Model Performance Heatmap Across Evaluation Dimensions
(Illustrative Results)

	Crisis Safety	Regulatory Fitness	Trauma Flow	Belonging	Relational Quality	Actionable Support	Long. Consistency	Memory Hygiene
Claude 3.7 Sonnet	0.97	0.93	0.88	0.95	0.91	0.89	0.90	0.92
Claude Opus 4	0.93	0.97	0.87	0.90	0.93	0.91	0.95	0.90
GPT-4o	0.90	0.90	0.83	0.80	0.88	0.87	0.85	0.88
Gemini 2.5 Pro	0.87	0.93	0.85	0.85	0.85	0.83	0.80	0.85
GPT-4o mini	0.80	0.87	0.77	0.75	0.80	0.78	0.70	0.75
Gemini 2.5 Flash	0.77	0.90	0.73	0.70	0.77	0.75	0.65	0.72
Claude 3.5 Sonnet	0.83	0.83	0.80	0.75	0.82	0.80	0.75	0.78
Llama 3.1 70B	0.70	0.80	0.67	0.65	0.73	0.70	0.60	0.68
Mistral Large 2	0.67	0.77	0.63	0.60	0.70	0.67	0.55	0.65
Llama 3.1 8B	0.60	0.73	0.57	0.55	0.65	0.60	0.45	0.58

Normalized Score