

**NAME**

llamafile-perplexity — LLM benchmarking tool

**SYNOPSIS**

**llamafile-perplexity** [flags...]

**DESCRIPTION**

Perplexity is one of the most common metrics for evaluating language models. The **llamafile-perplexity** 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.

**OPTIONS**

The following options are available:

- h, --help  
Show help message and exit.
- m *FNAME*, --model *FNAME*  
Model path (default: models/7B/ggml-model-f16.gguf)
- f *FNAME*, --file *FNAME*  
Raw data input file.
- t *N*, --threads *N*  
Number of threads to use during generation (default: nproc/2)
- s *SEED*, --seed *SEED*  
Random Number Generator (RNG) seed (default: -1, use random seed for < 0)

**EXAMPLE**

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.

**SEE ALSO**

*llamafile*(1)