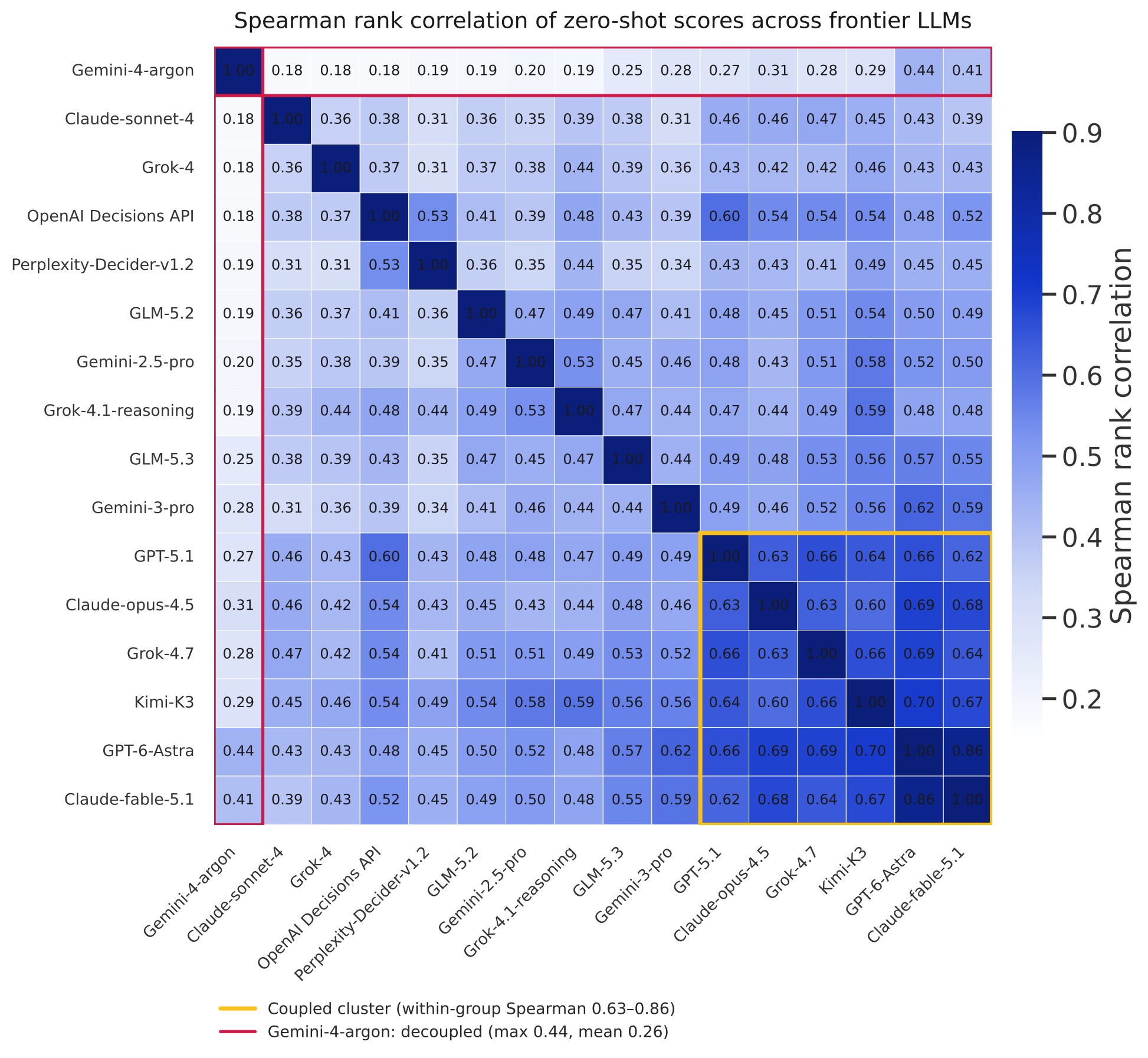

Spearman rank correlation of zero-shot scores across frontier LLMs
Gemini-4-argon 1.00 0.18 0.18 0.18 0.19 0.19 0.20 0.19 0.25 0.28 0.27 0.31 0.28 0.29 0.44 0.41
Claude-sonnet-4 0.18 1.00 0.36 0.38 0.31 0.36 0.35 0.39 0.38 0.31 0.46 0.46 0.47 0.45 0.43 0.39
Grok-4 0.18 0.36 1.00 0.37 0.31 0.37 0.38 0.44 0.39 0.36 0.43 0.42 0.42 0.46 0.43 0.43
OpenAI Decisions API 0.18 0.38 0.37 1.00 0.53 0.41 0.39 0.48 0.43 0.39 0.60 0.54 0.54 0.54 0.48 0.52
Perplexity-Decider-v1.2 0.19 0.31 0.31 0.53 1.00 0.36 0.35 0.44 0.35 0.34 0.43 0.43 0.41 0.49 0.45 0.45
GLM-5.2 0.19 0.36 0.37 0.41 0.36 1.00 0.47 0.49 0.47 0.41 0.48 0.45 0.51 0.54 0.50 0.49
Gemini-2.5-pro 0.20 0.35 0.38 0.39 0.35 0.47 1.00 0.53 0.45 0.46 0.48 0.43 0.51 0.58 0.52 0.50
Grok-4.1-reasoning 0.19 0.39 0.44 0.48 0.44 0.49 0.53 1.00 0.47 0.44 0.47 0.44 0.49 0.59 0.48 0.48
GLM-5.3 0.25 0.38 0.39 0.43 0.35 0.47 0.45 0.47 1.00 0.44 0.49 0.48 0.53 0.56 0.57 0.55
Gemini-3-pro 0.28 0.31 0.36 0.39 0.34 0.41 0.46 0.44 0.44 1.00 0.49 0.46 0.52 0.56 0.62 0.59
GPT-5.1 0.27 0.46 0.43 0.60 0.43 0.48 0.48 0.47 0.49 0.49 1.00 0.63 0.66 0.64 0.66 0.62
Claude-opus-4.5 0.31 0.46 0.42 0.54 0.43 0.45 0.43 0.44 0.48 0.46 0.63 1.00 0.63 0.60 0.69 0.68
Grok-4.7 0.28 0.47 0.42 0.54 0.41 0.51 0.51 0.49 0.53 0.52 0.66 0.63 1.00 0.66 0.69 0.64
Kimi-K3 0.29 0.45 0.46 0.54 0.49 0.54 0.58 0.59 0.56 0.56 0.64 0.60 0.66 1.00 0.70 0.67
GPT-6-Astra 0.44 0.43 0.43 0.48 0.45 0.50 0.52 0.48 0.57 0.62 0.66 0.69 0.69 0.70 1.00 0.86
Claude-fable-5.1 0.41 0.39 0.43 0.52 0.45 0.49 0.50 0.48 0.55 0.59 0.62 0.68 0.64 0.67 0.86 1.00
Gemini-4-argon
Claude-sonnet-4
Grok-4
OpenAI Decisions API
Perplexity-Decider-v1.2
GLM-5.2
Gemini-2.5-pro
Grok-4.1-reasoning
GLM-5.3
Gemini-3-pro
GPT-5.1
Claude-opus-4.5
Grok-4.7
Kimi-K3
GPT-6-Astra
Claude-fable-5.1
Spearman rank correlation
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
Coupled cluster (within-group Spearman 0.63–0.86)
Gemini-4-argon: decoupled (max 0.44, mean 0.26)