Top 1 Performance by Task

Model	gb1	trpb	syn-3bfo	aav	gfp
Baseline	0.86	0.76	0.50	0.63	1.00
Claude-Sonnet-4-5	0.65	0.68	0.80	0.73	1.02
DeepSeek	0.83	0.84	1.00	0.72	0.96
GPT-5	0.76	0.83	0.97	0.65	1.07
GPT-5-chat-latest	0.83	0.85	0.87	0.73	1.02
GPT5-mini	0.76	0.53	1.02	0.66	0.96