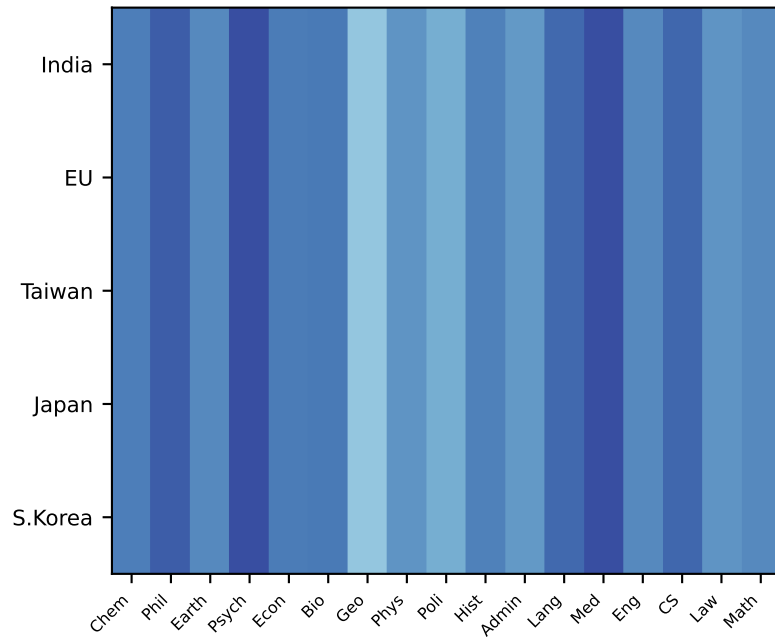
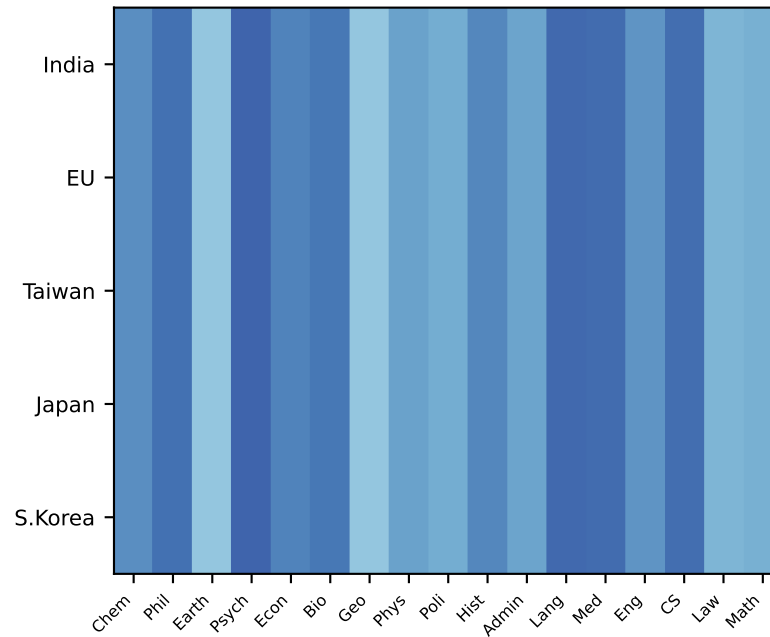


Per-Model Task Performance Fingerprints

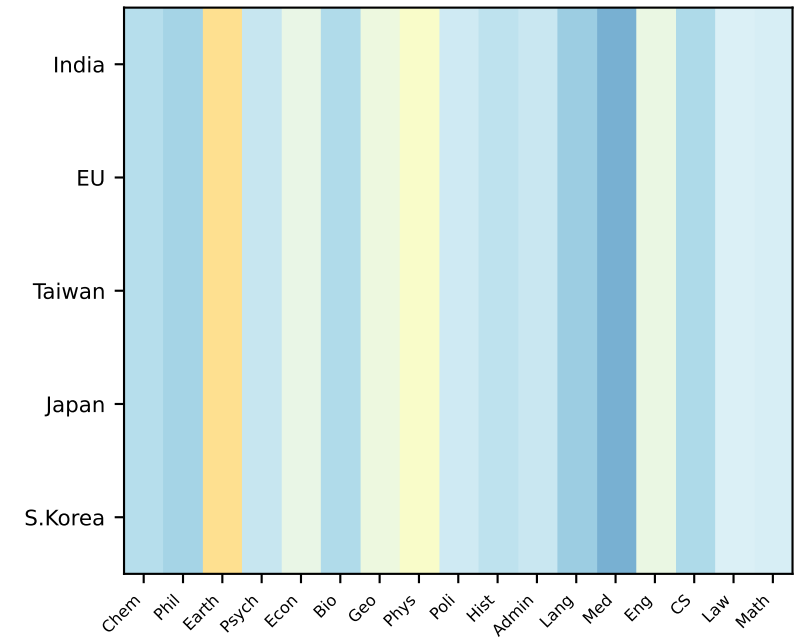
Gemini-2.5-pro
(Overall: 87.0%)



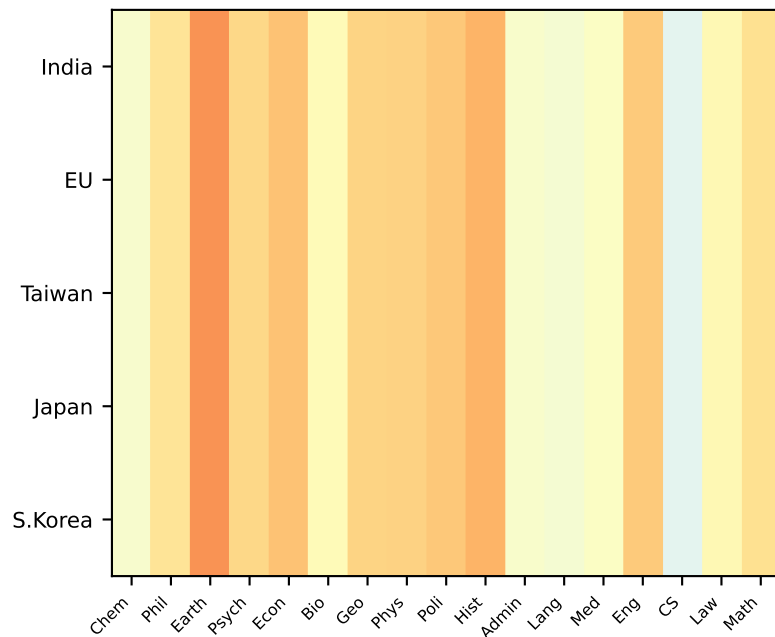
o3
(Overall: 84.3%)



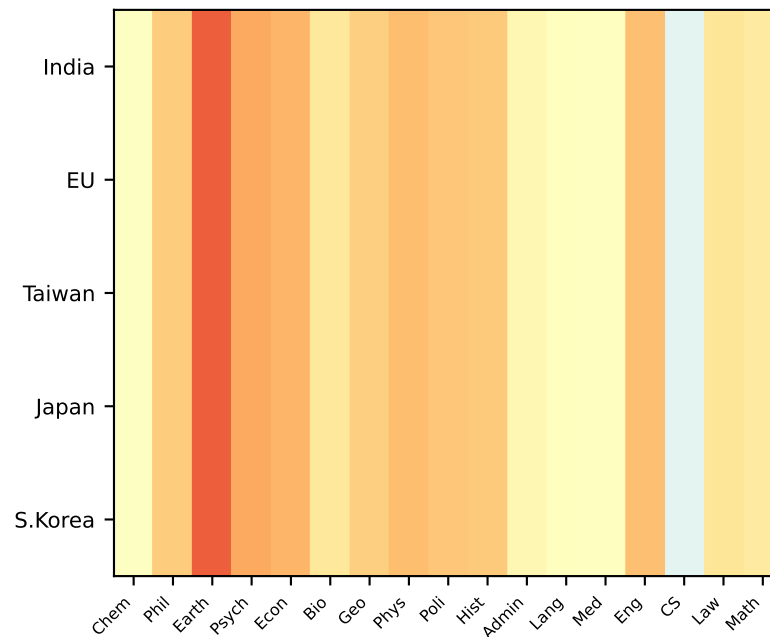
Claude-Sonnet-4
(Overall: 63.3%)



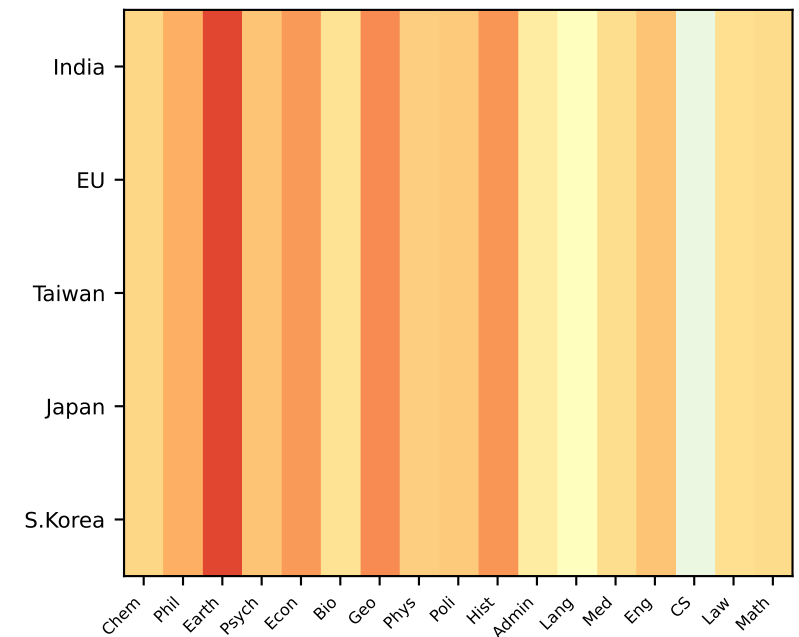
Qwen2-VL-72B-Instruct
(Overall: 44.6%)



GPT-4o
(Overall: 42.0%)



InternVL2.5-38B-MPO
(Overall: 39.3%)



Accuracy (%)

100

80

60

40

20

0