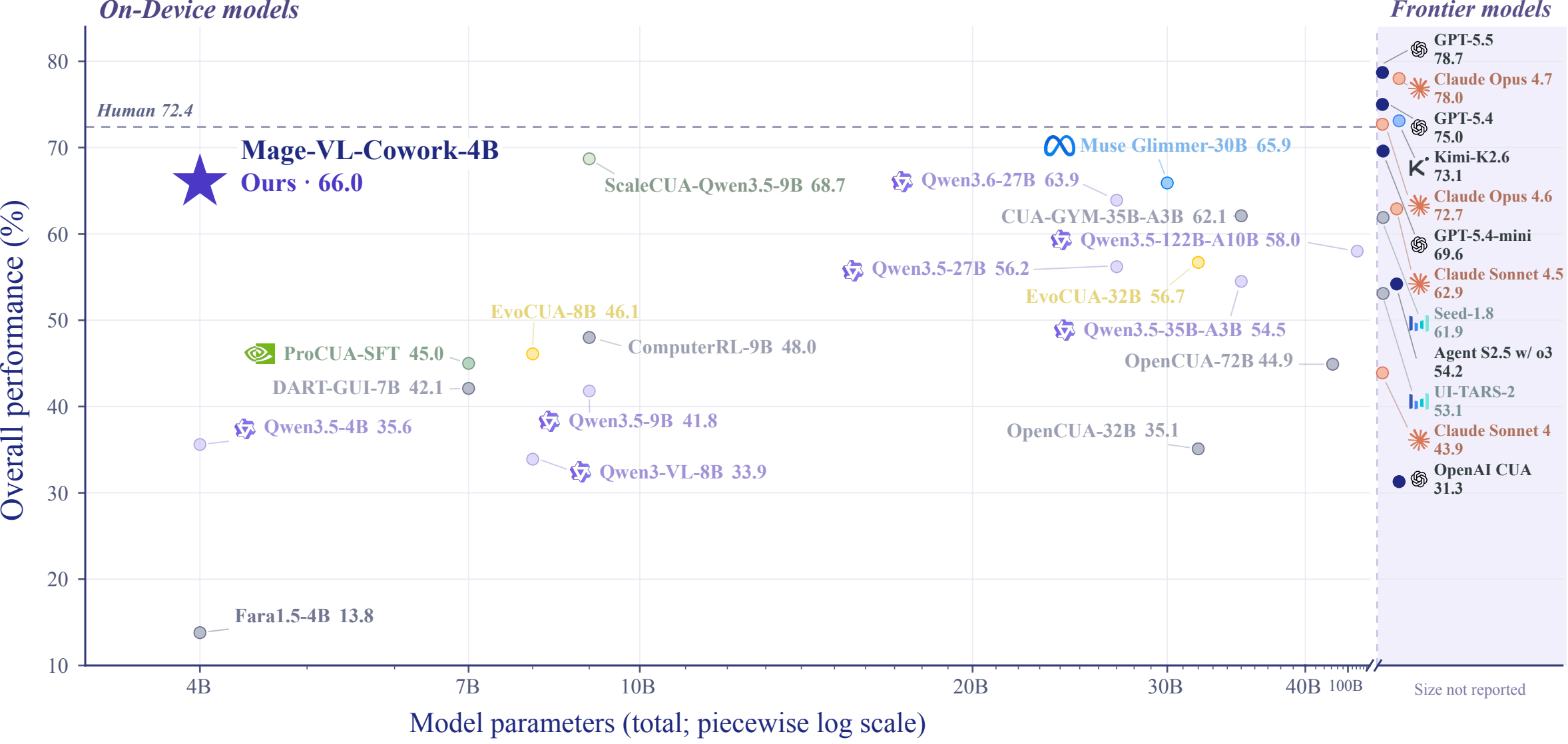
On-Device models
Frontier models
Overall performance (%)
Model parameters (total; piecewise log scale)
80
70
60
50
40
30
20
10
4B
7B
10B
20B
30B
40B
100B
Size not reported
Human 72.4
Mage-VL-Cowork-4B
Ours · 66.0
ScaleCUA-Qwen3.5-9B 68.7
Muse Glimmer-30B 65.9
Qwen3.6-27B 63.9
CUA-GYM-35B-A3B 62.1
Qwen3.5-122B-A10B 58.0
Qwen3.5-27B 56.2
EvoCUA-32B 56.7
Qwen3.5-35B-A3B 54.5
OpenCUA-72B 44.9
EvoCUA-8B 46.1
ComputerRL-9B 48.0
ProCUA-SFT 45.0
DART-GUI-7B 42.1
Qwen3.5-9B 41.8
Qwen3.5-4B 35.6
OpenCUA-32B 35.1
Qwen3-VL-8B 33.9
Fara1.5-4B 13.8
GPT-5.5
78.7
Claude Opus 4.7
78.0
GPT-5.4
75.0
Kimi-K2.6
73.1
Claude Opus 4.6
72.7
GPT-5.4-mini
69.6
Claude Sonnet 4.5
62.9
Seed-1.8
61.9
Agent S2.5 w/ o3
54.2
UI-TARS-2
53.1
Claude Sonnet 4
43.9
OpenAI CUA
31.3