LLM Performance vs Random Baseline
Statistical Comparison Summary

Model	Performance	Improvement	P-value
Gemini-2.0	61.4%	123×	< 0.001
DeepThink-R1	43.1%	86×	< 0.001
Mistral-L2	42.2%	84×	< 0.001
GPT-o1-Pro	41.8%	84×	< 0.001
GPT-o1	36.6%	73×	< 0.001
Claude-3.5	35.7%	71×	< 0.001
GPT-4o	35.7%	71×	< 0.001
Llama-3.1	32.2%	64×	< 0.001

Random Baseline: Mean = 0.5%, Max = 10.0%
All improvements statistically significant ($p < 0.001$)