Domain-Specific Model Performance
(Mean Time-to-Failure)

	STEM	Medical	Business	Legal	Humanities
CARG	7.8	7.6	7.5	7.9	7.7
Gemini-2.5	3.1	7.3	7.2	7.4	7.5
GPT-4	7.2	6.8	6.9	7.1	7.9
Qwen-Max	6.8	6.6	6.5	6.7	6.9
Claude-3.5	7.2	7.0	6.9	7.1	7.1

Models — Subject Domains

Performance (turns survived)