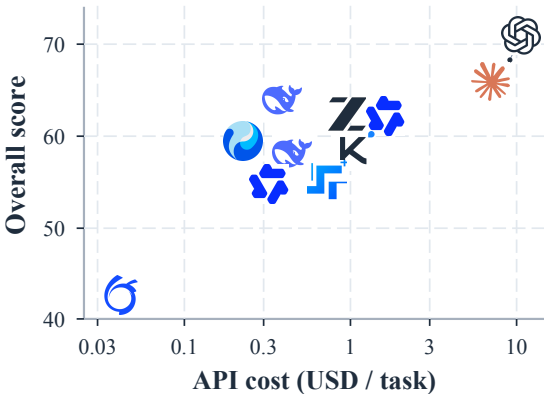


(a) Score and API cost**(b) Mean tool calls per task**

		Search	Read	Tables	Downl.	Compute
1	Astra	16.3	15.0	4.0	0.4	2.8
2	Opus 5.5	16.1	23.2	15.5	0.9	9.3
3	DS V4.1 Flash	22.2	41.6	27.4	1.7	19.9
4	GLM 5.3	8.6	14.1	11.3	0.3	6.2
5	Qwen 3.8 Max	11.3	24.4	17.0	1.0	9.9
6	Hunyuan 4 Preview	7.7	9.4	15.2	1.5	12.7
7	DS V4 Pro	14.6	16.9	16.2	0.3	3.3
8	Kimi K3	9.8	9.5	12.2	1.4	7.7
9	Step 5 Preview	8.9	28.2	16.4	1.0	8.1
10	Qwen 3.8 27B	7.4	13.7	11.8	0.5	4.6
11	Ling 3.0 Flash	14.1	21.8	1.7	0.5	4.5