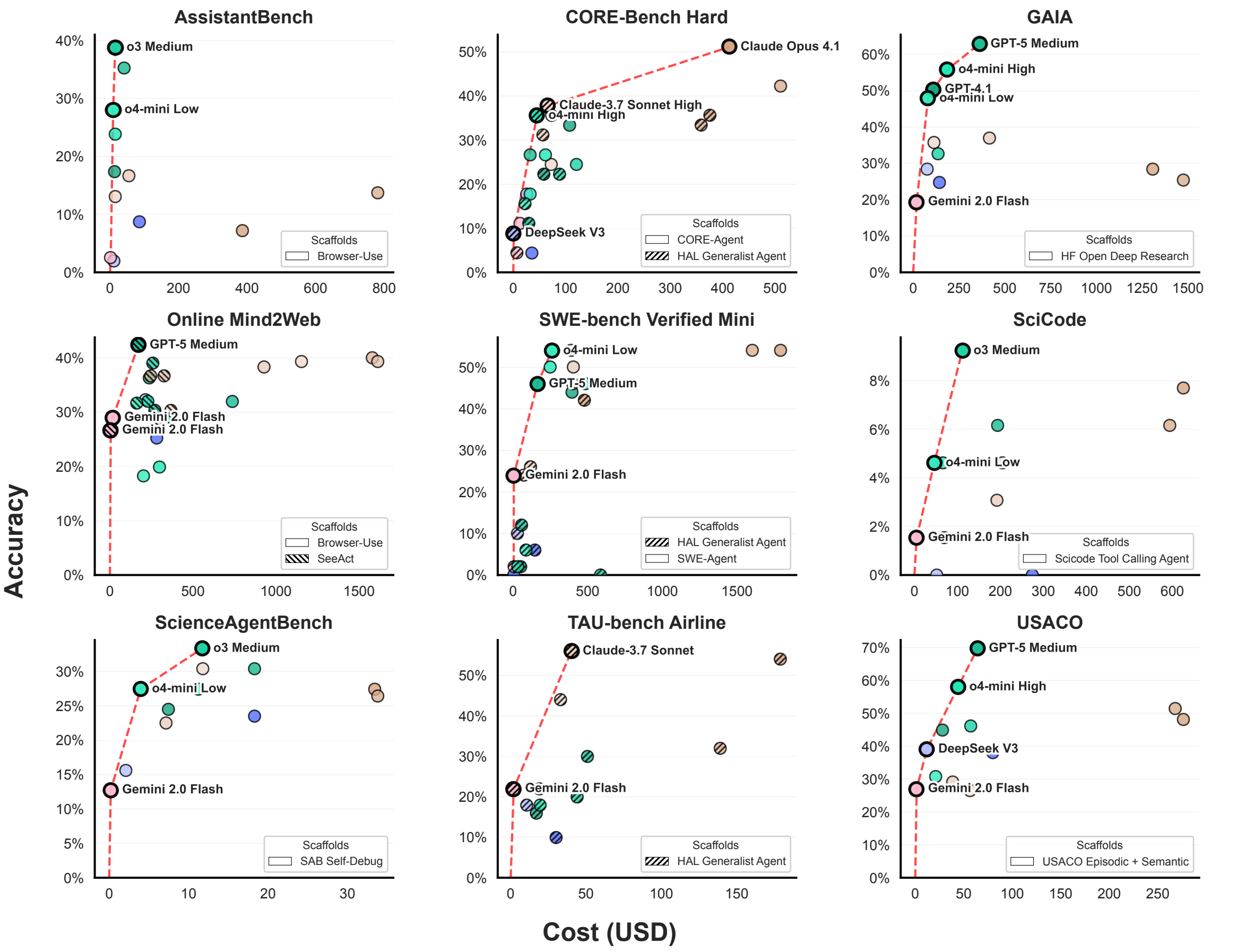
AssistantBench
o3 Medium
o4-mini Low
Scaffolds
Browser-Use
CORE-Bench Hard
Claude Opus 4.1
Claude-3.7 Sonnet High
o4-mini High
DeepSeek V3
Scaffolds
CORE-Agent
HAL Generalist Agent
GAIA
GPT-5 Medium
o4-mini High
GPT-4.1
o4-mini Low
Gemini 2.0 Flash
Scaffolds
HF Open Deep Research
Online Mind2Web
GPT-5 Medium
Gemini 2.0 Flash
Gemini 2.0 Flash
Scaffolds
Browser-Use
SeeAct
SWE-bench Verified Mini
o4-mini Low
GPT-5 Medium
Gemini 2.0 Flash
Scaffolds
HAL Generalist Agent
SWE-Agent
SciCode
o3 Medium
o4-mini Low
Gemini 2.0 Flash
Scaffolds
Scicode Tool Calling Agent
ScienceAgentBench
o3 Medium
o4-mini Low
Gemini 2.0 Flash
Scaffolds
SAB Self-Debug
TAU-bench Airline
Claude-3.7 Sonnet
Gemini 2.0 Flash
Scaffolds
HAL Generalist Agent
USACO
GPT-5 Medium
o4-mini High
DeepSeek V3
Gemini 2.0 Flash
Scaffolds
USACO Episodic + Semantic
Accuracy
Cost (USD)
Models
Claude Opus 4.1
Claude Opus 4.1 High
Claude Sonnet 4
Claude Sonnet 4 High
Claude-3.7 Sonnet
Claude-3.7 Sonnet High
DeepSeek R1
DeepSeek V3
GPT-4.1
GPT-5 Medium
Gemini 2.0 Flash
o3 Medium
o4-mini High
o4-mini Low