Table 1: **Performance comparison of different model categories on 9 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

The study evaluates 33 models—including large language models (LLMs), supervised algorithms, and unsupervised baselines—on synthetic regression benchmarks designed to test their ability to infer linear and non-linear relationships. All datasets are synthetically generated with controlled distributions to ensure reproducibility and avoid pre-exposure to LLMs during pre-training.

The linear tasks focus on sparse feature identification, such as datasets where only 1 out of 3 features is informative (e.g., $y = \beta x_1 + \epsilon$, with x_2, x_3 as noise). For non-linear tasks, the benchmarks include classical functions like Friedman#1 ($y = 10\sin(\pi x_0 x_1) + 20(x_2 - 0.5)^2 + 10x_3 + 5x_4 + \epsilon$) and custom-designed equations such as Original#1 ($y = x + 10\sin(5\pi x/100) + 10\cos(6\pi x/100)$), which combines linear trends with periodic oscillations.

Tab 1 which is summarized ranking results of, reveal that LLMs (e.g., Claude 3 Opus, GPT-4) achieve strong performance across tasks by leveraging in-context learning (ICL), even outperforming specialized supervised methods like Gradient Boosting in non-linear scenarios. For instance, Claude 3 Opus excels in approximating complex functional forms (e.g., Friedman#2: $y = \sqrt{x_1^2 + (x_2 x_3 - 1/(x_2 x_4))^2} + \epsilon$) without gradient updates, though it occasionally struggles with extreme outliers. Supervised models like deep MLPs dominate sparse linear tasks but falter on non-linear interactions, while Linear Regression with Polynomial Features shows competitive performance on certain synthetic curves. Unsupervised baselines (e.g., random guessing) consistently underperform, highlighting the sophistication of ICL.