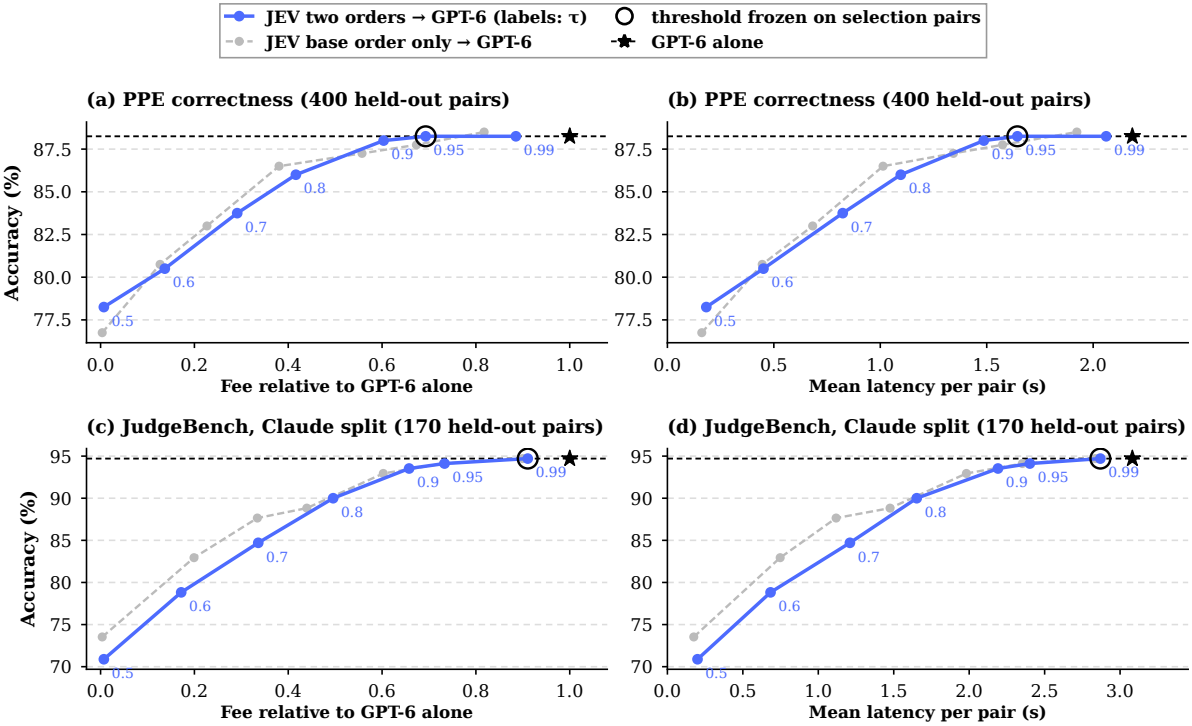

JEV two orders → GPT-6 (labels: τ)
JEV base order only → GPT-6
threshold frozen on selection pairs
GPT-6 alone
(a) PPE correctness (400 held-out pairs)
(b) PPE correctness (400 held-out pairs)
(c) JudgeBench, Claude split (170 held-out pairs)
(d) JudgeBench, Claude split (170 held-out pairs)
Accuracy (%)
Fee relative to GPT-6 alone
Mean latency per pair (s)
0.5
0.6
0.7
0.8
0.9
0.95
0.99