Perplexity is one of the most common metrics for evaluating language models. The program can be used to gauge the quality of an LLM implementation. It is defined as the exponentiated average negative log-likelihood of a sequence, calculated with exponent base e. Lower perplexity scores are better. The following options are available: Show help message and exit. Model path (default: models/7B/ggml-model-f16.gguf) Raw data input file. Number of threads to use during generation (default: nproc/2) Random Number Generator (RNG) seed (default: -1, use random seed for < 0) One dataset commonly used in the llama.cpp community for measuring perplexity is wikitext-2-raw. To use it when testing how well both your model and llamafile are performing you could run the following: wget https://cosmo.zip/pub/datasets/wikitext-2-raw/wiki.test.raw llamafile-perplexity -m model.gguf -f wiki.test.raw -s 31337 This can sometimes lead to surprising conclusions, like how Q5 weights might be better for a particular model than Q6.