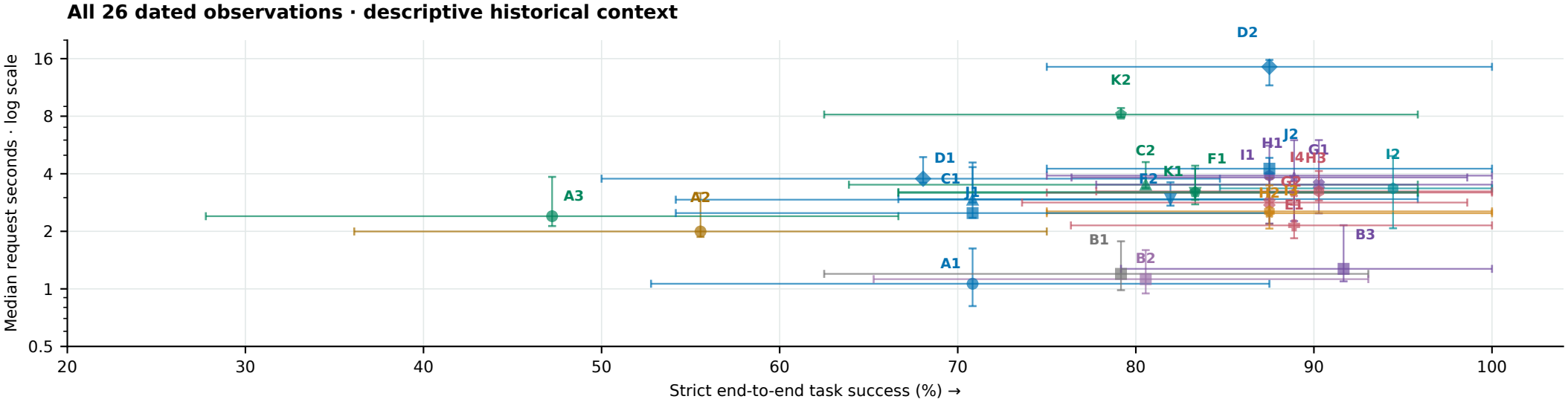
Task success, response time and model placement

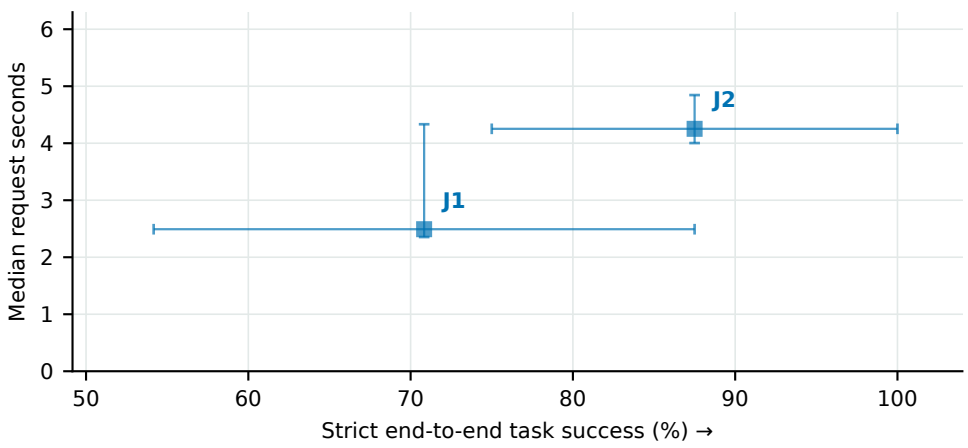
CERRA · updated 9 October 2026 · 24 primary tasks × 3 attempts per configuration

Cross-bars are separate marginal 95% task-cluster intervals. Failures remain in scores and observed timings.

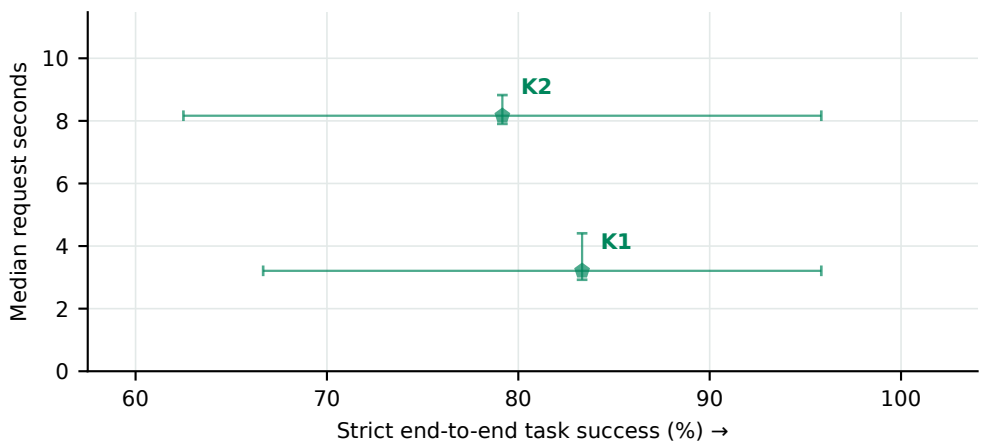
All 26 dated observations · descriptive historical context



Qwen · Oct 7 · full GPU residency



Gemma · Oct 9 · paired configurations



Code	Configuration	Run date	Strict pass	Median / P90 (s)
A1	Qwen3 8B	2026-09-18	51/72 (70.8%)	1.06 / 2.17
A2	Llama 3.1 8B	2026-09-18	40/72 (55.6%)	2.00 / 3.77
A3	Gemma 3 4B	2026-09-18	34/72 (47.2%)	2.40 / 4.18
B1	GPT-4.1 mini	2026-09-23	57/72 (79.2%)	1.20 / 2.59
B2	GPT-5.4 mini	2026-09-23	58/72 (80.6%)	1.12 / 2.76
B3	GPT-5.4	2026-09-23	66/72 (91.7%)	1.27 / 2.94
C1	Qwen3 14B	2026-09-24	51/72 (70.8%)	2.93 / 5.56
C2	Gemma 4 12B	2026-09-24	58/72 (80.6%)	3.51 / 6.71
D1	Qwen3 14B	2026-09-29	49/72 (68.1%)	3.77 / 7.20
D2	Qwen3.8 27B Q4_K_M	2026-09-29	63/72 (87.5%)	14.50 / 29.81
E1	GPT-6.1 Sol	2026-09-29	64/72 (88.9%)	2.15 / 4.88
F1	Gemma 4 12B	2026-09-29	60/72 (83.3%)	3.18 / 6.35
F2	Qwen3.5 9B Q4_K_M	2026-09-29	59/72 (81.9%)	2.95 / 6.19
G1	GPT-5.4	2026-09-30	65/72 (90.3%)	3.51 / 8.58
G2	GPT-6.1 Sol	2026-09-30	63/72 (87.5%)	2.83 / 5.97
H1	GPT-5.4	2026-09-30	64/72 (88.9%)	3.82 / 11.16
H2	GPT-6 Astra	2026-09-30	63/72 (87.5%)	2.49 / 4.92
H3	GPT-6.1 Sol	2026-09-30	64/72 (88.9%)	3.19 / 5.91
I1	GPT-5.4	2026-10-02	63/72 (87.5%)	3.91 / 11.35
I2	GPT-5.5	2026-10-02	68/72 (94.4%)	3.35 / 9.02
I3	GPT-6 Astra	2026-10-02	63/72 (87.5%)	2.54 / 5.06
I4	GPT-6.1 Sol	2026-10-02	65/72 (90.3%)	3.24 / 6.81
J1	Qwen3 14B Q4_K_M	2026-10-07	51/72 (70.8%)	2.49 / 4.76
J2	Qwen3.8 27B Q4_K_M	2026-10-07	63/72 (87.5%)	4.25 / 8.41
K1	Gemma 4 12B Q4_K_M	2026-10-09	60/72 (83.3%)	3.21 / 6.49
K2	Gemma 4 26B A4B Q8_0	2026-10-09	57/72 (79.2%)	8.16 / 15.94

Historical campaigns differ in application path, settings and load; the overview is not a controlled global ranking.
J: one benchmark model at a time, full observed residency on a shared RTX 5090; other GPU services stayed active.
G2 has 71 retained timings / 72 primary attempts. Vertical intervals estimate median uncertainty, not P90.
K compares model size, architecture and quantization together; it does not isolate an 8-bit quantization effect.