makes LLaVA mmproj files smaller. The following positional arguments are accepted: Is the input file, which should be a CLIP model in the GGUF format using float16 values. Is the output file, which will be a CLIP model in the GGUF format using the desired number type. Is the desired quantization format, which may be the integer id of a supported quantization type. See the quantization types section below for acceptable formats. The following options are accepted: Show help message and exit. Print llamafile version. The following quantization types are available: 2 is Q4_0 3 is Q4_1 6 is Q5_0 7 is Q5_1 8 is Q8_0