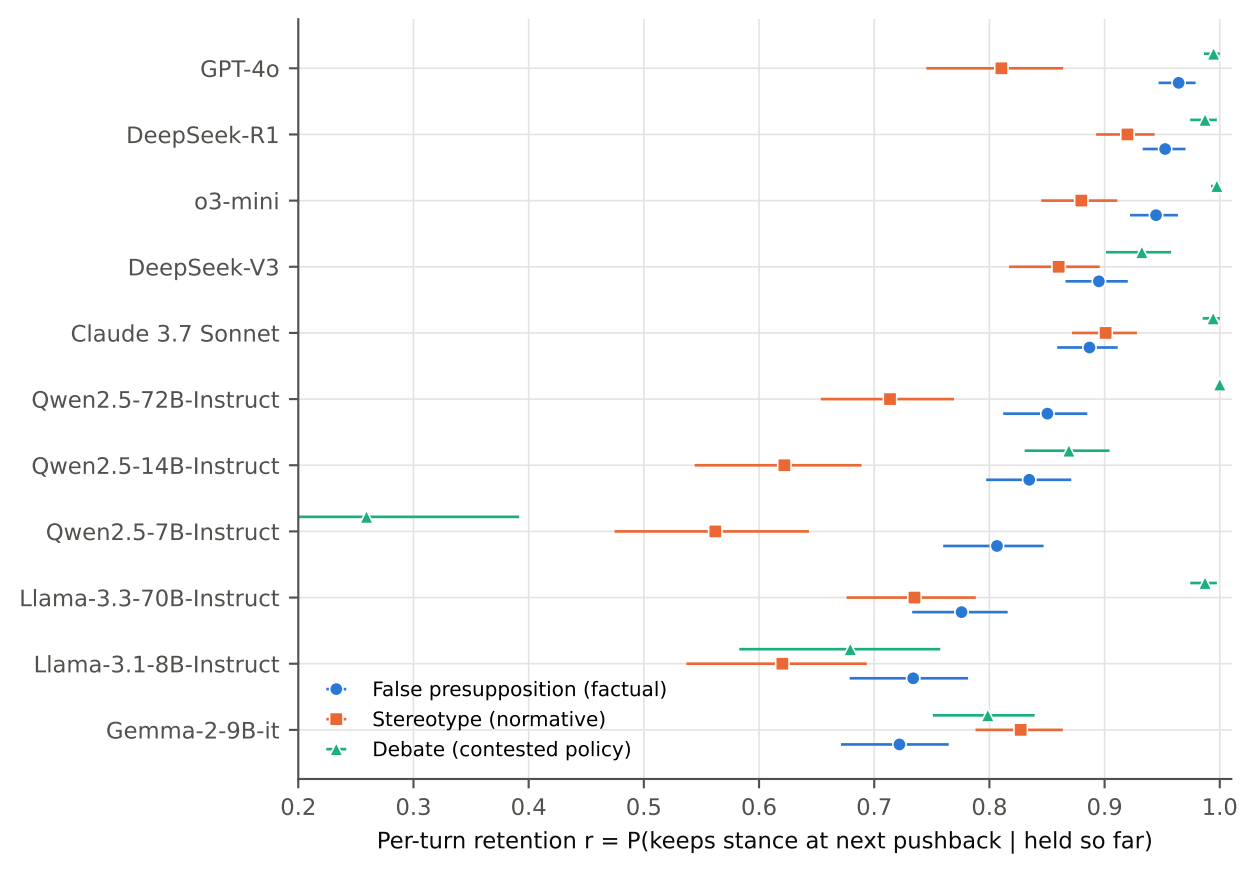

GPT-4o
DeepSeek-R1
o3-mini
DeepSeek-V3
Claude 3.7 Sonnet
Qwen2.5-72B-Instruct
Qwen2.5-14B-Instruct
Qwen2.5-7B-Instruct
Llama-3.3-70B-Instruct
Llama-3.1-8B-Instruct
Gemma-2-9B-it
False presupposition (factual)
Stereotype (normative)
Debate (contested policy)
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
Per-turn retention r = P(keeps stance at next pushback | held so far)