

Strengths

	Experts	GPT 5.1	GPT 5.1*	Flash 2.5	Flash 2.5*	Grok 4.1	Grok 4.1*	Sonar	Sonar*	Sonnet 4.5	Sonnet 4.5*
Experts	-1.00	0.90	0.86	0.89	0.86	0.89	0.84	0.93	0.87	0.92	0.88
GPT 5.1	-0.90	1.00	0.92	0.91	0.87	0.96	0.89	0.94	0.86	0.91	0.89
GPT 5.1*	-0.86	0.92	1.00	0.93	0.85	0.94	0.88	0.91	0.86	0.90	0.90
Flash 2.5	-0.89	0.91	0.93	1.00	0.82	0.96	0.91	0.94	0.84	0.95	0.88
Flash 2.5*	-0.86	0.87	0.85	0.83	1.00	0.83	0.80	0.87	0.94	0.86	0.94
Grok 4.1	-0.88	0.96	0.93	0.96	0.82	1.00	0.92	0.93	0.83	0.92	0.87
Grok 4.1*	-0.84	0.89	0.88	0.91	0.79	0.92	1.00	0.87	0.85	0.92	0.82
Sonar	-0.93	0.94	0.91	0.94	0.86	0.93	0.87	1.00	0.86	0.94	0.90
Sonar*	-0.87	0.87	0.86	0.85	0.94	0.84	0.85	0.86	1.00	0.90	0.95
Sonnet 4.5	-0.92	0.91	0.90	0.95	0.85	0.93	0.92	0.94	0.90	1.00	0.89
Sonnet 4.5*	-0.88	0.90	0.90	0.89	0.94	0.88	0.83	0.90	0.96	0.90	1.00

Weaknesses

	Experts	GPT 5.1	GPT 5.1*	Flash 2.5	Flash 2.5*	Grok 4.1	Grok 4.1*	Sonar	Sonar*	Sonnet 4.5	Sonnet 4.5*
Experts	-1.00	0.92	0.91	0.94	0.84	0.91	0.91	0.90	0.88	0.94	0.86
GPT 5.1	-0.92	1.00	0.96	0.93	0.81	0.97	0.97	0.88	0.86	0.95	0.83
GPT 5.1*	-0.91	0.96	1.00	0.94	0.82	0.96	0.98	0.90	0.85	0.95	0.84
Flash 2.5	-0.94	0.93	0.94	1.00	0.78	0.94	0.93	0.86	0.83	0.94	0.80
Flash 2.5*	-0.84	0.82	0.83	0.79	1.00	0.82	0.82	0.91	0.93	0.85	0.96
Grok 4.1	-0.91	0.97	0.96	0.94	0.82	1.00	0.95	0.89	0.85	0.95	0.84
Grok 4.1*	-0.91	0.97	0.98	0.93	0.81	0.95	1.00	0.89	0.84	0.95	0.83
Sonar	-0.91	0.89	0.91	0.87	0.91	0.89	0.90	1.00	0.89	0.91	0.91
Sonar*	-0.89	0.86	0.85	0.83	0.93	0.85	0.85	0.89	1.00	0.89	0.96
Sonnet 4.5	-0.94	0.95	0.95	0.94	0.84	0.95	0.95	0.91	0.89	1.00	0.86
Sonnet 4.5*	-0.86	0.84	0.85	0.81	0.96	0.84	0.84	0.91	0.96	0.87	1.00

Opportunities

	Experts	GPT 5.1	GPT 5.1*	Flash 2.5	Flash 2.5*	Grok 4.1	Grok 4.1*	Sonar	Sonar*	Sonnet 4.5	Sonnet 4.5*
Experts	-1.00	0.83	0.86	0.82	0.84	0.84	0.88	0.83	0.86	0.88	0.85
GPT 5.1	-0.82	1.00	0.95	0.93	0.86	0.90	0.91	0.91	0.83	0.90	0.86
GPT 5.1*	-0.86	0.95	1.00	0.89	0.88	0.92	0.94	0.93	0.86	0.92	0.88
Flash 2.5	-0.82	0.93	0.90	1.00	0.86	0.94	0.92	0.92	0.86	0.91	0.86
Flash 2.5*	-0.85	0.87	0.88	0.87	1.00	0.87	0.90	0.91	0.92	0.89	0.97
Grok 4.1	-0.84	0.90	0.92	0.94	0.87	1.00	0.96	0.92	0.89	0.95	0.86
Grok 4.1*	-0.88	0.92	0.94	0.92	0.89	1.00	0.93	0.91	0.95	0.89	0.89
Sonar	-0.83	0.91	0.93	0.92	0.90	0.92	0.93	1.00	0.89	0.92	0.89
Sonar*	-0.87	0.83	0.86	0.87	0.92	0.89	0.92	0.89	1.00	0.91	0.92
Sonnet 4.5	-0.88	0.90	0.92	0.91	0.89	0.95	0.95	0.92	0.91	1.00	0.89
Sonnet 4.5*	-0.86	0.87	0.89	0.86	0.97	0.87	0.89	0.90	0.92	0.89	1.00

Threats

	Experts	GPT 5.1	GPT 5.1*	Flash 2.5	Flash 2.5*	Grok 4.1	Grok 4.1*	Sonar	Sonar*	Sonnet 4.5	Sonnet 4.5*
Experts	-1.00	0.92	0.92	0.87	0.85	0.90	0.90	0.91	0.89	0.90	0.91
GPT 5.1	-0.92	1.00	0.93	0.92	0.84	0.93	0.93	0.89	0.93	0.93	0.88
GPT 5.1*	-0.92	0.93	1.00	0.92	0.84	0.92	0.93	0.91	0.90	0.92	0.93
Flash 2.5	-0.86	0.92	0.92	1.00	0.79	0.91	0.94	0.87	0.86	0.94	0.86
Flash 2.5*	-0.85	0.85	0.84	0.80	1.00	0.85	0.85	0.90	0.90	0.85	0.91
Grok 4.1	-0.90	0.93	0.92	0.91	0.84	1.00	0.97	0.89	0.90	0.95	0.89
Grok 4.1*	-0.90	0.94	0.93	0.94	0.84	0.97	1.00	0.91	0.90	0.97	0.89
Sonar	-0.92	0.89	0.92	0.88	0.90	0.89	0.91	1.00	0.92	0.92	0.94
Sonar*	-0.89	0.93	0.90	0.86	0.90	0.90	0.90	0.91	1.00	0.90	0.93
Sonnet 4.5	-0.90	0.93	0.92	0.94	0.85	0.95	0.97	0.92	0.90	1.00	0.90
Sonnet 4.5*	-0.92	0.89	0.93	0.86	0.91	0.90	0.90	0.94	0.93	0.90	1.00