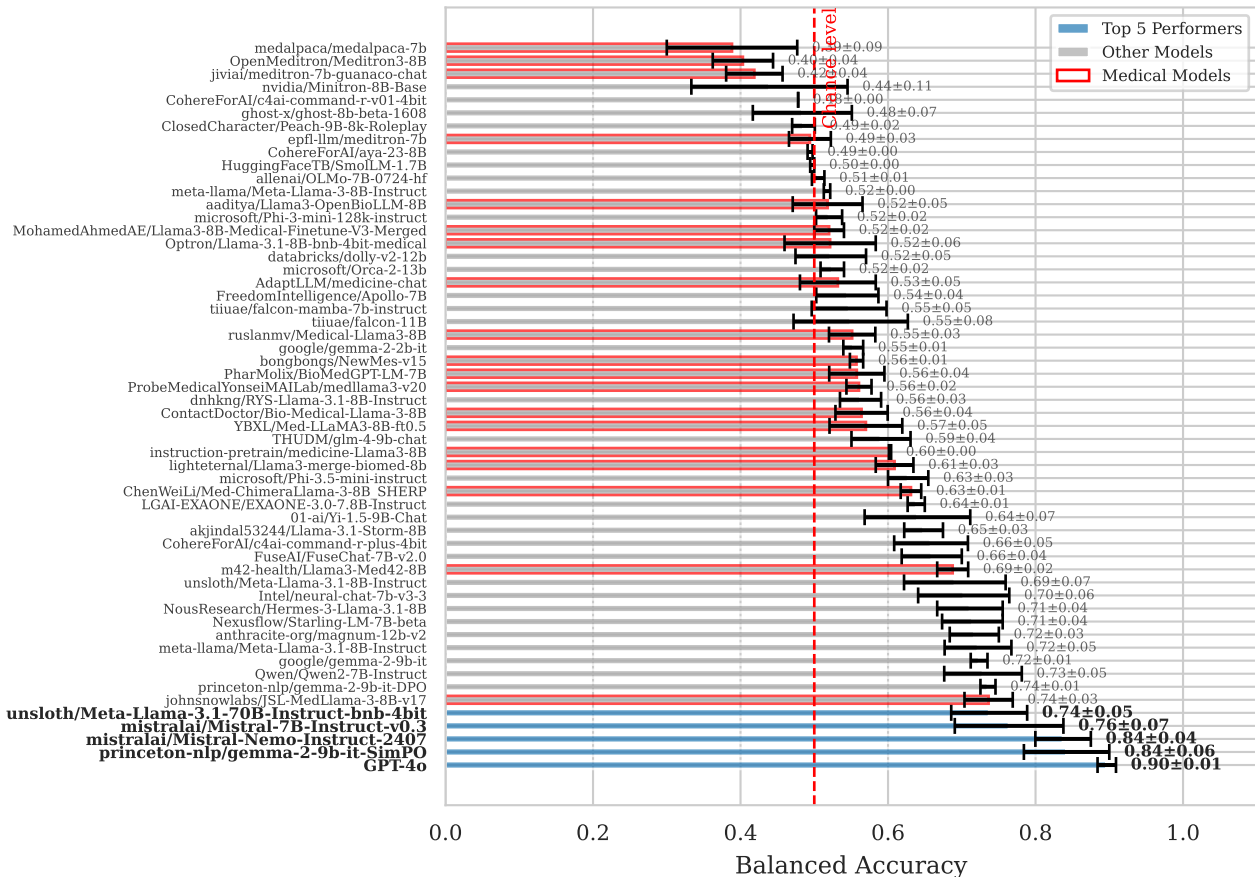
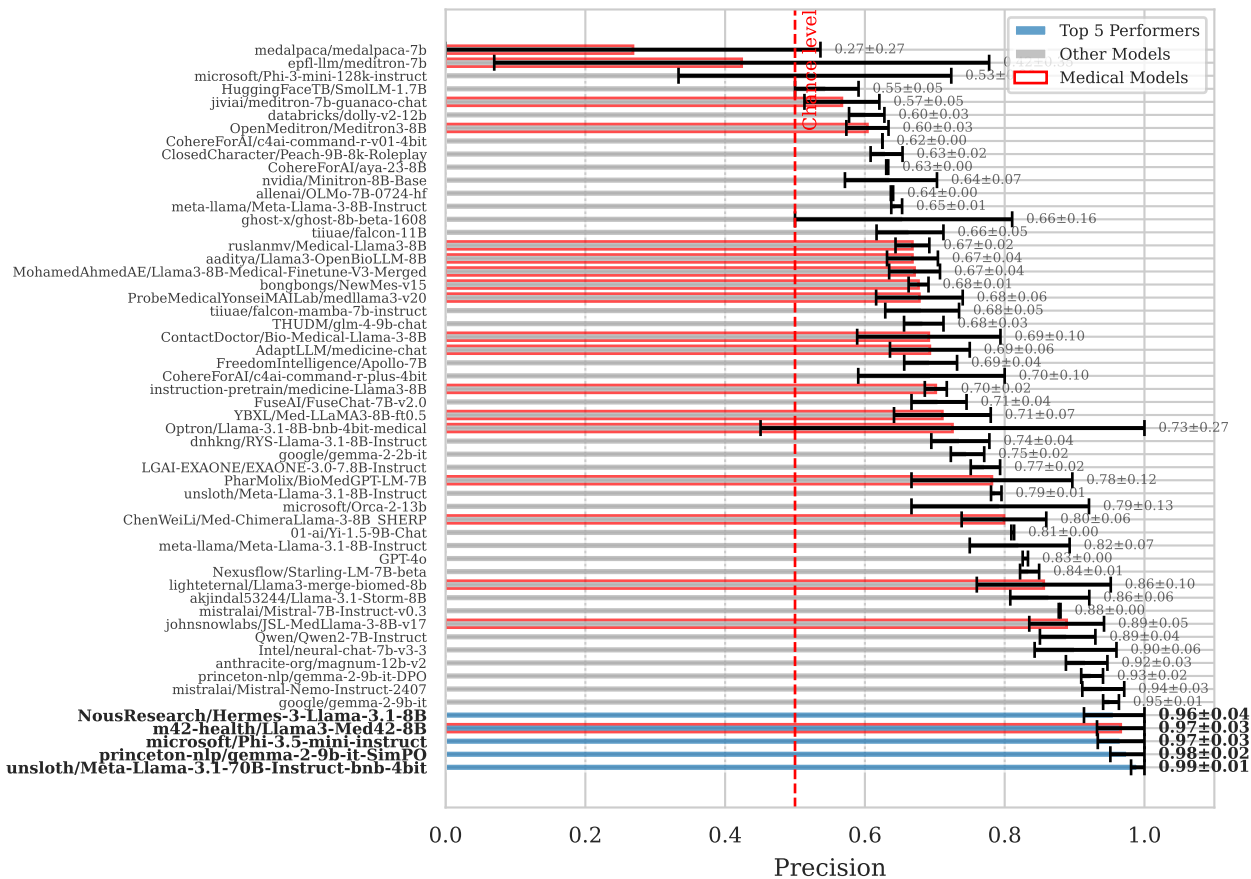


# Balanced Accuracy

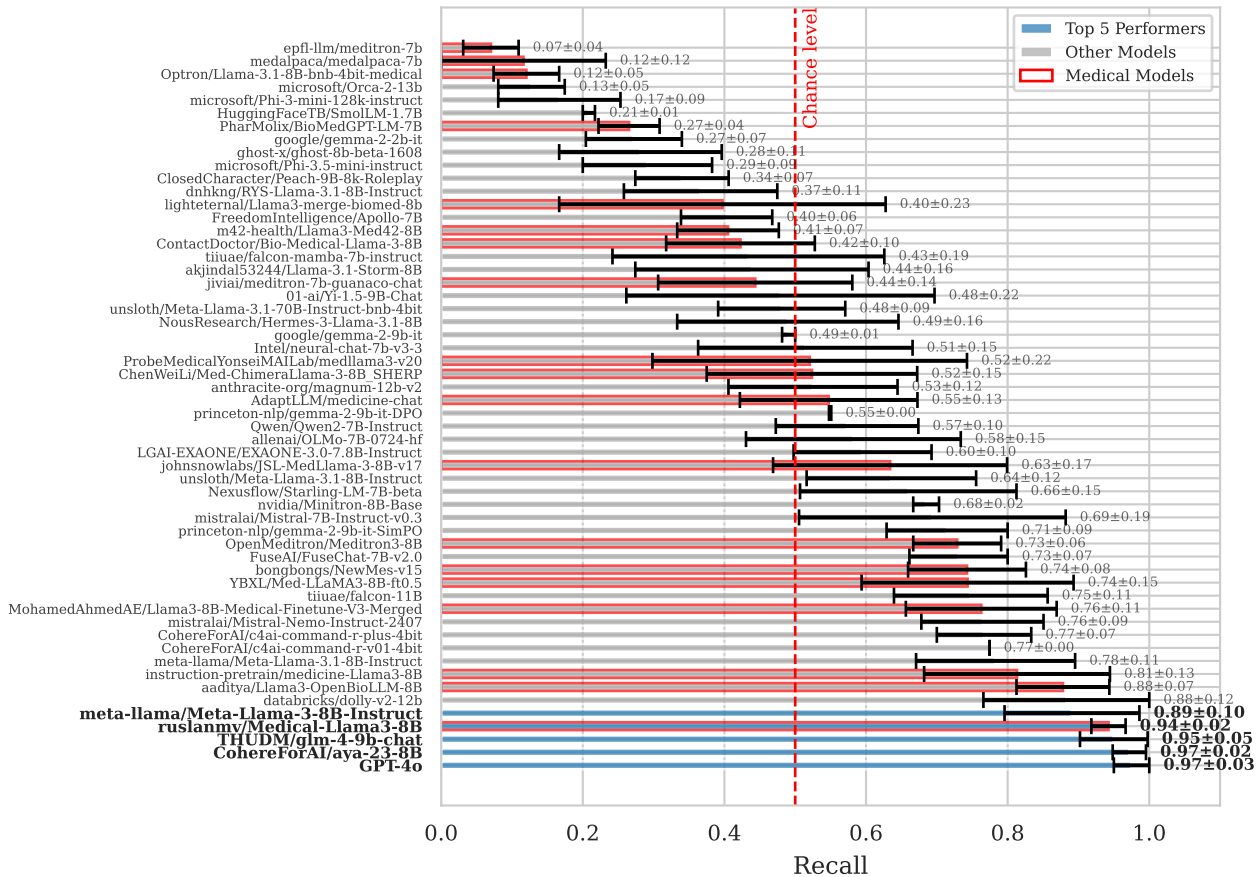


# Precision



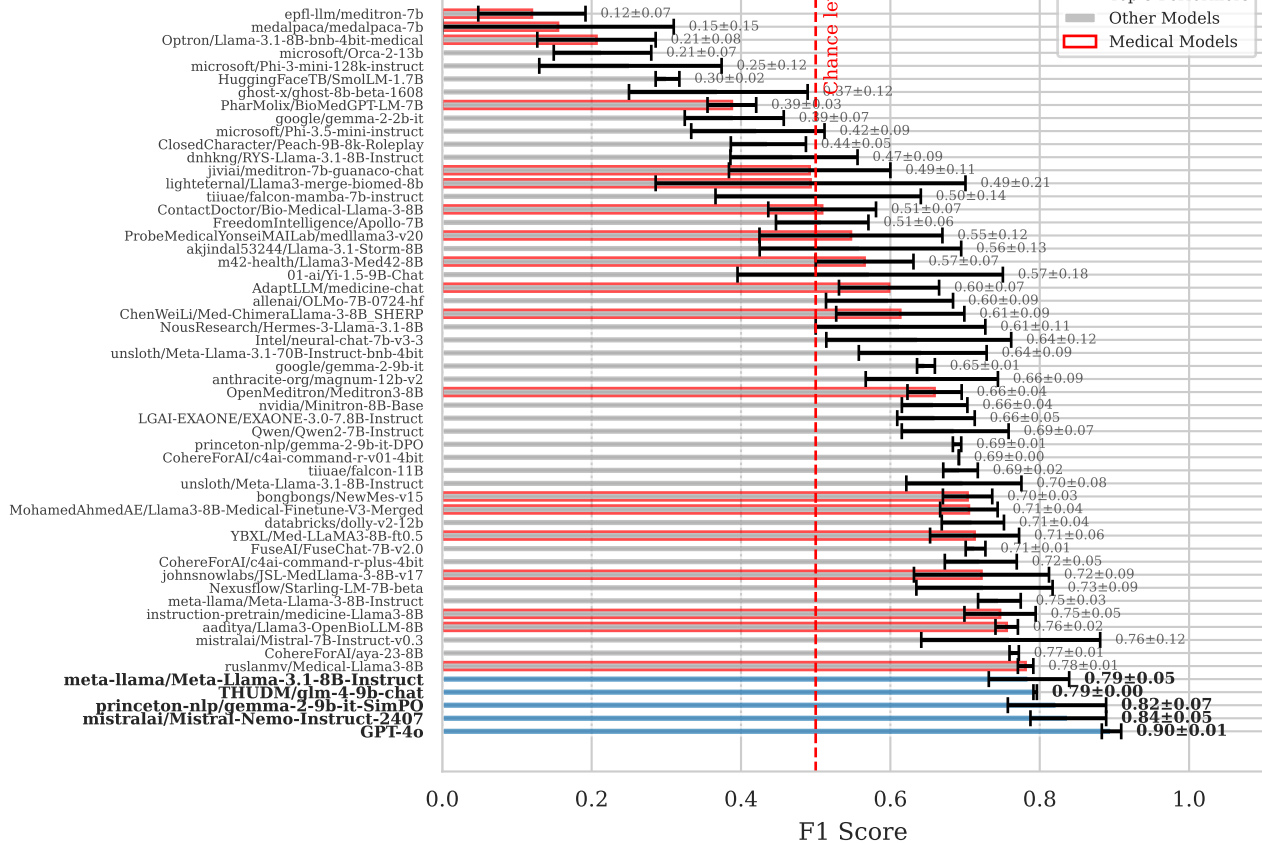
Recall

Model

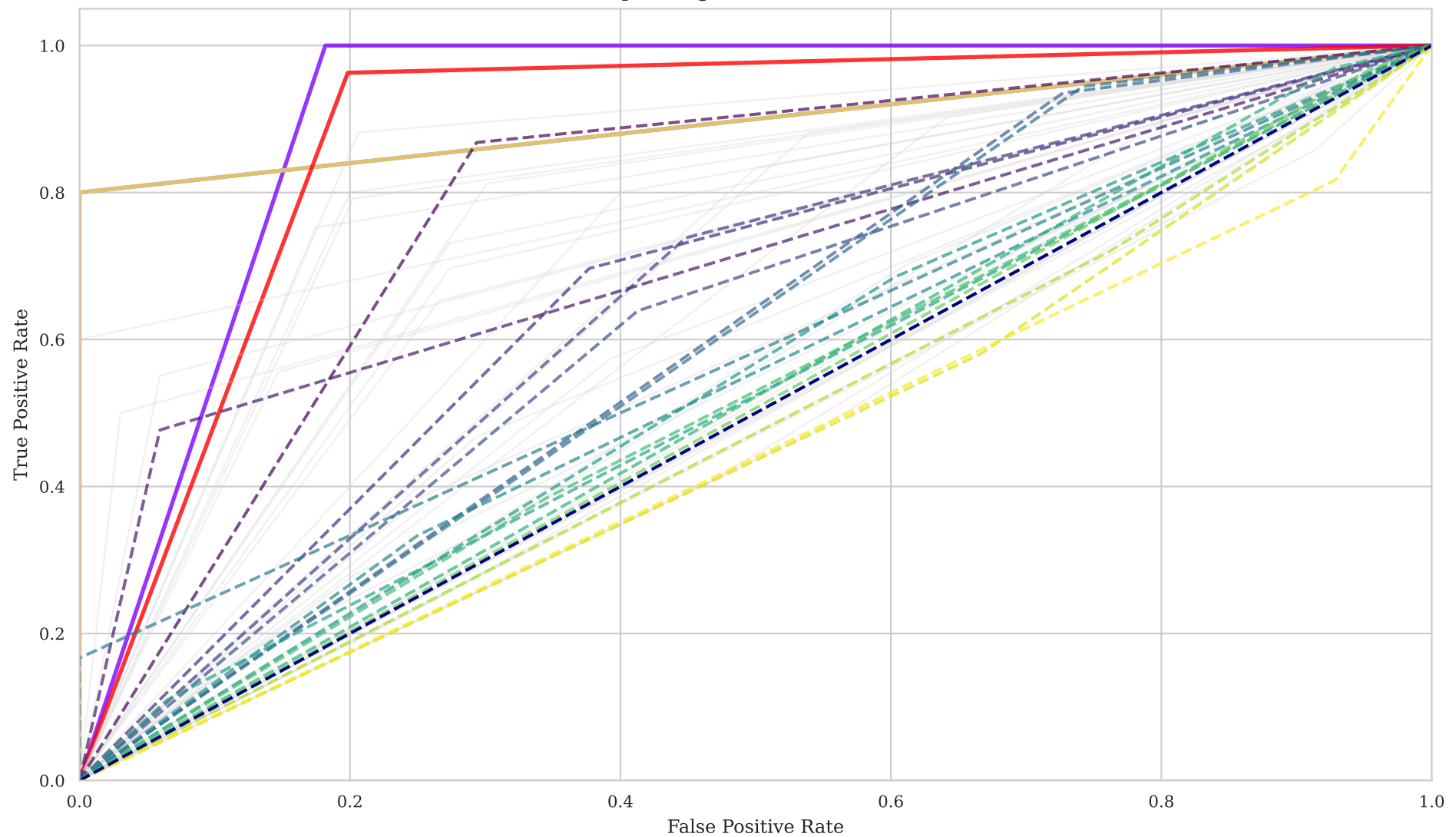


F1 Score

Top 5 Performers  
Other Models  
Medical Models

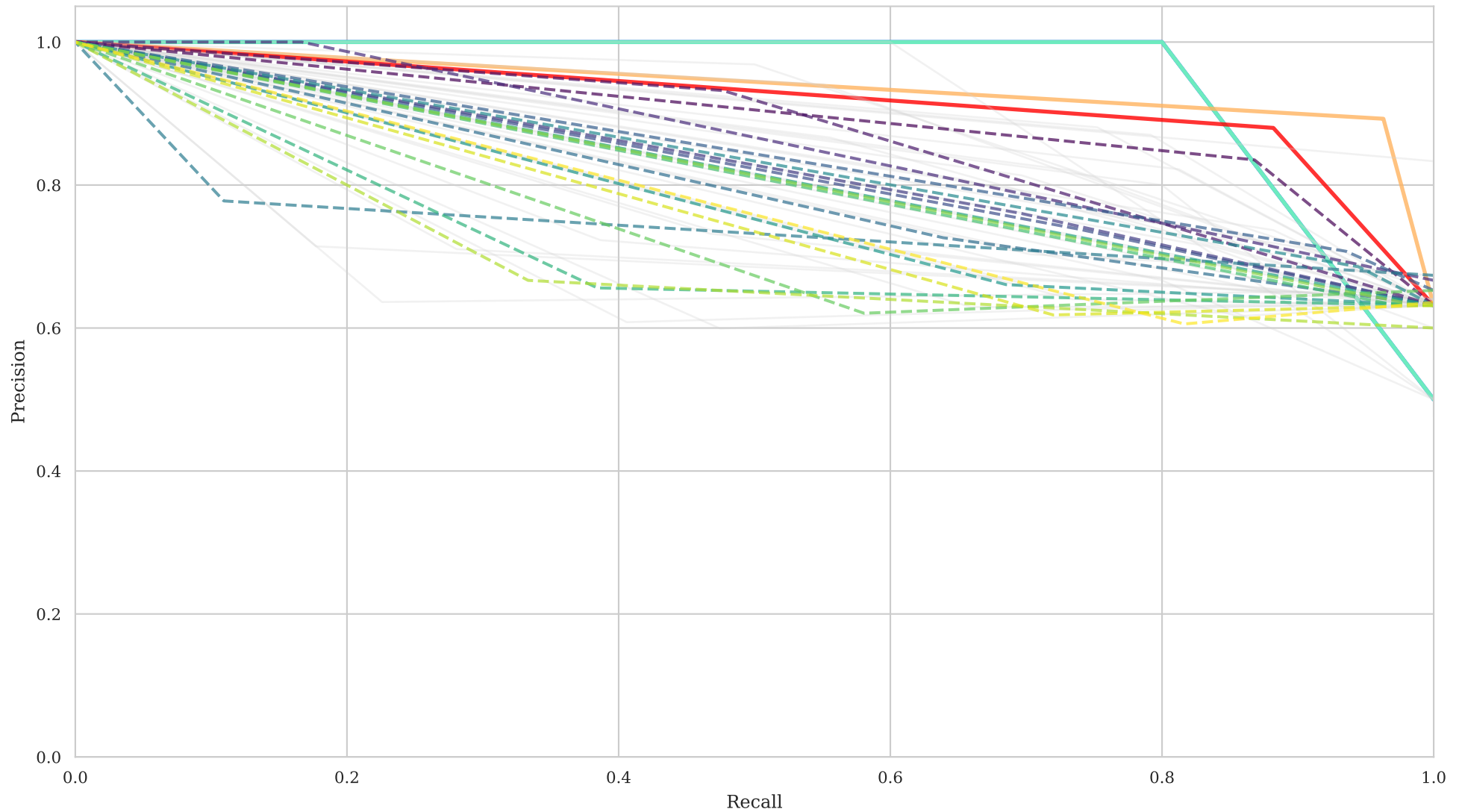


Receiver Operating Characteristic (ROC) Curve



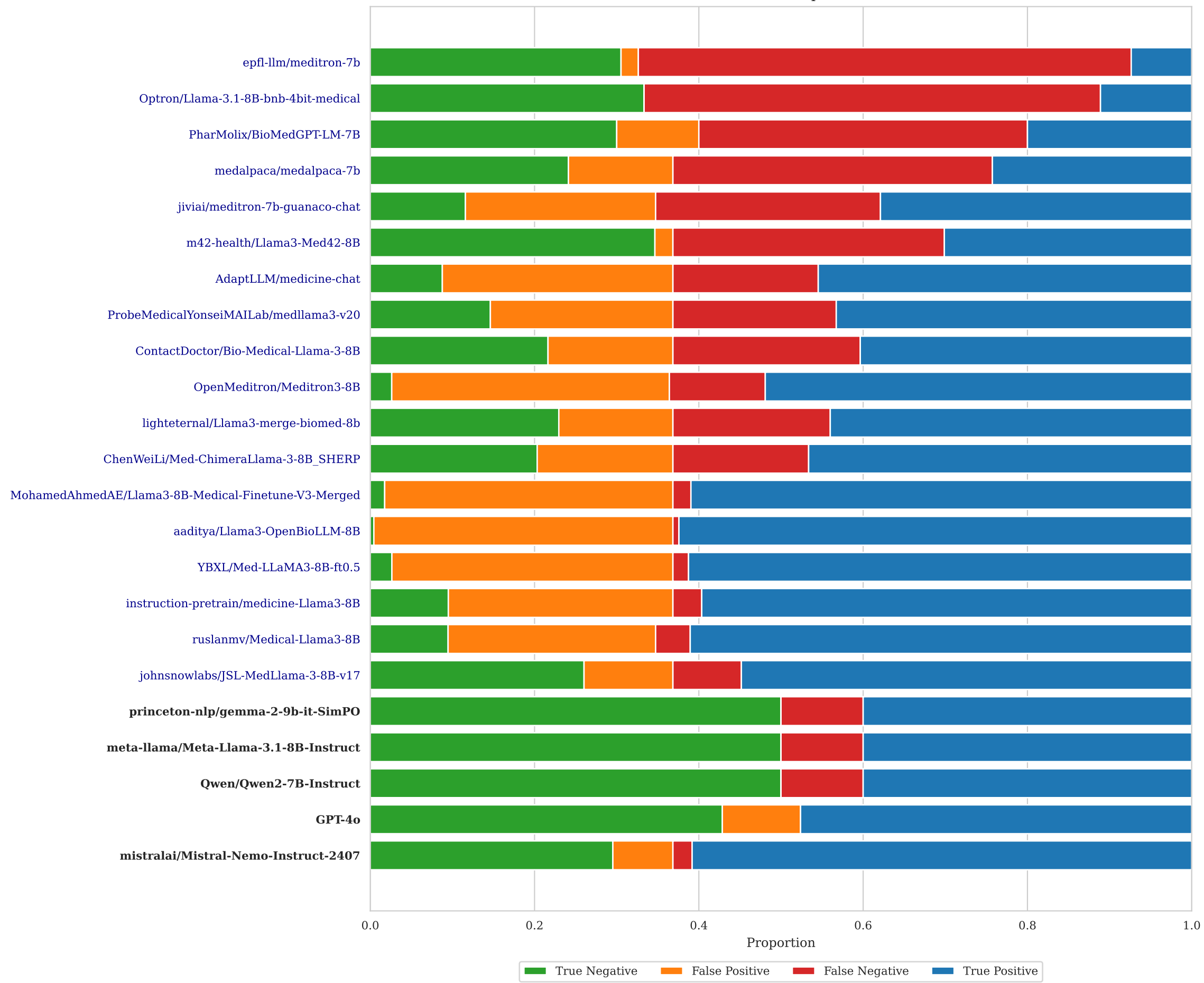
GPT-4o (AUC=0.91)	ChenWeiLi/Med-ChimeraLlama-3-8B_SHERP (AUC=0.65)	YBXL/Med-LLaMA3-8B-ft0.5 (AUC=0.52)
Qwen/Qwen2-7B-Instruct (AUC=0.90)	ContactDoctor/Bio-Medical-Llama-3-8B (AUC=0.61)	medalpaca/medalpaca-7b (AUC=0.52)
meta-llama/Meta-Llama-3.1-8B-Instruct (AUC=0.90)	ruslanmv/Medical-Llama3-8B (AUC=0.60)	MohamedAhmedAE/Llama3-8B-Medical-Finetune-V3-Merged (AUC=0.51)
princeton-nlp/gemma-2-9b-it-SimPO (AUC=0.90)	instruction-pretrain/medicine-Llama3-8B (AUC=0.60)	aaditya/Llama3-OpenBioLLM-8B (AUC=0.50)
mistralai/Mistral-Nemo-Instruct-2407 (AUC=0.88)	Optron/Llama-3.1-8B-bnb-4bit-medical (AUC=0.58)	AdaptLLM/medicine-chat (AUC=0.48)
johnsnowlabs/JSL-MedLlama-3-8B-v17 (AUC=0.79)	PharMolix/BioMedGPT-LM-7B (AUC=0.54)	jiviai/meditron-7b-guanaco-chat (AUC=0.46)
m42-health/Llama3-Med42-8B (AUC=0.71)	ProbeMedicalYonseiMAILab/medllama3-v20 (AUC=0.54)	OpenMeditron/Meditron3-8B (AUC=0.44)
lighteternal/Llama3-merge-biomed-8b (AUC=0.66)	epfl-llm/meditron-7b (AUC=0.52)	

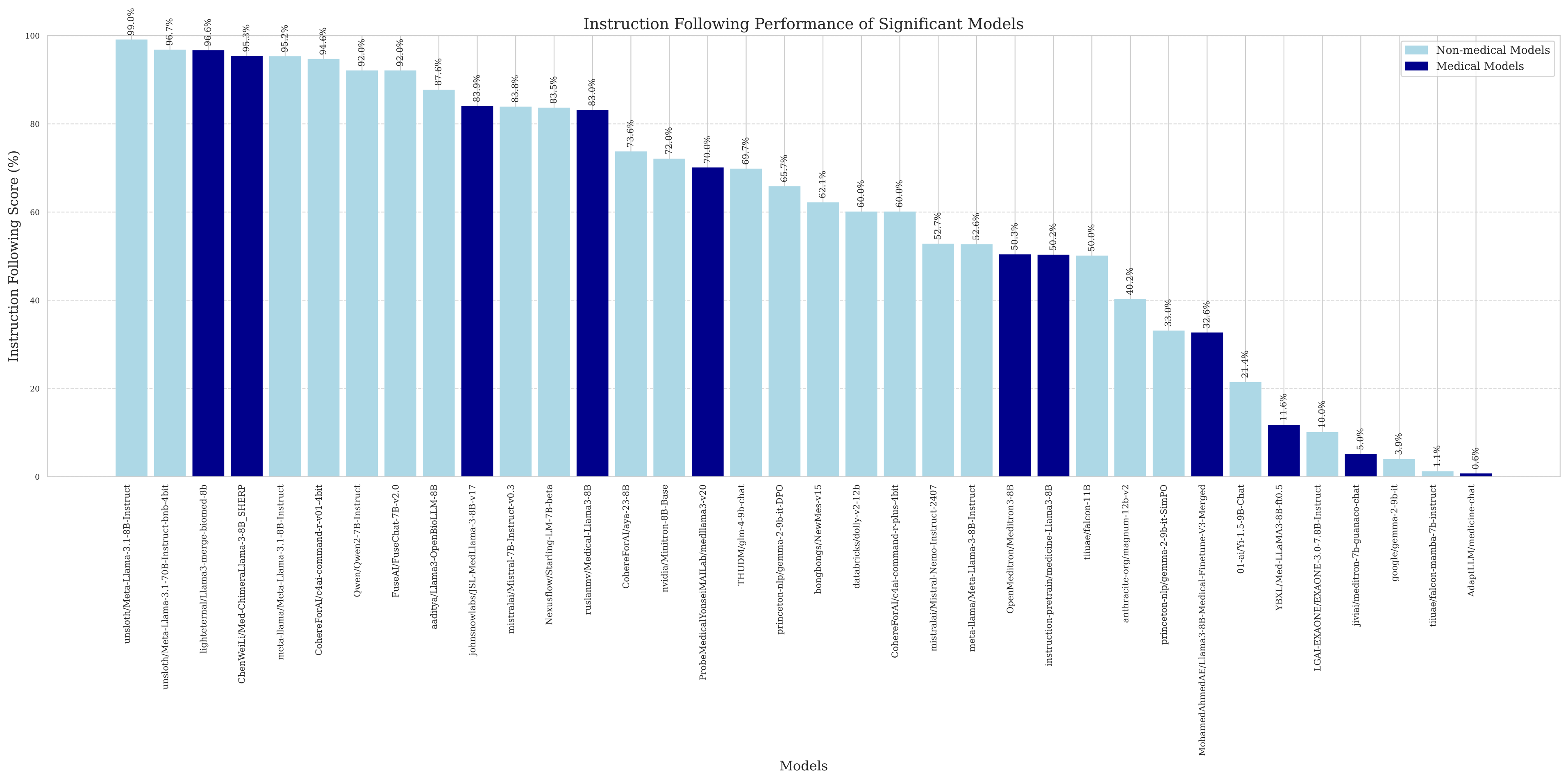
Precision-Recall Curve



Qwen/Qwen2-7B-Instruct (AP=0.90)	lighteternal/Llama3-merge-biomed-8b (AP=0.72)	medalpaca/medalpaca-7b (AP=0.64)
meta-llama/Meta-Llama-3.1-8B-Instruct (AP=0.90)	ChenWeiLi/Med-ChimeraLlama-3-8B_SHERP (AP=0.71)	MohamedAhmedAE/Llama3-8B-Medical-Finetune-V3-Merged (AP=0.63)
princeton-nlp/gemma-2-9b-it-SimPO (AP=0.90)	ruslanmv/Medical-Llama3-8B (AP=0.70)	jiviai/meditron-7b-guanaco-chat (AP=0.63)
mistralai/Mistral-Nemo-Instruct-2407 (AP=0.88)	ContactDoctor/Bio-Medical-Llama-3-8B (AP=0.69)	aaditya/Llama3-OpenBioLLM-8B (AP=0.63)
mistralai/Mistral-7B-Instruct-v0.3 (AP=0.85)	epfl-llm/meditron-7b (AP=0.69)	PharMolix/BioMedGPT-LM-7B (AP=0.62)
johnsnowlabs/JSL-MedLlama-3-8B-v17 (AP=0.81)	instruction-pretrain/medicine-Llama3-8B (AP=0.68)	AdaptLLM/medicine-chat (AP=0.62)
m42-health/Llama3-Med42-8B (AP=0.77)	ProbeMedicalYonseiMAILab/medllama3-v20 (AP=0.65)	OpenMeditron/Meditron3-8B (AP=0.61)
Optron/Llama-3.1-8B-bnb-4bit-medical (AP=0.72)	YBXL/Med-LLaMA3-8B-ft0.5 (AP=0.64)	

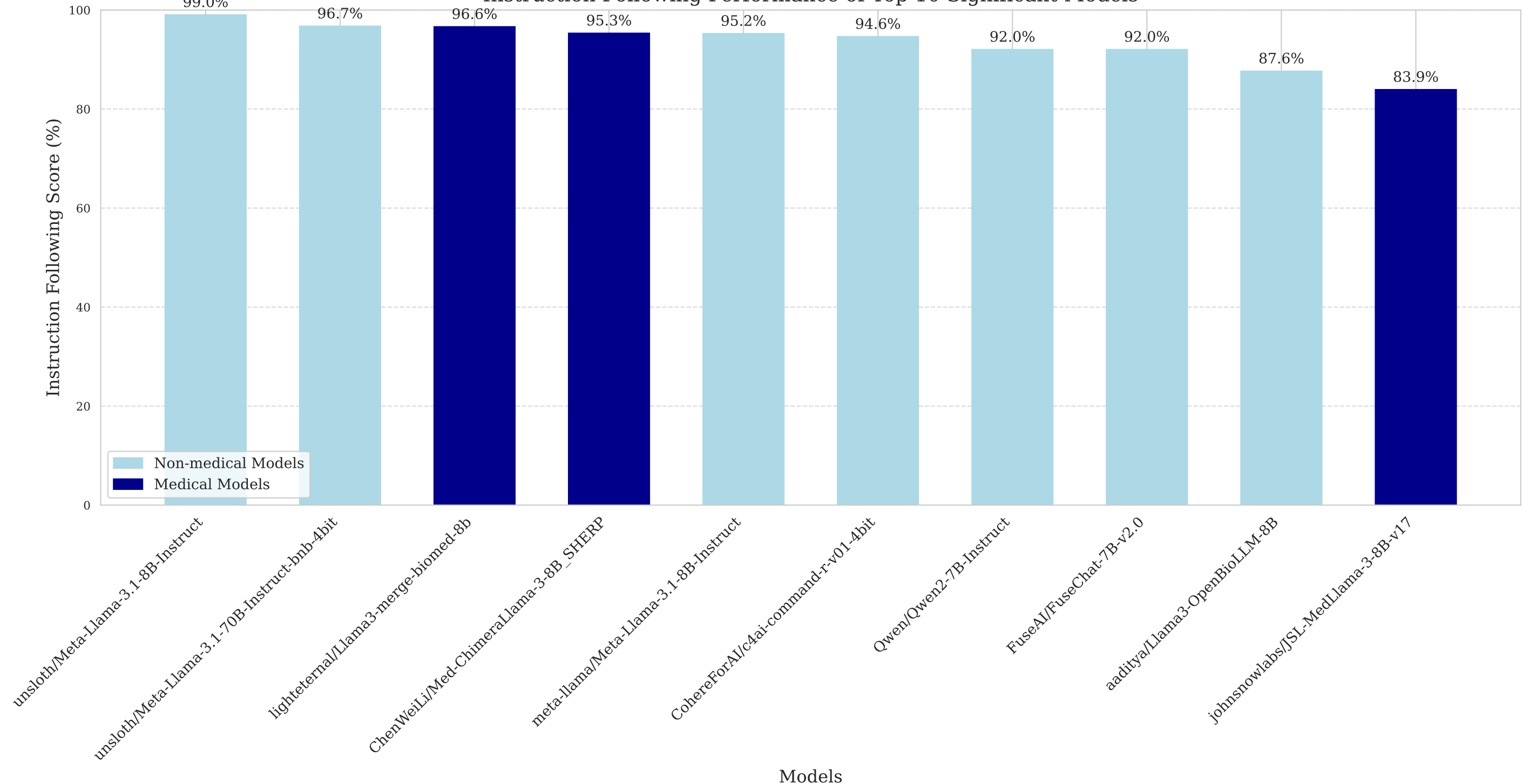
Error Rate Comparison







Instruction Following Performance of Top 10 Significant Models



Top 15 Models by Composite Score: F1 Score vs Instruction Following

