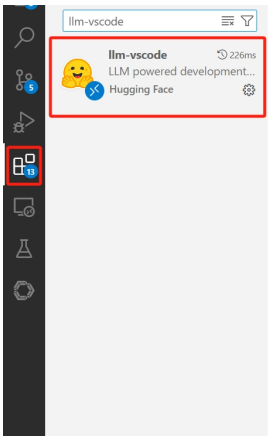
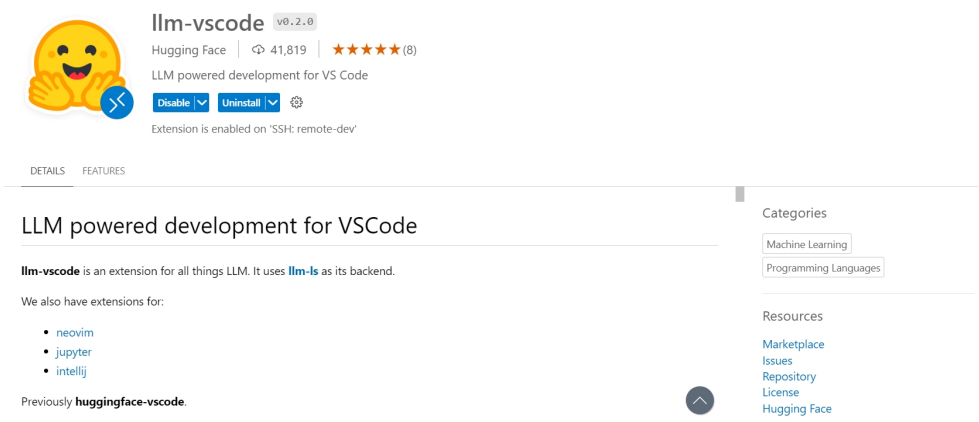
StarCoder VScode 代码生成插件

安装Visual Studio Code 及llm-vscode插件

请参考此指南安装Visual Studio Code。

完成VSCode安装后，在插件中搜索`llm-vscode`并进行安装，本次案例将基于此插件进行本地LLM部署。





环境配置

请事先安装conda并准备好Python 3.9环境。

请激活事先准备好的虚拟环境，执行如下命令，进行环境配置。

> 注：Linux和Windows系统的mindspore的安装命令不同，shell脚本中已列出两个系统的安装命令，请注意区分，保留正确的安装命令。

```
conda activate YOUR_ENV
bash run_environ.sh
```

端口执行案例

cd到指定路径下。

```
cd mindnlp/llm/inference/starcoder
```

启动端口。

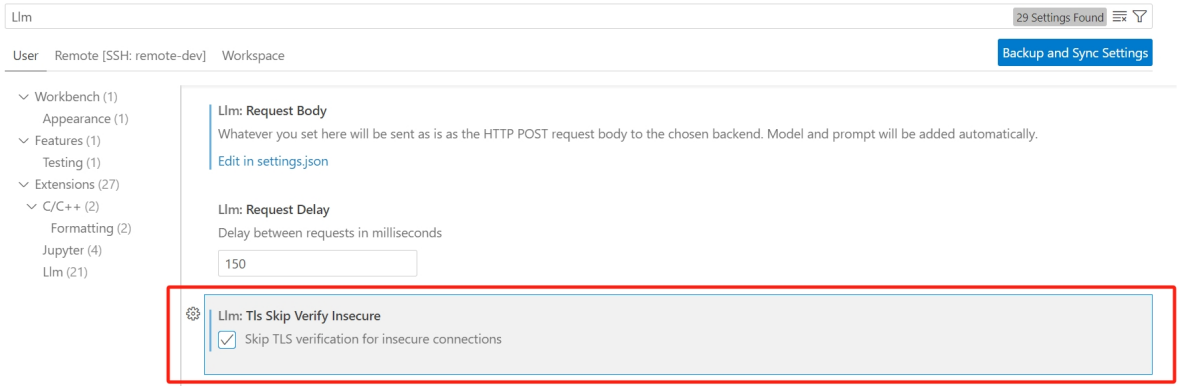
```
python main.py --pretrained bigscience/starcoder2-7b
# use mirror if you cannot access huggingface
python main.py --pretrained AI-ModelScope/starcoder2-7b --mirror modelscope
```

下载权重需要一些时间，如无法访问huggingface或下载权重的时间过长，可通过`--mirror`切换镜像。

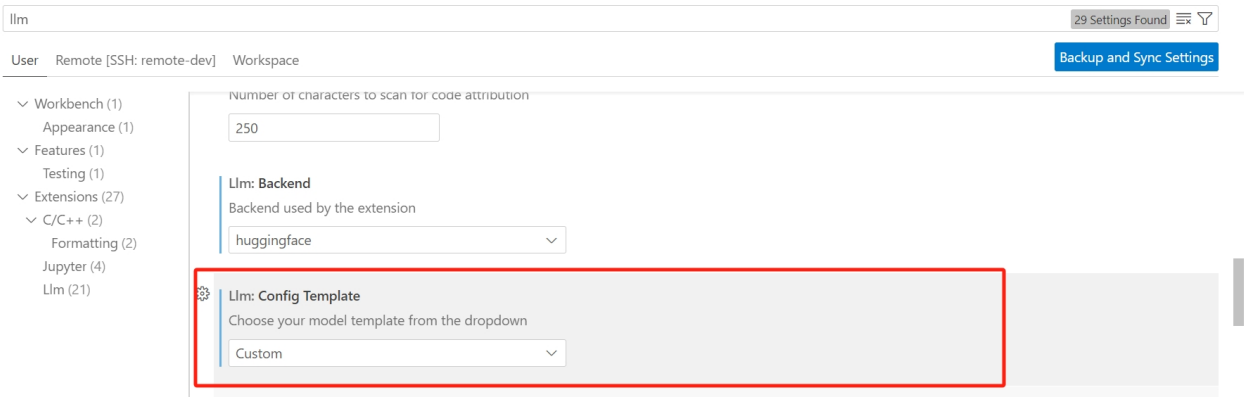
llm-vscode 配置

在VSCode配置中搜索`Llm`，并进行配置。

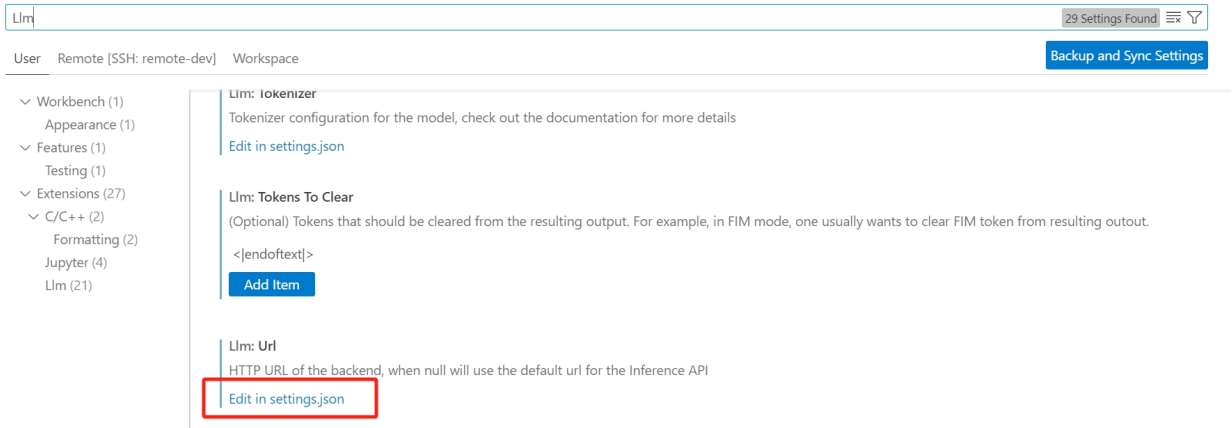
1. 将 `Skip TLS verification for insecure connections`打勾



2. 将`Llm:Backend`改为`tgi`， Llm:ConfigTemplate设置为`Custom`。



3. 在`Llm:Url`选项中点击`Edit in settings.json`。



对settings.json进行如下改动：

- 将`llm.url`设置为 `http://localhost:8000/api/generate/`
- 将`max_new_tokens`数值更换为20

```
    "githubPullRequests.pullBranch": "never",
    "Codegeex.EnableExtension": false,
    "llm.attributionEndpoint": "",
    "llm.backend": "tgi",
    "llm.url": "http://localhost:8000/api/generate/",
    "llm.fillInTheMiddle.enabled": true,
    "llm.fillInTheMiddle.prefix": "<fim_prefix>",
    "llm.fillInTheMiddle.middle": "<fim_middle>",
    "llm.fillInTheMiddle.suffix": "<fim_suffix>",
    "llm.requestBody": {
        "parameters": {
            "max_new_tokens": 20,
            "temperature": 0.2,
            "top_p": 0.95
        }
    },
```

启动代码自动补全

点击`File > New File > Python File`，在vscode界面新建一个Python文档。

可通过注释或输入代码片段启动自动补全，并通过`Tab`键接受自动生成内容。

```
1   def fibonacci(n):
2       if n == 0:
            return 0
        elif n == 1:
            return 1
        else:
            return fibonacci(n-1) + fibonacci(n-2)

    print(fibonacci(10))
```