Table 1: **Performance comparison of different model categories on linear and non-linear tasks.** The table shows the ranking performance of LLMs, supervised models, and unsupervised baselines across different task types. Lower ranks indicate better performance.

Model Category	Model Name	Linear Tasks	Non-linear Tasks	Overall Ranking
LLMs	Claude 3 Opus	4.75	3.0	3.88
	GPT-4	11.5	4.4	7.95
	DBRX	20.25	16.2	18.23
Supervised Models	MLP Deep 2	1.25	20.2	10.73
	MLP Deep 3	2.5	18.8	10.65
	Gradient Boosting	16.5	8.0	12.25
	Linear Reg + Poly	27.0	18.0	22.5
	Random Forest	5.0	21.4	13.2
Unsupervised Baselines	Average	30.25	22.0	26.13
	Random	13.5	34.0	23.75