Prompt 1 (P1): Accuracy

Model	Score
gemma-2-27b-it	0.846
llama-3.1-405b-instruct	0.742
llama-3.3-70b-instruct	0.735
command-r-08-2024	0.704
llama-4-maverick	0.678
phi-4	0.566
mistral-medium-3	0.564
deepseek-chat-v3.1	0.541
kimi-k2-0905	0.465
sonar	0.454
wizardlm-2-8x22b	0.449
claude-sonnet-4	0.416
gpt-4.1	0.416
qwen-2.5-72b-instruct	0.414
gpt-5.1	0.387
nova-lite-v1	0.353
mixtral-8x7b-instruct	0.337
gemini-2.5-flash	0.291
nova-micro-v1	0.246
grok-3-mini-beta	0.143
o4-mini-high	0.136
o1-mini	0.134
qwen2.5-vl-32b-instruct:free	0.000