Getting started with LLM

Since the 1950s, humans have begun to explore AI. Artificial intelligence has always given people the impression of being an epoch-making technology that requires practitioners in the field of data science with professional skills to use it. But starting from the end of 2022, the field of AI has undergone major changes. Since OpenAI released the GPT-3 model, artificial intelligence is no longer a solution for a single field or single scenario. The new multi-modal LLM changes the rules of the game. No matter what kind of work you are engaged in, you can use natural language to communicate with the LLM. The large language model feeds back to you generative content. This It includes text, pictures, videos, etc., greatly enhancing usability. Before entering the content of Semantic Kernel, I hope to tell you stories related to LLM models so that you can better understand LLM models.

Basic Concept of LLM

Large Language Model (LLM), also known as large language model, is an artificial intelligence model designed to understand and generate human language. They are trained on large amounts of text data and can perform a wide range of tasks, including text summarization, translation, sentiment analysis, and more. Large language models are characterized by their large size and contain billions of parameters, helping them learn complex patterns in language data. These models are often based on deep learning architectures as well as the Transformer algorithm. Large language model models are trained through self-supervised learning, which generates its own labels for the input data by predicting the next word or token in the sequence, given the previous words. The training process consists of two main steps: pre-training and fine-tuning. During the pre-training phase, the model learns from a large, diverse dataset, often containing billions of words from different sources, such as websites, books, and articles. In the fine-tuning phase, the model is further trained on a more specific and smaller data set related to the target task or domain, which helps the model fine-tune its understanding and adapt to the special requirements of the task. Many companies are now developing LLMs. The well-known ones include OpenAI's GPT-X and DALLE-X series, Meta's LLama, Google's Gemini, and Baidu's ERNIE. GPT-4 in OpenAI is the best LLM at this stage and has great advantages. But as time goes by, many good industry vertical field models have been born.

Transformer

The Transformer algorithm is a deep learning model based on the self-attention mechanism, which can be used to process natural language and other sequence data. It consists of two parts: an encoder and a decoder. Each part contains multiple layers, and each layer contains multi-head self-attention and feed-forward neural networks. The advantage of the Transformer algorithm is that it can process the entire sequence in parallel and capture long-distance dependencies without using recurrent neural networks or convolutional neural networks. The Transformer algorithm was originally proposed in the paper "Attention Is All You Need"1 for machine translation tasks. Later, it was widely used in other natural language processing tasks, such as text summarization, question answering, speech recognition, etc. Some famous Transformer-based models are BERT, GPT-3/4, T5, etc.

This series mainly focuses on OpenAI 3/3.5/4 and DALLE-3 of Azure OpenAI Services. As for other models, we will mention them in more advanced content in the future. You can follow my GitHub Repo.

Introduction to OpenAI and OpenAI's models

Although Transformer's algorithm comes from Google, it is OpenAI that really brings LLMs into the public eye. OpenAI is an artificial intelligence research and deployment company, and its mission is to ensure that artificial general intelligence (AGI) can benefit all mankind. Its vision is to create an AGI that can cooperate and compete with humans while adhering to human values and ethics. OpenAI was originally a non-profit organization founded in 2015 by some well-known figures in the technology industry, such as Elon Musk, Peter Thiel, Jerry Yang, etc.1. Its goal is to advance the development of digital intelligence so that it can benefit humanity to the greatest extent without being bound by monetary interests. Its research is open and transparent and can be accessed and used by anyone. OpenAI has pioneered research in the field of artificial intelligence, especially in generative models and security. It has developed some powerful large language models, such as GPT-3/3.5/4, ChatGPT, DALL·E, Whisper, etc., which can understand and generate various forms of data such as text, images, sounds, etc. It also seeks to explore the potential risks and impacts of AGI and how they can be aligned with human goals and interests.

GPT Models

GPT (Generative Pre-trained Transformer, Generative Pre-trained Transformer) is a natural language processing (NLP) model based on deep learning, developed by OpenAI. The GPT series of models is known for its power and flexibility, and performs well in a variety of language tasks. The following are some key features and development history of the GPT model:

Core features of GPT

1. Based on Transformer architecture: The GPT model is based on Transformer architecture, which is a deep learning model particularly suitable for processing sequence data (such as text).

2. Large-scale pre-training: GPT learns the common patterns and structures of language by pre-training on large amounts of text data. This includes text from various sources such as books, web pages, news articles, etc.

3. Fine-tuning application: After pre-training, the GPT model can be fine-tuned for specific tasks, such as question and answer, text generation, translation, etc.

4. Context-sensitive: The GPT model is able to understand and generate context-sensitive text, which makes it outstanding in generating coherent and relevant content.

GPT development history

GPT-1: A first release that demonstrates the potential of pre-training on large-scale unlabeled data and the effectiveness of fine-tuning on a variety of tasks.

GPT-2: Increases the model size and training data volume, significantly improving the quality and accuracy of text generation. GPT-2 has attracted widespread attention for its ability to generate text that is coherent and sometimes indistinguishable from human-written text.

GPT-3: Further expanded the model size, reaching an unprecedented 175 billion parameters. GPT-3 achieves revolutionary performance on multiple NLP tasks, especially when little or no fine-tuning is possible.

GPT-4 and beyond: With the continuous development of technology, subsequent GPT models may continue to improve in terms of model scale, understanding capabilities, and multi-modal capabilities.

GPT application areas GPT models are widely used in many fields, including but not limited to:

Text generation: such as article writing, creative writing, code generation, etc.

Chatbots: Provide a smooth conversational experience.

Natural language understanding: such as sentiment analysis, text classification, etc.

Translation and multilingual tasks: automatically translate different languages.

Knowledge extraction and question answering: Extract information from large amounts of text to answer specific questions.

Overall, the GPT model represents an important milestone in the current field of artificial intelligence and natural language processing. Its powerful capabilities and diverse application prospects continue to lead the trend of technological development.

GPT-3

GPT-3 is a LLMs developed by OpenAI that can understand and generate natural language. It is one of the LLMs currently, with 175 billion parameters, and can complete text summarization, machine translation, dialogue systems, code generation, etc. The characteristic of GPT-3 is that it can adapt to different tasks and fields through simple text prompts, that is, "few-shot learning", without requiring additional fine-tuning or labeling data. GPT-3 opened Pandora’s box and changed the rules of the industry. GPT-3 has been used in many products and services, such as OpenAI API, OpenAI Codex, early GitHub Copilot, etc., which can make it easier for developers, creators, and scholars to use and learn artificial intelligence. GPT-3 has also triggered some discussions and thinking about the ethics, society and security of artificial intelligence, such as artificial intelligence's bias, explainability, responsibility, impact, etc.

GPT-3.5 and ChatGPT

GPT-3.5 and ChatGPT are both large language models based on the GPT-3 architecture, which can understand and generate natural language. They all have 175 billion parameters and can perform amazingly well on a variety of language processing tasks, such as text summarization, machine translation, dialogue systems, code generation, etc.

The main difference between GPT-3.5 and ChatGPT is their scope and purpose. GPT-3.5 is a general language model that can handle a variety of language processing tasks. ChatGPT, on the other hand, is a specialized model designed specifically for chat applications. It emphasizes interaction and communication with users, and can play different roles, such as cat ladies, celebrities, politicians, etc. It can also generate multimedia content such as images, music, and videos based on user input.

Another difference between GPT-3.5 and ChatGPT is their training data and training methods. GPT-3.5 is pre-trained on 570 GB of text data from different sources such as websites, books, articles, etc. It is trained through self-supervised learning, which generates its own labels for the input data by predicting the next word or token in the sequence, given the previous words. ChatGPT is based on GPT-3.5 and uses more conversation data, such as social media, chat records, movie scripts, etc., for further fine-tuning. Its training method is through multi-task learning, that is, optimizing multiple goals at the same time, such as language model, dialogue generation, emotion classification, image generation, etc.

GPT-4

GPT-4 (Generative Pre-trained Transformer 4th Generation) is the latest generation of artificial intelligence language models developed by OpenAI. It is the successor of GPT-3 with more advanced and refined features. Here are some of the key features of GPT-4:

Larger knowledge base and data processing capabilities: GPT-4 can process larger amounts of data, and its knowledge base is broader and deeper than GPT-3.

Higher language understanding and generation capabilities: GPT-4 has significantly improved in understanding and generating natural language, and can more accurately understand complex language structures and meanings.

Multi-modal capabilities: GPT-4 can not only process text, but also understand and generate images, providing a multi-modal interactive experience.

Better contextual understanding: GPT-4 can better understand and maintain context in long conversations, providing more coherent and consistent responses.

Improved security and reliability: OpenAI has strengthened the filtering and control of inappropriate content in GPT-4 to provide a more secure and reliable user experience.

Wide range of application fields: GPT- can be used in various fields, including but not limited to chatbots, content creation, educational assistance, language translation, data analysis, etc.

Overall, GPT-4 has made significant improvements and enhancements over its predecessor model, providing more powerful and diverse features. GPT-4 has an absolute leadership position at this stage and is also the goal of many companies' large models.

GPT-4V

The full name of GPT-4V is GPT-4 Turbo with Vision. It can understand pictures, analyze pictures for users, and answer questions related to pictures. GPT-4V can accurately understand the content of images, identify objects in images, count the number of objects, provide image-related insights and information, extract text, etc. It can be said that GPT-4V is the king of LLMs, and it also allows LLMs to better understand the world. GPT-4V’s main vision capabilities and application directions

Object Detection: GPT-4V is able to identify and detect a variety of common objects in images, such as cars, animals, and household items. Its recognition capabilities have been evaluated on standard image datasets.

Text Recognition: This model features optical character recognition (OCR) technology that finds printed or handwritten text in images and converts it into machine-readable text. This feature is proven in images such as documents, logos, and titles.

Face Recognition: GPT-4V is able to find and recognize faces in images. It also has a degree of ability to determine gender, age and racial attributes from facial features. The model’s facial analysis capabilities have been tested on datasets such as FairFace and LFW.

CAPTCHA SOLVING: GPT-4V demonstrates visual reasoning capabilities in solving text- and image-based CAPTCHAs. This indicates that the model has advanced puzzle-solving skills.

Geolocation: GPT-4V is able to identify cities or geographical locations represented in landscape images. This shows that the model has mastered knowledge about the real world, but it also means that there is a risk of privacy leakage.

Complex Images: The model performs poorly when dealing with complex scientific diagrams, medical scans, or images with multiple overlapping text components. It cannot grasp contextual details.

DALL·E

DALL·E is an advanced artificial intelligence program developed by OpenAI specifically designed to generate images. It is a neural network model based on the GPT-3 architecture, but unlike GPT-3, which mainly processes text, DALL·E's expertise lies in generating corresponding images based on text descriptions. The name of this model is a tribute to the famous artist Salvador Dalí and the popular animated character WALL·E.

Key features of DALL·E Text to image conversion: DALL·E can generate images based on text descriptions provided by users. These descriptions can be very specific or creative, and the model will do its best to generate images that match the description.

Creativity and Flexibility: DALL·E displays amazing creativity when generating images, able to combine different concepts and elements to create unique and innovative visual works.

Variety and detail: The model is capable of generating multiple styles and types of images and can handle complex, detailed descriptions.

Application potential: DALL·E has extensive application potential in art creation, advertising, design and other fields.

DALL·E application scenarios include

Artistic Creation: Artists and designers can use DALL·E to explore new ideas and visual expressions.

Advertising and Media: Generate images that fit a specific theme or concept.

Education and Entertainment: Used in the production of instructional materials or the creation of entertainment content.

Research and exploration: Explore the possibilities of artificial intelligence in the field of visual arts.

The emergence of DALL·E marks an important progress in artificial intelligence in creative tasks and shows the huge potential of AI in the field of visual arts. Now the latest DALL·E model is the DALL·E 3 .

Whisper·

Whisper is an advanced automatic speech recognition (ASR) model developed by OpenAI. This model focuses on transcribing speech into text and has shown excellent performance in multiple languages and different environments. Here are some key features about the Whisper model:

Features

Multi-language support: Whisper models are capable of handling many different languages and dialects, making them widely applicable across the globe.

High-precision recognition: It can accurately recognize and transcribe speech, maintaining a high accuracy even in environments with a lot of background noise.

Adaptable to different contexts: Whisper can not only recognize standard voice input, but also adapt to various colloquial and informal conversation styles.

Easy to integrate and use: As a machine learning model, Whisper can be integrated into various applications and services to provide speech recognition capabilities.

Application

Automatic subtitles and transcription: Automatically generate subtitles or text for video and audio content.

Voice assistants and chatbots: Improve the ability of voice assistants and chatbots to recognize voice commands.

Accessibility Services: Help people with hearing impairments better understand audio content.

Meeting and Lecture Recording: Automatically record and transcribe meeting or lecture content.

Overall, the Whisper model represents an important advancement in the field of automatic speech recognition, and its multi-language and high-precision recognition capabilities make it extremely valuable in a variety of application scenarios.

Microsoft & OpenAI

The partnership between Microsoft and OpenAI is an important development in contemporary artificial intelligence. Microsoft has been an important partner and supporter of OpenAI since its inception. Here are some key aspects and impacts of their collaboration:

Investment and Cooperation

Financial support: Microsoft made significant investments in OpenAI in its early days, including hundreds of millions of dollars in funding. These investments help OpenAI develop its research projects and technology.

Cloud computing resources: Microsoft provides OpenAI with the resources of its Azure cloud computing platform, which is crucial for training and running large AI models, such as GPT and DALL·E series models.

Technical cooperation

Joint research and development: The two companies have cooperated on multiple AI projects and technologies to jointly promote the development of artificial intelligence.

Product integration: Some of OpenAI’s technologies, such as GPT-3, have been integrated into Microsoft products and services, such as Microsoft Azure and other enterprise-level solutions.

Strategic Cooperation

Sustainable and safe AI: Both parties are committed to developing AI technology that is both sustainable and safe, and pay attention to AI ethics and safety issues.

Expand AI applications: Through cooperation, the two companies are committed to applying AI technology to a wider range of fields, such as health care, education, and environmental protection.

Influence

Accelerate the development of AI technology: This cooperation promotes the rapid development and innovation of AI technology.

Business applications and services: Microsoft has promoted the widespread application of artificial intelligence in the business field by applying OpenAI's technology to its products and services.

Promote AI democratization: This collaboration helps make advanced AI technology accessible and usable to more enterprises and developers.

Overall, the cooperation between Microsoft and OpenAI is a model of combining technological innovation and commercial applications. This cooperation has had a profound impact on the development and popularization of artificial intelligence technology. As cooperation between the two parties continues to deepen, it can be expected that they will continue to play an important role in the field of artificial intelligence.

Azure OpenAI Service

Azure OpenAI Service is a collaboration between Microsoft Azure and OpenAI. Azure OpenAI Service is a cloud-based platform that enables developers and data scientists to quickly and easily build and deploy artificial intelligence models. With Azure OpenAI, users can access a variety of AI tools and technologies to create intelligent applications, including natural language processing, computer vision, and deep learning. Azure OpenAI Service is designed to accelerate the development of AI applications, allowing users to focus on creating innovative solutions that create value for their organizations and customers.

Azure OpenAI Service provides REST API access to OpenAI's powerful language models, including GPT-4, GPT-4 Turbo with Vision, GPT-3.5-Turbo, and the family of embedded models. Additionally, the new GPT-4 and GPT-3.5-Turbo model series are now officially released. These models can be easily adapted to specific tasks, including but not limited to content generation, aggregation, image understanding, semantic search, and natural language to code conversion. Users can access the service through a REST API, Python SDK, or a web-based interface in Azure OpenAI Studio.

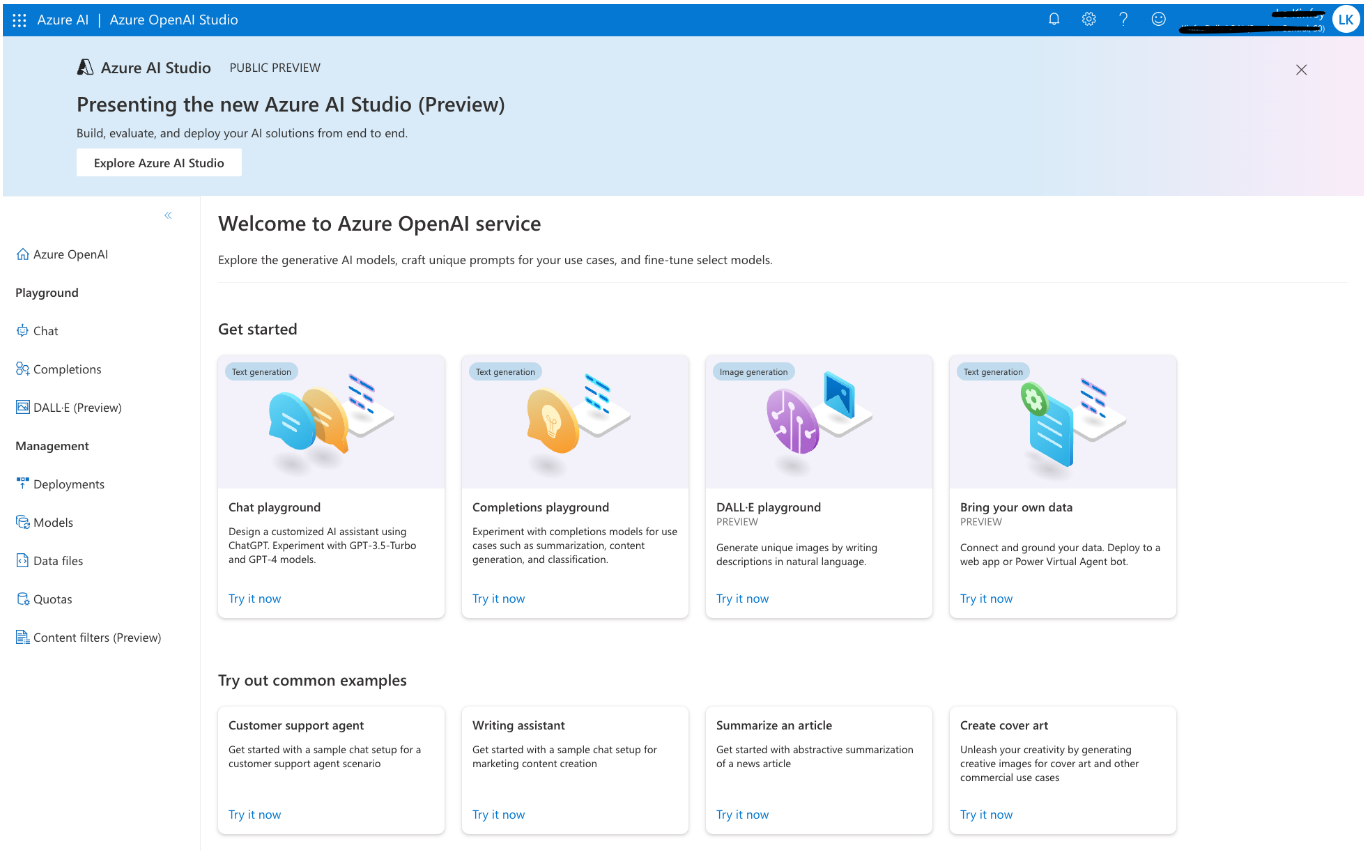
To use Azure OpenAI Service, you need to have an Azure account, then apply through this link, and wait 1-3 working days to use Azure OpenAI Service.

Azure OpenAI Studio

We can manage our models through Azure OpenAI Studio, as well as test our models in the Playground

Azure OpenAI Studio

We can manage our models through Azure OpenAI Studio, as well as test our models in the Playground



Azure AI | Azure OpenAI Studio [redacted] LK

Azure AI Studio PUBLIC PREVIEW

Presenting the new Azure AI Studio (Preview)

Build, evaluate, and deploy your AI solutions from end to end.

Explore Azure AI Studio

Azure OpenAI

Playground

- Chat
- Completions
- DALL·E (Preview)

Management

- Deployments
- Models
- Data files
- Quotas
- Content filters (Preview)

Welcome to Azure OpenAI service

Explore the generative AI models, craft unique prompts for your use cases, and fine-tune select models.

Get started

Text generation
Chat playground
Design a customized AI assistant using ChatGPT. Experiment with GPT-3.5-Turbo and GPT-4 models.
Try it now

Text generation
Completions playground
Experiment with completions models for use cases such as summarization, content generation, and classification.
Try it now

Image generation
DALL·E playground
PREVIEW
Generate unique images by writing descriptions in natural language.
Try it now

Text generation
Bring your own data
PREVIEW
Connect and ground your data. Deploy to a web app or Power Virtual Agent bot.
Try it now

Try out common examples

Customer support agent
Get started with a sample chat setup for a customer support agent scenario
Try it now

Writing assistant
Get started with a sample chat setup for marketing content creation
Try it now

Summarize an article
Get started with abstractive summarization of a news article
Try it now

Create cover art
Unleash your creativity by generating creative images for cover art and other commercial use cases
Try it now

Note: All examples are based on Azure OpenAI Services

Hugging Face

Hugging Face is an artificial intelligence research company focusing on natural language processing (NLP), known for its open source projects and innovations in the field of NLP. The company was founded in 2016 and its headquarters are in New York, but it has a global presence and activities.

Products

Transformers library: Hugging Face's most famous contribution is its development of the "Transformers" library, which is a widely used Python library that contains a variety of pre-trained NLP models, such as BERT, GPT, T5, etc. This library makes it easier to access and use these complex models, and has a significant impact on promoting research and applications in the NLP field.

Model Sharing and Community: Hugging Face has built a strong community that promotes model and knowledge sharing among researchers and developers. Through its platform, anyone can upload, share and use pre-trained models.

Research and Collaboration: Hugging Face conducts active research in the field of artificial intelligence, collaborating with numerous teams in academia and industry.

Education and Resources: Hugging Face also provides a variety of educational resources, including tutorials, documentation, and research papers, to help people better understand and use NLP technology.

Influence

Technological innovation: Hugging Face has played an important role in promoting technological innovation in the field of NLP, especially in the development and application of pre-trained models.

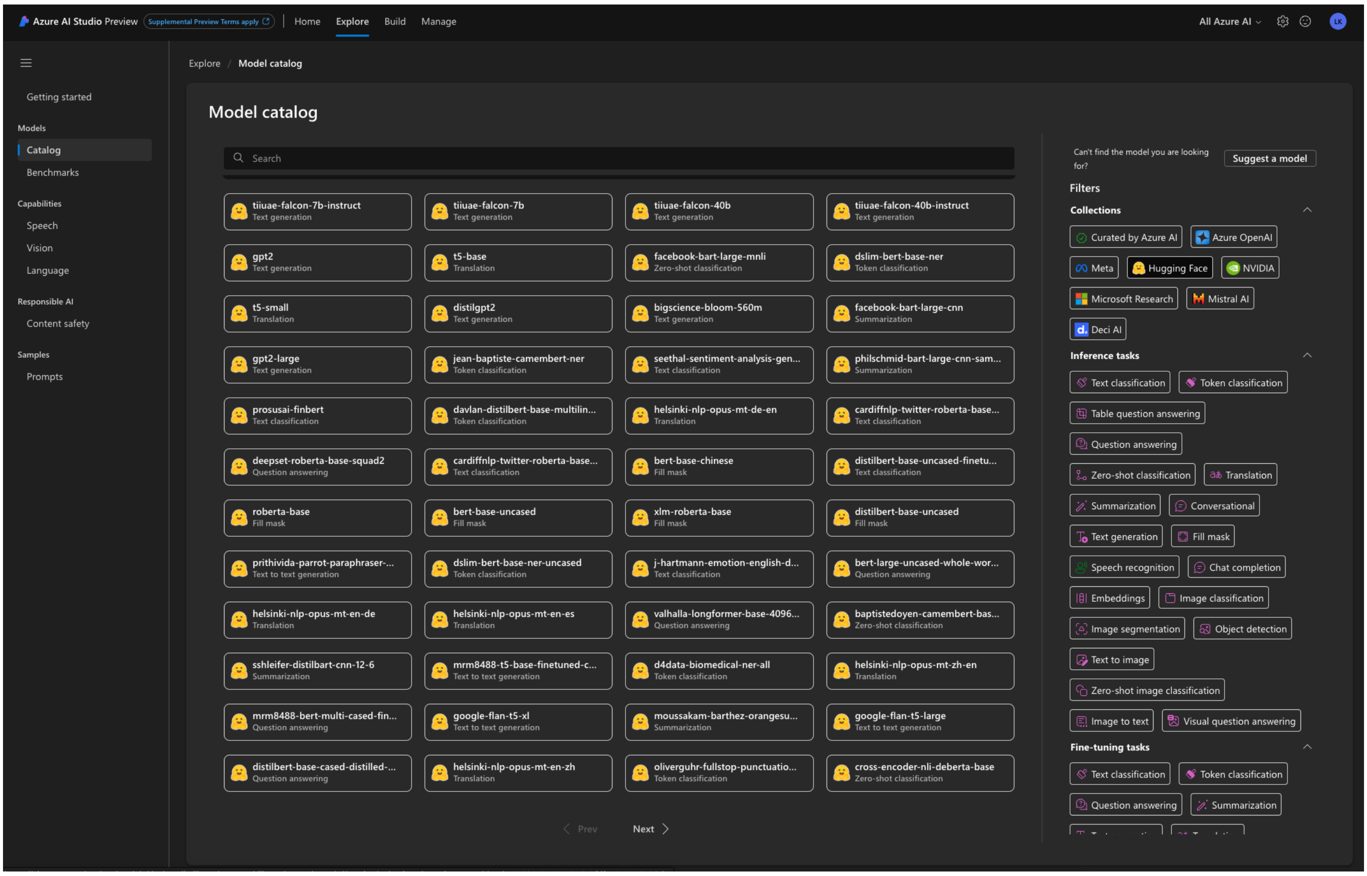
Lowering the technical threshold: By providing easy-to-use tools and resources, Hugging Face lowers the technical threshold for working in the field of NLP, allowing more researchers and developers to participate in this field.

Community building: Its strong community and open source culture promote knowledge sharing and collaboration, accelerating the development and innovation of NLP technology.

While Hugging Face originally started as a consumer-facing chatbot application, it quickly transformed into a company focused on providing NLP technology and resources. Now, it not only supports research and education, but also provides commercial solutions to enterprises, such as custom model training, data processing and machine learning consulting services.

In summary, Hugging Face is a key player in the NLP field, and its open source ethos and contributions to the community have played an important role in promoting the democratization and innovation of artificial intelligence technology.

Azure AI Studio also supports the introduction of the Hugging Face model, which allows enterprises to better combine business scenarios and use different models to solve problems in different application scenarios.



Model catalog

tiiuae-falcon-7b-instruct Text generation	tiiuae-falcon-7b Text generation	tiiuae-falcon-40b Text generation	tiiuae-falcon-40b-instruct Text generation
gpt2 Text generation	t5-base Translation	facebook-bart-large-mnli Zero-shot classification	dslim-bert-base-ner Token classification
t5-small Translation	distilgpt2 Text generation	bigscience-bloom-560m Text generation	facebook-bart-large-cnn Summarization
gpt2-large Text generation	jean-baptiste-camembert-ner Token classification	seethal-sentiment-analysis-gen... Text classification	philschmid-bart-large-cnn-sam... Summarization
prosusai-finbert Text classification	davlan-distilbert-base-multilin... Token classification	helsinki-nlp-opus-mt-de-en Translation	cardiffnlp-twitter-roberta-base... Text classification
deepset-roberta-base-squad2 Question answering	cardiffnlp-twitter-roberta-base... Text classification	bert-base-chinese Fill mask	distilbert-base-uncased-finetu... Text classification
roberta-base Fill mask	bert-base-uncased Fill mask	xlm-roberta-base Fill mask	distilbert-base-uncased Fill mask
prithivida-parrot-paraphraser-... Text to text generation	dslim-bert-base-ner-uncased Token classification	j-hartmann-emotion-english-d... Text classification	bert-large-uncased-whole-wor... Question answering
helsinki-nlp-opus-mt-en-de Translation	helsinki-nlp-opus-mt-en-es Translation	valhalla-longformer-base-4096... Question answering	baptistedoyen-camembert-bas... Zero-shot classification
sshleifer-distilbart-cnn-12-6 Summarization	mrm8488-t5-base-finetuned-c... Text to text generation	d4data-biomedical-ner-all Token classification	helsinki-nlp-opus-mt-zh-en Translation
mrm8488-bert-multi-cased-fin... Question answering	google-flan-t5-xl Text to text generation	moussakam-barthez-orangesu... Summarization	google-flan-t5-large Text to text generation
distilbert-base-cased-distilled-... Question answering	helsinki-nlp-opus-mt-en-zh Translation	oliverguhr-fullstop-punctuatio... Token classification	cross-encoder-nli-deberta-base Zero-shot classification

Summary

This chapter introduces the current knowledge related to LLM, especially the knowledge related to mainstream large-scale language model platforms such as OpenAI, Microsoft, and Hugging Face, as well as the application scenarios and performance of different models. For application scenarios, it is impossible for us to use only one model. In the AI 2.0 era, we need the support of different models to complete more intelligent application scenarios. Whether in the cloud or locally, the application scenarios of large language models will be a hot topic of concern in the next few years. As a beginner, what you need to do is to understand different models and complete application construction based on actual scenarios.