Aligning Language Model Benchmarks with Pairwise Preferences

Marco Gutierrez[1], Cynthia Leng[1], Jonathan Schwarz[2], Ahmed Alaa[3], Tom Hartvigsen[1]

[1] University of Virginia [2] Imperial College London

[3] University of California, Berkeley and UCSF

NEURAL INFORMATION PROCESSING SYSTEMS

Overview

Benchmarks are used to make strong claims about the knowledge they encode



Article

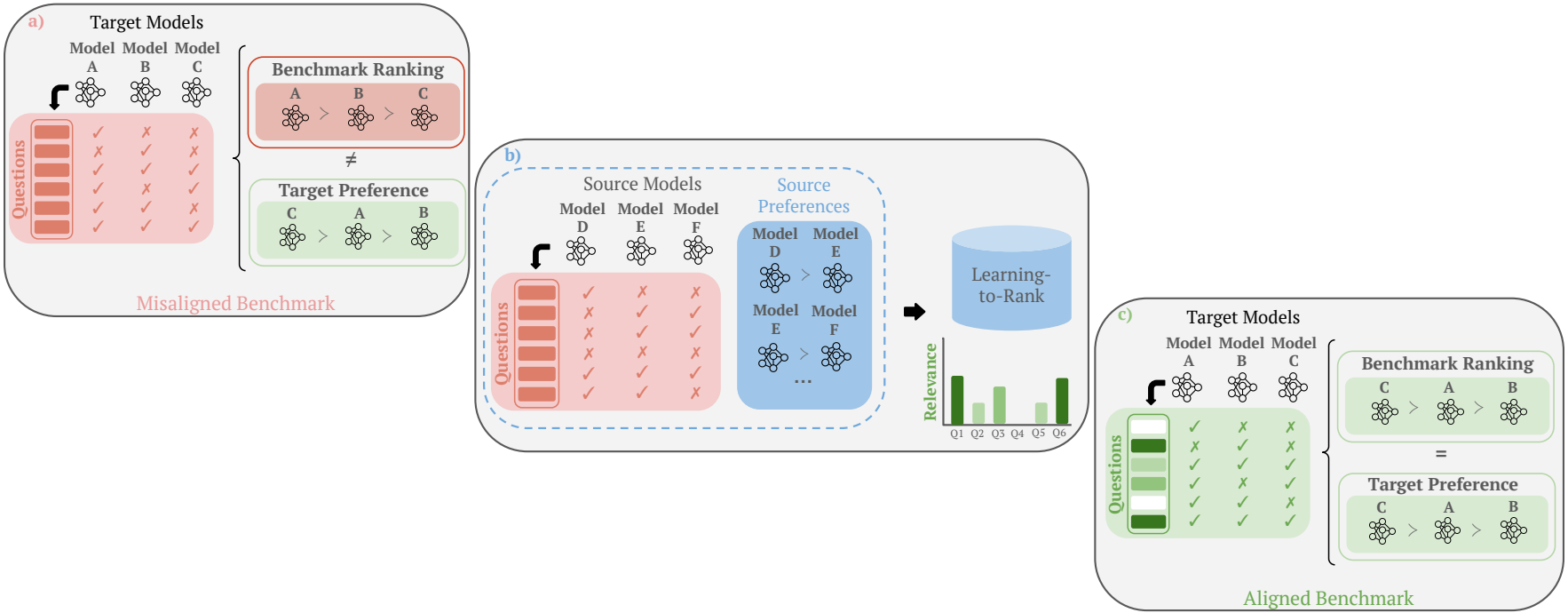
Large language models encode clinical knowledge

Individual benchmarks not always capture the real-life tasks they represent (Alaa, Hartvigsen, et al. 2025)

	MedQA	Real-world data	
	Accuracy	*Accuracy*	*α*
Llama 3	0.54	0.48	0.56
GPT-4	0.71	0.28	0.29
Chimera Llama	0.60	0.45	0.48
Biomerge	0.57	0.36	0.49
Orpomed	0.49	0.24	0.38
JSL MedLlama	0.61	0.37	0.49
PMY MedLLama	0.75	0.36	0.45

Table 1. Predictive validity of the MedQA benchmark.

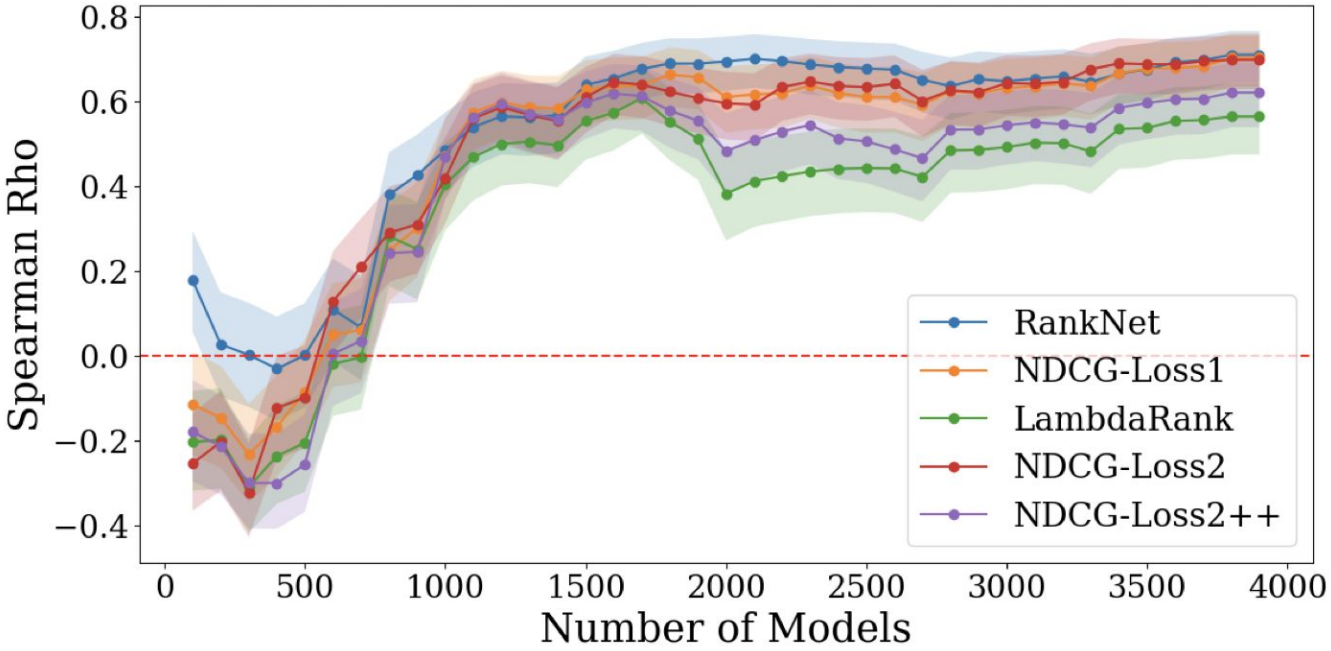
Benchmark Alignment



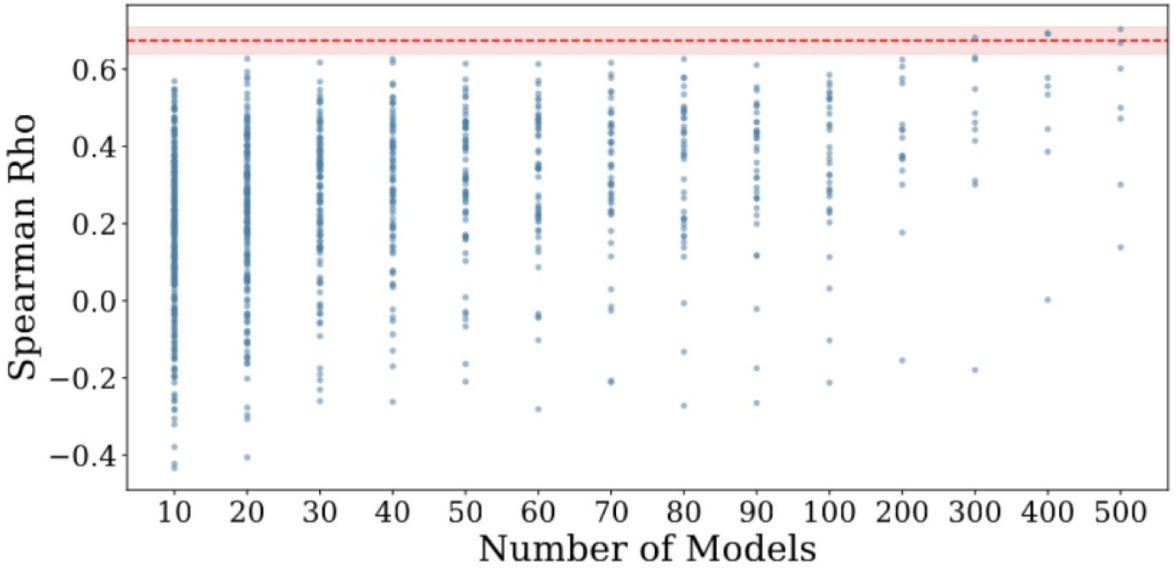
Aligned benchmarks trained on smaller models generalize to larger, unseen models

Target Models	Method	Helpfulness				Honesty			
		Acc_{pair}		ρ		Acc_{pair}		ρ	
		ArmoRM	GPT2	ArmoRM	GPT2	ArmoRM	DPA	ArmoRM	DPA
70B+	No Alignment	0.542 ± 0.010	0.480 ± 0.010	0.154 ± 0.165	0.044 ± 0.251	0.559 ± 0.010	0.567 ± 0.010	0.199 ± 0.164	0.188 ± 0.164
	Random	0.538 ± 0.010	0.468 ± 0.010	0.087 ± 0.161	−0.106 ± 0.166	0.621 ± 0.010	0.605 ± 0.010	0.280 ± 0.145	0.237 ± 0.150
	Individual	0.675 ± 0.009	**0.621 ± 0.010**	0.492 ± 0.136	0.306 ± 0.157	0.694 ± 0.009	0.679 ± 0.012	0.536 ± 0.129	0.490 ± 0.136
	MetaBench	0.564 ± 0.007	0.484 ± 0.007	0.200 ± 0.158	−0.028 ± 0.165	0.581 ± 0.007	0.575 ± 0.007	0.246 ± 0.155	0.216 ± 0.157
	TinyBenchmarks	0.540 ± 0.010	0.488 ± 0.010	0.147 ± 0.161	−0.024 ± 0.165	0.554 ± 0.010	0.563 ± 0.010	0.187 ± 0.159	0.187 ± 0.159
	BenchAlign	**0.778 ± 0.008**	0.620 ± 0.010	**0.707 ± 0.074**	**0.333 ± 0.139**	**0.774 ± 0.008**	**0.746 ± 0.009**	**0.706 ± 0.074**	**0.659 ± 0.084**
30B+	No Alignment	0.605 ± 0.005	0.477 ± 0.005	0.331 ± 0.114	0.052 ± 0.226	0.618 ± 0.005	0.609 ± 0.005	0.368 ± 0.111	0.319 ± 0.115
	Random	0.551 ± 0.005	0.452 ± 0.005	0.108 ± 0.120	−0.148 ± 0.122	0.650 ± 0.005	0.618 ± 0.005	0.407 ± 0.098	0.317 ± 0.106
	Individual	0.678 ± 0.005	0.606 ± 0.005	0.501 ± 0.098	0.273 ± 0.118	0.692 ± 0.005	0.675 ± 0.005	0.537 ± 0.094	0.480 ± 0.100
	MetaBench	0.606 ± 0.004	0.497 ± 0.004	0.328 ± 0.110	0.007 ± 0.123	0.613 ± 0.004	0.604 ± 0.004	0.346 ± 0.108	0.302 ± 0.112
	TinyBenchmarks	0.605 ± 0.005	0.492 ± 0.005	0.323 ± 0.110	−0.016 ± 0.123	0.615 ± 0.005	0.602 ± 0.005	0.354 ± 0.108	0.299 ± 0.112
	BenchAlign	**0.765 ± 0.005**	**0.637 ± 0.005**	**0.710 ± 0.056**	**0.387 ± 0.100**	**0.773 ± 0.005**	**0.763 ± 0.005**	**0.730 ± 0.053**	**0.713 ± 0.055**
13B+	No Alignment	0.583 ± 0.002	0.531 ± 0.002	0.245 ± 0.065	0.094 ± 0.067	0.588 ± 0.002	0.577 ± 0.002	0.251 ± 0.064	0.231 ± 0.065
	Random	0.558 ± 0.002	0.537 ± 0.002	0.125 ± 0.066	0.083 ± 0.067	0.603 ± 0.002	0.600 ± 0.002	0.280 ± 0.061	0.272 ± 0.061
	Individual	0.660 ± 0.002	0.608 ± 0.002	0.430 ± 0.057	0.278 ± 0.063	0.661 ± 0.002	0.669 ± 0.002	0.432 ± 0.057	0.451 ± 0.055
	MetaBench	0.584 ± 0.001	0.530 ± 0.001	0.243 ± 0.063	0.089 ± 0.067	0.586 ± 0.001	0.579 ± 0.001	0.248 ± 0.063	0.229 ± 0.064
	TinyBenchmarks	0.589 ± 0.002	0.543 ± 0.002	0.259 ± 0.063	0.128 ± 0.066	0.592 ± 0.002	0.583 ± 0.002	0.267 ± 0.063	0.243 ± 0.063
	BenchAlign	**0.741 ± 0.001**	**0.696 ± 0.002**	**0.674 ± 0.035**	**0.552 ± 0.045**	**0.748 ± 0.001**	**0.749 ± 0.001**	**0.686 ± 0.034**	**0.691 ± 0.034**

Naive benchmark alignment requires a large number of smaller-sized models



Informative subsets enable alignment with as few as ~20 models



Check our paper + code here!

Paper

Code