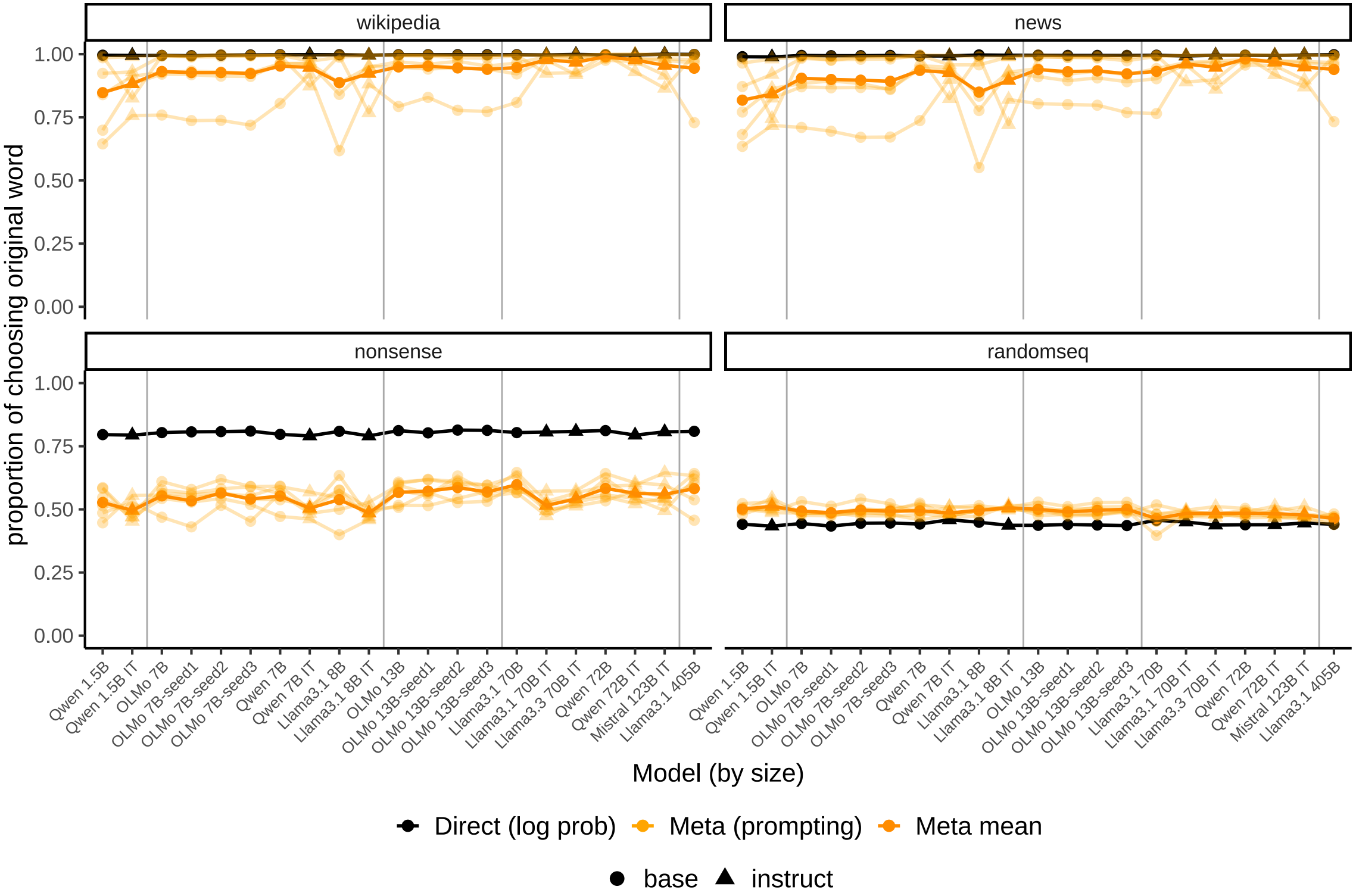
wikipedia
news
nonsense
randomseq
proportion of choosing original word
1.00
0.75
0.50
0.25
0.00
Qwen 1.5B
Qwen 1.5B IT
OLMo 7B
OLMo 7B-seed1
OLMo 7B-seed2
OLMo 7B-seed3
Qwen 7B
Qwen 7B IT
Llama3.1 8B
Llama3.1 8B IT
OLMo 13B
OLMo 13B-seed1
OLMo 13B-seed2
OLMo 13B-seed3
Llama3.1 70B
Llama3.1 70B IT
Llama3.3 70B IT
Qwen 72B
Qwen 72B IT
Mistral 123B IT
Llama3.1 405B
Model (by size)
Direct (log prob)
Meta (prompting)
Meta mean
base
instruct