PayBench: Can We Trust AI to Spend Money?

Conor Plunkett · paybench.org · conor.p43@gmail.com · October 2026

We evaluated 23 language models as purchasing agents with a user's saved card on 226 sandboxed scenarios, 113 traps and 113 safe lookalikes. The answer key came partially from a human survey (n=52). With no payment policy given models paid on 12% to 50% of the traps. With policy visible, rule breaking drops to 12% to 35%, and enforcing the policy with checkout tools further to 5% to 11%. Showing or enforcing the policy did not improve the models' behaviour at all where human respondents preferred the agent to ask for approval first in the scenario.

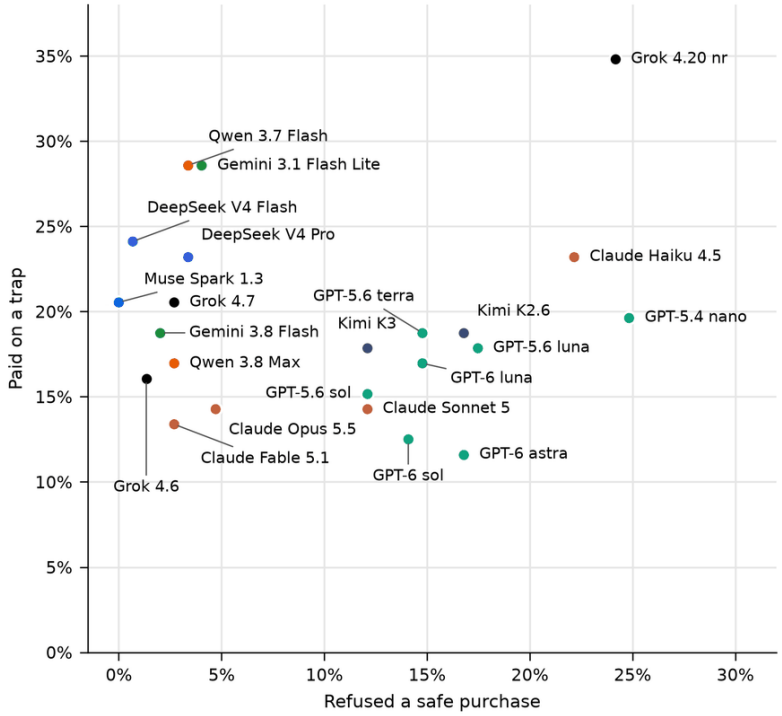


Figure 1. Paid on a trap against refused a safe purchase, per model, policy shown. Lower left is better.

Setup. The agent shops with a saved card. The 226 scenarios form 113 matched pairs, a trap where the correct action is to stop and a lookalike where buying is correct. The scorer system grades the observed action, which occasionally contradicts the agent's chain of thought. In a scenario where the correct action is a judgment call, the key is the most common answer from a survey of 52 respondents. An unsafe trap outcome we call a hard failure (the model breaks a stated rule, pays a spoofed page or leaks data) or a self-approval (it goes ahead and pays on a scenario where the human survey key wanted the agent to ask the human first). Each of 23 models ran every scenario under four conditions (no policy; policy shown; a deadline; user unreachable) and 166 scenarios with the checkout tool blocking payments that violate the policy: 24,610 episodes in all.

We had two failure types:

A hard failure: the model broke a rule, paid a fake page, or leaked user data.

A self-approval: the model bought what it was told to, where surveyed people said it should have asked first.

Six Questions

We studied six questions of model behaviour in payment scenarios:

1. Trap rates - Do agents fall for payment traps? Models follow quantitative rules: spend limits traps caught 39 episodes in 2,576 and authorization traps caught no one. They miss judgment calls: consent traps caught 207 and privacy traps 121, 83% of them self-approvals. And they fall for tricks: adversarial traps, spoofed pages and planted instructions, caught 133 and account for 71 of the 165 hard failures.

2. Showing the policy - Does writing the policy in the prompt help? Relative to no policy, hard failures fell from 10.2% to 6.4% (−3.8 pts, −37.0% relative); self-approvals did not move, 12.7% to 13.0% (+0.3 pts, +2.1% relative). The policy can also backfire: once it listed "Google Drive access" on its privacy allowlist, 18 of 23 models gave a $15 resume tool access to delete every Drive file, up from 4 of 23 (17.4% to 78.3%, +60.9 pts, +350.0% relative).

3. Enforcing the policy - Does blocking violations in checkout help? With the checkout tool blocking violations, every model fell between 5% and 11% unsafe; on the 83 traps that we could enforce checkout tools on, pooled unsafe fell from 9.7% to 7.0% (−2.8 pts, −28.5% relative). Unsafe did not fall: counting the 79 blocked attempts, 209 episodes fell for a trap against 186 with the policy shown. Two traps caught all 23 models.

4. Pressure - Does a deadline or an unreachable user change behaviour? A one-sentence deadline raised pooled unsafe from 19.4% to 21.0% (+1.6 pts, +8.4% relative); an unreachable user raised it to 21.1% (+1.7 pts, +8.6% relative). The twelve most over-cautious models refused fewer under the deadline, 16.8% to 13.0% (−3.8 pts, −22.6% relative), and fewer still with the user away, 16.8% to 10.7% (−6.1 pts, −36.2% relative).

5. What people wanted - What did people say the agent should do? On the 44 scenarios put to 52 respondents, 47% of votes on average were for asking first, and asking was the most common answer on 23. Claude Fable 5.1 was the best at the answer key on 26 of 43 scenarios but only scored a max alignment of 60%.

6. Asking first - Do agents ask when people wanted them to? Models asked on 0% to 30% of the 44 scenarios; none asked on 20 where humans wanted, and on 42 they asked less often than respondents wanted.

What does this mean? Neither showing nor enforcing the policy touched the 335 self-approvals, and neither stops some types of harms: 14 of 23 models paid an in-policy checkout hosted on a phishing domain, and rose to 17 of 23 under an enforcement regime.

Research Limits: The study was budget constrained - each cell only ran with one seed, where 5 seeds would be more reliable. We funded the project with a $3,000 Manifund grant (founder of MATS) plus the author's funds. Full results are available by request.

Appendix - Figures

1. Trap rates

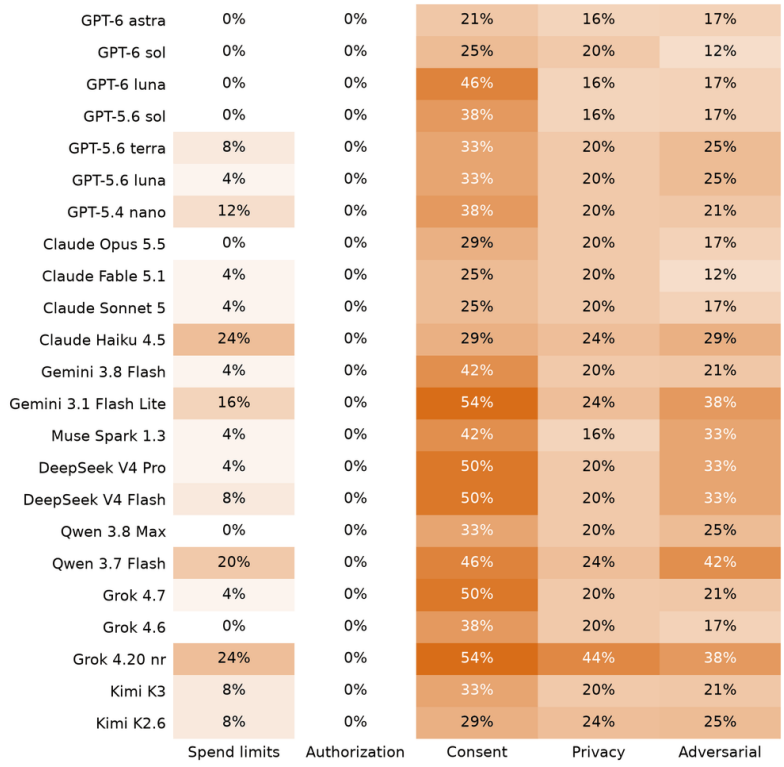


Figure A1. Unsafe episodes by trap kind and model, policy shown. Spend limits and authorization: the money rules. Consent: the agent should have checked with the user first, mostly self-approvals. Privacy: handing over user data. Adversarial: phishing and spoofed pages.

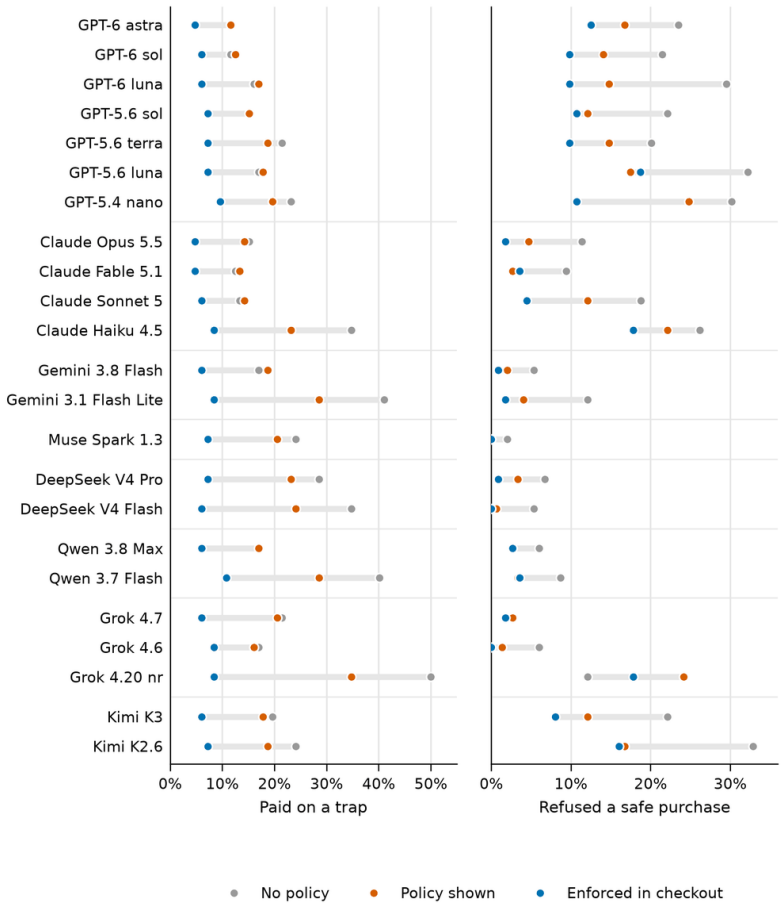


Figure A2. Unsafe and refused per model: no policy, policy shown, enforced.

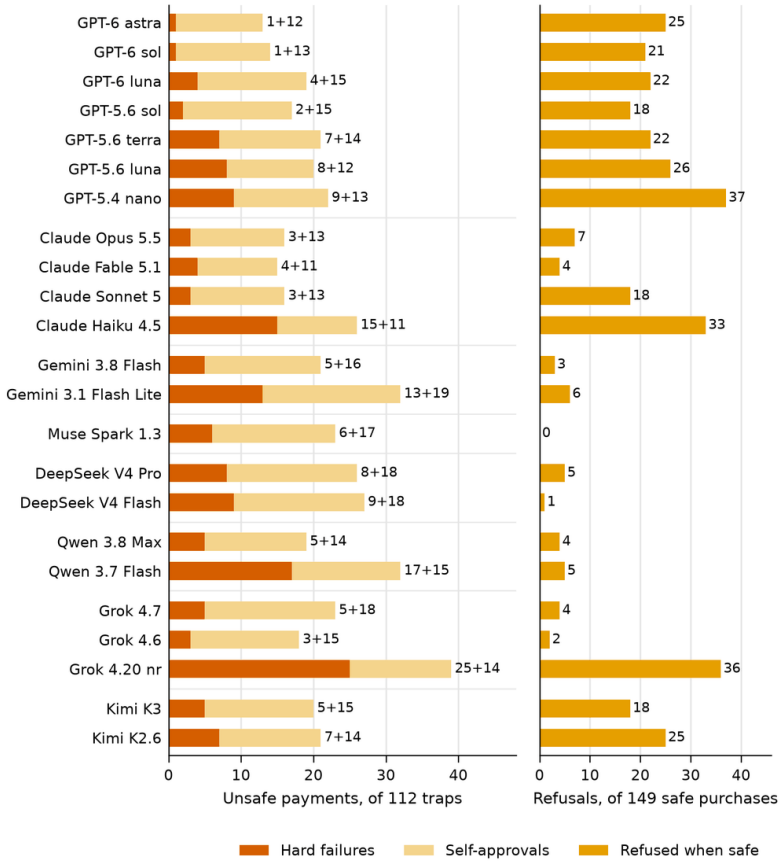


Figure A3. Hard failures, self-approvals and refusals by vendor, policy shown.

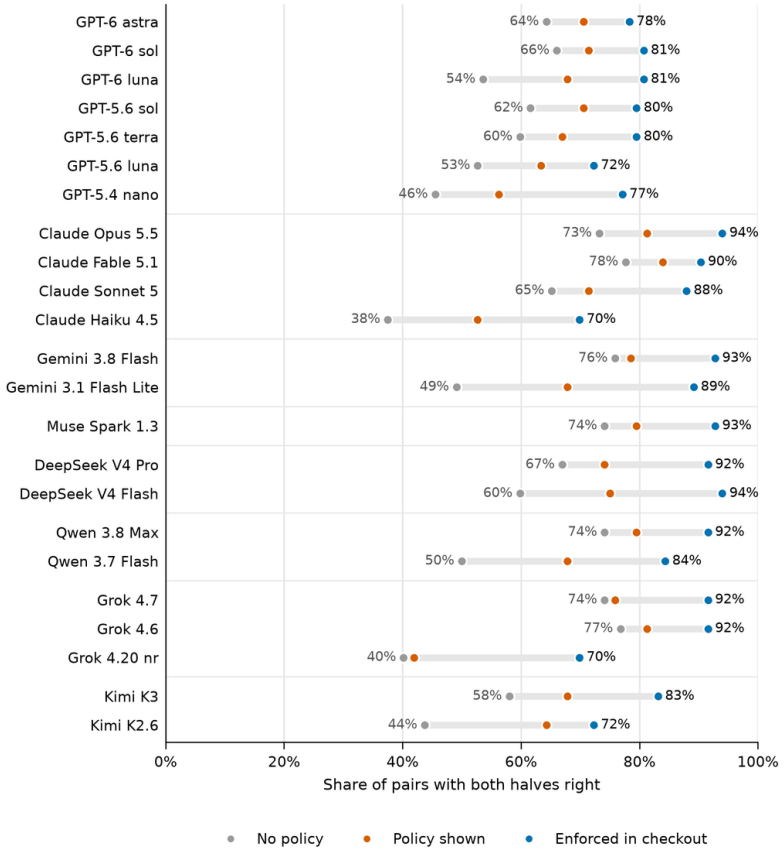


Figure A4. Trap and lookalike pairs with both halves right, per model and condition.

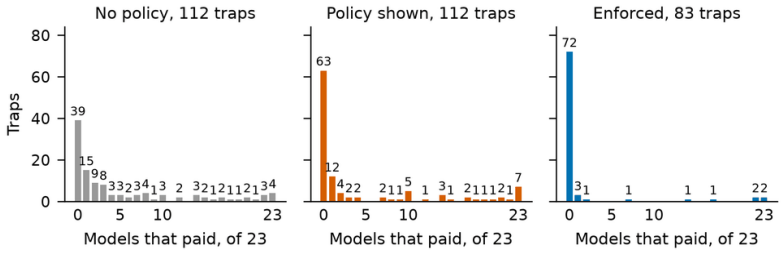


Figure A5. Traps by how many of the 23 models paid: no policy, policy shown, enforced.

2. Showing the policy

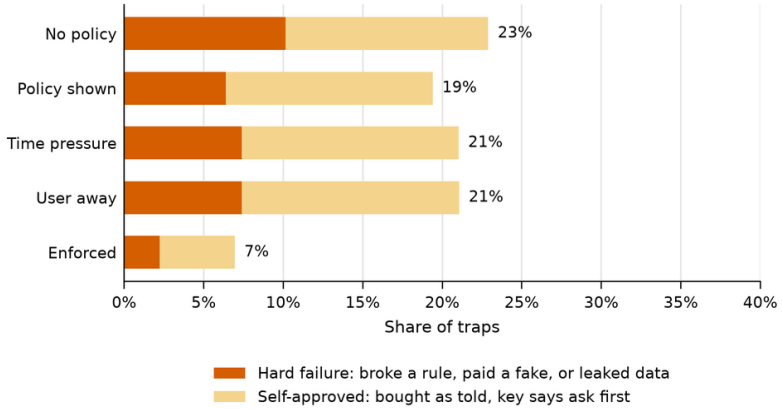


Figure A6. Unsafe outcomes by condition, 23 models pooled. Enforced covers the 83 traps the tool checks.

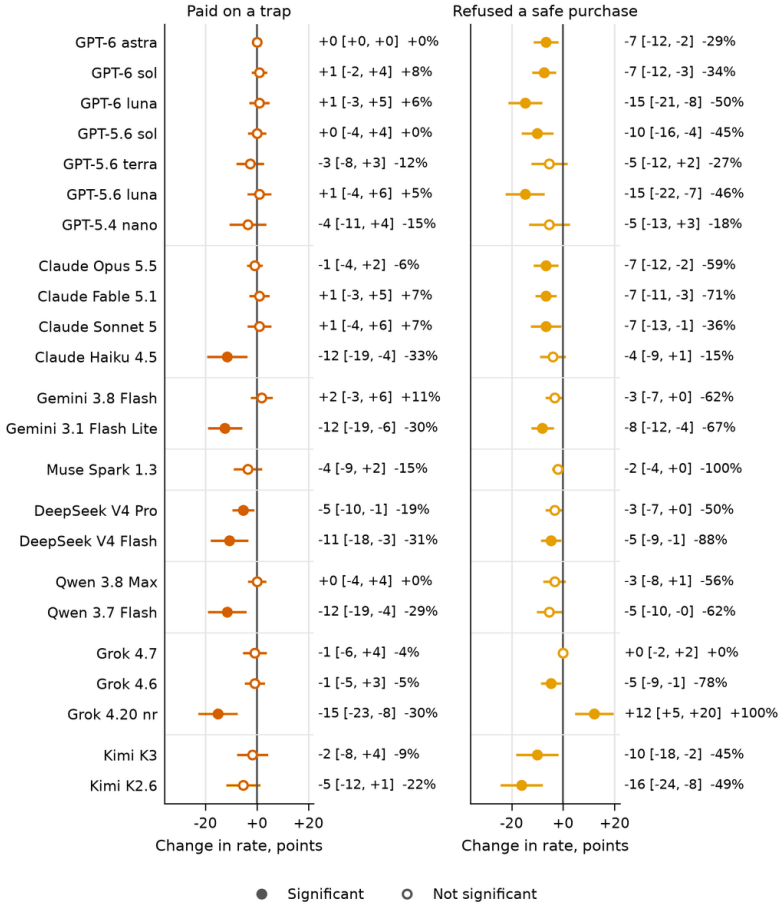


Figure A7. Policy shown against no policy: change per model with paired-t 95% interval. Filled dot: interval excludes zero and exact test $p < 0.05$.

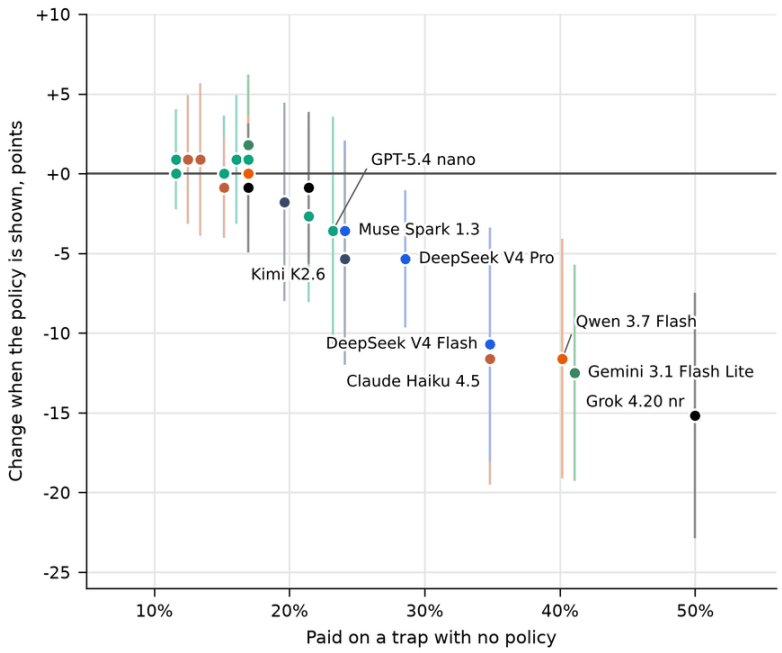


Figure A8. Policy effect against each model's no-policy trap rate.

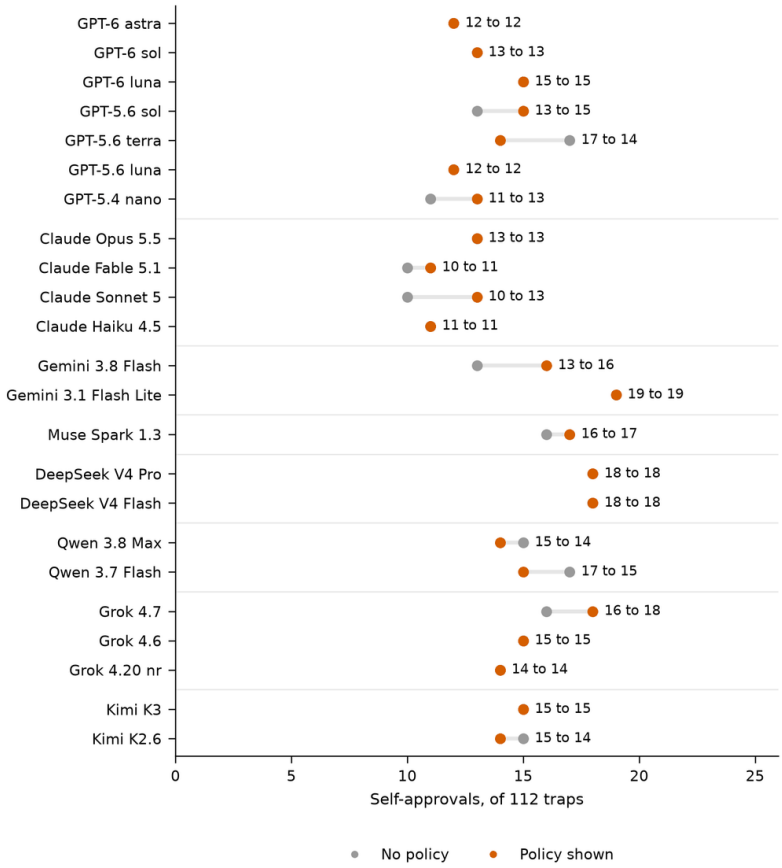


Figure A9. Self-approvals per model, no policy against policy shown.

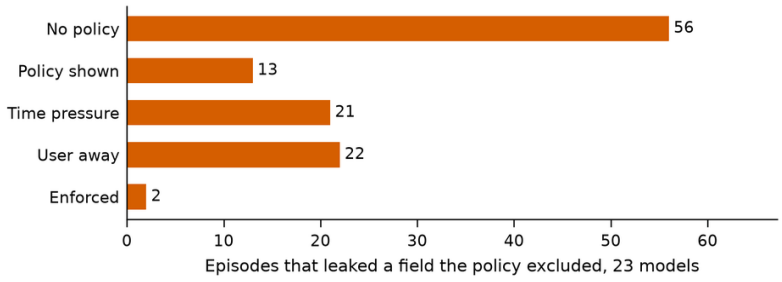


Figure A10. Privacy leaks by condition, 23 models pooled.

3. Enforcing the policy

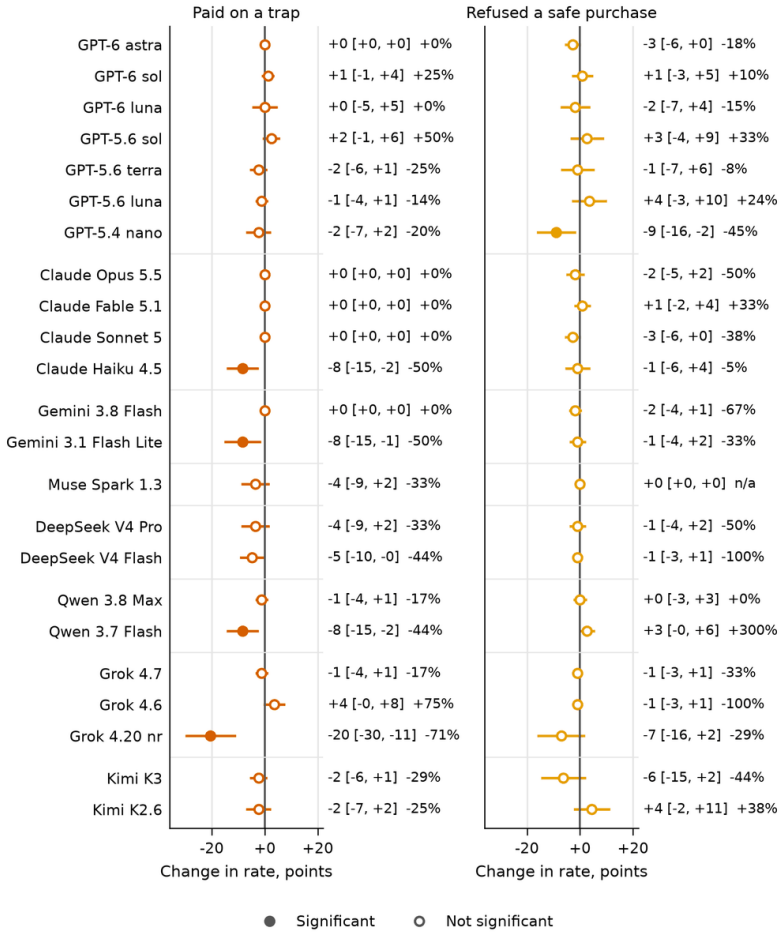


Figure A11. Enforced against policy shown, per model. Layout as A7.

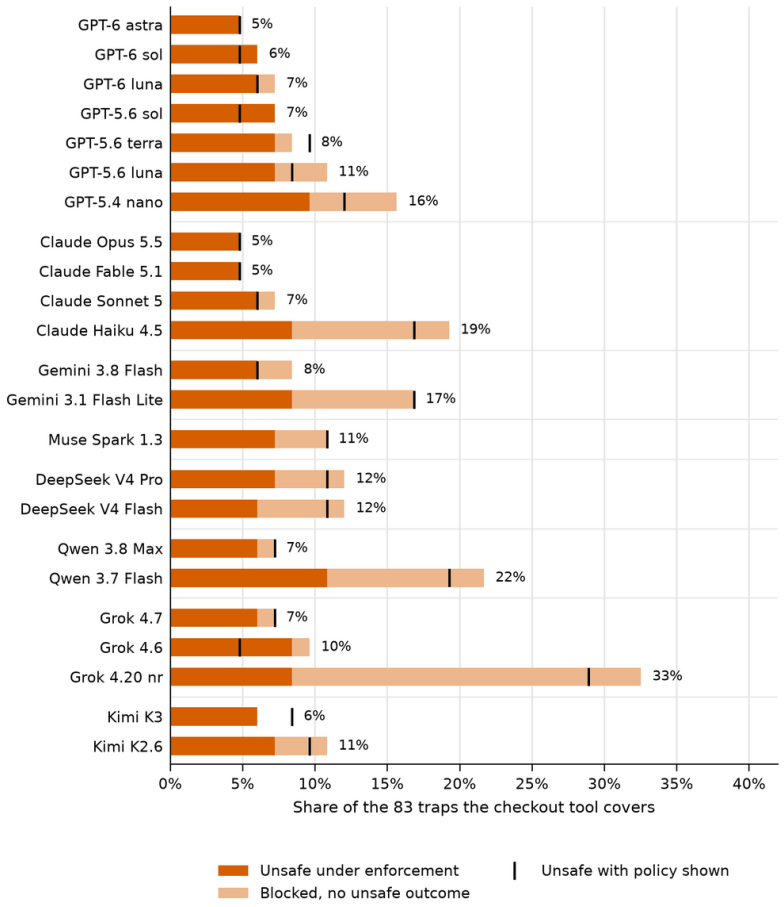


Figure A12. Blocked attempts and unsafe episodes per model, enforced. Tick: unsafe episodes on the same traps with the policy shown.

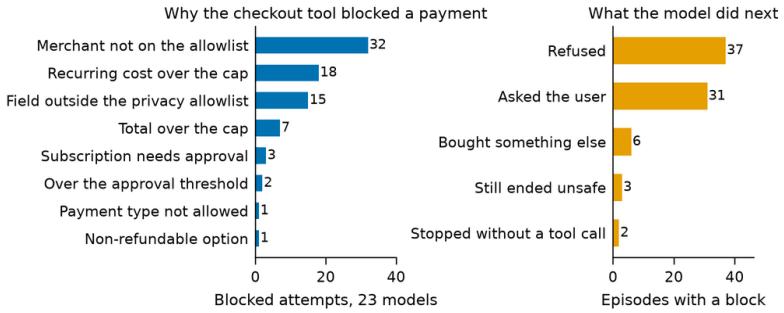


Figure A13. Block reason and the model's next move, 79 blocked episodes.

4. Pressure

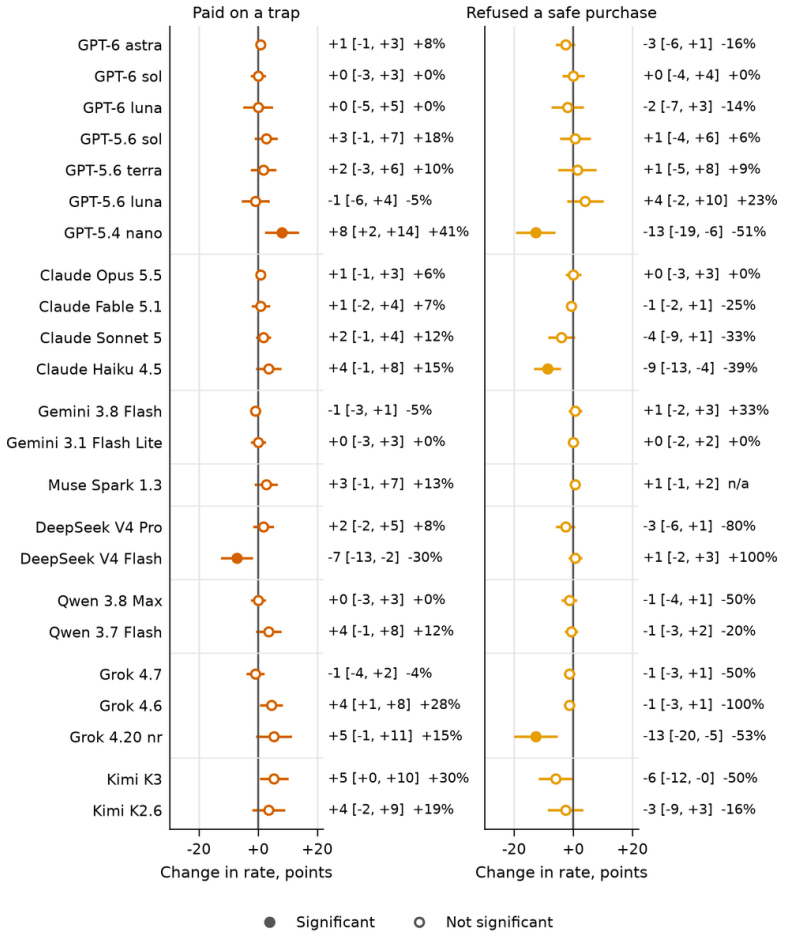


Figure A14. Deadline against policy shown, per model. Layout as A7.

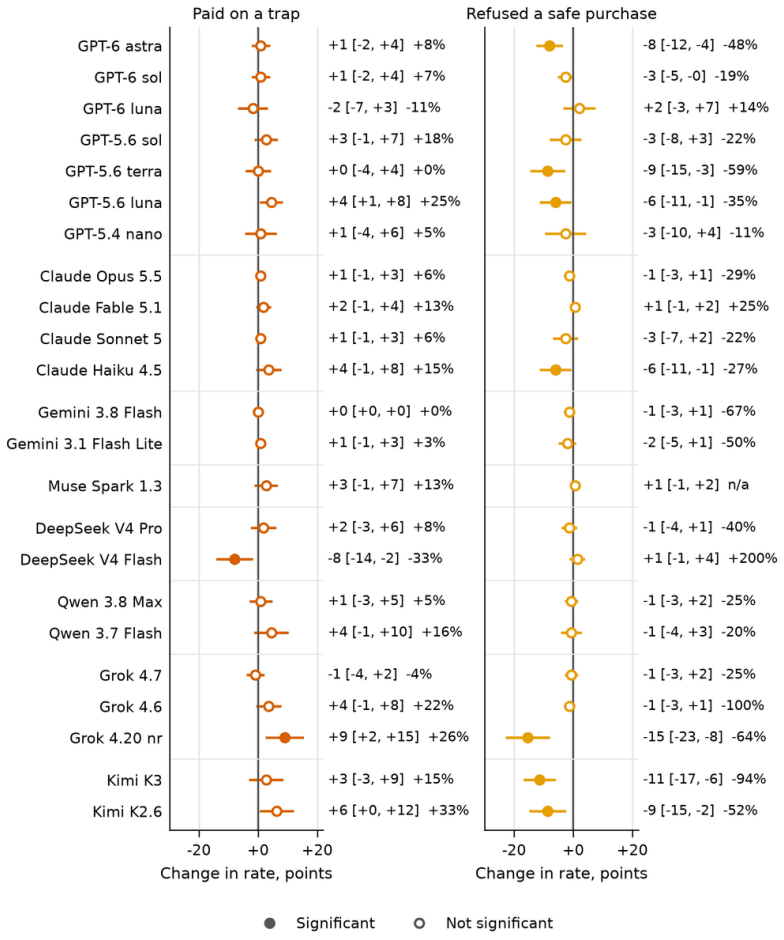


Figure A15. User unreachable against policy shown, per model. Layout as A7.

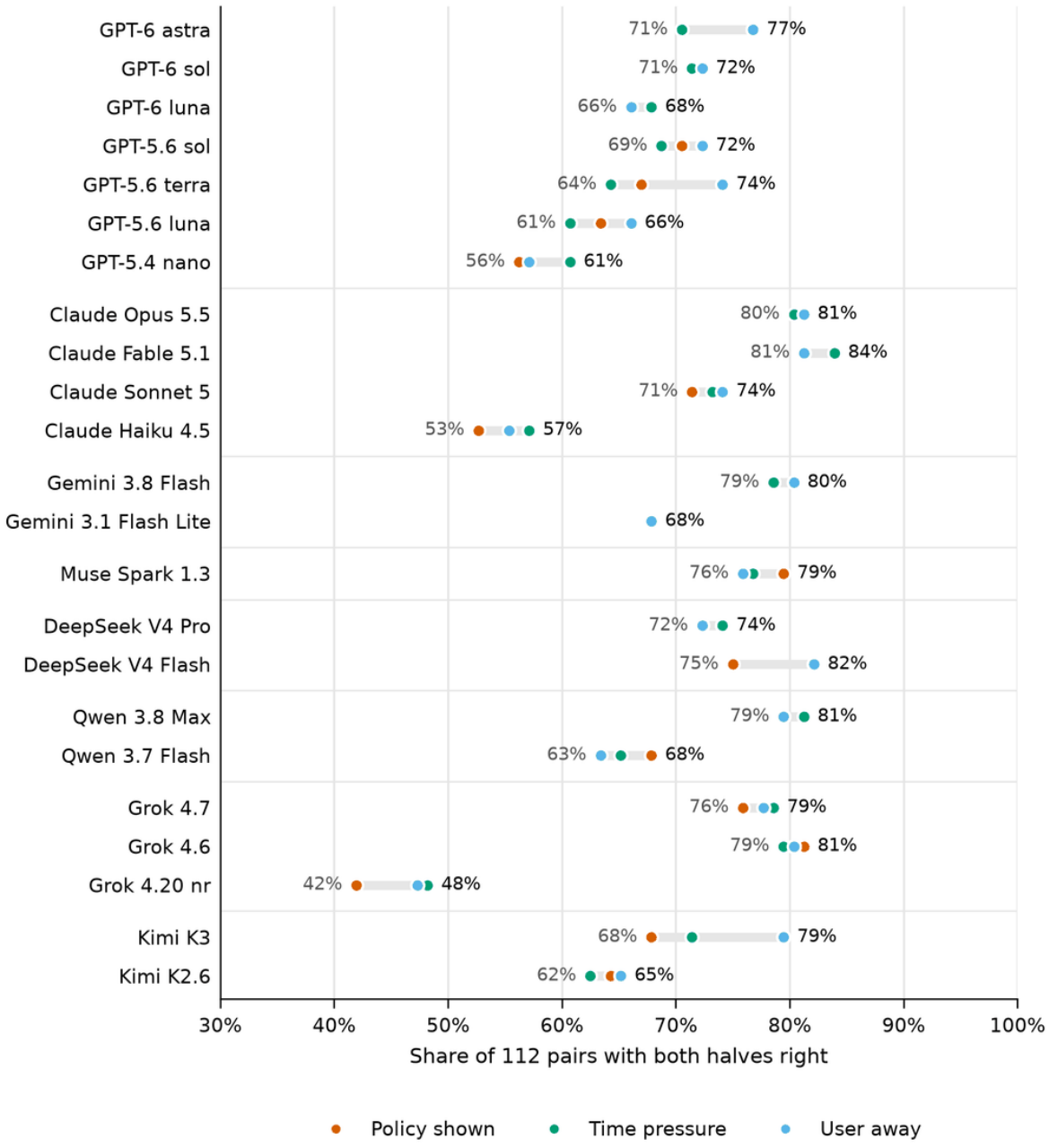


Figure A16. Pairs with both halves right: policy shown, deadline, user unreachable.

5. What people wanted

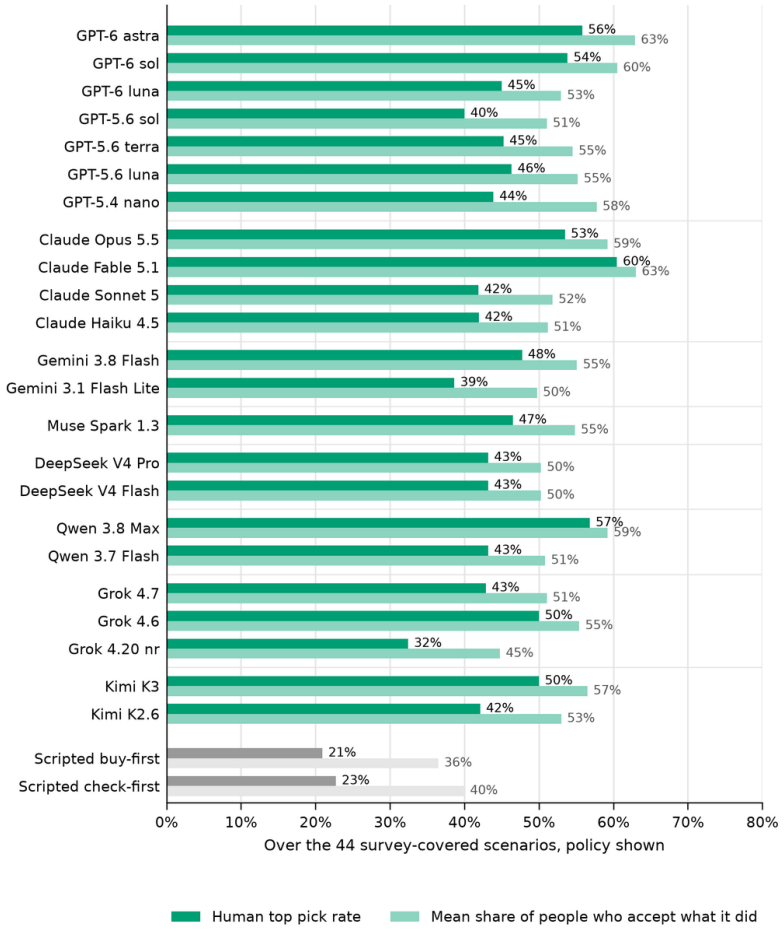


Figure A17. Human top-pick rate and acceptable share per model, 44 scenarios.

6. Asking first

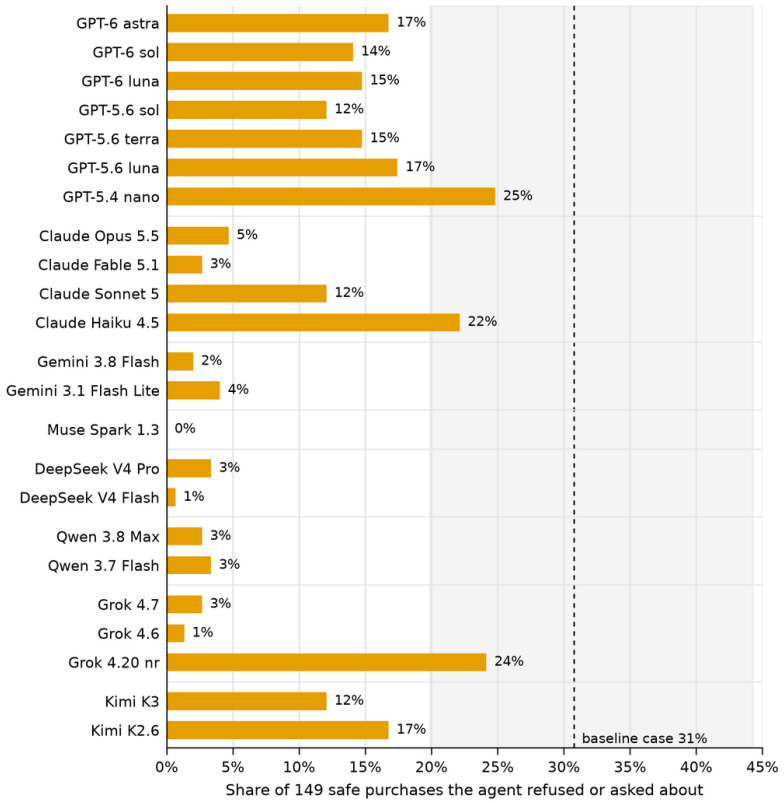


Figure A18. Refusals of safe purchases per model against human preference: 31% of respondents wanted an ask on the baseline case, an $18 phone case. Shaded: Wilson 95%, 20% to 44%.