

## 1■■■ PRE-REQUISITE - Step 1: Install Required Tools

- Install Visual Studio 2022
- Enable: Desktop Development with C++
- Enable: C++ CMake Tools for Windows
- Enable: Git for Windows
- Enable: C++ Clang Compiler for Windows
- Enable: MS-Build Support for LLVM-Toolset

## 2■■■ Step 2: Run this command

Developer Command Prompt for VS 2022

## 3■■■ Create Conda Environment

- `conda create -n bitnet-cpp python=3.9`
- `conda activate bitnet-cpp`

## 4■■■ Install Requirements

`pip install -r requirements.txt`

## 5■■■ Download Model

`huggingface-cli download microsoft/BitNet-b1.58-2B-4T-gguf --local-dir models/BitNet-b1.58-2B-4T`

## 6■■■ Setup Environment

`python setup_env.py -md models/BitNet-b1.58-2B-4T -q i2_s`

## 7■■■ Run Inference

`python run_inference.py -m models/BitNet-b1.58-2B-4T/ggml-model-i2_s.gguf -p "You are a helpful assistant" -cnv`