# Introduction

This workshop consists of 3 sections.

* LM Studio: Selecting and chatting with a local model  
  This workshop does not require much technical skills. Can be done independantly.
* Code completion in VSCode using a local model hosted with Ollama  
  This workshop is mostly interesting to developers
* Implementing sementic search  
  This depends on environment set-up as performed in the code completion workshop.

# LM Studio. Selecting and chatting with a local model

Download and install LM Studio: <https://lmstudio.ai/>

It is recommended to install LM Studio not in a virtualized environment. I encountered issues when running in a Windows Sandbox environment.

Download a model. You can choose one from the main screen, for example Llama3 8B Instruct. Once a download is started you can monitor progress:

Afbeelding met tekst, schermopname, software, Multimediasoftware

Automatisch gegenereerde beschrijving

In the search screen you can see how many model downloads there have been for a specific model (more is usually better) and whether a model is likely to run on your computer.

Afbeelding met tekst, schermopname, software, Multimediasoftware

Automatisch gegenereerde beschrijving

Once a model is downloaded, you can first load i tand then chat with it using the Chat TAB

Afbeelding met schermopname, tekst, software, Multimediasoftware

Automatisch gegenereerde beschrijving

Before using the model update the preset tot he left;

Afbeelding met tekst, schermopname, software, Multimediasoftware

Automatisch gegenereerde beschrijving

Please note the following:

* 7B models are small and not ‘smart’. Can only be used for simple tasks.  
  70B models are often ok but require around 64Gb RAM to run.
* The model does not have an out of the box way to access the internet such as is usual with ChatGPT and Copilot. The model runs in a sandbox and knowledge is pretrained.

LM Studio can host models and provide an OpenAI compliant API to program against similar to ollama. LM Studio is a closed source product.

# Code completion in VSCode using a local model hosted with ollama

## Preparation

You can choose to install tools locally or use a Windows Sandbox environment which already has several things prepared.

### Sandbox

Enable Windows Sandbox

More information on: <https://learn.microsoft.com/en-us/windows/security/application-security/application-isolation/windows-sandbox/windows-sandbox-overview>

Start Powershell with Administrative privileges

Execute the following

Enable-WindowsOptionalFeature -FeatureName "Containers-DisposableClientVM" -All -Online

Install Git: https://git-scm.com/downloads

c:

cd \

git clone <https://github.com/AMISConclusion/Workshop_LocalLLM> sandbox

Execute vscode.wsb

This will create a sandbox environment with VSCode started with the workspace loaded and required files for this workshop in the Downloads folder.

Continue with ‘Install required extensions’

### Locally

Check whether you do not already have

Install VSCode: <https://code.visualstudio.com/>

Install Git: https://git-scm.com/downloads

Create a folder for your repositories. Start a command prompt in that folder and do:

git clone <https://github.com/AMISConclusion/Workshop_LocalLLM> sandbox

Open the folder Projects\demo in VSCode and trust the authors

Afbeelding met tekst, schermopname, software, Multimediasoftware

Automatisch gegenereerde beschrijving

Install Python

curl -L https://www.python.org/ftp/python/3.12.4/python-3.12.4-amd64.exe --output python.exe

python.exe -Wait /PrependPath=1 /quiet InstallAllUsers=1 TargetDir="C:\python" Include\_test=0

Download ollama

curl -L "https://ollama.com/download/OllamaSetup.exe" --output OllamaSetup.exe

### Install required extensions

A prompt in VSCode will pop up asking to install recommended extensions. Click Install.

Afbeelding met tekst, schermopname, software, Multimediasoftware

Automatisch gegenereerde beschrijving

### Install Ollama

Afbeelding met tekst, schermopname, software, Computerpictogram

Automatisch gegenereerde beschrijving

Click Install

Afbeelding met tekst, schermopname, Lettertype, Webpagina

Automatisch gegenereerde beschrijving

Start a command prompt and download local models

For 16Gb RAM (sandbox)

ollama pull deepseek-coder:6.7b-base

Afbeelding met tekst, software, Lettertype, Multimediasoftware

Automatisch gegenereerde beschrijving

### Configure VSCode

Configure the Continue plugin

Afbeelding met tekst, schermopname, software

Automatisch gegenereerde beschrijving

Choose Local model

Afbeelding met tekst, schermopname, software, Lettertype

Automatisch gegenereerde beschrijving

Select the downloaded model: Deepseek Coder 6.7B. Click continu a couple of times to exit the wizard and update the Continu extension configuration.

Afbeelding met tekst, software, Multimediasoftware, Grafische software

Automatisch gegenereerde beschrijving

{

"models": [

{

"title": "Ollama",

"provider": "ollama",

"model": "deepseek-coder:6.7b-base"

}

],

"customCommands": [

{

"name": "test",

"prompt": "{{{ input }}}\n\nWrite a comprehensive set of unit tests for the selected code. It should setup, run tests that check for correctness including important edge cases, and teardown. Ensure that the tests are complete and sophisticated. Give the tests just as chat output, don't edit any file.",

"description": "Write unit tests for highlighted code"

}

],

"tabAutocompleteModel": {

"title": "Ollama",

"provider": "ollama",

"model": "deepseek-coder:6.7b-base"

},

"allowAnonymousTelemetry": true,

"embeddingsProvider": {

"provider": "transformers.js"

}

}

## Code completion and content generation

Open the test.py Python file

Afbeelding met schermopname, tekst, software, Multimediasoftware

Automatisch gegenereerde beschrijving

Press CTRL-I and ask the question: create python code to open a webbrowser and go to www.google.com

Afbeelding met schermopname, Multimediasoftware, tekst, software

Automatisch gegenereerde beschrijving

This generated:

import webbrowser

webbrowser.open('https://www.google.com')

Accept the generated code and run it

Afbeelding met schermopname, tekst, software, Computerpictogram

Automatisch gegenereerde beschrijving

Try out the TAB completion by typing on a second line: webbrowser. A second line could be something like:

webbrowser.open\_new\_tab('https://www.nu.nl')

Try out the new code.

Select some tekst and press CTRL-I. Ask to generate documentation.

Afbeelding met tekst, schermopname, software, Multimediasoftware

Automatisch gegenereerde beschrijving

Accept the suggested change

# Implementing simple semantic search using a local model

Will be added