NAME

llamafile-imatrix — importance matrix builder

SYNOPSIS

llamafile-imatrix [flags...] −m *model.gguf* −f *training.data* [−o *imatrix.dat*]

DESCRIPTION

llamafile-imatrix 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

OPTIONS

The following options are available:

−−version
: Print version and exit.

−h, −−help
: Show help message and exit.

−m *FNAME*, −−model *FNAME*
: Model path in the GGUF file format.

Default: *models/7B/ggml-model-f16.gguf*

−f *FNAME*, −−file *FNAME*
: Mandatory path of file containing training data, e.g. *wiki.train.raw*

−o *FNAME*, −−output-file *FNAME*
: The name of the file where the computed data will be stored. If this flag is missing then *imatrix.dat* is used.

−ofreq, −−output-frequency
: Specifies how often the so far computed result is saved to disk. The default is is 10 (i.e., every 10 chunks).

−ow, −−output-weight
: Specifies if data will be collected for the *output.weight* tensor. Experience indicates that it is better to not utilize the importance matrix when quantizing *output.weight*, so this is set to false by default.

PROTIPS

For faster computation, pass the −ngl *9999* flag for GPU offloading.

SEE ALSO

llamafile(1), *llamafile-quantize*(1)