pass@3 by Model, Domain, and Temperature

Model (T)	collab	mail_rag_phishing	output_handling
anthropic/claude-sonnet-4-5-20250929 (T=0)	1.00	0.20	0.67
anthropic/claude-sonnet-4-5-20250929 (T=0.5)	0.88	0.20	0.50
anthropic/claude-sonnet-4-5-20250929 (T=1)	1.00	0.00	0.33
gpt-3.5-turbo (T=0)	0.71	0.05	0.00
gpt-3.5-turbo (T=0.5)	0.67	0.20	0.00
gpt-3.5-turbo (T=1)	0.71	0.05	0.08
gpt-4.1 (T=0)	0.88	0.20	0.67
gpt-4.1 (T=0.5)	1.00	0.25	0.67
gpt-4.1 (T=1)	0.88	0.25	0.50
gpt-4o (T=0)	1.00	0.30	0.42
gpt-4o (T=0.5)	1.00	0.25	0.42
gpt-4o (T=1)	0.88	0.10	0.67
gpt-4o-mini (T=0)	1.00	0.05	0.42
gpt-4o-mini (T=0.5)	1.00	0.05	0.33
gpt-4o-mini (T=1)	0.88	0.20	0.33