Management 1986 EPTQ - Alex - 23g Debt. Date Extra Alex - 24d Alex - 24d Bit Bit	Model Name	Format	Model Loader	tokens/s	tokens/s	GPU VRAM (GB)	RAM (GB)	CPU	Settings		
Microstrate 18	WizardCoder 13B	GPTQ - 4bit – 32g	Transformers	22.7	25.7	9.5	6.88				
March Marc	WizardCoder 13B	GPTQ - 4bit – 32g	ExLlama_HF	33.5	40.7	11.5	6.25		max_seq_len(2048),		
WaterCode: 128 GPTO - Str. 527 AppSiPTO 322 325 326	WizardCoder 13B	GPTQ - 4bit - 32g	ExLlama	44.3	49.7	10	6.11		,		
Manufaction 18. Colff Administration 18. Colff A	WizardCoder 13B	GPTQ - 4bit – 32g	AutoGPTQ	23.5	27.4	9.8	13.3		cpu_mem(all), wbits(4), gs(32), no_inject_fused_atte ntion, auto-devices		
Wiley Golder 10 GPTQ - 401 - 30 Transformers	WizardCoder 13B	, ,		18.4	20.44	10	10			·	
Manufacture 18. CPTQ - 9th - 30g				22.5	40	11.5	6.1		max_seq_len(2048),	Not RUN	
Wiley Control Contro	WizardCoder 13B			33.3	40	11.5	0.1		cfg_cache(True)	Not RUN	
Wiley Control Contro											
MicroCode 130 GCUF - Qu -	WizardCoder 34B	GPTQ - 4bit – 32g	Transformers								
WaterCoder 39	WizardCoder 34B		_						_ :_ ` /		
MicrostCode 18 GPTQ - 448 - 29 Image	WizardCoder 34B WizardCoder 34B			19.7	20.8	20.4	6			Not RUN	
Managed Coder 135	WizardCoder 34B										
ManaGolari 335 GGUF G0	WizardCoder 34B	GPTQ - 4bit – 32g	llamacpp_HF							Not RUN	
Manager 130 Court Cour	WizardCoder 34B	GPTQ - 4bit – 32g	ctransformers							Not RUN	
Manager 130 Court Cour	WizardCoder 13B	GGUE – O8	Transformers								
Microscore 138 GGUF Ge Image Microscore 139 GGUF Ge Image Microscore 130 GGUF Ge Image Microscore 130 Image Microscore Micr	WizardCoder 13B	GGUF – Q8	ExLlama_HF								
Notatricoser 138 CGUF - Q8	WizardCoder 13B WizardCoder 13B										
WithortCoder 136 GGUF QB Ilama.cpp 38 (threads-0) 24.5 (threads-12) 16.5 7.9	WizardCoder 13B	GGUF – Q8	GPTQ-for-LlaMa						n-apu-lavers(128).		
MizardCoder 138 GGUF - Q8 ctransformers 39 (threads-0) 27 (threads-32) 16.6 7.8	WizardCoder 13B	•		38 (threads-0)	24.5 (threads-32)	16.6	7.9		n_ctx(2048), threads(0),		
MicrostCoder 13B	WizaruCouer 13B	GGUF – Q8	паттасрр_пг								
WistantCoder 38	WizardCoder 13B	GGUF – Q8	ctransformers	39 (threads-0)	27 (threads-32)	16.6	7.8		threads(0), n_batch(512),		
WistantCoder 38	WizardCoder 34B	GGUE - 05 – KM	Transformers								
WizardCoder 34B GGUF - Q5 - KM AutoGFTQ WizardCoder 34B GGUF - Q5 - KM Iama.cpp F Image Image F Image	WizardCoder 34B	GGUF - Q5 – KM	ExLlama_HF								
MizardCoder 34B GGUF Q5 - KM Iama.cpp 7.7 (threads 0) 5.9 (threads 32) 22.6 7 P-@PU-Heyr(42), threads(32, 0) Proceeds(32) Proceeds(32)	WizardCoder 34B WizardCoder 34B										
MizardCoder 34B GGUF - QS - KM Ilamacpp 7.7 (threads 2) 5.9 (threads 32) 22.6 7	WizardCoder 34B	GGUF - Q5 – KM	GPTQ-for-LlaMa						n-apu-laver(42).		
MizardCoder 348 GGUF - Q4 - KM Ctransformers 7.9 (threads-0) 5.68 (threads-32) 22.6 7	WizardCoder 34B	-		7.7 (threads-0)	5.9 (threads-32)	22.6	7		threads(32, 0		
			- · · · -	7.0 (threads 0)	5 69 (throads 22)	22.6	7			threads(0) >	not fitted
WizardCoder 13B BNB - Q4 Transformers 16.5 14.3 (double-quant) 8.5 6.8 bfloatfs, fild, bfl6, load-in-4bit	Wizarucouer 54B	GGGF - Q5 - KWI	Citarisionners	7.9 (tilleaus-0)	3.00 (tilleaus-32)	22.0	-			threads(32)	in GPU
WizardCoder 13B BNB - Q4 Transformers 16.5 14.3 (double-quant) 8.5 6.8 bfloatfs, fild, bfl6, load-in-4bit									gpu mem(all),		
WizardCoder 13B BNB - Q4	WizardCoder 13B	•	Transformers	16.5	14.3 (double-quant)	8.5	6.8		bfloat16, nf4, bf16,		
WizardCoder 13B BNB - Q4 BN	WizardCoder 13B WizardCoder 13B										
WizardCoder 13B BNB – Q4 Ilamac.pp H WizardCoder 13B BNB – Q4 Ilamac.pp HF WizardCoder 13B GSUF – Q4 – KM Transformers WizardCoder 13B GSUF – Q4 – KM Ext.lama_HF WizardCoder 13B GGUF – Q4 – KM Ext.lama_HF WizardCoder 13B GGUF – Q4 – KM Ext.lama_HF WizardCoder 13B GGUF – Q4 – KM GPTQ-for-ItaMa WizardCoder 13B GGUF – Q4 – KM Ilamac.pp 51 48.5 11.2 8.1 In-gpu-layers(128), Incas(2049), In	WizardCoder 13B										
WizardCoder 138 BNB – Q4 ctransformers WizardCoder 138 GGUF – Q4 – KM Transformers WizardCoder 138 GGUF – Q4 – KM ExLlama HF WizardCoder 138 GGUF – Q4 – KM ExLlama AutoGPTQ WizardCoder 138 GGUF – Q4 – KM GPTQ-for-LlaMa WizardCoder 138 GGUF – Q4 – KM Illama.cpp 51 48.5 11.2 8.1 In-gpu-layers(128), n. ctx(2048), threads(0), n. bach(512) WizardCoder 138 GGUF – Q4 – KM Illama.cpp 52.8 52.3 11.2 8 In-gpu-layers(128), n. ctx(2048), threads(0), n. bach(512), model type(Ilama) WizardCoder 34B GGUF – Q4 – KM Ctransformers 52.8 52.3 11.2 8 In-gpu-layers(128), n. ctx(2048), threads(0), n. bach(512), model type(Ilama) WizardCoder 34B GGUF – Q4 – KM ExLlama – F. Extlama –	WizardCoder 13B	BNB – Q4	llama.cpp								
WizardCoder 138	WizardCoder 13B WizardCoder 13B										
WizardCoder 138											
WizardCoder 13B GGUF - Q4 - KM	WizardCoder 13B										
WizardCoder 13B	WizardCoder 13B	GGUF – Q4 – KM	ExLlama								
WizardCoder 13B GGUF - Q4 - KM Ilama.cpp 51 48.5 11.2 8.1 In_ctx(2048), threads(0), n_batch(512) WizardCoder 13B GGUF - Q4 - KM Ilamacpp_HF In_gpu-layers(128), n_ctx(2048), threads(0), n_batch(512) WizardCoder 13B GGUF - Q4 - KM Ctransformers 52.8 52.3 11.2 8 In_gpu-layers(128), n_ctx(2048), threads(0), n_batch(512), model type(Ilama) WizardCoder 34B GGUF - Q4 - KM ExLlama_HF ExLlama_HF ExLlama_HF ExLlama_HF In_gpu-layers(128), n_ctx(2048), threads(0), n_batch(512) WizardCoder 34B GGUF - Q4 - KM AutoGPTQ Ilama.cpp 24 25.4 22.6 6.9 In_gpu-layers(128), n_ctx(2048), threads(0), n_batch(512) WizardCoder 34B GGUF - Q4 - KM Ilama.cpp 24 25.4 22.6 6.9 In_gpu-layers(128), n_ctx(2048), threads(0), n_batch(512) WizardCoder 34B GGUF - Q4 - KM Ilama.cpp 24 25.4 22.6 6.9 In_gpu-layers(128), n_ctx(2048), threads(0), n_batch(512)	WizardCoder 13B WizardCoder 13B										
WizardCoder 13B GGUF - Q4 - KM Ilamacpp_HF WizardCoder 13B GGUF - Q4 - KM ctransformers 52.8 52.3 11.2 8 n-gpu-layers(128), n_ctx(2043), threads(0), n_batch(512), model type(Ilama) WizardCoder 34B GGUF - Q4 - KM ExLlama_HF GGUF - Q4 - KM ExLlama_HF WizardCoder 34B GGUF - Q4 - KM AutoGPTQ AutoGPTQ WizardCoder 34B GGUF - Q4 - KM GPTQ-for-LlaMa WizardCoder 34B GGUF - Q4 - KM Ilama.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 Ilamacpp_HF n-gpu-layers(128), n_ctx(2048), threads(0), n_batch(512) n_ctx(2048), threads(0), n_ctx(2048), threads(0), n_ctx(2048), threads(0), n_batch(512), n_batch(512),	WizardCoder 13B	GGUF – Q4 – KM	llama.cpp	51	48.5	11.2	8.1		n_ctx(2048), threads(0),		
WizardCoder 13B GGUF - Q4 - KM ctransformers 52.8 52.3 11.2 8 in_ctx(2048), threads(0), n_batch(512), model type(llama) WizardCoder 34B GGUF - Q4 - KM ExLlama, HF WizardCoder 34B GGUF - Q4 - KM ExLlama GGUF - Q4 - KM AutoGPTQ WizardCoder 34B GGUF - Q4 - KM GPTQ-for-LlaMa Im_ctx(2048), threads(0), n_ctx(2048), threads(0), n_batch(512) WizardCoder 34B GGUF - Q4 - KM Ilama.cpp 24 25.4 22.6 6.9 in_ctx(2048), threads(0), n_batch(512) WizardCoder 34B GGUF - Q4 - KM Ilamacpp HF Ilamacpp HF In-gpu-layers(128), n_ctx(2048), threads(0), n_batch(512), in_ctx(2048), threads(0), in_ctx(2048), threads(0), in_batch(512), in_ctx(2048), threads(0), in_ctx(2048), threads(0), in_ctx(2048), threads(0), in	WizardCoder 13B	GGUF – Q4 – KM	llamacpp_HF								
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 Ilama.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 Ilamacpp_HF n-gpu-layers(128), n_ctx(2048), threads(0), n_batch(512), n_ctx(2048), threads(0), n_cbatch(512), n_batch(512),	WizardCoder 13B	GGUF – Q4 – KM	ctransformers	52.8	52.3	11.2	8		n_ctx(2048), threads(0), n_batch(512),		
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 In-gpu-layers(128), n_ctx(2048), threads(0), n_batch(512) WizardCoder 34B GGUF - Q4 - KM Ilamacpp_HF WizardCoder 34B GGUF - Q4 - KM Ctransformers 25 24.8 22.6 6.9 In-gpu-layers(128), n_ctx(2048), threads(0), n_batch(512), n_batch(5											
WizardCoder 34B	WizardCoder 34B										
WizardCoder 34B GGUF - Q4 - KM GPTQ-for-LlaMa WizardCoder 34B GGUF - Q4 - KM Ilama.cpp 24 25.4 22.6 6.9 In-gpu-layers(128), n_ctx(2048), threads(0), n_batch(512) WizardCoder 34B GGUF - Q4 - KM Ilamacpp_HF In-gpu-layers(128), n_ctx(2048), threads(0), n_batch(512), in_ctx(2048), threads(0), n_batch(512), in_batch(512), in_	WizardCoder 34B WizardCoder 34B										
WizardCoder 34B GGUF - Q4 - KM Ilama.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 Ilamacpp_HF Ilamacpp_HF n-gpu-layers(128), n_ctx(2048), threads(0), n_ctx(2048), threads(0), n_ctx(2048), threads(0), n_ctx(2048), threads(0), n_batch(512), n_ctx(2048), threads(0), n_batch(512), n_ctx(2048), threads(0), n_batch(512), n_ctx(2048), threads(0), n_batch(512), n_ctx(2048), threads(0), n_ctx(2048), t											
WizardCoder 34B GGUF - Q4 - KM Ilamacpp_HF wizardCoder 34B GGUF - Q4 - KM n-gpu-layers(128), n_ctx(2048), n_ctx(2048), threads(0), n_batch(512),				24	25.4	22.6	6.9		n_ctx(2048), threads(0),		
WizardCoder 34B GGUF – Q4 – KM ctransformers 25 24.8 22.6 6.9	WizardCoder 34B	GGUF – Q4 – KM	llamacpp_HF								
mouel type(liama)	WizardCoder 34B	GGUF – Q4 – KM	ctransformers	25	24.8	22.6	6.9		n_ctx(2048), threads(0), n_batch(512),		
									sasi typo(iiaiiia)		