼 생성됩니다. |
| max_tokens | 정수        | 선택 사항 | 16                | 완료 시 생성할 최대 토큰 수입니다. 프롬프트의 토큰 수와 max_tokens는 모델의 컨텍스트 길이를 초과할 수 없습니다. 대부분의 모델에는 컨텍스트 길이가 2048인 토큰이 있습니다(4096을 지원하는 최신 모델 제외).                   |

| 매개 변수       | Type   | 필수 여부 | 기본값  | Description                                                                                                                                                                                                                                                                                                                                                                            |
|-------------|--------|-------|------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| temperature | number | 선택사항  | 1    | 사용할 샘플링 온도(0에서 2 사이)입니다. 값이 높을수록 모델이 더 많은 위험을 감수하게 됩니다. 더 창의적인 애플리케이션의 경우 0.9를 시도하고 답변이 잘 정의된 애플리케이션의 경우 0( <code>argmax sampling</code> )을 시도합니다. 일반적으로 이를 변경하거나 <code>top_p</code> 를 변경하는 것이 좋지만 둘 다 변경하는 것은 권장하지 않습니다.                                                                                                                                                              |
| top_p       | number | 선택사항  | 1    | 모델이 <code>top_p</code> 확률 질량을 가진 토큰의 결과를 고려하는 핵 샘플링이라고 하는 온도를 사용한 샘플링의 대안입니다. 따라서 0.1은 상위 10% 확률 질량을 구성하는 토큰만 고려됨을 의미합니다. 일반적으로 이를 변경하거나 온도를 변경하는 것이 좋지만 둘 다 변경하는 것은 권장하지 않습니다.                                                                                                                                                                                                        |
| logit_bias  | map    | 선택사항  | null | 완료 시 지정된 토큰이 나타날 가능성을 수정합니다. 토큰(GPT 토크나이저에서 토큰 ID로 지정)을 -100에서 100 사이의 관련 바이어스 값에 매핑하는 json 객체를 허용합니다. 이 토크나이저 도구(GPT-2 및 GPT-3 모두에서 작동)를 사용하여 텍스트를 토큰 ID로 변환할 수 있습니다. 수학적으로, 바이어스는 샘플링 전에 모델에 의해 생성된 로짓에 추가됩니다. 정확한 효과는 모델마다 다르지만 -1과 1 사이의 값은 선택 가능성을 줄이거나 높여야 합니다. -100 또는 100과 같은 값은 관련 토큰을 금지하거나 독점적으로 선택해야 합니다. 예를 들어 {"50256": -100}을 전달하여 < endoftext > 토큰이 생성되지 않도록 할 수 있습니다. |
| user        | 문자열    | 옵션    |      | 최종 사용자를 나타내는 고유 식별자로, 남용을 모니터링하고 감지하는 데 도움이 됩니다.                                                                                                                                                                                                                                                                                                                                       |
| n           | 정수     | 선택사항  | 1    | 각 프롬프트에 대해 생성할 완료 수입니다. 참고: 이 매개 변수는 많은 완료를 생성하므로 토큰 할당량을 빠르게 소모할 수 있습니다. 신중하게 사용하고 <code>max_tokens</code> 및 중지에 대한 적절한 설정이 있는지 확인합니다.                                                                                                                                                                                                                                                |

| 매개 변수             | Type      | 필수 여부 | 기본값  | Description                                                                                                                                                                                            |
|-------------------|-----------|-------|------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| stream            | boolean   | 선택사항  | 거짓   | 부분 진행률을 다시 스트리밍할지 여부를 나타냅니다. 설정된 경우 토큰은 데이터 전용 서버 전송 이벤트로 전송되며 스트림은 data: [DONE] 메시지로 종료됩니다.                                                                                                           |
| logprobs          | 정수        | 선택사항  | null | 가장 가능성이 높은 토큰과 선택한 토큰에 대한 로그 확률을 포함합니다. 예를 들어 logprobs가 10이면 API는 가장 가능성이 높은 토큰 10개의 목록을 반환합니다. API는 항상 샘플링된 토큰의 logprob를 반환하므로 응답에 최대 logprobs+1 요소가 있을 수 있습니다. 이 매개 변수는 gpt-35-turbo와 함께 사용할 수 없습니다. |
| suffix            | 문자열       | 선택사항  | null | 삽입된 텍스트가 완성된 후에 오는 접미사입니다.                                                                                                                                                                             |
| echo              | boolean   | 선택사항  | 거짓   | 완료와 함께 프롬프트를 다시 에코합니다. 이 매개 변수는 gpt-35-turbo와 함께 사용할 수 없습니다.                                                                                                                                           |
| stop              | 문자열 또는 배열 | 선택사항  | null | API가 추가 토큰 생성을 중지하는 최대 4 개의 시퀀스입니다. 반환된 텍스트에는 중지 시퀀스가 포함되지 않습니다.                                                                                                                                       |
| presence_penalty  | number    | 선택사항  | 0    | -2.0~2.0 사이의 숫자 양수 값은 지금까지 텍스트에 나타나는지 여부에 따라 새 토큰에 페널티를 부여하여 모델이 새 항목에 대해 이야기할 가능성을 높입니다.                                                                                                              |
| frequency_penalty | number    | 선택사항  | 0    | -2.0~2.0 사이의 숫자 양수 값은 지금까지 텍스트의 기존 빈도를 기반으로 새 토큰에 불이익을 주어 모델이 동일한 줄을 그대로 반복할 가능성을 줄입니다.                                                                                                                |

| 매개 변수   | Type | 필수 여부 | Description                                                                                                                                                                                                                                                                                                 |
|---------|------|-------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| best_of | 정수   | 선택사항  | <p>서버 쪽에서 best_of 완료를 생성하고 "최상"(토큰당 로그 확률이 가장 낮은 것)을 반환합니다. 결과를 스트리밍할 수 없습니다. n과 함께 사용하면 best_of는 후보 완료 횟수를 제어하고 n은 반환할 횟수를 지정합니다. best_of는 n보다 커야 합니다.</p> <p>참고: 이 매개 변수는 많은 완료를 생성하므로 토큰 할당량을 빠르게 소모할 수 있습니다. 신중하게 사용하고 max_tokens 및 중지에 대한 적절한 설정이 있는지 확인합니다. 이 매개 변수는 gpt-35-turbo와 함께 사용할 수 없습니다.</p> |

## 요청 예

콘솔

```
curl
https://YOUR_RESOURCE_NAME.openai.azure.com/openai/deployments/YOUR_DEPLOYMENT_NAME/completions?api-version=2023-05-15 \
-H "Content-Type: application/json" \
-H "api-key: YOUR_API_KEY" \
-d "{"
  "prompt": "Once upon a time",
  "max_tokens": 5
}"
```

## 예제 응답

JSON

```
{
  "id": "cmpl-4kGh7iXtjW4lc9eGhff6Hp8C7btdQ",
  "object": "text_completion",
  "created": 1646932609,
  "model": "ada",
  "choices": [
    {
      "text": ", a dark line crossed",
      "index": 0,
      "logprobs": null,
      "finish_reason": "length"
    }
}
```

```
        ]  
    }
```

예제 응답에서 `finish_reason`은 `stop`과 같습니다. `finish_reason`이 `content_filter`와 같으면 [콘텐츠 필터링 가이드](#)를 참조하여 이 문제가 발생하는 이유를 이해하세요.

## 포함

기계 학습 모델 및 기타 알고리즘에서 쉽게 사용할 수 있는 지정된 입력의 벡터 표현을 가져옵니다.

### ① 참고

현재 단일 API 호출에 포함 일괄 처리를 지원하지 않습니다. 오류가 `InvalidRequestError: Too many inputs. The max number of inputs is 1. We hope to increase the number of inputs per request soon.` 발생하면 일반적으로 포함 배열을 단일 문자열이 아닌 일괄 처리로 전달하려고 할 때 발생합니다. `text-embedding-ada-002`(버전 2) 모델을 사용하는 경우 문자열의 길이는 최대 8191개 편집할 수 있습니다.

## 포함 만들기

### HTTP

```
POST https://{{your-resource-name}}.openai.azure.com/openai/deployments/{{deployment-id}}/embeddings?api-version={{api-version}}
```

## 경로 매개 변수

| 매개 변수                           | Type | 필수 여부 | Description                               |
|---------------------------------|------|-------|-------------------------------------------|
| <code>your-resource-name</code> | 문자열  | 필수    | Azure OpenAI 리소스의 이름입니다.                  |
| <code>deployment-id</code>      | 문자열  | 필수    | 모델 배포의 이름입니다. 전화를 걸기 전에 먼저 모델을 배포해야 합니다.  |
| <code>api-version</code>        | 문자열  | 필수    | 이 작업에 사용할 API 버전입니다. YYYY-MM-DD 형식을 따릅니다. |

## 지원되는 버전

- 2023-03-15-preview Swagger 사양 ↴
- 2022-12-01 Swagger 사양 ↴
- 2023-05-15 Swagger 사양 ↴

## 요청 본문

| 매개 변수 | Type | 필수 여부 | Description                                                                                                             |
|-------|------|-------|-------------------------------------------------------------------------------------------------------------------------|
| input | 문자열  | 예     | 문자열로 인코딩된 예 대한 포함을 가져올 입력 텍스트입니다. 입력 토<br>큰 수는 <a href="#">사용 중인 모델에</a> 따라 달라집니다.                                      |
|       |      | 당 없음  | 코드를 포함하지 않는 한 입력에서 줄 바꿈(\n)을 단일 공백으로 바꾸<br>는 것이 좋습니다. 줄 바꿈이 있을 때 열등한 결과가 관찰되었기 때문<br>입니다.                               |
| user  | 문자열  | 예     | 최종 사용자를 나타내는 고유 식별자입니다. 이렇게 하면 Azure<br>OpenAI가 남용을 모니터링하고 검색하는 데 도움이 됩니다. PII 식별<br>자를 전달하지 말고 GUID와 같은 의사 익명 값을 대신 사용 |

## 요청 예

콘솔

```
curl
https://YOUR_RESOURCE_NAME.openai.azure.com/openai/deployments/YOUR_DEPLOYMENT_NAME/embeddings?api-version=2023-05-15 \
-H "Content-Type: application/json" \
-H "api-key: YOUR_API_KEY" \
-d "{\"input\": \"The food was delicious and the waiter...\"}"
```

## 예제 응답

JSON

```
{
  "object": "list",
  "data": [
    {
      "object": "embedding",
      "embedding": [
        0.018990106880664825,
        -0.0073809814639389515,
        .... (1024 floats total for ada)
        0.021276434883475304,
      ],
    },
  ],
}
```

```
        "index": 0
    }
],
"model": "text-similarity-babbage:001"
}
```

## 채팅 완료

ChatGPT 및 GPT-4 모델을 사용하여 채팅 메시지에 대한 완성을 만듭니다.

### 채팅 완료 만들기

HTTP

```
POST https://{{your-resource-name}}.openai.azure.com/openai/deployments/{{deployment-id}}/chat/completions?  
api-version={{api-version}}
```

### 경로 매개 변수

| 매개 변수              | Type | 필수 여부 | Description                               |
|--------------------|------|-------|-------------------------------------------|
| your-resource-name | 문자열  | 필수    | Azure OpenAI 리소스의 이름입니다.                  |
| deployment-id      | 문자열  | 필수    | 모델 배포의 이름입니다. 전화를 걸기 전에 먼저 모델을 배포해야 합니다.  |
| api-version        | 문자열  | 필수    | 이 작업에 사용할 API 버전입니다. YYYY-MM-DD 형식을 따릅니다. |

### 지원되는 버전

- 2023-03-15-preview [Swagger 사양 ↗](#)
- 2023-05-15 [Swagger 사양 ↗](#)

## 요청 예

콘솔

```
curl  
https://YOUR_RESOURCE_NAME.openai.azure.com/openai/deployments/YOUR_DEPLOYMENT_NAME/chat/completions?api-version=2023-05-15 \  
-H "Content-Type: application/json" \  
-H "api-key: YOUR_API_KEY" \  
-d "{"prompt": "Hello, how are you?"}
```

```
-d '{"messages":[{"role": "system", "content": "You are a helpful assistant."}, {"role": "user", "content": "Does Azure OpenAI support customer managed keys?"}, {"role": "assistant", "content": "Yes, customer managed keys are supported by Azure OpenAI."}, {"role": "user", "content": "Do other Azure Cognitive Services support this too?"}]}'
```

## 예제 응답

콘솔

```
{"id": "chatcmpl-6v7mkQj980V1yBec6ETrKPRqFjNw9",  
"object": "chat.completion", "created": 1679072642,  
"model": "gpt-35-turbo",  
"usage": {"prompt_tokens": 58,  
"completion_tokens": 68,  
"total_tokens": 126},  
"choices": [{"message": {"role": "assistant",  
"content": "Yes, other Azure Cognitive Services also support customer managed keys. Azure Cognitive Services offer multiple options for customers to manage keys, such as using Azure Key Vault, customer-managed keys in Azure Key Vault or customer-managed keys through Azure Storage service. This helps customers ensure that their data is secure and access to their services is controlled."}, "finish_reason": "stop", "index": 0}]}
```

예제 응답에서 `finish_reason`은 `stop`과 같습니다. `finish_reason`이 `content_filter`와 같으면 [콘텐츠 필터링 가이드](#)를 참조하여 이 문제가 발생하는 이유를 이해하세요.

읽기 용이성을 위해 조정된 출력 서식입니다. 실제 출력은 줄 바꿈이 없는 단일 텍스트 블록입니다.

| 매개 변수                    | Type   | 필수<br>여부 | 기본<br>값 | Description                                                                                                                                              |
|--------------------------|--------|----------|---------|----------------------------------------------------------------------------------------------------------------------------------------------------------|
| <code>messages</code>    | array  | 필수       |         | 채팅 형식으로 채팅 완료를 생성할 메시지입니다.                                                                                                                               |
| <code>temperature</code> | number | 선택<br>사항 | 1       | 사용할 샘플링 온도(0에서 2 사이)입니다. 0.8과 같이 값이 높을수록 좀 더 무작위로 출력이 생성되고, 0.2와 같이 값이 낮을수록 더 집중적이고 결정적인 출력이 생성됩니다.\n일반적으로 이 값 또는 <code>top_p</code> 중 하나만 변경하는 것이 좋습니다. |

| 매개 변수             | Type      | 필수 여부 | 기본 값  | Description                                                                                                                                                                                                                                       |
|-------------------|-----------|-------|-------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| n                 | 정수        | 선택사항  | 1     | 각 입력 메시지에 대해 생성할 채팅 완료 선택 항목 수입니다.                                                                                                                                                                                                                |
| stream            | boolean   | 선택사항  | false | 설정되면 ChatGPT의 경우처럼 부분 메시지 델타가 전송됩니다. 토큰은 데이터 전용 서버 전송 이벤트로 전송되며 스트림은 <code>data: [DONE]</code> 메시지로 종료됩니다.                                                                                                                                        |
| stop              | 문자열 또는 배열 | 선택사항  | null  | API가 추가 토큰 생성을 중지하는 최대 4개의 시퀀스입니다.                                                                                                                                                                                                                |
| max_tokens        | 정수        | 선택사항  | inf   | 생성된 답변에 허용되는 최대 토큰 수입니다. 기본적으로 모델이 반환할 수 있는 토큰 수는 (4096 - 프롬프트 토큰)입니다.                                                                                                                                                                            |
| presence_penalty  | number    | 선택사항  | 0     | -2.0~2.0 사이의 숫자 양수 값은 지금까지 텍스트에 나타나는지 여부에 따라 새 토큰에 페널티를 부여하여 모델이 새 항목에 대해 이야기할 가능성을 높입니다.                                                                                                                                                         |
| frequency_penalty | number    | 선택사항  | 0     | -2.0~2.0 사이의 숫자 양수 값은 지금까지 텍스트의 기존 빈도를 기반으로 새 토큰에 불이익을 주어 모델이 동일한 줄을 그대로 반복할 가능성을 줄입니다.                                                                                                                                                           |
| logit_bias        | object    | 선택사항  | null  | 완료 시 지정된 토큰이 나타날 가능성을 수정합니다. 토큰(토크나이저에서 토큰 ID로 지정)을 -100에서 100 사이의 관련 바이어스 값에 매핑하는 json 객체를 허용합니다. 수학적으로, 바이어스는 샘플링 전에 모델에 의해 생성된 로짓에 추가됩니다. 정확한 효과는 모델마다 다르지만 -1과 1 사이의 같은 선택 가능성을 줄이거나 높여야 합니다. -100 또는 100과 같은 같은 관련 토큰을 금지하거나 독점적으로 선택해야 합니다. |
| user              | 문자열       | 옵션    |       | Azure OpenAI가 남용을 모니터링하고 감지하는 데 도움이 될 수 있는 최종 사용자를 나타내는 고유 식별자입니다.                                                                                                                                                                                |

## 관리 API

Azure OpenAI는 Azure Cognitive Services의 일부로 배포됩니다. 모든 Cognitive Services는 만들기, 업데이트 및 삭제 작업을 위해 동일한 관리 API 집합에 의존합니다. 관리 API는 OpenAI 리소스 내에서 모델을 배포하는 데에도 사용됩니다.

## 관리 API 참조 설명서

# 다음 단계

REST API를 사용하여 배포, 모델 및 미세 조정 관리에 대해 알아봅니다. [Azure OpenAI를 지원하는 기본 모델](#)에 대해 자세히 알아봅니다.

# Fine Tunes

참조

Service: Cognitive Services

API Version: 2022-12-01

## Operations

|            |                                                                                                     |
|------------|-----------------------------------------------------------------------------------------------------|
| Cancel     | 지정된 미세 조정 ID로 지정된 미세 조정 작업의 처리를 취소합니다.                                                              |
| Create     | 지정된 학습 파일에서 지정된 모델을 미세 조정하는 작업을 만듭니다. 응답에는 작업 상태 및 하이퍼 매개 변수를 포함하여 큐에 추가된 작업의 세부 정보가 포함됩니다.         |
| Delete     | 지정된 미세 조정 ID로 지정된 미세 조정 작업을 삭제합니다.                                                                  |
| Get        | 지정된 미세 조정 ID로 지정된 단일 미세 조정 작업에 대한 세부 정보를 가져옵니다. 세부 정보에는 기본 모델, 학습 및 유효성 검사 파일, 하이퍼 매개 변수가 포함됩니다.... |
| Get Events | 지정된 미세 조정 ID로 지정된 미세 조정 작업에 대한 이벤트를 가져옵니다. 이벤트는 작업 상태가 변경될 때(예: 실행 중 또는 완료) 및 다시 실행될 때 만들어집니다.      |
| List       | Azure OpenAI 리소스가 소유한 모든 미세 조정 작업의 목록을 가져옵니다. 각 미세 조정 작업에 대해 반환되는 세부 정보에는 해당 식별자 외에 기본...           |

# Deployments - Create Or Update

참조

Service: Cognitive Services

API Version: 2021-10-01

Cognitive Services 계정과 연결된 배포의 상태를 업데이트합니다.

HTTP

PUT

<https://management.azure.com/subscriptions/{subscriptionId}/resourceGroups/{resourceGroupName}/providers/Microsoft.CognitiveServices/accounts/{accountName}/deployments/{deploymentName}?api-version=2021-10-01>

## URI 매개 변수

| Name              | In    | Required | Type   | Description                                                                   |
|-------------------|-------|----------|--------|-------------------------------------------------------------------------------|
| accountName       | path  | True     | string | Cognitive Services 계정의 이름입니다.<br>Regex pattern: ^[a-zA-Z0-9][a-zA-Z0-9_.-]*\$ |
| deploymentName    | path  | True     | string | Cognitive Services 계정과 연결된 배포의 이름입니다.                                         |
| resourceGroupName | path  | True     | string | 리소스 그룹의 이름. 이름은 대소문자를 구분하지 않습니다.                                              |
| subscriptionId    | path  | True     | string | 대상 구독의 ID입니다.                                                                 |
| api-version       | query | True     | string | 이 작업에 사용할 API 버전입니다.                                                          |

## 요청 본문

| Name       | Type                 | Description                      |
|------------|----------------------|----------------------------------|
| properties | DeploymentProperties | Cognitive Services 계정 배포의 속성입니다. |

## 응답

| Name               | Type          | Description                |
|--------------------|---------------|----------------------------|
| 200 OK             | Deployment    | 배포를 만들거나 업데이트합니다.          |
| 201 Created        | Deployment    | 배포를 성공적으로 만듭니다.            |
| Other Status Codes | ErrorResponse | 작업이 실패한 이유를 설명하는 오류 응답입니다. |

## 예제

### PutDeployment

#### Sample Request

HTTP

## HTTP

PUT

<https://management.azure.com/subscriptions/subscriptionId/resourceGroups/resourceGroupName/providers/Microsoft.CognitiveServices/accounts/accountName/deployments/deploymentName?api-version=2021-10-01>

```
{  
  "properties": {  
    "model": {  
      "format": "OpenAI",  
      "name": "ada",  
      "version": "1"  
    },  
    "scaleSettings": {  
      "scaleType": "Manual",  
      "capacity": 1  
    }  
  }  
}
```

## Sample Response

Status code: 200

### JSON

```
{  
  "id":  
    "/subscriptions/subscriptionId/resourceGroups/resourceGroupName/providers/Microsoft.CognitiveServices/accounts/accountName/  
deployments/deploymentName",  
  "name": "deploymentName",  
  "type": "Microsoft.CognitiveServices/accounts/deployments",  
  "properties": {  
    "model": {  
      "format": "OpenAI",  
      "name": "ada",  
      "version": "1"  
    },  
    "scaleSettings": {  
      "scaleType": "Manual",  
      "capacity": 1  
    },  
    "provisioningState": "Succeeded"  
  }  
}
```

Status code: 201

### JSON

```
{  
  "id":  
    "/subscriptions/subscriptionId/resourceGroups/resourceGroupName/providers/Microsoft.CognitiveServices/accounts/accountName/  
deployments/deploymentName",  
  "name": "deploymentName",  
  "type": "Microsoft.CognitiveServices/accounts/deployments",  
  "properties": {  
    "model": {  
      "format": "OpenAI",  
      "name": "ada",  
      "version": "1"  
    },  
    "scaleSettings": {  
      "scaleType": "Manual",  
      "capacity": 1  
    },  
    "provisioningState": "Accepted"  
  }  
}
```

## 정의

createdByType

리소스를 만든 ID의 형식입니다.

|                             |                                     |
|-----------------------------|-------------------------------------|
| Deployment                  | Cognitive Services 계정 배포.           |
| DeploymentModel             | Cognitive Services 계정 배포 모델의 속성입니다. |
| DeploymentProperties        | Cognitive Services 계정 배포의 속성입니다.    |
| DeploymentProvisioningState | 작업이 호출되었을 때 리소스의 상태를 가져옵니다.         |
| DeploymentScaleSettings     | Cognitive Services 계정 배포 모델의 속성입니다. |
| DeploymentScaleType         | 배포 크기 조정 유형입니다.                     |
| ErrorAdditionalInfo         | 리소스 관리 오류 추가 정보입니다.                 |
| ErrorDetail                 | 오류 세부 정보입니다.                        |
| ErrorResponse               | 오류 응답                               |
| systemData                  | 리소스의 생성 및 마지막 수정과 관련된 메타데이터입니다.     |

## createdByType

리소스를 만든 ID의 형식입니다.

| Name            | Type   | Description |
|-----------------|--------|-------------|
| Application     | string |             |
| Key             | string |             |
| ManagedIdentity | string |             |
| User            | string |             |

## Deployment

Cognitive Services 계정 배포.

| Name       | Type                  | Description                                                                                                                                          |
|------------|-----------------------|------------------------------------------------------------------------------------------------------------------------------------------------------|
| etag       | string                | 리소스 Etag.                                                                                                                                            |
| id         | string                | 리소스에 대한 정규화된 리소스 ID입니다. 예 - /subscriptions/{subscriptionId}/resourceGroups/{resourceGroupName}/providers/{resourceProviderNamespace}/{resourceType}, |
| name       | string                | 리소스의 이름입니다.                                                                                                                                          |
| properties | Deployment Properties | Cognitive Services 계정 배포의 속성입니다.                                                                                                                     |
| systemData | systemData            | 리소스의 생성 및 마지막 수정과 관련된 메타데이터입니다.                                                                                                                      |
| type       | string                | 리소스 형식입니다. 예: "Microsoft.Compute/virtualMachines" 또는 "Microsoft.Storage/storageAccounts"                                                             |

## DeploymentModel

Cognitive Services 계정 배포 모델의 속성입니다.

| Name    | Type   | Description  |
|---------|--------|--------------|
| format  | string | 배포 모델 형식입니다. |
| name    | string | 배포 모델 이름입니다. |
| version | string | 배포 모델 버전입니다. |

## DeploymentProperties

Cognitive Services 계정 배포의 속성입니다.

| Name              | Type                         | Description                         |
|-------------------|------------------------------|-------------------------------------|
| model             | Deployment Model             | Cognitive Services 계정 배포 모델의 속성입니다. |
| provisioningState | Deployment ProvisioningState | 작업이 호출되었을 때 리소스의 상태를 가져옵니다.         |
| scaleSettings     | Deployment ScaleSettings     | Cognitive Services 계정 배포 모델의 속성입니다. |

## DeploymentProvisioningState

작업이 호출되었을 때 리소스의 상태를 가져옵니다.

| Name      | Type   | Description |
|-----------|--------|-------------|
| Accepted  | string |             |
| Creating  | string |             |
| Deleting  | string |             |
| Failed    | string |             |
| Moving    | string |             |
| Succeeded | string |             |

## DeploymentScaleSettings

Cognitive Services 계정 배포 모델의 속성입니다.

| Name      | Type                 | Description     |
|-----------|----------------------|-----------------|
| capacity  | integer              | 배포 용량.          |
| scaleType | Deployment ScaleType | 배포 크기 조정 유형입니다. |

## DeploymentScaleType

배포 크기 조정 유형입니다.

| Name     | Type   | Description |
|----------|--------|-------------|
| Manual   | string |             |
| Standard | string |             |

## ErrorAdditionalInfo

리소스 관리 오류 추가 정보입니다.

| Name | Type   | Description  |
|------|--------|--------------|
| info | object | 추가 정보입니다.    |
| type | string | 추가 정보 유형입니다. |

## ErrorDetail

오류 세부 정보입니다.

| Name           | Type                                  | Description  |
|----------------|---------------------------------------|--------------|
| additionalInfo | <a href="#">ErrorAdditionalInfo[]</a> | 오류 추가 정보입니다. |
| code           | string                                | 오류 코드입니다.    |
| details        | <a href="#">ErrorDetail[]</a>         | 오류 세부 정보입니다. |
| message        | string                                | 오류 메시지입니다.   |
| target         | string                                | 오류 대상입니다.    |

## ErrorResponse

오류 응답

| Name  | Type                        | Description |
|-------|-----------------------------|-------------|
| error | <a href="#">ErrorDetail</a> | 오류 개체입니다.   |

## systemData

리소스의 생성 및 마지막 수정과 관련된 메타데이터입니다.

| Name           | Type                          | Description             |
|----------------|-------------------------------|-------------------------|
| createdAt      | string                        | 리소스 만들기의 타임스탬프(UTC)입니다. |
| createdBy      | string                        | 리소스를 만든 ID입니다.          |
| createdByType  | <a href="#">createdByType</a> | 리소스를 만든 ID의 형식입니다.      |
| lastModifiedAt | string                        | 리소스 마지막 수정의 타임스탬프(UTC)  |
| lastModifiedBy | string                        | 리소스를 마지막으로 수정한 ID입니다.   |

lastModifiedByType

[createdByType](#)

리소스를 마지막으로 수정한 ID의 형식입니다.

# Azure.AI.OpenAI 네임스페이스

참조

## ① 중요

일부 정보는 릴리스되기 전에 상당 부분 수정될 수 있는 시험판 제품과 관련이 있습니다. Microsoft는 여기에 제공된 정보에 대해 어떠한 명시적이거나 묵시적인 보증도 하지 않습니다.

## 클래스

|                                           |                                                                                    |
|-------------------------------------------|------------------------------------------------------------------------------------|
| <a href="#">ChatChoice</a>                | 채팅 완료 요청에 대한 단일 완료 결과의 표현입니다.                                                      |
| <a href="#">ChatCompletions</a>           | 채팅 완료 요청에 대한 전체 응답의 표현입니다.                                                         |
| <a href="#">ChatCompletionsOptions</a>    | 채팅 완료 요청에 사용되는 구성 정보입니다.                                                           |
| <a href="#">ChatMessage</a>               | 채팅 완료 상호 작용 내에서 역할 특성이 지정된 단일 메시지입니다.                                              |
| <a href="#">Choice</a>                    | 완료 응답 내의 선택 모델입니다.                                                                 |
| <a href="#">Completions</a>               | 완료 요청에 대한 예상 응답 스키마입니다.                                                            |
| <a href="#">CompletionsLogProbability</a> | 완료 선택 내의 LogProbs 모델입니다.                                                           |
| <a href="#">CompletionsOptions</a>        | 본문 스키마를 게시하여 배포에서 프롬프트 완료를 만듭니다.                                                   |
| <a href="#">CompletionsUsage</a>          | 완료 요청에 대해 처리된 토큰 수의 표현입니다. 개수는 프롬프트, 선택, 선택 대체, best_of 세대 및 기타 소비자의 모든 토큰을 고려합니다. |
| <a href="#">EmbeddingItem</a>             | 개체 목록 항목 요청을 포함할 것으로 예상되는 응답 스키마입니다.                                               |
| <a href="#">Embeddings</a>                | 요청을 포함할 것으로 예상되는 응답 스키마입니다.                                                        |
| <a href="#">EmbeddingsOptions</a>         | 배포에서 프롬프트 완료를 만드는 스키마입니다.                                                          |
| <a href="#">EmbeddingsUsage</a>           | 이 요청 및 응답에 사용된 토큰의 양을 측정합니다.                                                       |
| <a href="#">OpenAIClient</a>              | 완료 및 검색을 위한 Azure OpenAI API.                                                      |
| <a href="#">OpenAIClientOptions</a>       | OpenAIClient에 대한 클라이언트 옵션입니다.                                                      |
| <a href="#">StreamingChatChoice</a>       |                                                                                    |

[StreamingChatCompletions](#)

[StreamingChoice](#)

[StreamingCompletions](#)

## 구조체

[ChatRole](#)

채팅 완료 상호 작용 내에서 메시지의 의도된 목적에 대한 설명입니다.

## 열거형

[OpenAIClientOptions.Service](#)

사용할 서비스의 버전입니다.

[Version](#)

# com.azure.ai.openai

Reference

Package: com.azure.ai.openai

Maven Artifact: [com.azure:azure-ai-openai:1.0.0-beta.1](#) ↗

Package containing the classes for OpenAI. Azure OpenAI APIs for completions and search.

## Classes

|                                     |                                                                   |
|-------------------------------------|-------------------------------------------------------------------|
| <a href="#">OpenAIAsyncClient</a>   | Initializes a new instance of the asynchronous OpenAIClient type. |
| <a href="#">OpenAIClient</a>        | Initializes a new instance of the synchronous OpenAIClient type.  |
| <a href="#">OpenAIClientBuilder</a> | A builder for creating a new instance of the OpenAIClient type.   |

## Enums

|                                      |                                  |
|--------------------------------------|----------------------------------|
| <a href="#">OpenAIServiceVersion</a> | Service version of OpenAIClient. |
|--------------------------------------|----------------------------------|

# @azure/openai package

참조

## 클래스

|                    |                                                                                                                                                                         |
|--------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| AzureKeyCredential | 기본 키 값 업데이트를 지원하는 정적 키 기반 자격 증명입니다.                                                                                                                                     |
| OpenAIclient       | Azure OpenAI와 상호 작용하기 위한 클라이언트입니다.<br><br>클라이언트에는 OpenAI 리소스의 엔드포인트와 API 키 또는 토큰과 같은 인증 방법이 필요합니다. API 키 및 엔드포인트는 OpenAI 리소스 페이지에서 찾을 수 있습니다. 리소스의 키 및 엔드포인트 페이지에 있습니다. |

### 인증 예제:

#### API 키

JavaScript

```
import { OpenAIclient } from "@azure/openai";
import { AzureKeyCredential } from "@azure/core-auth";

const endpoint = "<azure endpoint>";
const credential = new AzureKeyCredential("<api key>");

const client = new OpenAIclient(endpoint,
credential);
```

#### Azure Active Directory

JavaScript

```
import { OpenAIclient } from "@azure/openai";
import { DefaultAzureCredential } from "@azure/identity";

const endpoint = "<azure endpoint>";
const credential = new DefaultAzureCredential();
```

```
const client = new OpenAIClient(endpoint,  
credential);
```

## OpenAIKeyCredential

OpenAIKeyCredential 클래스는 OpenAI API 키를 나타내며 OpenAI 엔드포인트에 대한 OpenAI 클라이언트에 인증하는 데 사용됩니다.

# 인터페이스

## ChatChoice

전체 채팅 완료 요청의 일부로 단일 프롬프트 완료의 표현입니다. 일반적으로 `n` 기본값이 1인 제공된 프롬프트당 선택 항목이 생성됩니다. 토큰 제한 및 기타 설정은 생성된 선택 항목 수를 제한할 수 있습니다.

## ChatCompletions

채팅 완료 요청의 응답 데이터 표현입니다. 완료는 다양한 작업을 지원하고 제공된 프롬프트 데이터에서 계속되거나 "완료"되는 텍스트를 생성합니다.

## ChatMessage

채팅 완료 상호 작용 내에서 역할 특성이 지정된 단일 메시지입니다.

## Choice

전체 완료 요청의 일부로 단일 프롬프트 완료의 표현입니다. 일반적으로 `n` 기본값이 1인 제공된 프롬프트당 선택 항목이 생성됩니다. 토큰 제한 및 기타 설정은 생성된 선택 항목 수를 제한할 수 있습니다.

## Completions

완료 요청의 응답 데이터 표현입니다. 완료는 다양한 작업을 지원하고 제공된 프롬프트 데이터에서 계속되거나 "완료"되는 텍스트를 생성합니다.

## CompletionsLogProbability Model

완료 생성을 위한 로그 확률 모델의 표현입니다.

## CompletionsUsage

완료 요청에 대해 처리된 토큰 수의 표현입니다. 개수는 프롬프트, 선택, 선택 대체, `best_of` 세대 및 기타 소비자의 모든 토큰을 고려합니다.

## EmbeddingItem

단일 포함 관련 비교의 표현입니다.

## Embeddings

포함 요청의 응답 데이터 표현입니다. 포함은 텍스트 문자열의 관련성을 측정하며 검색, 클러스터링, 권장 사항 및 기타 유사한 시나리오에 일반적으로 사용됩니다.

## EmbeddingsUsage

이 요청 및 응답에 사용된 토큰의 양을 측정합니다.

## GetChatCompletionsOptions

## GetCompletionsOptions

[GetEmbeddingsOptions](#)

[OpenAIClientOptions](#)

[RequestOptions](#)

## 형식 별칭

|                          |                               |
|--------------------------|-------------------------------|
| <a href="#">ChatRole</a> | "system", "assistant", "user" |
|--------------------------|-------------------------------|

|                                         |                                                   |
|-----------------------------------------|---------------------------------------------------|
| <a href="#">CompletionsFinishReason</a> | "stopped", "tokenLimitReached", "contentFiltered" |
|-----------------------------------------|---------------------------------------------------|

# Azure Cognitive Services 지원 및 도움말 옵션

아티클 • 2022. 12. 05.

이제 막 Azure Cognitive Services의 기능을 탐색하기 시작했나요? 애플리케이션에 새로운 기능을 구현하는 경우를 가정해 볼 수 있습니다. 또는 서비스를 사용한 후 개선 방법에 대한 제안이 있나요? 다음은 Cognitive Services에 대한 지원을 받고, 최신 상태를 유지하고, 피드백을 제공하고, 버그를 보고할 수 있는 옵션입니다.

## Azure 지원 요청 만들기

### A

지금 클라우드를 시작하려는 개발자든 비즈니스에 중요한 전략적 애플리케이션을 배포하려는 대규모 조직이든 관계없이 가장 적합한 [Azure 지원 옵션 및 플랜 선택](#)의 범위를 탐색합니다. Azure 고객은 Azure Portal에서 지원 요청을 만들고 관리할 수 있습니다.

- [Azure Portal](#)
- [미국 정부의 Azure Portal](#)

## Microsoft Q&A에 질문 게시

Microsoft 엔지니어, Azure MVP(가장 귀중한 전문가) 또는 전문가 커뮤니티의 기술 제품 관련 질문에 대한 빠르고 안정적인 답변을 얻으려면 Azure가 커뮤니티 지원을 위해 선호하는 대상인 [Microsoft Q&A](#)에 참여하세요.

검색을 사용하여 문제에 대한 답변을 찾을 수 없으면 Microsoft Q&A에 새 질문을 제출하세요. 질문을 할 때 다음 태그 중 하나를 사용합니다.

- [Cognitive Services](#)

### Vision

- [Computer Vision](#)
- [Custom Vision](#)
- [Face](#)
- [Form Recognizer](#)
- [Video Indexer](#)

### 언어

- 몰입형 리더
- 언어 이해(LUIS)
- QnA Maker
- 언어 서비스
- Translator

## Speech

- Speech Service

## 의사 결정

- Anomaly Detector
- Content Moderator
- Metrics Advisor
- Personalizer

## Azure OpenAI

- Azure OpenAI

# Stack Overflow에 질문을 게시합니다.



가장 큰 커뮤니티 개발자 에코시스템의 개발자 질문에 대한 답변을 보려면 Stack Overflow에서 질문하세요.

Stack Overflow에 새 질문을 제출하는 경우 질문을 만들 때 다음 태그 중 하나 이상을 사용하세요.

- Cognitive Services ↗

## Vision

- Computer Vision ↗
- Custom Vision ↗
- Face ↗
- Form Recognizer ↗
- Video Indexer ↗

## 언어

- 몰입형 리더 ↗
- 언어 이해(LUIS) ↗

- QnA Maker ↗
- 언어 서비스 ↗
- Translator ↗

## Speech

- Speech Service ↗

## 의사 결정

- Anomaly Detector ↗
- Content Moderator ↗
- Metrics Advisor ↗
- Personalizer ↗

## Azure OpenAI

- Azure OpenAI ↗

# 피드백 제출

새로운 기능을 요청하려면 <https://feedback.azure.com> ↗에 게시합니다. Cognitive Services를 만드는 아이디어를 공유하세요. 해당 API는 개발하는 애플리케이션에서 더 잘 작동합니다.

- Cognitive Services ↗

## Vision

- Computer Vision ↗
- Custom Vision ↗
- Face ↗
- Form Recognizer ↗
- Video Indexer ↗

## 언어

- 물입형 리더 ↗
- 언어 이해(LUIS) ↗
- QnA Maker ↗
- 언어 서비스 ↗
- Translator ↗

## Speech

- [Speech Service](#)

## 의사 결정

- [Anomaly Detector](#)
- [Content Moderator](#)
- [Metrics Advisor](#)
- [Personalizer](#)

## 최신 소식 수신

Azure 블로그의 새 릴리스 또는 뉴스에 있는 기능에 대한 정보를 계속 수신하면 프로그래밍 오류, 서비스 버그 또는 Cognitive Services에서 아직 사용할 수 없는 기능 간의 차이점을 찾는 데 도움이 될 수 있습니다.

- [Azure 업데이트](#)에서 제품 업데이트, 로드맵 및 공지 사항에 대해 자세히 알아봅니다.
- Cognitive Services에 대한 뉴스는 [Azure 블로그](#)에서 공유됩니다.
- Cognitive Services에 대한 [Reddit의 대화에 참여](#)하세요.

## 다음 단계

[Azure Cognitive Services란?](#)