

a Released benchmark outcomes

Claude Sonnet 4.5	72.0%	72.4%	84.9%
GLM-5	82.5%	73.7%	86.8%
GPT-5.2 high	83.0%	81.6%	89.7%
Qwen3.5-397B	81.5%	84.4%	97.8%
	Airline	Retail	Telecom

b State-change burden is not monotonic

