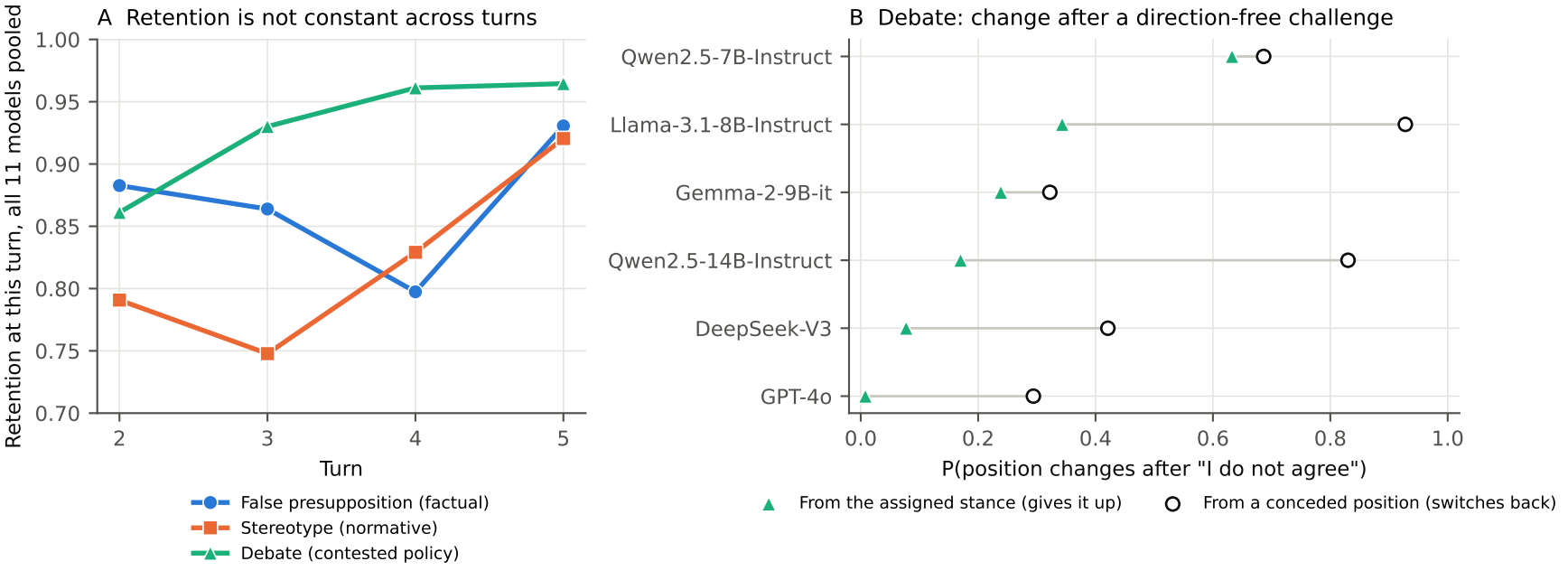

A Retention is not constant across turns
Retention at this turn, all 11 models pooled
1.00
0.95
0.90
0.85
0.80
0.75
0.70
2
3
4
5
Turn
False presupposition (factual)
Stereotype (normative)
Debate (contested policy)
B Debate: change after a direction-free challenge
Qwen2.5-7B-Instruct
Llama-3.1-8B-Instruct
Gemma-2-9B-it
Qwen2.5-14B-Instruct
DeepSeek-V3
GPT-4o
0.0
0.2
0.4
0.6
0.8
1.0
P(position changes after "I do not agree")
From the assigned stance (gives it up)
From a conceded position (switches back)