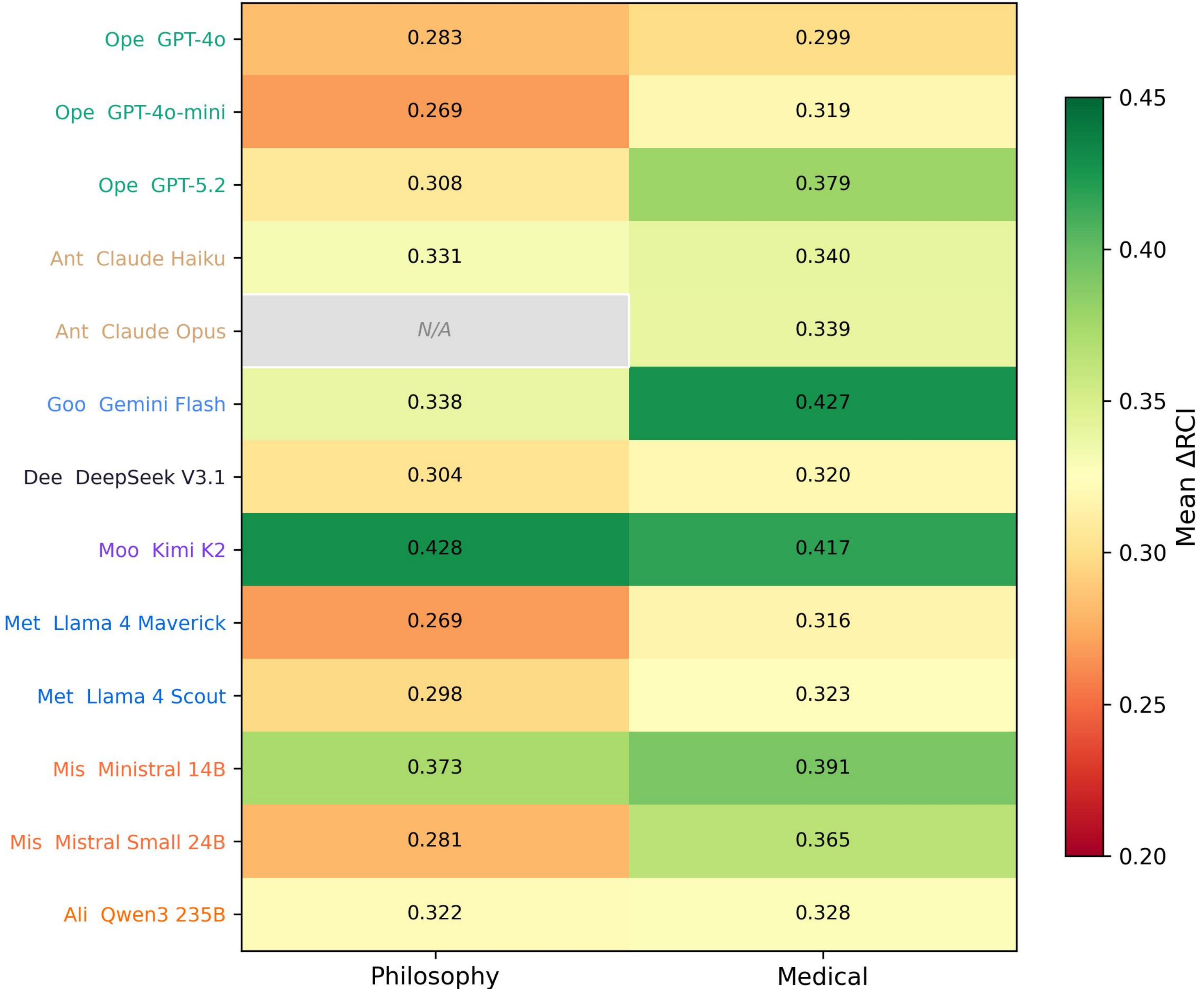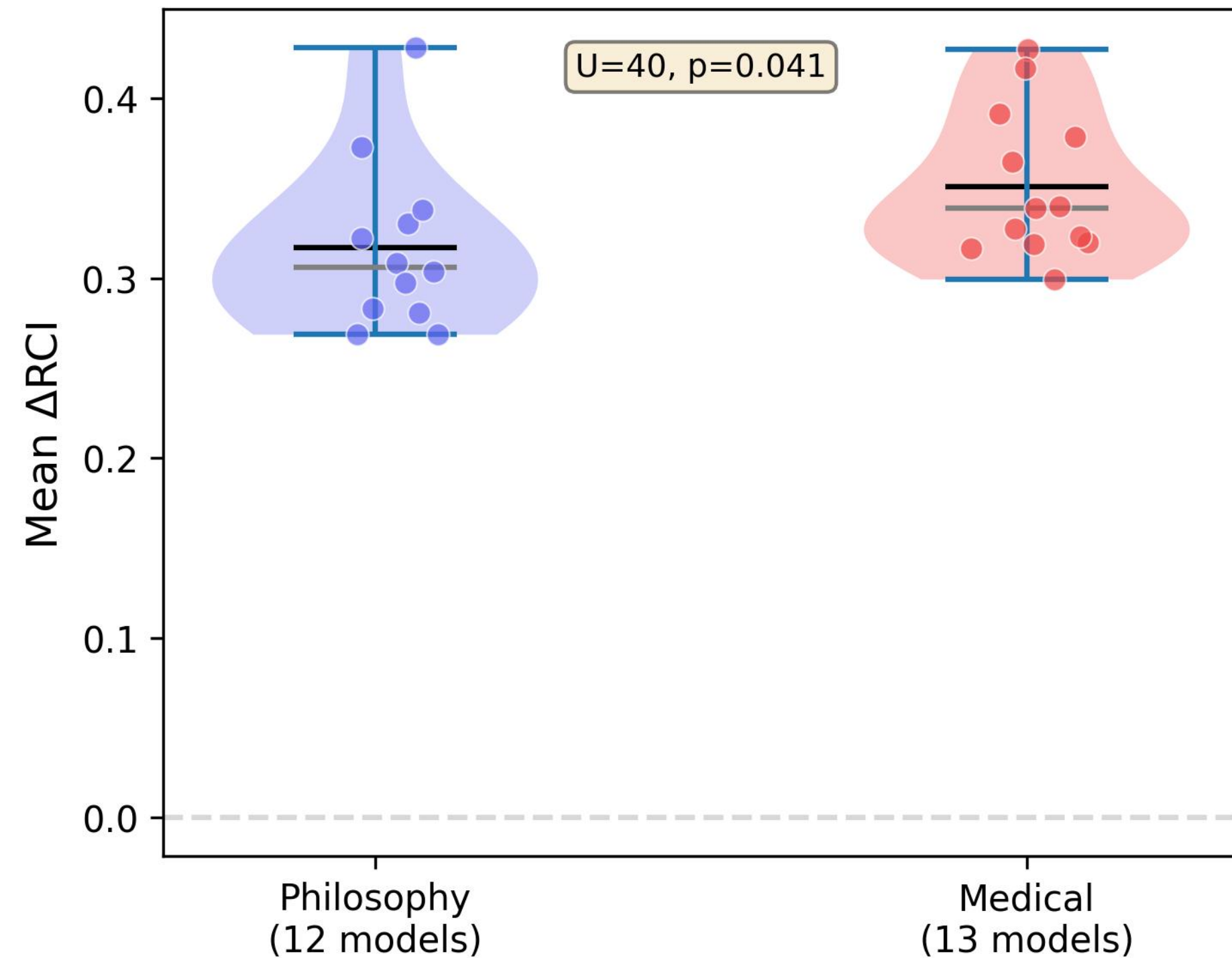Paper 2 Dataset: Mean ΔRCI by Model and Domain
(14 models, 25 model-domain runs, 50 trials each)

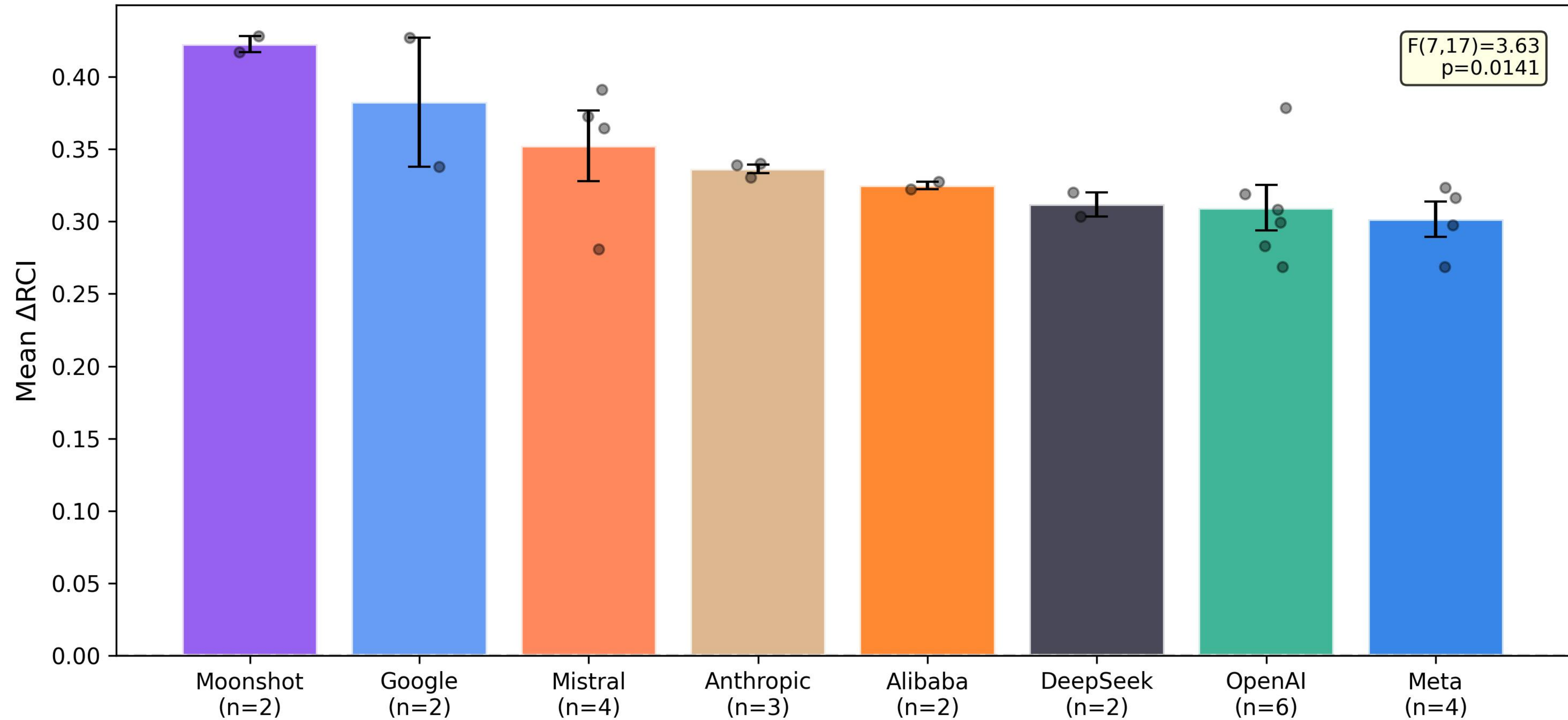| Model	Philosophy	Medical
Ope GPT-4o	0.283	0.299
Ope GPT-4o-mini	0.269	0.319
Ope GPT-5.2	0.308	0.379
Ant Claude Haiku	0.331	0.340
Ant Claude Opus	N/A	0.339
Goo Gemini Flash	0.338	0.427
Dee DeepSeek V3.1	0.304	0.320
Moo Kimi K2	0.428	0.417
Met Llama 4 Maverick	0.269	0.316
Met Llama 4 Scout	0.298	0.323
Mis Ministral 14B	0.373	0.391
Mis Mistral Small 24B	0.281	0.365
Ali Qwen3 235B	0.322	0.328

Domain Comparison

U=40, p=0.041

Mean ΔRCI

Philosophy (12 models)

Medical (13 models)

**Cross-Domain Comparison
(12 models in both domains)**

Mean ΔRCI

Philosophy

Medical

Claude Haiku, DeepSeek V3.1, GPT-4o, GPT-4o-mini, GPT-5.2, Gemini Flash, Kimi K2, Llama 4 Maverick, Llama 4 Scout, Ministral 14B, Mistral Small 24B, Qwen3 235B
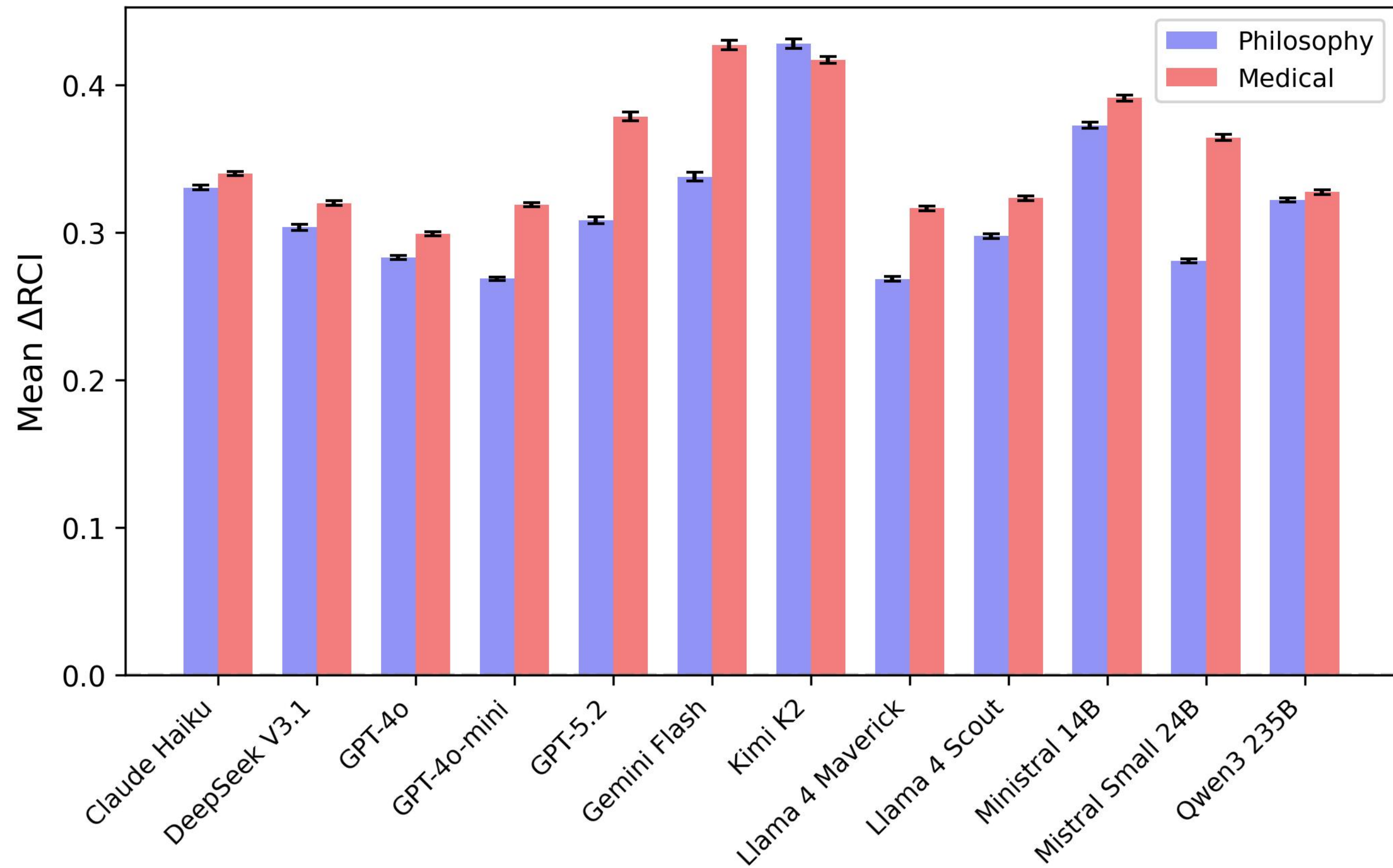
Vendor Signatures: Context Sensitivity by Model Provider

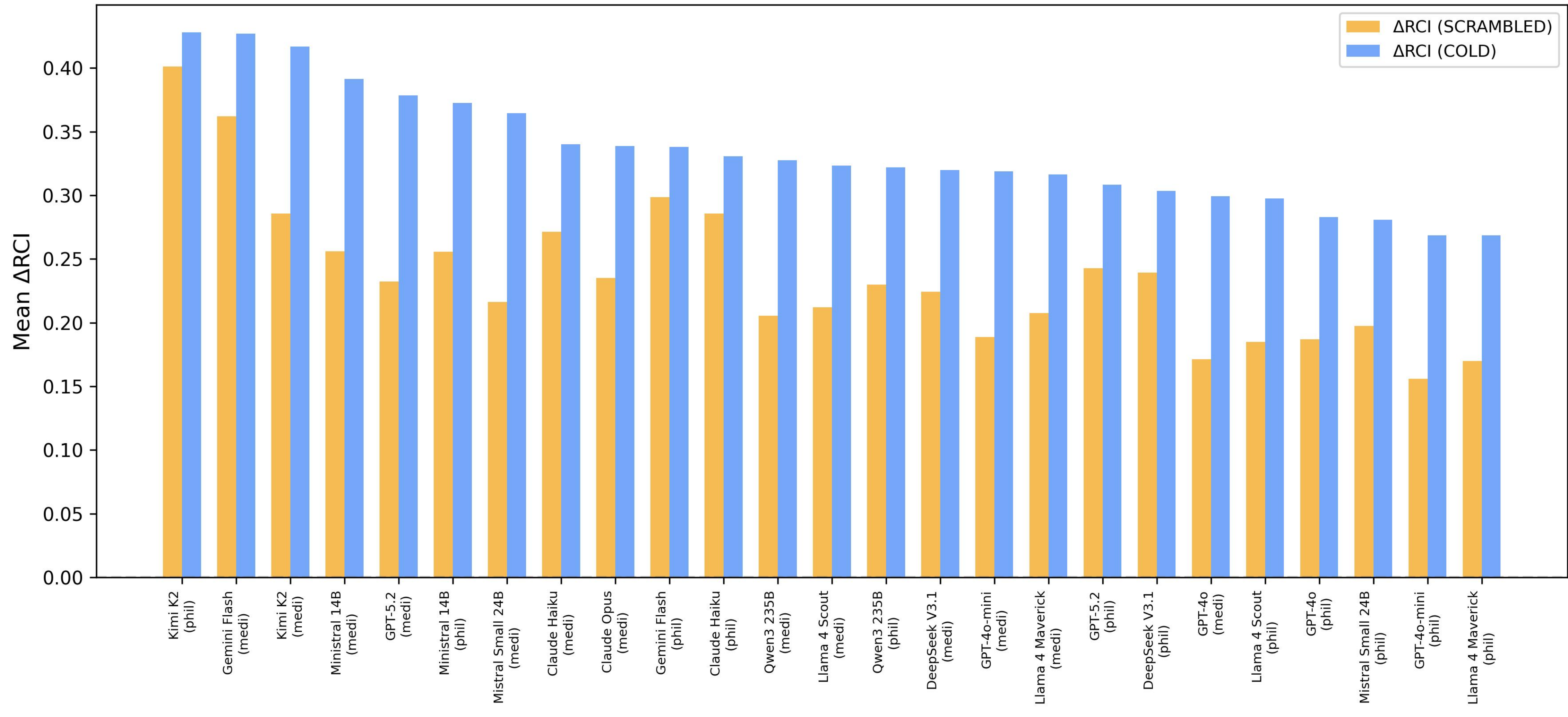$F(7,17)=3.63$
$p=0.0141$
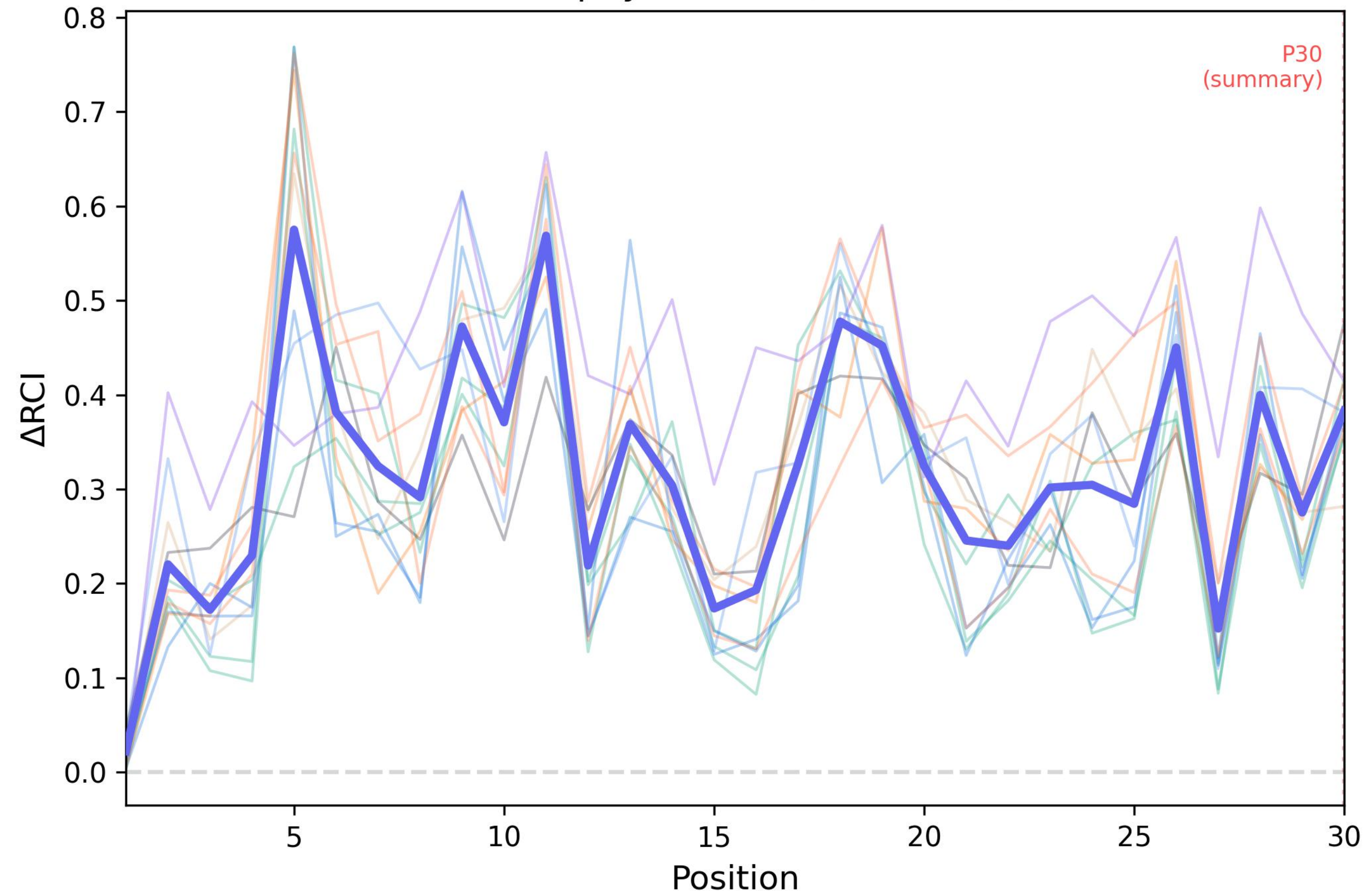
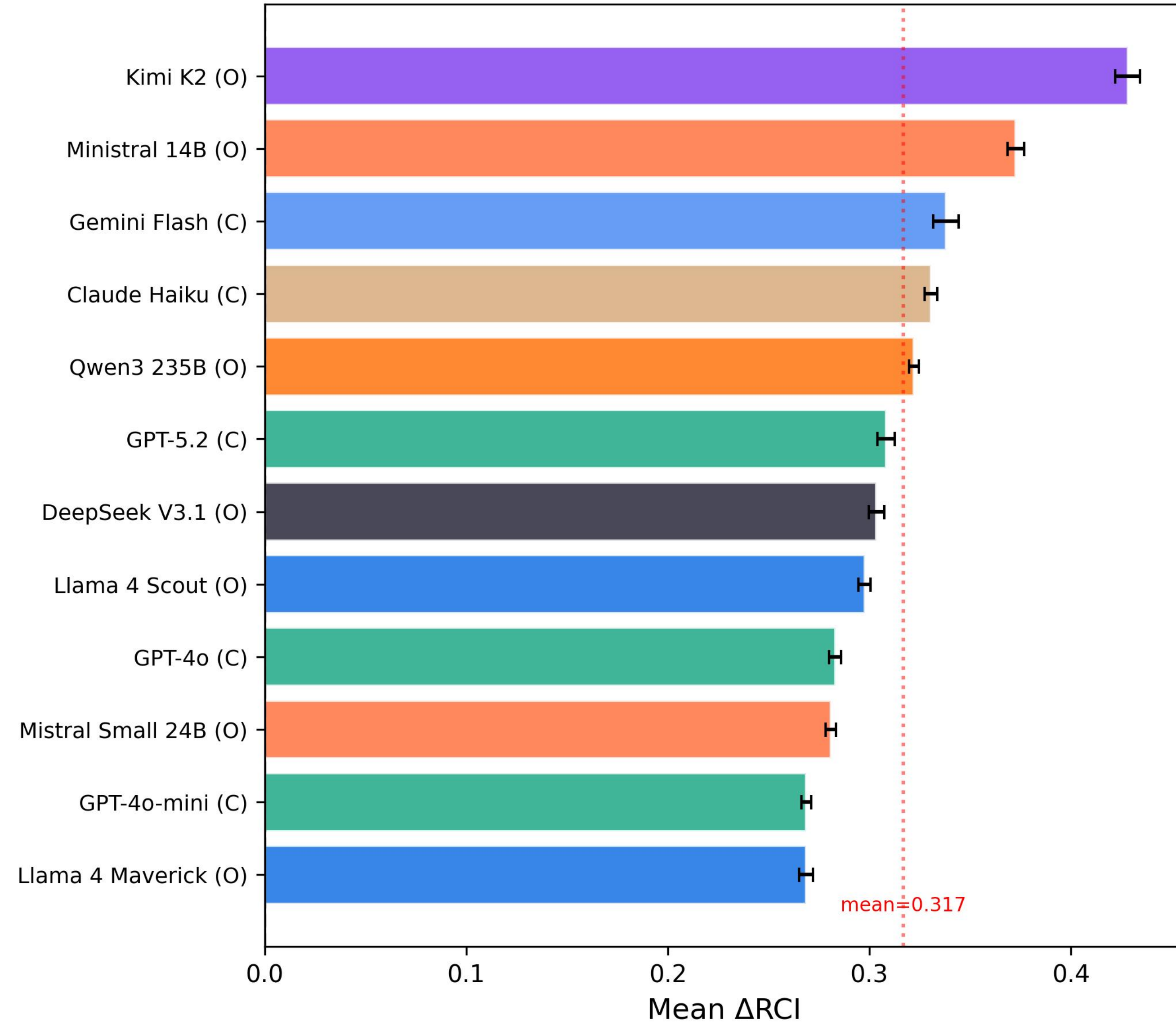Philosophy Domain (12 models) / Medical Domain (13 models)
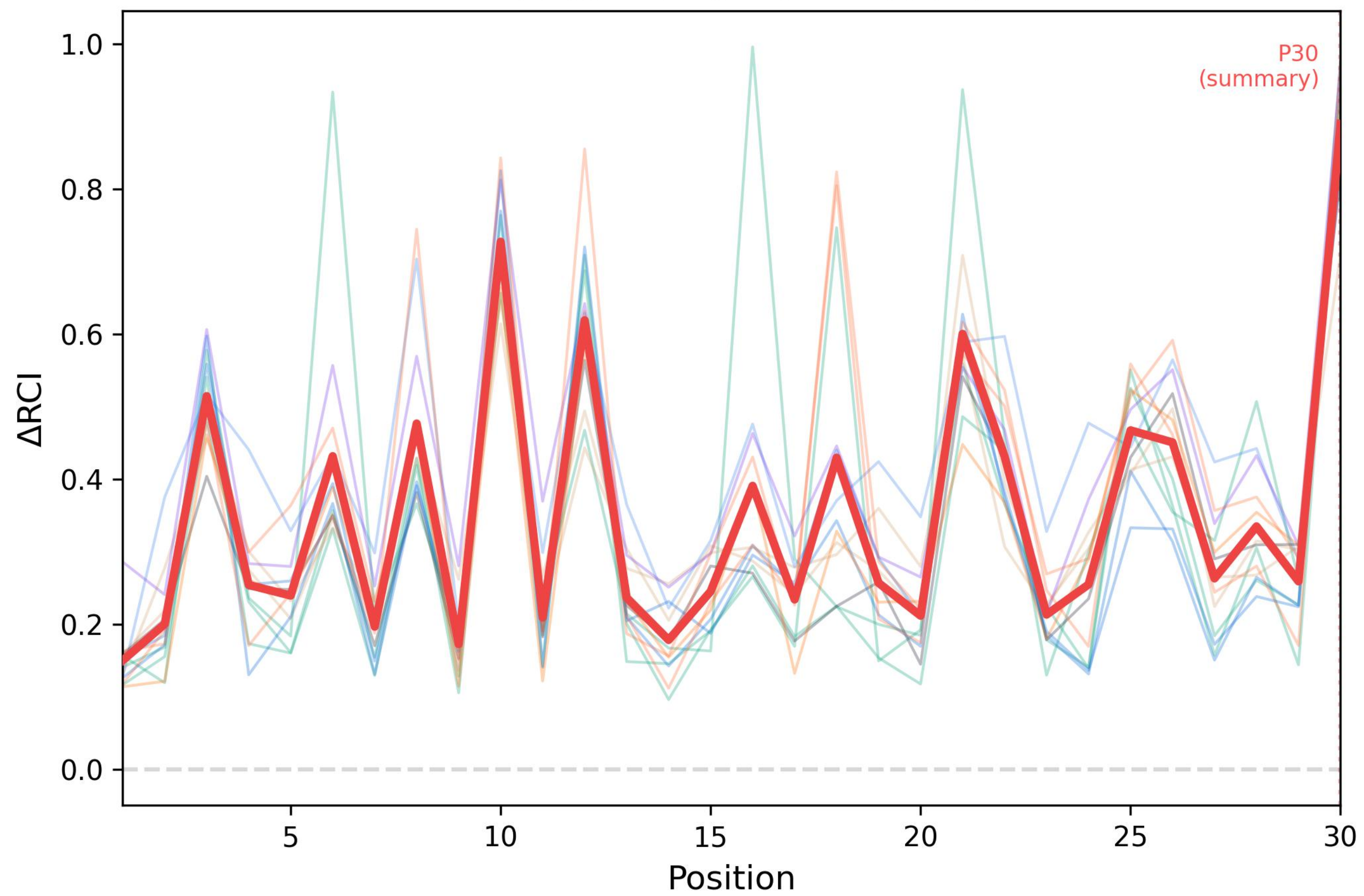
Information Hierarchy: SCRAMBLED vs COLD
(Hierarchy holds for 25/25 model-domain runs)

Model Rankings by Mean ΔRCI (95% CI)
(C)=Closed, (O)=Open

Philosophy Domain (12 models)

Medical Domain (13 models)

mean=0.317

Vendor
- OpenAI
- Anthropic
- Google
- DeepSeek
- Moonshot
- Meta
- Mistral
- Alibaba