

Model	Non-Emb. Params	Tokens	Total
MINITRON 4B-instruct	2.6B	90B	5.61
Phi-2	2.5B	1.4T	4.29
Qwen-1.5 Chat	1.2B	N/A	5.29
Gemma-2B-IT	2B	6T	5.19
StableLM 2 Chat	1.6B	2T	5.42
TinyLlama v1.0 Chat	1.1B	3T	3.46

Table 6: Evaluation results on MT-Bench.

Model	Prompt-level Acc. (strict)	Prompt-level Acc. (loose)	Instruction-level Acc. (loose)
MINITRON 4B-instruct	31.6	35.49	44.84
Gemma-2B-IT	-	28.70	40.50
Qwen2-1.5B-Instruct	29	-	-

Table 7: Evaluation results on IFEval.

Model	Avg
MINITRON 4B-instruct	41.11
Gemma-2B-IT	33.31

Table 8: Evaluation results on ChatRAG-Bench.

Model	Avg.
MINITRON 4B-instruct	64.23
Gemma-2B-IT	47.00
Llama-3-8B-instruct	58.78

Table 9: Evaluation results on BFCL.

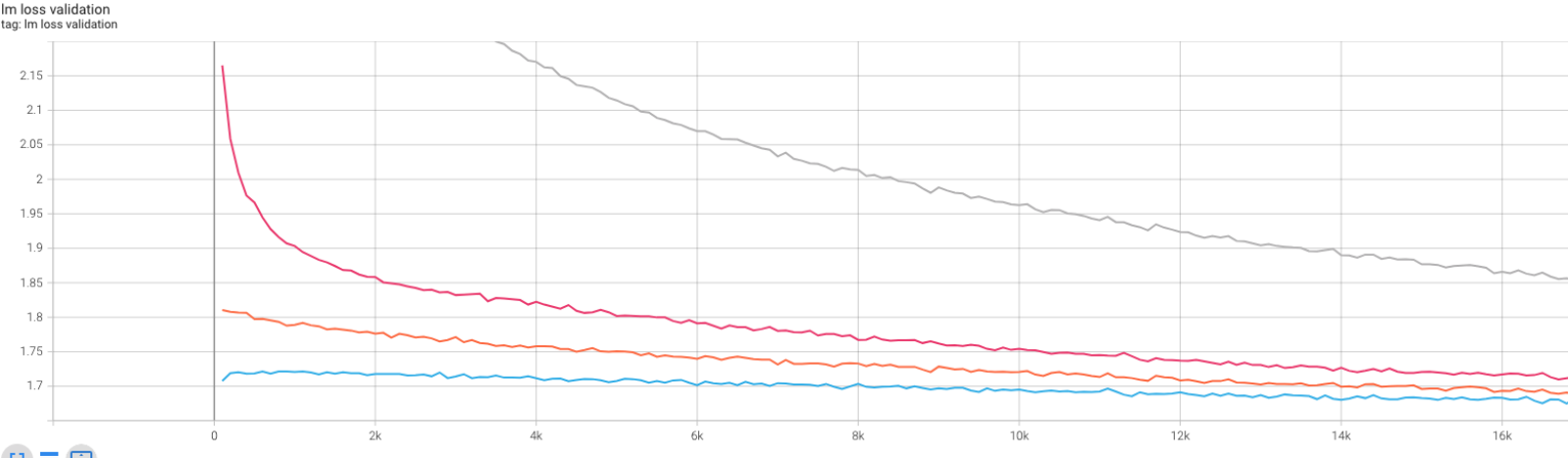


Figure 1