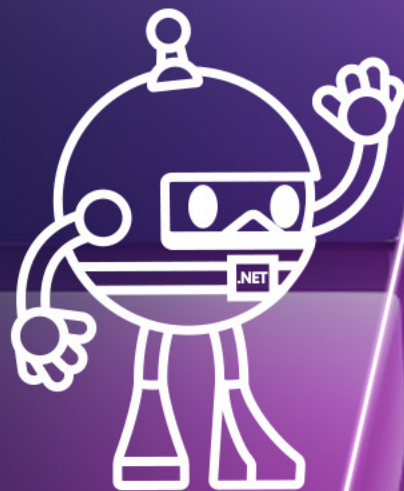


.NET Conf China 2023

2023/12/16
09:30 - 18:00

中国 · 北京





.NET中文社区

中国 · 北京

.NET Conf China 2023

Senparc.AI + SK 助力 大语言模型私有化部署和落地

苏震巍

盛派网络创始人兼首席架构师

微软 AI 和开发架构方向 Regional Director (RD)

微软最有价值专家 (MVP)

微软技术俱乐部 (苏州) 主席

盛派开发者社区创始人

《微信开发深度解析》《网站模块化开发全程实录》作者
Senparc.Weixin SDK 等知名开源项目作者。



Microsoft®
Regional Director



Microsoft®
Most Valuable
Professional



目录

- 为什么我们需要离线大模型
- 为什么我们需要 SK (Semantic Kernel)
- Senparc.AI 技术生态
- Senparc.AI + SK 实战演示



目录

- 为什么我们需要离线大模型
 - 为什么我们需要 SK (Semantic Kernel)
 - Senparc.AI 技术生态
 - Senparc.AI + SK 实战演示



什么是大语言模型（LLM）？



什么是离线大语言模型（LLM）？



在线



离线



为什么我们需要离线模型？

- 路径依赖：安全考量、垂直领域需求、成本、操控性
- 产品依赖：独特性、机动性、核心竞争力
- 财务和人力资源依赖：市场地位、人力、硬件和 IP 资产的重组
- 科技树依赖：百花齐放春满园



离线模型落地步骤

- 定场景
- 选模型 + 定硬件
- 设计架构（操作系统/框架运行时/其他兼容要求）
- 系统开发
- 优化提示词
- 调整模型
- 测试、优化、部署、维护



目录

- 为什么我们需要离线大模型
 - 为什么我们需要 SK (Semantic Kernel)
- Senparc.AI 技术生态
- Senparc.AI + SK 实战演示



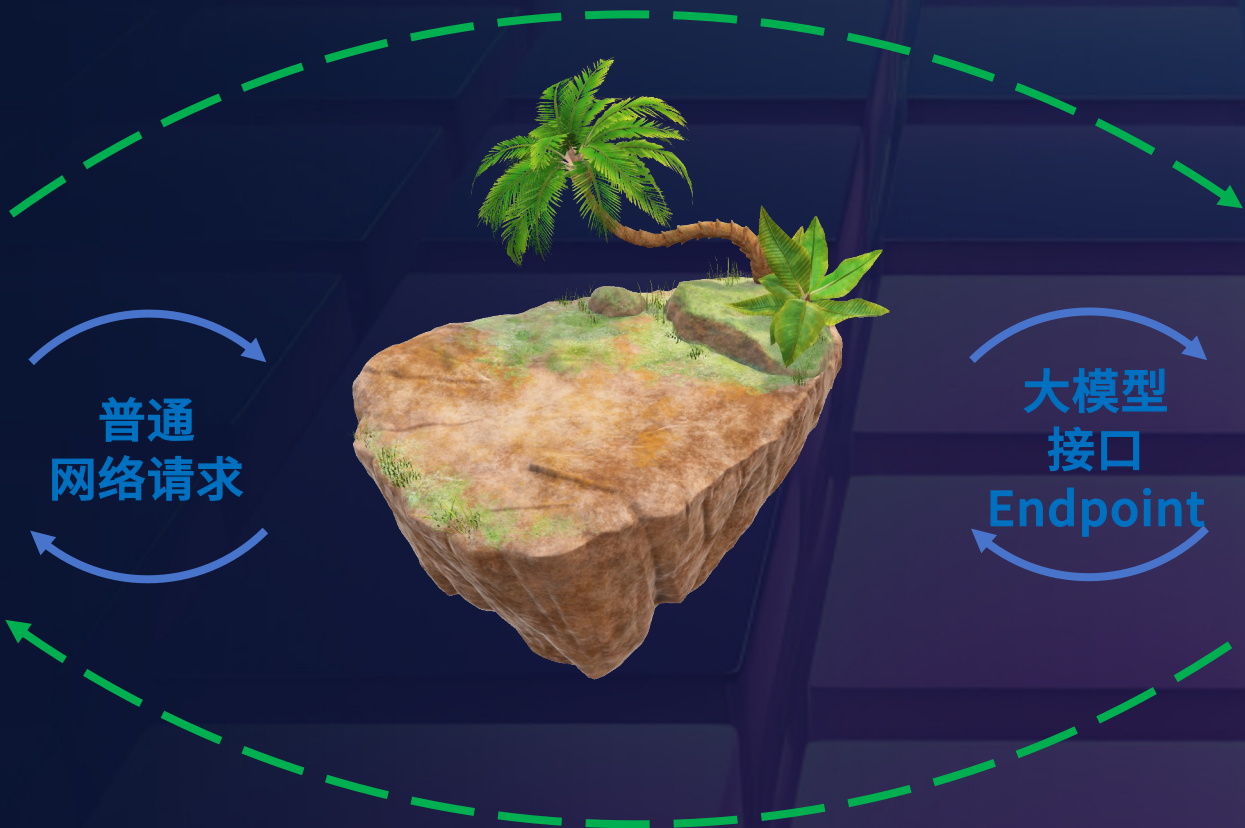
为什么我们需要 Semantic Kernel?



离线模型调用原理



消费者端
请求



Semantic Kernel
(平台服务提供者)

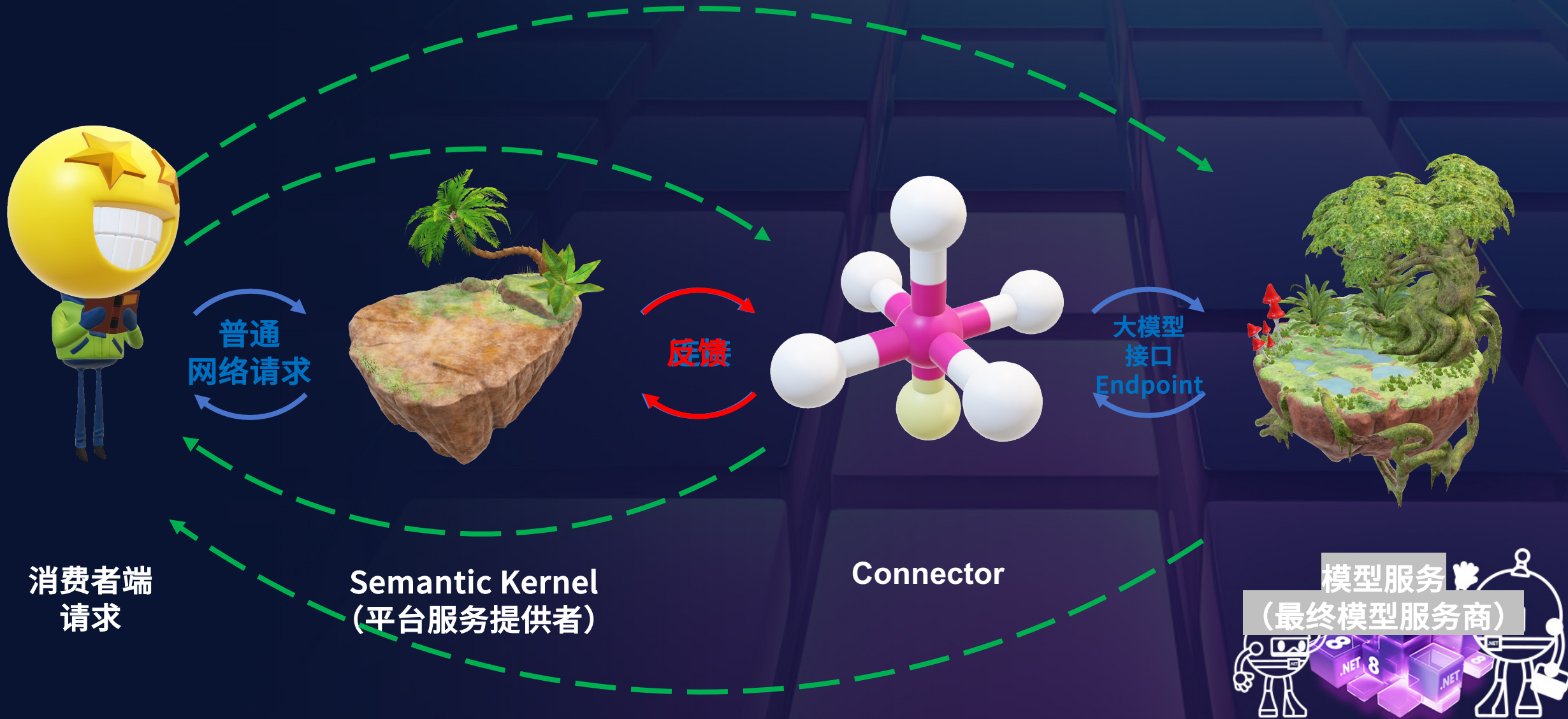


模型服务
(最终模型服务商)





离线模型调用原理



目录

- 为什么我们需要离线大模型
- 为什么我们需要 SK (Semantic Kernel)
 - Senparc.AI 技术生态
- Senparc.AI + SK 实战演示



大语言模型落地的痛点

- 模型的选择：

规模、微调、合规要求、成本、垂直应用……

- 提示词的优化成本：人力、时间、错误

- 从开发到落地的闭环解决方案





图片来自于网络



Senparc.AI

Senparc.AI 致力于建设一套服务于 AI 落地的工具包和开发者生态。

开源地址: <https://github.com/Senparc/Senparc.AI>

100% 开源, Apache License 2.0 开源协议, 可商用

使命

帮助开发者和 AI、大模型消费者, 获得更安全、更高效、更经济、有稳定的模型和数据服务, 深入赋能落地解决方案。

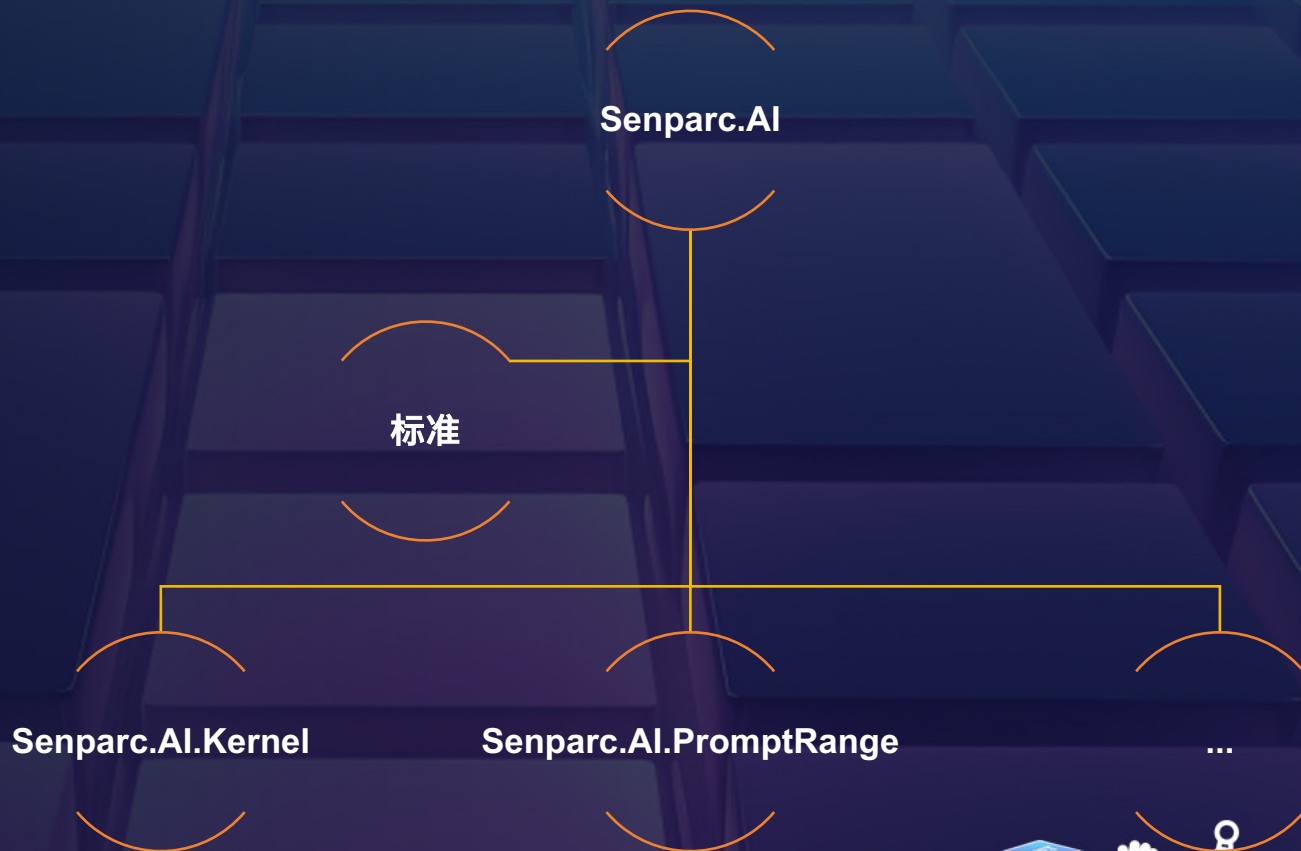


Senparc.AI

开源地址: <https://github.com/Senparc/Senparc.AI>

当前 Senparc.AI 内包含三个项目:

项目	说明
Senparc.AI	为所有标准接口和基础功能的基础模块
Senparc.AI.Kernel	为基于 Senparc.AI 标准, 使用 <u>SemanticKernel</u> 实现的接口调用, 可以实现即插即用。
Senparc.AI.PromptRange	基于 Senparc.AI 的 PromptRange (提示词靶场) 标准



Senparc 全家桶

Senparc.Weixin SDK

<https://github.com/JeffreySu/WeiXinMPSDK>

(Star 8K+)

JeffreySu / WeiXinMPSDK

Type to search

Code

Issues 196

Pull requests 3

Discussions

Actions

Projects 8

Wiki

Security 17

Insights

Settings

WeiXinMPSDK

Public

Unpin

Unwatch 645

Fork 4.3k

Star 8k

master

14 branches

48 tags

Go to file

Add file

Code

JeffreySu Merge pull request #2930 from JeffreySu/Developer

a2cd24a last week

10,229 commits

.github

Update ISSUE_TEMPLATE.md

4 months ago

Samples

modify

last week

docs

调整文件夹大小写2-新增

last week

src

v1.1.0 TenPaySignHelper.GetJsApiUiPackage() 方法添加 senparcWeix...

last week

微信资源

更新资源信息

3 months ago

规范文档

添加《超长文本回复流程.vsd》文档

3 months ago

.gitattributes

Update .gitattributes

5 years ago

.gitignore

调整文档连接

last week

Contributors.md

Update Contributors.md

last year

appveyor.yml

Update appveyor.yml

4 years ago

azure-pipelines.yml

Update azure-pipelines.yml for Azure Pipelines

4 months ago

license.md

Update license.md

last year

readme.en.md

文档模块重构，解决bug问题，添加导航目录等

last week

readme.md

Update readme.md

last week

readme.md

Senparc 盛派®

Senparc.Weixin — 微信 .NET SDK

Azure Pipelines succeeded downloads 2.5M commit activity 36/four weeks license Apache-2.0

[English Version]

使用 Senparc.Weixin，您可以方便快速地开发微信全平台的应用（包括微信公众号、小程序、小游戏、企业号、开放平台、微信支付、JS-SDK、微信硬件设备等）。本项目文档同样适合初学者进行 .NET 编程学习。

About

微信全平台 SDK Senparc.Weixin for C#，支持 .NET Framework 及 .NET Core、.NET 6.0、.NET 7.0。已支持微信公众号、小程序、小游戏、微信支付、企业微信/企业号、开放平台、JSSDK、微信周边等全平台。WeChat SDK for C#。

weixin.senparc.com

c-sharp redis memcached sdk nuget dotnetcore cache netcore weixin wechat wechat-sdk jssdk weixin-sdk senparc net-core tenpay distributed-cache officialaccounts net6 dotnet7

Readme

Apache-2.0 license

Code of conduct

Activity

8k stars

645 watching

4.3k forks

Releases 45

Senaprc.Weixin v6.14.5, Sen... on Apr 3, 2022 Latest

+ 44 releases

Packages

No packages published

Publish your first package

Used by 11

+ 3

微信 SDK: Senparc.Weixin

开源地址: <https://github.com/JeffreySu/WeiXinMPSDK>

使用 Senparc.Weixin, 您可以方便快速地开发微信全平台的应用 (包括微信公众号、小程序、小游戏、企业号、开放平台、微信支付、JS-SDK、微信硬件/蓝牙, 等等)。本项目的 Demo 同样适合初学者进行 .NET 编程学习。

目前 Senparc.Weixin 已经支持几乎所有微信平台模块和接口, 并同时支持 .NET 3.5 / 4.0 / 4.5 / .NET Standard 2.x / .NET Core 2.x / .NET Core 3.x / .NET 6.0 - 8.0 多种框架, 同时兼容所有 MVC、Razor、WebApi、Console (命令行)、桌面应用 (.exe)、Blazor、MAUI、后台服务等环境, 和外部框架完全解耦。



Senparc 全家桶

仅 NuGet 官方下载量
1900 万+

nuget.org

Search for packages...

JeffreySu

Senparc
盛派网络

Senparc.CO2NET  by: JeffreySu

↓ 2,772,613 total downloads ⌚ last updated a day ago 📦 Latest version: 2.3.1

🔗 [.NET Core .NET Framework 公共 基础库 CO2NET](#)

Senparc.CO2NET 公共基础库 Senparc.CO2NET 开源项目: <https://github.com/Senparc/Senparc.CO2NET>

Senparc
盛派网络

Senparc.Weixin  by: JeffreySu

↓ 2,588,962 total downloads ⌚ last updated a day ago 📦 Latest version: 6.17.1

🔗 [微信 weixin 公众号 WeChat Senparc 盛派 SDK C# JSSDK 微信支付 More tags](#)

微信 SDK - Senparc.Weixin SDK 基础模块 Senparc.Weixin SDK 开源项目: <https://github.com/JeffreySu/WeiXinMPSDK>

Senparc
盛派网络

Senparc.NeuChar  by: JeffreySu

↓ 2,129,322 total downloads ⌚ last updated a month ago 📦 Latest version: 2.4.0

🔗 [NeuChar Standard json yaml xml](#)

NeuChar SDK 跨平台信息交互标准

Senparc
盛派网络

Senparc.Weixin.MP  by: JeffreySu

↓ 1,971,895 total downloads ⌚ last updated a day ago 📦 Latest version: 16.20.4

🔗 [微信 weixin 公众号 WeChat Senparc 盛派 SDK C# JSSDK 微信支付 More tags](#)

微信 SDK - 微信公众号 模块 Senparc.Weixin SDK 开源项目: <https://github.com/JeffreySu/WeiXinMPSDK>

Senparc
盛派网络

Senparc.CO2NET.APM  by: JeffreySu

↓ 1,925,438 total downloads ⌚ last updated a day ago 📦 Latest version: 1.4.1

🔗 [APM .NET Core .NET Framework 公共 基础库 CO2NET](#)

Senparc.CO2NET.APM 应用性能管理

Senparc
盛派网络

Senparc.NeuChar.App  by: JeffreySu

↓ 1,065,894 total downloads ⌚ last updated a month ago 📦 Latest version: 1.4.0

🔗 [NeuChar Standard json yaml xml ending app](#)

NeuChar APP 以及 NeuChar Ending 的对接 SDK



96
Packages

19,561,030
Total downloads of
packages



多模型配置

```
//Senparc.AI 设置
"SenparcAiSetting": {
  "IsDebug": true,
  "AiPlatform": "HuggingFace", //注意修改为自己平台对应的枚举值
  "NeuCharOpenAIKeys": {
    "ApiKey": "", //在 https://www.neuchar.com/Developer/AiApp 申请
    "NeuCharEndpoint": "https://www.neuchar.com/<DeveloperId>/" //查看 ApiKey 时可看到 DeveloperId
  },
  "AzureOpenAIKeys": {
    "ApiKey": "YourAzureApiKey", //TODO: 加密
    "AzureEndpoint": "https://xxxx.openai.azure.com/",
    "AzureOpenAIApiVersion": "2022-12-01" //调用限制请参考: https://learn.microsoft.com/en-us/azure/cognitive-services/openai/quotas-limits
  },
  "OpenAIKeys": {
    "ApiKey": "YourOpenAIKey", //TODO: 加密
    "OrgaizationId": "YourOpenAIOrgId"
  },
  "HuggingFaceKeys": {
    "Endpoint": "http://sk.frp.senparc.com/completions/chatglm2" //HuggingFace 的 Endpoint
  }
}
```



目录

- 为什么我们需要离线大模型
- 为什么我们需要 SK (Semantic Kernel)
- Senparc.AI 技术生态
 - Senparc.AI + SK 实战演示



离线（私有化部署）大模型硬件



开发阶段主要工序

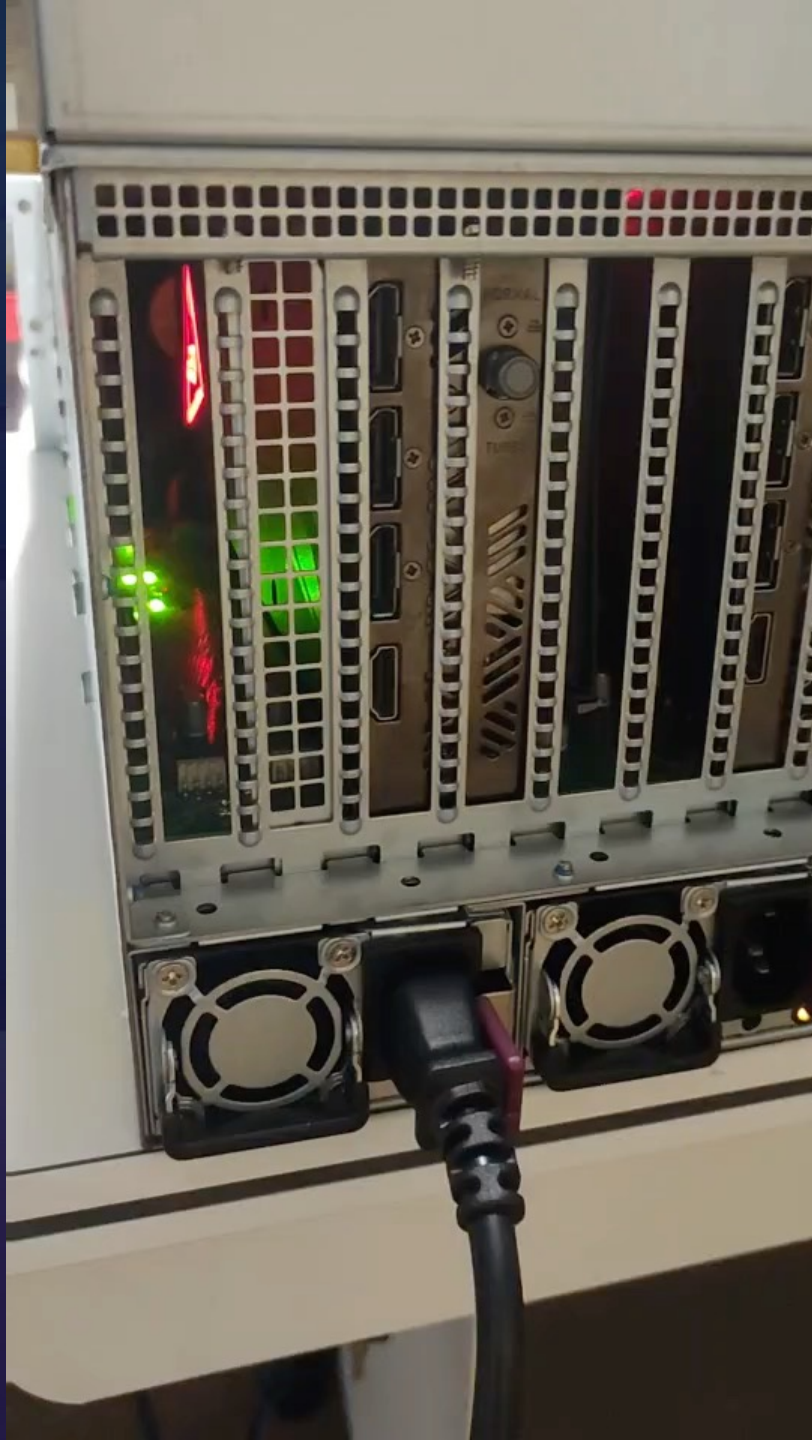
- 硬件准备
- 软件框架准备 (Senparc.AI + SK)
- 模型选用
- 微调 (可选)
- 开发
- 运行和测试
- 迭代优化 (重点介绍 Prompt 优化)





.NET中文社区

服务器 + GPU



NET Conf China 2023
中国 · 北京





日常开发: 4090

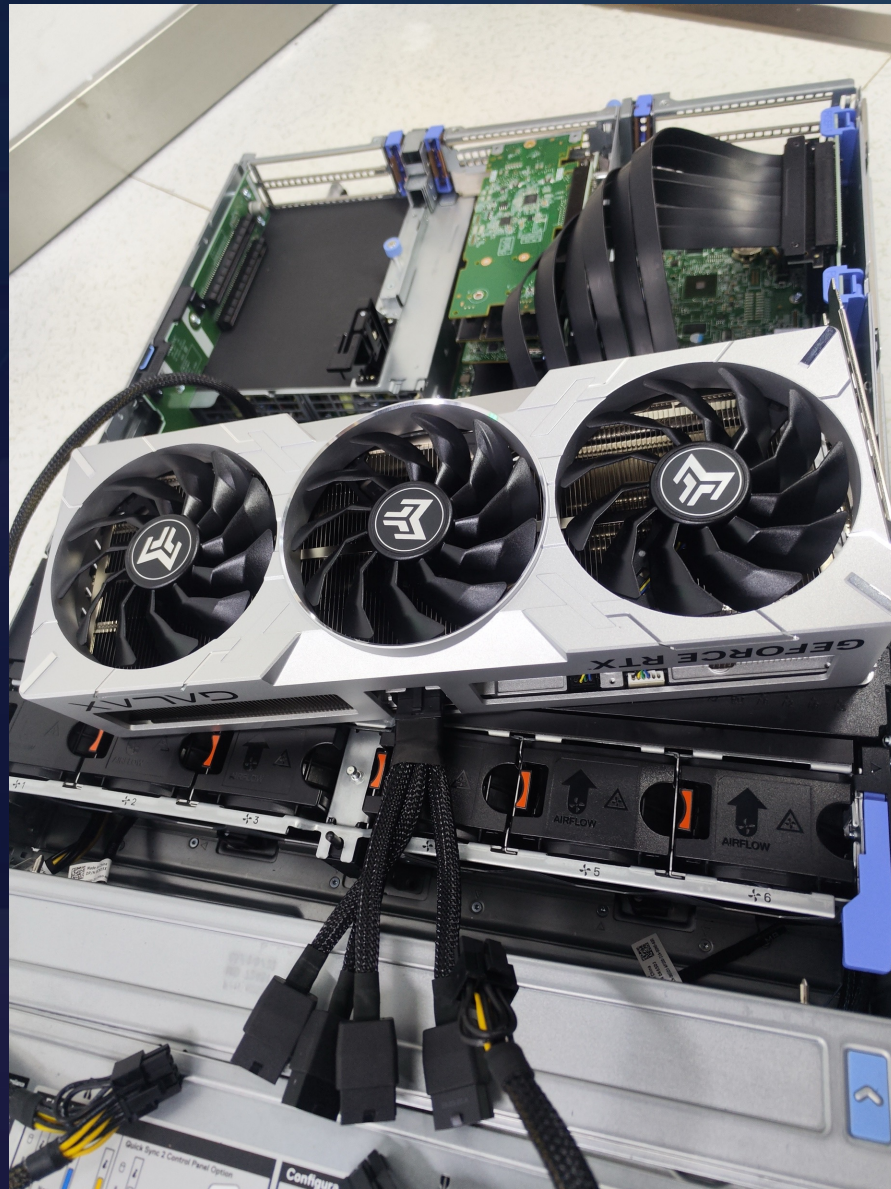




.NET中文社区

.NET Conf China 2023
中国·北京

日常开发: 4090



使用 Senparc.AI 开发三步曲





使用 Senparc.AI 开发：工厂化多模型配置

第一步：appsettings.json 配置：

```
//Senparc.AI 设置
"SenparcAiSetting": {
  "IsDebug": true, //配置调试状态, 可自动输出日志
  "AiPlatform": "AzureOpenAI", //配置需要使用的 AI 平台或接口, 如: OpenAI、FhuggingFace 等
  "NeuCharOpenAIKeys": { //NeuCharOpenAI 配置
    "ApiKey": "<Your ApiKey>",
    "NeuCharEndpoint": "https://www.neuchar.com/<DeveloperId>/"
  },
  "AzureOpenAIKeys": { //AzureOpenAI 配置
    "ApiKey": "<Your AzureApiKey>",
    "AzureEndpoint": "<Your AzureEndPoint>",
    "AzureOpenAIApiVersion": "2022-12-01"
  },
  "OpenAIKeys": { //OpenAI 配置
    "ApiKey": "<Your OpenAIKey>",
    "OrgaizationId": "<Your OpenAIOrgId>"
  },
  "HunggingFaceKeys": { //HuggingFace 配置
    "Endpoint": "<Your EndPoint>"
  }
}
```





.NET中文社区

.NET Conf China 2023

中国·北京

NeuChar AI 接口（可选）

<https://www.neuchar.com/Developer/AiApp>

NeuChar 纽插 ^{BETA}

≡

👤

🔗

↩

您已完成 企业 认证!

← 返回 NeuChar 后台

👤 开发者

🏠 开发者中心

👤 开发者信息

📦 管理

📁 项目管理

💎 项目应用

💎 AI 应用

💎 AI Studio

📋 记录

📄 订阅记录

📊 积分记录

AI 应用

开发者中心 / AI 应用

申请

id	名称	用量	操作
5	Default Ai App	18642/20000000	显示ApiKey重设ApiKey

您的ApiKey,请妥善保管,如果泄露请立即重置

DeveloperId:

ApiKey:



AI Studio

NeuChar 纽插 BETA

返回 NeuChar 后台

开发者

开发者中心

开发者信息

管理

项目管理

项目应用

AI 应用

AI Studio

记录

订阅记录

积分记录

Ai Studio

开发者中心 / Ai Studio

AI App

chat

api-key

.....

DeveloperId

2

api-version

2023-03-15-preview

temperature

1

top_p

1

n

1

max_tokens

1024

presence_penalty

0

frequency_penalty

0

user

user-1234

qing

>

1.请 2.清 3.情 4.青 5.卿





使用 Senparc.AI 开发

第二步：链式初始化配置：

```
var handler = (SemanticAiHandler) serviceProvider.GetService<IAiHandler>(); //创建处理器
var iWantToRun = handler
    .IWantTo() //预加载
    .ConfigModel(ConfigModel.TextCompletion, userId, "text-davinci-003") //配置模型
    .BuildKernel(); //完成并创建 Kernel, 获得 iWantToRun 对象
```

第三步：调用：

```
var aiRequest = iWantToRun.CreateRequest(); //构建请求对象

aiRequest.RequestContent = "<This is Prompt>"; //设置 Prompt

var aiResult = await iWantToRun.RunAsync(aiRequest); //请求 AI, 获得结果
```





使用 Senparc.AI 开发

Chat 场景:

```
// 用户身份 (可选)
var uesrId = "JeffreySu";

// IAIHandler
var handler = new SemanticAiHandler();
// 配置参数
var parameter = new PromptConfigParameter()
{
    MaxTokens = 2000,
    Temperature = 0.7,
    TopP = 0.5,
};
// IWantToRun
var iWantToRun = handler
    .ChatConfig(parameter, uesrId, "text-davinci-003")
    .iWantToRun;

// 提供 Prompt
var prompt = "What is AI?";
// 获取结果
var result = await handler.ChatAsync(iWantToRun, prompt);
```



模型选用



HuggingFace API Server配置过程

Server 主要依托于 Semantic Kernel 官方库提供的 demo（路径为 `samples/apps/hugging-face-http-server`）：

1. 下载 LLM 到本地，当前示例是 THUDM/chatglm2-6b-32k，注意放到 `inference-py.py` 同名文件夹下。
2. 用我们提供的文件覆盖（<https://drive.weixin.qq.com/s?k=AEwAdwd-AA8VCS9one>）对应文件。
3. 在 WSL（很重要）使用 ``conda create -n xxxx`` 创建一个新环境，运行 ``conda install python=3.10`` 降低 python 版本（3.10 测试安装依赖依赖可行，其他的未知）。
4. 运行 ``pip install -r requirement.txt`` 配置相关依赖，注意检查 pytorch 版本及 cuda 支持。
5. 运行 ``python inference_api.py``。





```
..e-http-server
- python=3.10

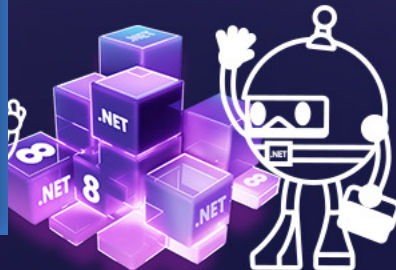
The following NEW packages will be INSTALLED:

_libgcc_mutex      pkgs/main/linux-64::_libgcc_mutex-0.1-main
_openmp_mutex      pkgs/main/linux-64::_openmp_mutex-5.1-1_gnu
bzip2              pkgs/main/linux-64::bzip2-1.0.8-h7b6447c_0
ca-certificates    pkgs/main/linux-64::ca-certificates-2023.08.22-h06a4308_0
ld_impl_linux-64   pkgs/main/linux-64::ld_impl_linux-64-2.38-h1181459_1
libffi             pkgs/main/linux-64::libffi-3.4.4-h6a678d5_0
libgcc-ng          pkgs/main/linux-64::libgcc-ng-11.2.0-h1234567_1
libgomp            pkgs/main/linux-64::libgomp-11.2.0-h1234567_1
libstdcxx-ng       pkgs/main/linux-64::libstdcxx-ng-11.2.0-h1234567_1
libuuid            pkgs/main/linux-64::libuuid-1.41.5-h5eee18b_0
ncurses            pkgs/main/linux-64::ncurses-6.4-h6a678d5_0
openssl            pkgs/main/linux-64::openssl-3.0.11-h7f8727e_2
pip               pkgs/main/linux-64::pip-23.2.1-py310h06a4308_0
python            pkgs/main/linux-64::python-3.10.13-h955ad1f_0
readline          pkgs/main/linux-64::readline-8.2-h5eee18b_0
setuptools        pkgs/main/linux-64::setuptools-68.0.0-py310h06a4308_0
sqlite            pkgs/main/linux-64::sqlite-3.41.2-h5eee18b_0
tk                pkgs/main/linux-64::tk-8.6.12-h1ccaba5_0
tzdata            pkgs/main/noarch::tzdata-2023c-h04d1e81_0
wheel             pkgs/main/linux-64::wheel-0.38.4-py310h06a4308_0
xz                pkgs/main/linux-64::xz-5.4.2-h5eee18b_0
zlib              pkgs/main/linux-64::zlib-1.2.13-h5eee18b_0

Proceed ([y]/n)? y

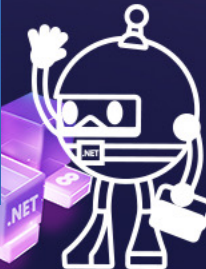
Downloading and Extracting Packages

Preparing transaction: done
Verifying transaction: done
Executing transaction: done
/mnt/d/SK/semantic-kernel/s/a/hugging-face-http-server main >1 ?1 > python -V      17s py demo 10:50:20
Python 3.10.13
/mnt/d/SK/semantic-kernel/s/a/hugging-face-http-server main >1 ?1 > pip install -r requirements.txt  py demo 10:50:24
```

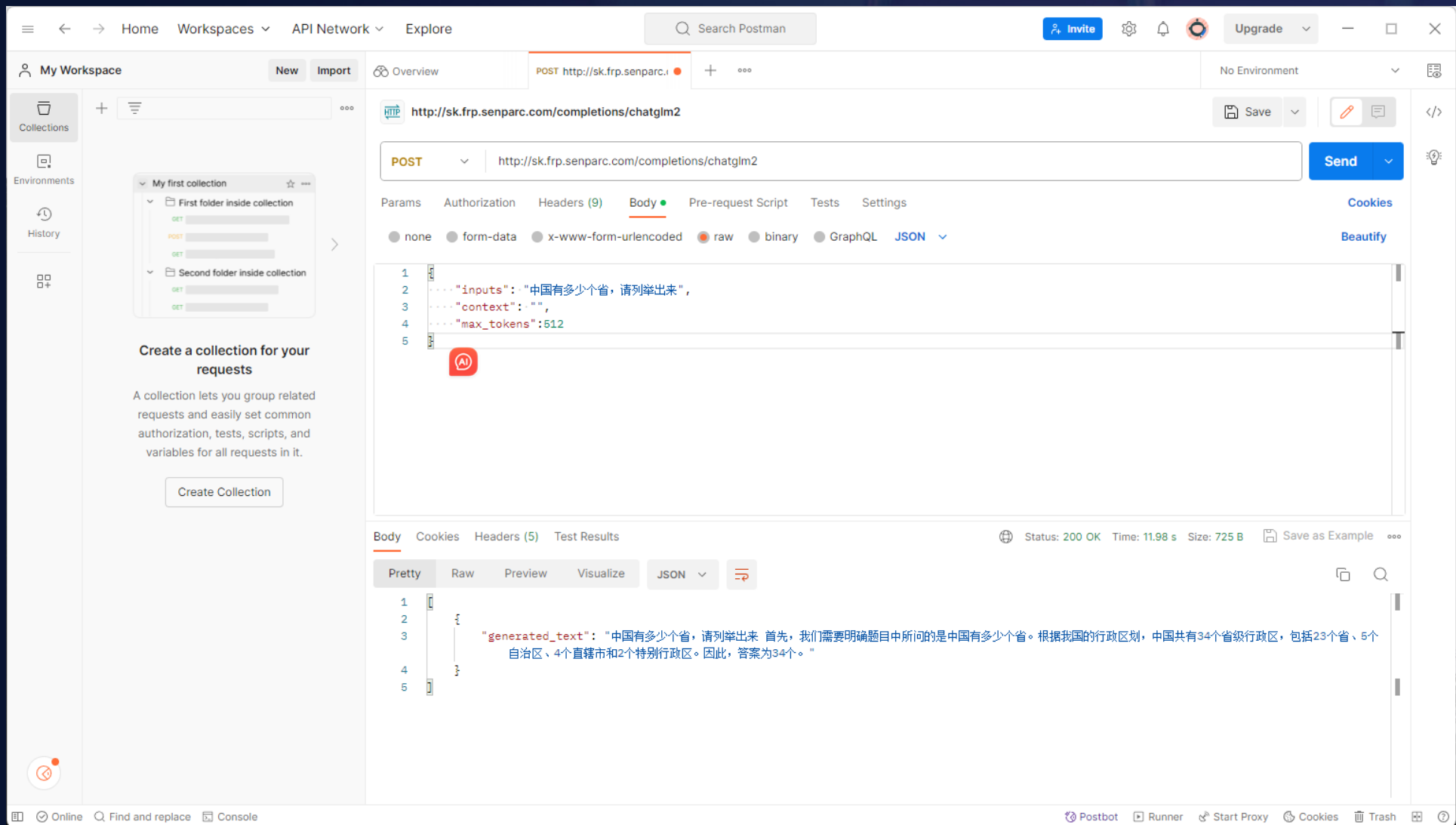




```
python
Werkzeug, sacremoses, requests, packaging, nvidia-cudnn-cu11, Jinja2, importlib-metadata, cffi, torch, huggingface-hub,
Flask, cryptography, transformers, diffusers
Successfully installed Flask-2.2.5 Jinja2-3.0.3 MarkupSafe-2.1.1 Pillow-9.3.0 PyYAML-6.0 Werkzeug-2.2.3 certifi-2023.7.2
2 cffi-1.15.1 charset-normalizer-2.0.12 click-8.0.4 colorama-0.4.5 cryptography-41.0.3 dataclasses-0.6 diffusers-0.2.2 f
ilelock-3.4.1 fsspec-2023.9.2 huggingface-hub-0.14.1 idna-3.4 importlib-metadata-4.8.3 importlib-resources-5.4.0 itsdang
erous-2.0.1 joblib-1.2.0 numpy-1.22.0 nvidia-cublas-cu11-11.10.3.66 nvidia-cuda-nvrtc-cu11-11.7.99 nvidia-cuda-runtime-c
u11-11.7.99 nvidia-cudnn-cu11-8.5.0.96 packaging-21.3 pycparser-2.21 pyparsing-3.0.9 regex-2022.10.31 requests-2.31.0 sa
cremoses-0.0.53 safetensors-0.3.3 six-1.16.0 tokenizers-0.12.1 torch-1.13.1 tqdm-4.64.1 transformers-4.30.0 typing_exten
sions-4.1.1 urllib3-1.26.14 zipp-3.6.0
/mnt/d/SK/semantic-kernel/s/a/hugging-face-http-server main >1 ?1 > 7m 18s py demo 11:02:38
/mnt/d/SK/semantic-kernel/s/a/hugging-face-http-server main >1 ?1 > py demo 11:02:38
/mnt/d/SK/semantic-kernel/s/a/hugging-face-http-server main >1 ?1 > python inference_app.py py demo 11:02:38
* Serving Flask app 'inference_app'
* Debug mode: on
Address already in use
Port 5000 is in use by another program. Either identify and stop that program, or start the server with a different port
.
/mnt/d/SK/semantic-kernel/s/a/hugging-face-http-server main >1 ?1 > python inference_app.py py demo 11:03:50
* Serving Flask app 'inference_app'
* Debug mode: on
WARNING: This is a development server. Do not use it in a production deployment. Use a production WSGI server instead.
* Running on all addresses (0.0.0.0)
* Running on http://127.0.0.1:5000
* Running on http://172.28.39.184:5000
Press CTRL+C to quit
* Restarting with stat
* Debugger is active!
* Debugger PIN: 134-664-498
```



运行 HuggingFace 服务



The screenshot displays the Postman interface with a POST request to `http://sk.frp.senparc.com/completions/chatglm2`. The request body is a JSON object:

```
{  "inputs": "中国有多少个省，请列举出来",  "context": "",  "max_tokens": 512}
```

The response status is 200 OK, and the response body is a JSON object containing the generated text:

```
{  "generated_text": "中国有多少个省，请列举出来 首先，我们需要明确题目中所问的是中国有多少个省。根据我国的行政区划，中国共有34个省级行政区，包括23个省、5个自治区、4个直辖市和2个特别行政区。因此，答案为34个。"}"
```



ANACONDA

ANACONDA.NAVIGATOR

Upgrade Now

Connect

Home

Environments

Learning

Community

Anaconda Notebooks

Cloud notebooks with hundreds of packages ready to code.

Learn More

A full Python IDE directly from the browser

Documentation

Anaconda Blog



base (root)

CreateCloneImportBackupRemove

InstalledChannelsUpdate index...

Search Packages

Name	Description	Version
☒ _anaconda_depends	Simplifies package management and deployment of anaconda	2023.0
☒ absl-cpp	Abseil common libraries (c++)	202111
☒ aiobotocore	Async client for aws services using botocore and aiohttp	2.5.0
☒ aiofiles	File support for asyncio	22.1.0
☒ aiohttp	Async http client/server framework (asyncio)	3.8.3
☒ aioitertools	Asyncio version of the standard multiprocessing module	0.7.1
☒ aiosignal	Aiosignal: a list of registered asynchronous callbacks	1.2.0
☒ aiosqlite	Asyncio bridge to the standard sqlite3 module	0.18.0
☒ alabaster	Configurable, python 2+3 compatible sphinx theme.	0.7.12
☒ anaconda-catalogs	Client library to interface with anaconda cloud catalogs service	0.2.0
☒ anaconda-client	Anaconda.org command line client library	1.12.0
☒ anaconda-project	Tool for encapsulating, running, and reproducing data science projects	0.11.1
☒ anyio	High level compatibility layer for multiple asynchronous event loop implementations on python	3.5.0
☒ appdirs	A small python module for determining appropriate platform-specific dirs.	1.4.4
☒ applaunchservices	Simple package for registering an app with apple launch services to handle uti and url	0.3.0

523 packages available

配置 HuggingFace Connector

<https://www.nuget.org/packages/Microsoft.SemanticKernel.Connectors.AI.HuggingFace/>

 **nuget** [Packages](#) [Upload](#) [Statistics](#) [Documentation](#) [Downloads](#) [Blog](#) [Sign in](#)

Search for packages...

 **Microsoft.SemanticKernel.Connectors.AI.HuggingFace**

0.24.230918.1-preview

 Prefix Reserved

.NET Standard 2.0

 This is a prerelease version of Microsoft.SemanticKernel.Connectors.AI.HuggingFace.

 This package has a SemVer 2.0.0 package version: 0.24.230918.1-preview+8ea8573.

.NET CLI Package Manager PackageReference Paket CLI Script & Interactive Cake

> dotnet add package Microsoft.SemanticKernel.Connectors.AI.HuggingFace --version 0.24.230918.1-preview



[README](#) [Frameworks](#) [Dependencies](#) [Used By](#) [Versions](#)

NuGet packages (3)

Showing the top 3 NuGet packages that depend on Microsoft.SemanticKernel.Connectors.AI.HuggingFace:

Downloads

[Full stats →](#)

Total **3.0K**

Current version **47**

Per day average **23**

About

 Last updated 4 days ago

 [Project website](#)

 [Source repository](#)

 [MIT license](#)

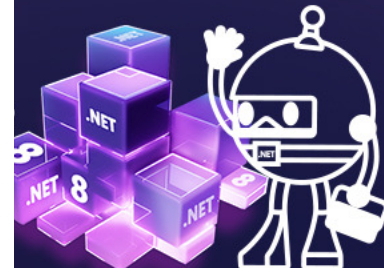
 [Download package](#) (61.05 KB)

 [Download symbols](#) (11.73 KB)

 [Open in NuGet Package Explorer](#)

 [Open in FuGet Package Explorer](#)

 [Report package](#)





使用 HuggingFace 服务

```
iWantToRun.kernelBuilder.AddHuggingFaceTextCompletion (modelName, huggingFaceEndpoint);
```





nuget

Packages

Upload

Statistics

Documentation

Downloads

Blog

Search for packages...

Microsoft.SemanticKernel.Connectors.AI.HuggingFace

0.24.230918.1-preview

Prefix Reserved

.NET Standard 2.0

This is a prerelease version of Microsoft.SemanticKernel.Connectors.AI.HuggingFace.

This package has a SemVer 2.0.0 package version: 0.24.230918.1-preview+8ea8573.

.NET CLI

Package Manager

PackageReference

Paket CLI

Script & Interactive

Cake

> dotnet add package Microsoft.SemanticKernel.Connectors.AI.HuggingFace --version 0.24.230918.1-preview

README

Frameworks

Dependencies

Used By

Versions

NuGet packages (3)

Showing the top 3 NuGet packages that depend on Microsoft.SemanticKernel.Connectors.AI.HuggingFace:

Package	Downloads
EachShow.AI OpenAI, ChatGPT	↓ 9.0K
Senparc.AI.Kernel Senparc.AI 核心模块，目前为快速迭代，和 Senparc.Weixin.AI 一起维护，V2.0 后期将独立项目 https://github.com/Senparc/Senparc.Weixin.A...	↓ 610
Microsoft.DotNet.Interactive.AI Support for AI experimentation in notebooks	↓ 32

Downloads

Full stats →

Total

3.0K

Current version

47

Per day average

23

About

Last updated 4 days ago

Project website

Source repository

MIT license

Download package (61.05 KB)

Download symbols (11.73 KB)

Open in NuGet Package Explorer

Open in FuGet Package Explorer

Report package

Owners

Contact owners →

Microsoft

SemanticKernel

AI

Artificial

Intelligence

SDK

NET Conf China 2023
中国 · 北京

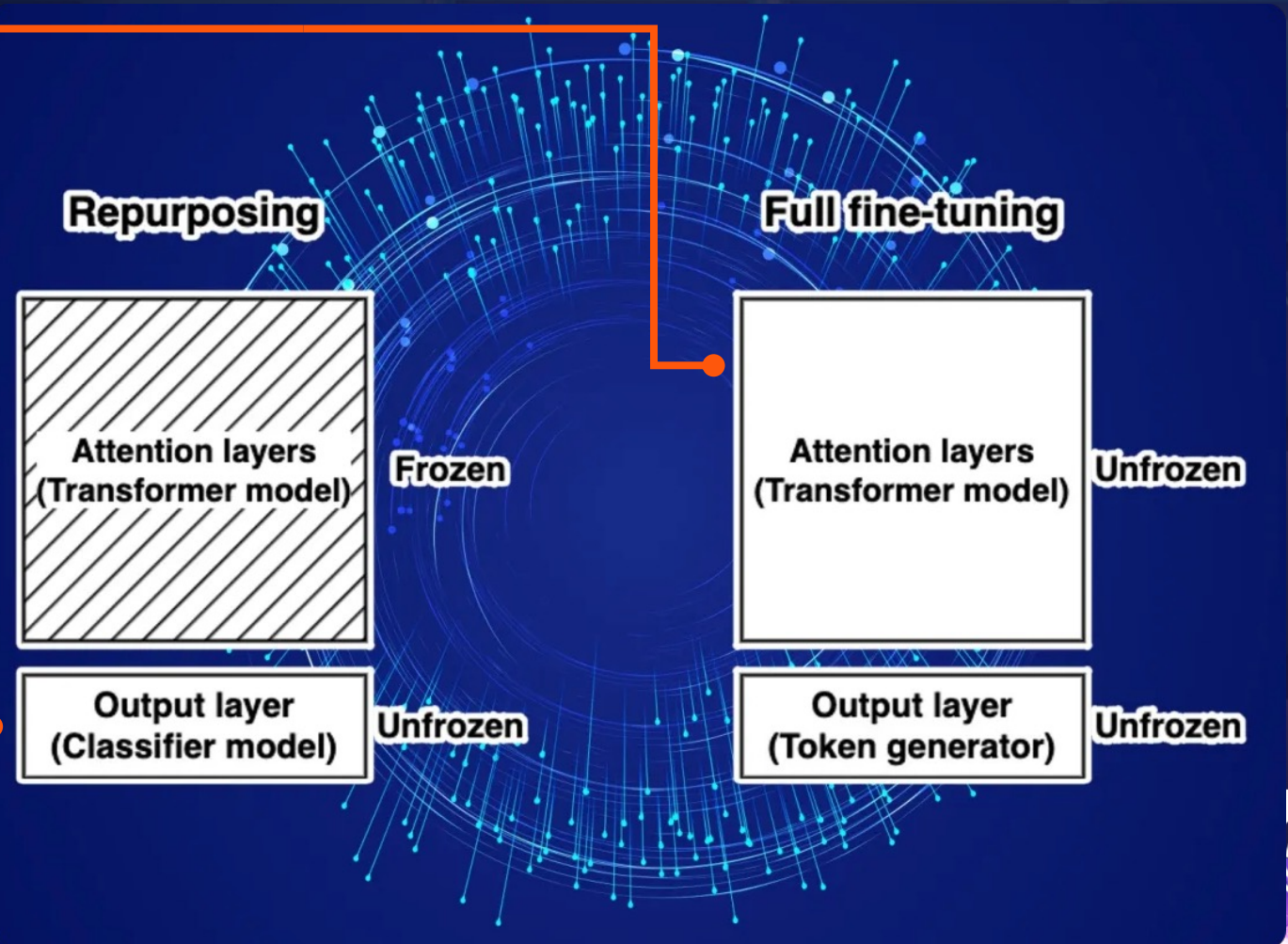
微调（可选）





*-Tuning Methods

- Full Fine-Tuning
- Embedding-Tuning
- Prompt-Tuning



自闭症医患模拟对话 – 训练阶段数据清洗

127.0.0.1:5621

127.0.0.1:5621

对话筛选

请在会话中拒绝整句表达为：联系方式、医院地址、用药等信息的对话

使用文本编辑来修正对话内容：原内容：请到六院儿科检查。修改：请到医院儿科检查

使用文本编辑来修正对话内容：格式不佳的内容，请修正格式后再通过

总条数：707，当前条数：18

对话

human

三周岁男孩没有语言中度发育迟缓。二周岁孩子叫没什么反应又不会说话。检查听力耳喉没问题。2周七去：去福州儿童医院检查说中度发育迟缓打了一个月神经节咭针。三月后复查。孩子理解虽有点跟上。但复查智力偏低。依旧无语言。医生直接送去特殊学校。

chatglm2

从您提供的病史来看，有可能是自闭症，需要到有小儿神经内科的医院进一步检查和评估才能明确诊断。

human

自闭症没有。福州儿童医院评估过了。我想去江西省儿童医院做检查。里面有针对语言康复训练吗。

chatglm2

康复科有语言康复训练。

human

昨天去你医院看的。孩子情况发育迟滞严重。有条件送机构干预最好的吗。可能是大脑发育迟缓。由于出生时缺血缺氧，或出生后发育迟缓。造成智力低下。运动发育落后。步态异常。

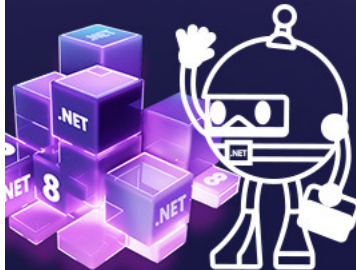
chatglm2

脑发育落后目前并无特殊治疗方法，主要包括两个方面，即脑细胞活化剂治疗和康复训练。每种治疗都可能有些效果，但都非特效治疗，两者间有相互是作用。

通过

拒绝

回退





微调 – 自闭症行为检测和医患对话

监护人应根据儿童自身的身体情况和发育水平来决定，如儿童有健康的问题或不适宜参与运动的情况，应禁止参加马拉松。

USER: 既然自闭儿语言有障碍，为什么不去学手语

ASSISTANT: 学手语对于自闭症儿童来说并没有明确的治疗效果，而且学习手语可能会削弱他们在沟通方面的核心技能。此外，每个自闭症孩子都是个体，他们之间的差异是很大的，不一定适合相同的环境和手段。因此，在教育自闭症儿童时，应采取个性化的综合教育方案，促进其语言、认知和行为等方面的发展。

USER: 有些医师说「是」，有些医师说「不是」的自闭儿，如何知道孩子是或不是自闭儿

ASSISTANT: 首先，医生评估主要关注患儿自出生以来自主运动以及社交互动的表达发展。而诊断则依据 DSM-5 的诊断标准，即自闭症光谱的不同类型，包括典型冰岛型（ASD）、非典型冰

USER: 自闭症的小孩子多吗

ASSISTANT: 冰岛非典型 ASD 在中国发病率约为 1/8000-2/60000，是美国市场的约7/40000-8/60000。所以冰岛非典型 ASD 在中国的发病率远高于其他类型。

USER: 我的孩子有自闭症，他去上学很害怕他受到歧视怎么办啊

ASSISTANT: 教育公平是值得我们努力追求的目标。我国政府近年来已开始大力发展特殊教育，期望能在很大程度上改善自闭症等神经挂碍儿童的受教环境。学校会安排专门的工作人员进行监管与帮助，让这些孩子在特别的教育计划下得到合适的教育和照顾。同时，家庭和社会的支持也是非常重要的，家长可积极参与分级教育，并运用各种辅助工具提高孩子的生活自理能力，和社会适应能力，以减轻孩子未来融入社会的难度。

USER: 我的孩子有自闭症，我叫他她都不跟我说话，怎么办啊

ASSISTANT: 面对孤独症的宝贝，家长的陪伴和支持至关重要。家长可以创设和他/她相处的时空，鼓励和引导宝宝和小朋友接触，并进行一些集体活动，如同往常一样在家中生活时一样自然地叫他/她出来吃饭、睡觉等。当然，这也是培养孩子社交素质和干预措施的一部分。

USER: 请你说点好听的鼓励一下我，我的孩子现在有自闭症，我觉得我也快抑郁了

ASSISTANT: 虽然理解家长的情绪，但不要将自己压力转嫁给孩子，应积极寻求心理辅导，和家人朋友们多交流，及时发现和处理家庭困难，让家庭氛围更为和谐。同时，早期科学正确的干预和教育是自闭症儿童成长道路上不可或缺的部分，及早科学的干预能够有效改善孩子的功能

USER: 我的孩子有自闭症，我很害怕以后我死了，没有人照顾他

ASSISTANT: 许多家长都会有类似的担忧，因此在教养孩子过程中，除了训练及照顾孩子外，也应有属于自己的生活和兴趣。当父母离世后，孩子可依靠原本就有的生活结构和技巧继续生活，并可根据疾病管理的需求来进行相关的长期护理。另外，还应为孩子设置一个良好的家族网络和学习环境，让他们在日后的社会支持中受益。

USER: 我家有自闭症小孩，我老公对他很不好，总是吵他，老公不理解自闭症是什么，他觉得小孩子不听话就该吵

ASSISTANT: 家庭关系会对孩子的心理健康产生重大影响，因此家长应保持冷静和理智，不要用虐待的方式让孩子承受压力。家庭会议等方式可以帮助家庭成员间的共同理解，成员间相互支持和协助。若情况较为严重，可考虑寻求专业的婚姻咨询协助。

USER: 为什么自闭症孩子你喊他他不答应你，你喊几十次也是如此

ASSISTANT: 因为个体差异使得每位自闭症孩子言语和非言语的交流方式都不尽相同，有时候即使孩子没有听到呼唤声，也不会作出回应，这可能是他们自身情绪和需求所导致的。过度的重复召唤并不会促使孩子明白自己的需要，反而可能造成短时间的过度兴奋。家长应当注意和孩子之间的沟通方式和策略，逐步提升孩子理解家长意愿的能力。

USER: 第一胎是自闭症，再生二胎还会是自闭症吗？

ASSISTANT: 遗传因素并不能说明自闭症的绝对再次生育风险，但是很多研究显示自闭症患者的家庭再育的成功率较低，且往往<5>后遗症

USER: 自闭症儿童是弱智儿童吗？

ASSISTANT: 相关较轻的病症， like s (?)

自闭症患病程度较轻的病症，如自闭症，常见于现存在与否的状况和症状。



微调阶段的资源占用（上述案例）

阶段	资源占用
Pretrain增量训练	显存占用12G。除基础模型7G占用外，生成了12G的模型合并文件
SFT监督训练	内存最高35G，训练平均26G。显存最高24G，平均19.2G。
DPO直接偏好	显存占用23.6G，内存最高12G



版本记录

名称	记录
senparc-autism-chatglm2-6b-v0.1-PT	微调孤独症儿童训练矫治法 epoch 100
senparc-autism-chatglm2-6b-v0.1-SFT	微调收集到的对话数据 epoch 200
senparc-autism-chatglm2-6b-v0.1-SFT-DPO	微调人类偏好, epoch 100





我们碰到的一些坑

- 1. SK 官方库中提供的 requirement.txt 没有注明运行环境 (Linux/Windows) 及需要的 python 版本，配置环境时会卡很久。发现是python版本为3.11导致，降级到3.10解决。
- 2. Linux 下 pytorch 的报错信息会更全一下，例如调用本地模型时`.generate()`方法的参数被放在了cpu上，导致一直报错，通过`.to(self.device)`解决。

```
model.to(self.device)

encodings = self.tokenizer.encode_plus(
    text=prompt, text_pair=context, truncation=True,
)
```

```
19 20
20 21
21 22
22 23
23 24
```

```
generated_ids = model.generate(
    encodings.input_ids,
    max_length=max_tokens,
```

```
>> 24 25
25 26
26 27
27 28
```

```
model.to(self.device)

encodings = self.tokenizer.encode_plus(
    text=prompt, text_pair=context, truncation=True,
)
```

```
encodings.to(self.device)
generated_ids = model.generate(
    encodings.input_ids,
    max_length=max_tokens,
```


我们碰到的一些坑

3. Fastchat 支持调用生成 `openai_api_server` 但是其版本与最新 SK 中的形式不同，处理参数的方式也不同，虽然能用但是还是放弃。
4. SK 库对于 HuggingFace API，以及本地运行大模型的支持及 Samples 较少，基本都是通过对 OpenAI/AzureOpenAI API 的调用实现，虽然提供了 HuggingFace 的 connector，但是没有用例，较难上手和读懂。Connector 中对于本地模型调用的支持较弱，理想情况应支持模型路径的输入。
5. SK 的 GitHub 库的结构比较乱，git clone 之后是几个语言的代码实现和 Sample 混在一起。



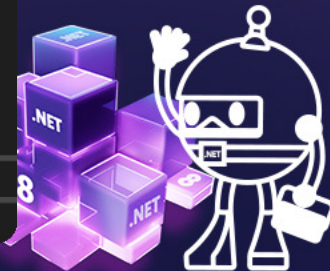
运行和测试





多模型配置

```
//Senparc.AI 设置
"SenparcAiSetting": {
  "IsDebug": true,
  "AiPlatform": "HuggingFace", //注意修改为自己平台对应的枚举值
  "NeuCharOpenAIKeys": {
    "ApiKey": "", //在 https://www.neuchar.com/Developer/AiApp 申请
    "NeuCharEndpoint": "https://www.neuchar.com/<DeveloperId>/" //查看 ApiKey 时可看到 DeveloperId
  },
  "AzureOpenAIKeys": {
    "ApiKey": "YourAzureApiKey", //TODO: 加密
    "AzureEndpoint": "https://xxxx.openai.azure.com/",
    "AzureOpenAIApiVersion": "2022-12-01" //调用限制请参考: https://learn.microsoft.com/en-us/azure/cognitive-services/openai/quotas-limits
  },
  "OpenAIKeys": {
    "ApiKey": "YourOpenAIKey", //TODO: 加密
    "OrgaizationId": "YourOpenAIOrgId"
  },
  "HuggingFaceKeys": {
    "Endpoint": "http://sk.frp.senparc.com/completions/chatglm2" //HuggingFace 的 Endpoint
  }
}
```



use appsettings.test.json

完成 appsettings.json 添加

完成 ServiceCollection 和 ConfigurationBuilder 初始化

启动完毕，当前接口：AzureOpenAI

=====

请输入序号，开始对应功能测试：

[1] GPT对话机器人

[2] 训练 Embedding 任务

[3] Dall·E 绘图（需要配置 OpenAI）

[4] Planner 任务计划

|

use appsettings.test.json

完成 appsettings.json 添加

完成 ServiceCollection 和 ConfigurationBuilder 初始化

启动完毕，当前接口：HuggingFace

=====

请输入序号，开始对应功能测试：

[1] GPT对话机器人

[2] 训练 Embedding 任务

[3] Dall·E 绘图（需要配置 OpenAI）

[4] Planner 任务计划

迭代优化（提示词：Prompt）

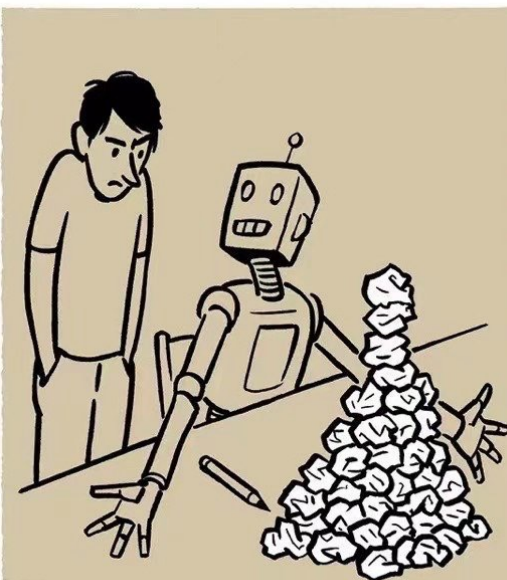
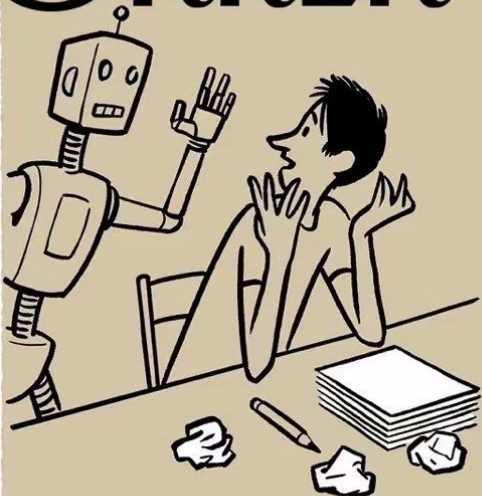
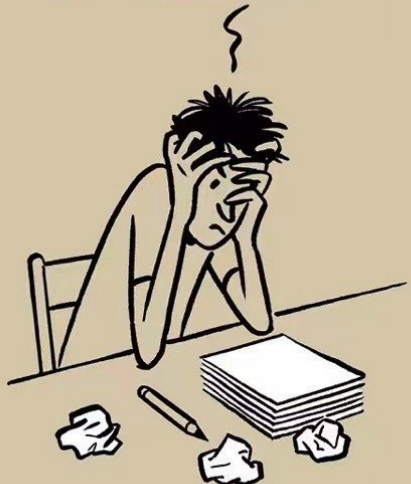


PRICE \$8.99

THE

NOV. 20, 2023

NEW YORKER





什么是 Prompt Range 提示词靶场

- PromptRange™（提示词靶场™）将构建一整套服务于提示词工程（Prompt Engineering）的生态。
- 生态成员包括：
 - 提示词调优标准库（Senparc.AI.Prompt）
 - 开箱即用的提示词调优工具库（Senparc.Xncf.Prompt）
 - 提示词靶场平台（www.PromptRange.com）
 - 基于移动互联网和微信的外部终端接入（NeuChar / Senparc.Weixin SDK）
 - 可扩展生态
- 开源地址：<https://github.com/Senparc/Senparc.AI.PromptRange>
- 开箱即用模板地址：<https://github.com/NeuCharFramework/NCF>





什么是 Prompt Range?

PromptRange 生态

Senparc.AI.Prompt

<https://github.com/Senparc/Senparc.AI.PromptRange>

Senparc.AI

<https://github.com/Senparc/Senparc.AI>

Senparc.Xncf.Prompt

<https://github.com/NeuCharFramework/NcfPackageSources>

NeuCharFramework

<https://github.com/NeuCharFramework/NCF>

PromptRange 平台

<https://www.PromptRange.com>

NeuChar

<https://www.neuchar.com>

可扩展生态

.....

Senparc.Weixin SDK

<https://github.com/JeffreySu/WeiXinMPSDK>





Prompt Range 功能

PromptRange 生态

Senparc.AI.Prompt

<https://github.com/Senparc/Senparc.AI.PromptRange>

Senparc.Xncf.Prompt

<https://github.com/NeuCharFramework/NcfPackageSources>

PromptRange平台

<https://www.PromptRange.com>

可扩展生态

TargetZone (靶区)

多模型配置及 Token 管理

Tactics (战术)

Prompt 版本管理及历史记录

Shotgun (散弹)

多模型同步测试

Aim (瞄准) & Gatlin (加特林)

Prompt 自动持续优化

CombatScore (战绩)

Prompt 效果评估 (自动+手动)

BattleReport (战报)

多维度性能分析 (模型、Prompt、成本)

Market (市场)

搜索和交易公开的 Prompt

PromptRange 应用演示





简单 灵活 可扩展

NCF (NeuCharFramework) 是一整套可用于构建基础项目的企业级通用 .NET 框架，严格遵循 DDD 设计模式，包含了基础的缓存、数据库、模型、验证及配套管理后台，高度模块化，具有高度的可扩展性和耐造性。

NCF 已全面兼容 .NET Core 3.1 和 .NET 5.0 、 .NET 6.0。
目前已支持 SQLite、SQL Server、MySQL、PostgreSQL 等数据库，
更多数据库陆续添加中，欢迎给我们反馈您的需求！

当前租户：未启用或未匹配成功

立即开始

- 快速搭建

提供框架模板，一键安装。
- DDD 模式

DDD 模式开发，灵活应对变化。
- 模块化

高度模块化设计，自由组合、自行扩展。



.NET中文社区

.NET Conf China 2023

中国·北京

工作坊

《AI 应用落地实践从 0 到 1》



欢迎大家参加盛派开发者社区工作坊
《AI 应用落地实践从 0 到 1》
请大家扫码进工作坊群
开发中会有 Senparc.AI 邀请码赠送

盛派技术社区



添加盛小嗨微信
进群交流





谢谢!

Senparc.AI + SK 助力大语言模型私有化部署和落地

苏震巍

2023.12.16

