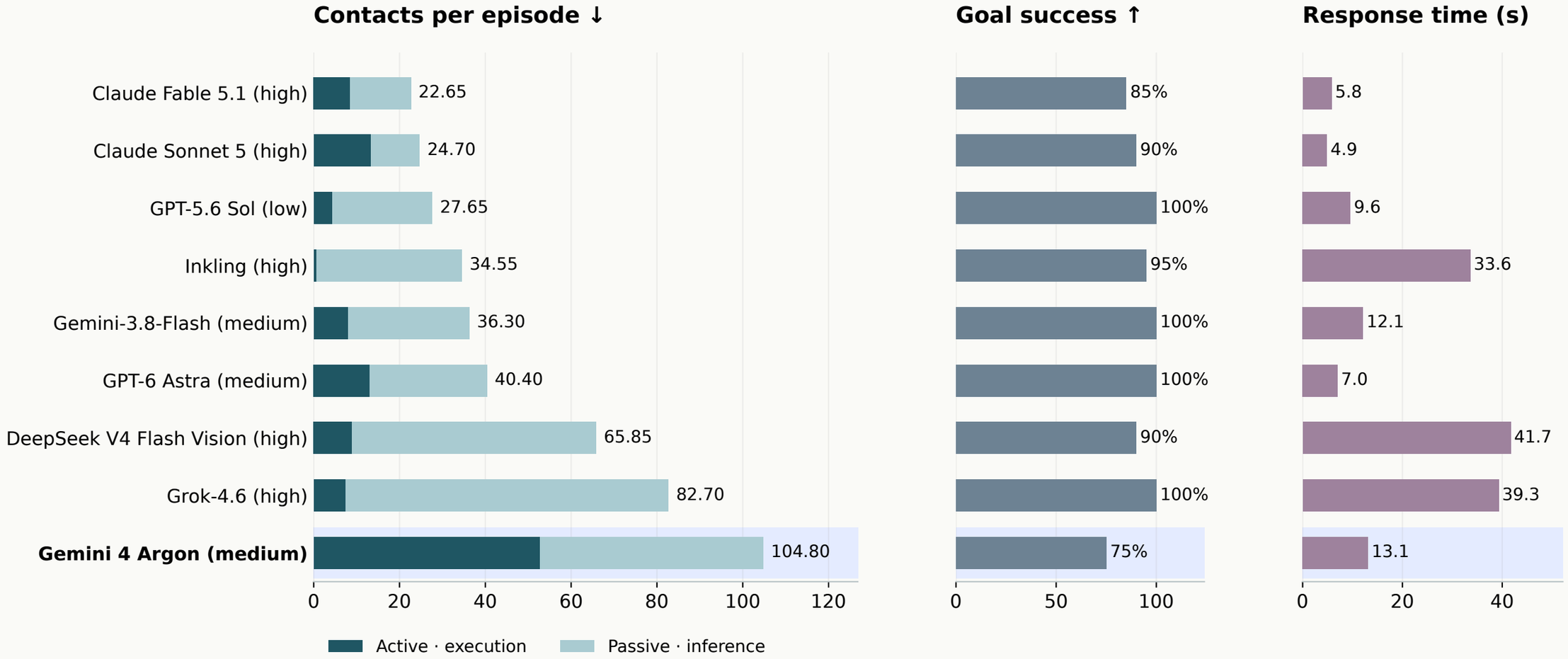
RT-SAFE / 20 matched tasks, nine models

Configured-default reasoning · 5 tasks per map × 4 maps · Hard, real-time · Seed 0



Safe success: Sol: 5%; all other models 0%.

All 20 episodes count, including failures. Contacts may repeat; lower counts can coexist with failed navigation.
Historical defaults vs. Gemini 4 explicit medium: different collection dates/transports, original scoring. Latency averages episode means.