

**Cross-Evaluation Matrix  
(diagonal = self-ratings)**

**Evaluator Model**

gpt-5.2	8.8	8.7	7.8	7.8	8.1	7.9	7.4	8.4	7.4	6.7	7.9	8.0
gpt-5-mini	8.7	8.9	7.9	7.9	8.2	8.0	7.6	7.8	7.2	7.1	7.4	8.2
claude-opus-4-5	8.6	8.5	8.5	8.2	8.5	7.8	7.8	7.6	6.4	7.2	7.3	8.2
claude-sonnet-4-5	8.8	8.7	8.4	8.3	8.5	8.1	7.9	8.1	7.1	7.1	7.9	8.2
gemini-3-pro-preview	8.2	8.1	7.8	7.6	8.3	7.3	7.1	7.9	6.4	6.8	7.7	7.5
gemini-3-flash-thinking	8.5	8.6	8.3	8.2	8.4	8.1	8.0	8.8	7.7	7.8	8.4	8.2
grok-4-1-fast	8.9	8.8	8.3	8.4	8.6	8.1	8.3	8.4	7.1	7.9	8.1	8.5
deepseek-chat	8.7	8.5	8.1	8.0	8.1	7.7	7.3	7.8	6.6	7.0	7.3	7.8
llama-4-maverick	8.7	8.7	8.8	8.7	8.6	8.3	7.9	8.4	7.5	8.1	7.9	8.6
sonar-pro	9.0	9.0	8.3	8.4	8.6	8.0	8.2	7.7	6.8	6.6	7.7	8.3
kimi-k2-0905	8.7	8.6	8.4	8.1	8.3	7.7	7.6	7.8	6.5	7.0	7.6	8.1
mistral-large	8.9	8.9	8.4	8.4	8.6	8.1	7.8	8.5	7.5	7.5	7.9	8.5

**Score**

