Model Name	Format	Model Loader	tokens/s	tokens/s	GPU VRAM (GB)	RAM (GB)	CPU	Settings		
WizardCoder 13B	GPTQ - 4bit – 32g	Transformers	22.7	25.7	9.5	6.88		GPU_mem(all), CPU_mem(all)		
WizardCoder 13B	GPTQ - 4bit – 32g	ExLlama_HF	33.5	40.7	11.5	6.25		max_seq_len(2048), cfg_cache(True)		
WizardCoder 13B	**GPTQ - 4bit – 32g**	**ExLlama**	**44.3**	**49.7**	**10**	**6.11**				
WizardCoder 13B	GPTQ - 4bit – 32g	AutoGPTQ	23.5	27.4	9.8	13.3		gpu_mem(all), cpu_mem(all), wbits(4), gs(32), no_inject_fused_attention, auto-devices		
WizardCoder 13B	GPTQ - 4bit – 32g	GPTQ-for-LlaMa	18.4	20.44	10	10		wbits(4), gs(32), model_type(llama)	output numbers	
WizardCoder 13B	GPTQ - 4bit – 32g	llama.cpp							Not RUN	
WizardCoder 13B	GPTQ - 4bit – 32g	llamacpp_HF	33.5	40	11.5	6.1		max_seq_len(2048), cfg_cache(True)		
WizardCoder 13B	GPTQ - 4bit – 32g	ctransformers							Not RUN	
WizardCoder 34B	GPTQ - 4bit – 32g	Transformers								
WizardCoder 34B	GPTQ - 4bit – 32g	ExLlama_HF	17	18.7	20.8	6.2		max_seq_len(2048), cfg_cache(True)		
WizardCoder 34B	**GPTQ - 4bit – 32g**	**ExLlama**	**19.7**	**20.8**	**20.4**	**6**				
WizardCoder 34B	GPTQ - 4bit – 32g	AutoGPTQ							Not RUN	
WizardCoder 34B	GPTQ - 4bit – 32g	GPTQ-for-LlaMa							Not RUN	
WizardCoder 34B	GPTQ - 4bit – 32g	llama.cpp							Not RUN	
WizardCoder 34B	GPTQ - 4bit – 32g	llamacpp_HF							Not RUN	
WizardCoder 34B	GPTQ - 4bit – 32g	ctransformers							Not RUN	
WizardCoder 13B	GGUF – Q8	Transformers								
WizardCoder 13B	GGUF – Q8	ExLlama_HF								
WizardCoder 13B	GGUF – Q8	ExLlama								
WizardCoder 13B	GGUF – Q8	AutoGPTQ								
WizardCoder 13B	GGUF – Q8	GPTQ-for-LlaMa								
WizardCoder 13B	GGUF – Q8	llama.cpp	38 (threads-0)	24.5 (threads-32)	16.6	7.9		n-gpu-layers(128), n_ctx(2048), threads(0), n_batch(512)		
WizardCoder 13B	GGUF – Q8	llamacpp_HF								
WizardCoder 13B	**GGUF – Q8**	**ctransformers**	**39 (threads-0)**	**27 (threads-32)**	**16.6**	**7.8**		**n-gpu-layers(128), n_ctx(2048), threads(0), n_batch(512), model_type(llama)**		
WizardCoder 34B	GGUF - Q5 – KM	Transformers								
WizardCoder 34B	GGUF - Q5 – KM	ExLlama_HF								
WizardCoder 34B	GGUF - Q5 – KM	ExLlama								
WizardCoder 34B	GGUF - Q5 – KM	AutoGPTQ								
WizardCoder 34B	GGUF - Q5 – KM	GPTQ-for-LlaMa								
WizardCoder 34B	GGUF - Q5 – KM	llama.cpp	7.7 (threads-0)	5.9 (threads-32)	22.6	7		n-gpu-layer(42), threads(32, 0 prefered-faster)	threads(0) > threads(32)	not fitted in GPU
WizardCoder 34B	GGUF - Q5 – KM	llamacpp_HF								
WizardCoder 34B	**GGUF - Q5 – KM**	**ctransformers**	**7.9 (threads-0)**	**5.68 (threads-32)**	**22.6**	**7**		**n-gpu-layer(42), threads(32, 0 prefered-faster)**	**threads(0) > threads(32)**	**not fitted in GPU**
WizardCoder 13B	**BNB – Q4**	**Transformers**	**16.5**	**14.3 (double-quant)**	**8.5**	**6.8**		**gpu_mem(all), bfloat16, nf4, bf16, load-in-4bit**		
WizardCoder 13B	BNB – Q4	ExLlama_HF								
WizardCoder 13B	BNB – Q4	ExLlama								
WizardCoder 13B	BNB – Q4	AutoGPTQ								
WizardCoder 13B	BNB – Q4	GPTQ-for-LlaMa								
WizardCoder 13B	BNB – Q4	llama.cpp								
WizardCoder 13B	BNB – Q4	llamacpp_HF								
WizardCoder 13B	BNB – Q4	ctransformers								
WizardCoder 13B	GGUF – Q4 – KM	Transformers								
WizardCoder 13B	GGUF – Q4 – KM	ExLlama_HF								
WizardCoder 13B	GGUF – Q4 – KM	ExLlama								
WizardCoder 13B	GGUF – Q4 – KM	AutoGPTQ								
WizardCoder 13B	GGUF – Q4 – KM	GPTQ-for-LlaMa								
WizardCoder 13B	GGUF – Q4 – KM	llama.cpp	51	48.5	11.2	8.1		n-gpu-layers(128), n_ctx(2048), threads(0), n_batch(512)		
WizardCoder 13B	GGUF – Q4 – KM	llamacpp_HF								
WizardCoder 13B	**GGUF – Q4 – KM**	**ctransformers**	**52.8**	**52.3**	**11.2**	**8**		**n-gpu-layers(128), n_ctx(2048), threads(0), n_batch(512), model_type(llama)**		
WizardCoder 34B	GGUF – Q4 – KM	Transformers								
WizardCoder 34B	GGUF – Q4 – KM	ExLlama_HF								
WizardCoder 34B	GGUF – Q4 – KM	ExLlama								
WizardCoder 34B	GGUF – Q4 – KM	AutoGPTQ								
WizardCoder 34B	GGUF – Q4 – KM	GPTQ-for-LlaMa								
WizardCoder 34B	**GGUF – Q4 – KM**	**llama.cpp**	**24**	**25.4**	**22.6**	**6.9**		**n-gpu-layers(128), n_ctx(2048), threads(0), n_batch(512)**		
WizardCoder 34B	GGUF – Q4 – KM	llamacpp_HF								
WizardCoder 34B	GGUF – Q4 – KM	ctransformers	25	24.8	22.6	6.9		n-gpu-layers(128), n_ctx(2048), threads(0), n_batch(512), model_type(llama)		