Compute an importance matrix for a model and given text dataset. Can be used during quantization to enchance the quality of the quantum models. More information is available here: https://github.com/ggerganov/llama.cpp/pull/4861 The following options are available: Print version and exit. Show help message and exit. Model path in the GGUF file format. Default: Mandatory path of file containing training data, e.g. The name of the file where the computed data will be stored. If this flag is missing then is used. Specifies how often the so far computed result is saved to disk. The default is is 10 (i.e., every 10 chunks). Specifies if data will be collected for the tensor. Experience indicates that it is better to not utilize the importance matrix when quantizing so this is set to false by default. For faster computation, pass the flag for GPU offloading.