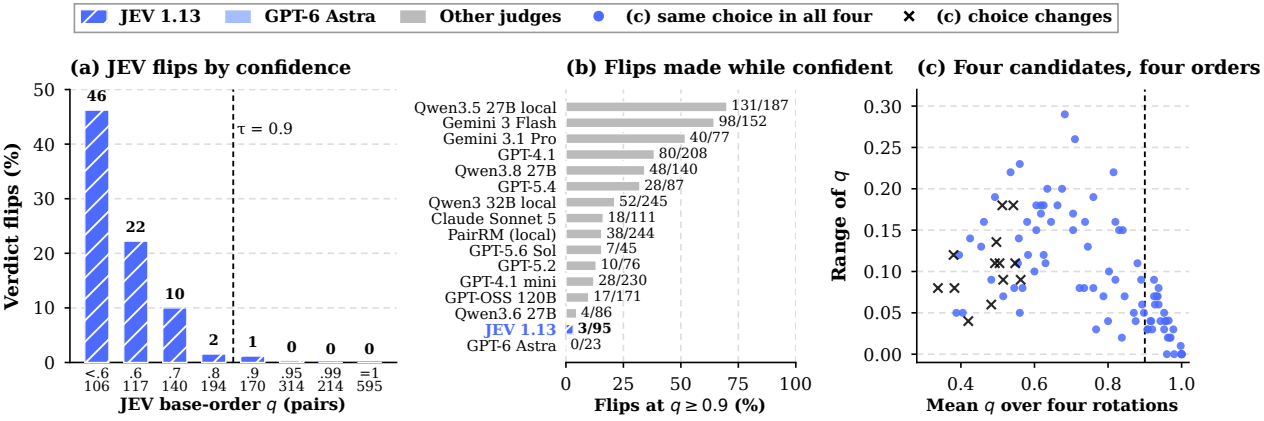

JEV 1.13
GPT-6 Astra
Other judges
(c) same choice in all four
(c) choice changes
(a) JEV flips by confidence
Verdict flips (%)
50
40
30
20
10
0
46
22
10
2
1
0
0
0
τ = 0.9
<.6
106
.6
117
.7
140
.8
194
.9
170
.95
314
.99
214
=1
595
JEV base-order q (pairs)
(b) Flips made while confident
Qwen3.5 27B local 131/187
Gemini 3 Flash 98/152
Gemini 3.1 Pro 40/77
GPT-4.1 80/208
Qwen3.8 27B 48/140
GPT-5.4 28/87
Qwen3 32B local 52/245
Claude Sonnet 5 18/111
PairRM (local) 38/244
GPT-5.6 Sol 7/45
GPT-5.2 10/76
GPT-4.1 mini 28/230
GPT-OSS 120B 17/171
Qwen3.6 27B 4/86
JEV 1.13 3/95
GPT-6 Astra 0/23
0
25
50
75
100
Flips at q ≥ 0.9 (%)
(c) Four candidates, four orders
Range of q
0.30
0.25
0.20
0.15
0.10
0.05
0.00
0.4
0.6
0.8
1.0
Mean q over four rotations