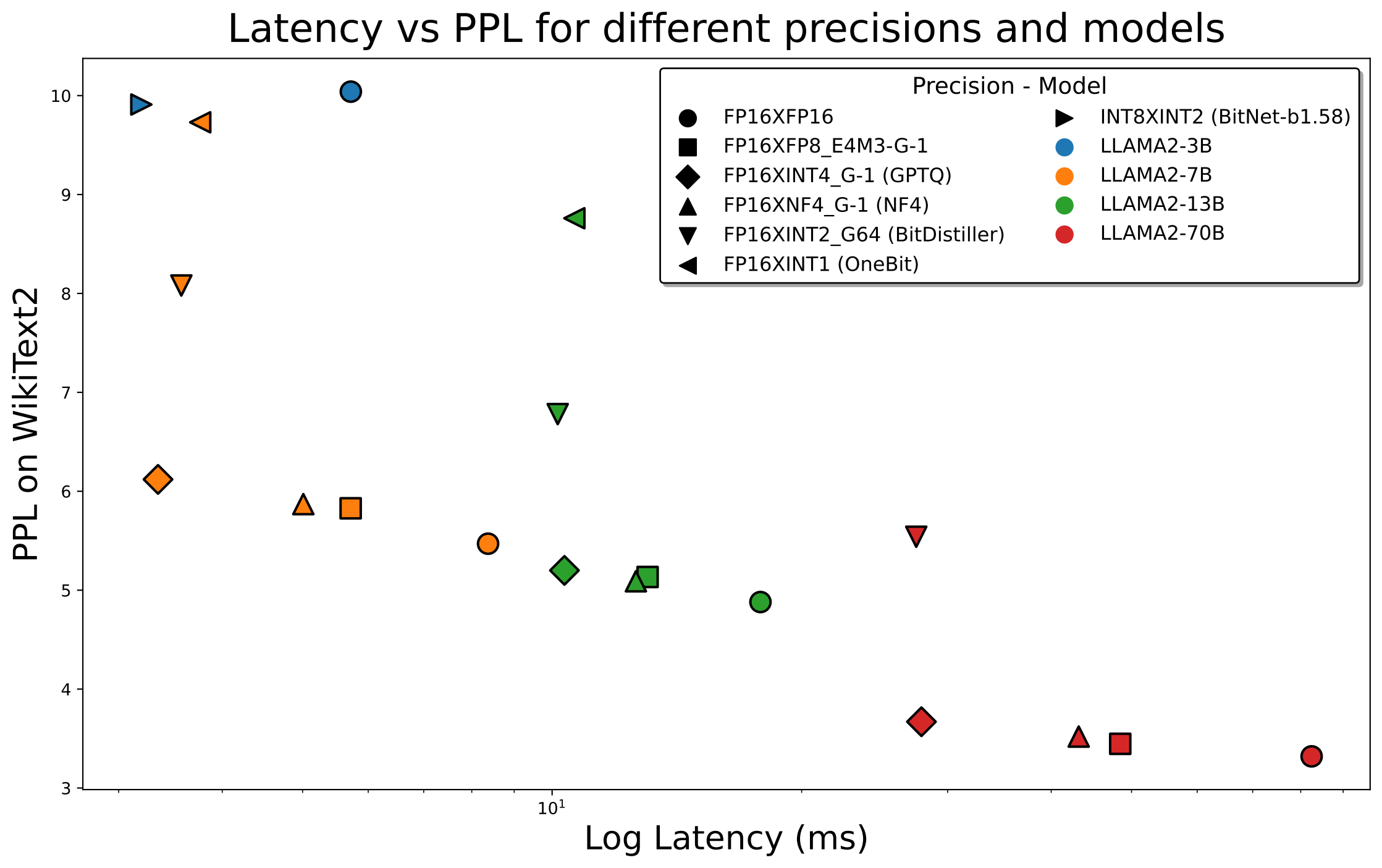
Latency vs PPL for different precisions and models
Precision - Model
FP16XFP16
FP16XFP8_E4M3-G-1
FP16XINT4_G-1 (GPTQ)
FP16XNF4_G-1 (NF4)
FP16XINT2_G64 (BitDistiller)
FP16XINT1 (OneBit)
INT8XINT2 (BitNet-b1.58)
LLAMA2-3B
LLAMA2-7B
LLAMA2-13B
LLAMA2-70B
PPL on WikiText2
Log Latency (ms)
10
9
8
7
6
5
4
3
10^1