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

Step 2: Run this command

Developer Command Prompt for VS 2022

Create Conda Environment

• conda create -n bitnet-cpp python=3.9

• conda activate bitnet-cpp

Install Requirements

pip install -r requirements.txt

Download Model

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

Add the following line:

```
#include <chrono>
```

In these files:

• 3rdparty\llama.cpp\common\common.cpp

• 3rdparty\llama.cpp\common\log.cpp

• 3rdparty\llama.cpp\examples\imatrix\imatrix.cpp

• 3rdparty\llama.cpp\examples\perplexity\perplexity.cpp

Setup Environment

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

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