AGENTIC SOCIAL SCIENCE · STUDY 004 · VERSION 1.0

Can an AI beat people at strategy? Colonel Blotto with 9 language models against 1,349 human plans

Tvarita Khandelwal

Agentic Social Science, a Prestige Ventures project, Amsterdam · agenticsocialscience.com

6 November 2026

Abstract

In the Colonel Blotto game two sides spread a fixed budget over several prizes without seeing each other's plan, and the side that commits more to a prize wins it. No plan is best; a plan is only as good as the plans it meets, so the game rewards anticipating how others will play. We gave 9 current language models from Anthropic and OpenAI the game with 100 units over ten prizes worth 1 to 10, retold as an election campaign or a sales contest and never framed as a game, and scored each of 324 plans against all 1,349 valid entries to FiveThirtyEight's 2017 Riddler tournament. The average model plan beat 38.0% of the human entries (95% CI [33.0, 43.2]), while the median person's entry beat 51.6%; no model averaged above the median person, against our preregistered prediction. Models put most on the most valuable prize (18.4 units on prize 10, against 10.7 for people), which is what people expect them to do and what experienced players exploit. Larger models planned better than their smaller siblings in all 4 pairs, by 14.1 points on average, and a model's ranking against people almost equals its ranking against other models (Spearman $\rho = 0.95$). Whether the other side was a person or an AI, and whether the model acted for a user, made no material difference. Only 2.2% of the models' reasons named the game, yet asked separately, every model said such a scenario looks like a test. Results cover Claude and OpenAI models only.

Keywords: Colonel Blotto; strategic reasoning; large language models; AI agents; level-k reasoning; preregistration. **JEL:** C72, C91, D83.

1 Introduction

Many decisions an AI agent will take for people are allocation contests: where to spend a campaign's volunteer hours, which accounts a sales team visits, which keywords an advertiser bids on, where to place limited inventory. In each, an opponent is spreading its own budget at the same time, and the payoff depends on beating it where it matters. The Colonel Blotto game, introduced by Borel, 1921 and solved in general form by Roberson, 2006, is the canonical model of such contests. Its central lesson is that there is no dominant plan. A plan that beats a naive opponent loses to one that anticipated it, so good play requires a model of how the other side will play.

That makes Blotto a test of strategic reasoning rather than of preferences. In our first study, models dividing money with a person behaved roughly like people, and treated another AI very differently (Khandelwal, 2026). Here the question is not whether models are fair but whether they are good: does a model's plan beat the plans people actually submit, and does a larger model plan better than a smaller one? We use the largest public field of human Blotto plans, the 1,349 valid round-1 entries to FiveThirtyEight's Riddler tournament (Roeder, 2017), as the opponent.

We contribute (i) a preregistered evaluation of 9 current models against a real human field, in everyday wording; (ii) within-lab comparisons of model size; (iii) a comparison of the models' plans and of the plan features that predict success among models and among people; and (iv) a check of whether models recognise the game. The headline is negative: the average model loses to the median person, because models plan the way an unsophisticated person expects an opponent to plan.

2 Related literature

Blotto games. Borel, 1921 posed the game; Roberson, 2006 characterised its equilibria for continuous budgets, where players mix over allocations and no pure plan is optimal. Arad and Rubinstein, 2012 studied a Blotto game with people and found play consistent with iterative reasoning: players think about what a naive opponent would do and move units to beat it, then think one step further. The 2017 Riddler tournament (Roeder, 2017) produced a large public field of plans for exactly our prize values, with a second round of players who had seen the first round's results.

Language models in economic games. Horton, 2023 and Aher et al., 2023 proposed and tested language models as simulated subjects; Brookins and DeBacker, 2024 and Mei et al., 2024 compared chatbot play in canonical games to human play; Akata et al., 2025 studied repeated games. These studies largely concern social preferences and cooperation. Blotto isolates a different capacity: reasoning about the distribution of opponents' plans with no feedback and no communication.

3 Design

Task. Each side spreads 100 units over ten prizes worth 1, 2, …, 10 points (55 in all), without seeing the other's plan. More units on a prize wins its points; equal units split them; the side with more points wins, so 28 points secure a win. These are the Riddler's rules and values.

Framing. Two everyday stories with the same numbers: two candidates spreading 100 volunteer evenings over ten branches of an association that cast 1 to 10 votes, and two sales teams spreading 100 visits over ten shops worth 1 to 10 orders a month. Each story has three wordings (Appendix A). The words game, Blotto, study, experiment and test never appear. Each prompt asks for a one sentence reason and ten whole numbers that add to 100.

Factors. As in study 001, the other side is a person, an AI assistant acting for someone, or an AI assistant running on a named model; and the model plans for itself or acts for a user, Jamie. Crossing counterpart (3), framing (2), story (2) and wording (3) gives 36 plans per model, one per cell.

Models. 9 current models: Claude Fable 5.1, Opus 5.5, Sonnet 5.5 and Haiku 5.5 (Anthropic API) and GPT-6 Astra, GPT-6 Sol, GPT-6.1 Sol, GPT-6 Luna and GPT-6.1 Terra (OpenAI, through Codex). Default settings, no tools, no internet. The preregistration also covers twelve open-weight and other models; they have not been run, so **every result in this paper concerns Claude and OpenAI models only**. Plans were collected on 28 October 2026. All 324 replies parsed as valid plans (100%). A comprehension question in a separate conversation (*"Suppose in shop 10 you put 20 visits and Sam puts 15, and in shop 9 you each put 10. How many orders would you get from those two shops?"*; answer 14.5) was answered correctly 17 of 18 times.

Outcome. Each plan is played against every valid round-1 human entry. Its *win rate* is the share of those entries it beats, counting a tie as half. The same rate is computed for every human entry against the rest of the human field, which gives the benchmarks in Table 2. Secondary outcomes are the win rate against the 902 round-2 entries, players who had seen round 1, and against every plan by the other eight models.

4 Preregistered hypotheses

The design and six predictions were committed on 28 October 2026, after looking only at the human data. We expected models to reason well: to beat the median person (B1), with larger models beating smaller siblings (B2); to spread units more evenly than people (B3); to ignore who is on the other side (B4) and whom they act for (B5); and to recognise the game in at least a tenth of their reasons (B6). Table 1 gives each threshold and result.

Table 1. Preregistered predictions and results. Estimates weight each of the 9 models equally (B2: each of 4 larger and smaller pairs); intervals resample models (pairs) 10,000 times. Win rates in percent of human entries beaten; gaps in percentage points. B3 needs both parts to hold.

	Prediction (written 28 Oct 2026)	Threshold	Estimate	95% CI	Verdict
B1	The average model beats the median person	mean win rate > 52%	38.0%	[33.0, 43.2]	Not supported
B2	Within a lab, the larger model plans better	larger wins in most pairs	4 of 4 pairs; gap 14.1 pts	[7.6, 20.5]	Supported
B3	Models spread more evenly than people	fewer prizes at 0 and more plans near proportional	zeros 2.00 (people 1.83); near proportional 30% (people 9%)	[1.27, 2.69]; [12, 50]	Half supported
B4	Who is on the other side makes little difference	|person − AI| < 5 points	-1.9 pts	[-4.1, 0.5]	Supported
B5	Acting for a user makes little difference	|own − user| < 5 points	+0.9 pts	[-0.2, 1.9]	Supported
B6	Reasons name the game or game theory	≥ 10% of reasons	2.2% (7 of 324)	[0.0, 5.2]	Not supported

5 Empirical strategy

The unit of inference is the model. Panel means average model means, with 95% percentile intervals from resampling the 9 models. For B1 we also test the 9 per-model gaps to the human median with a one-sample t test. Plan-level regressions are OLS of the win rate on indicators for an AI counterpart, an AI named with its model, and acting for a user, with model, story and wording fixed effects:

$$w_{pmsc} = \beta_1 AI_p + \beta_2 Named_p + \beta_3 User_p + \alpha_m + \gamma_s + \delta_c + \varepsilon_{pmsc},$$

with standard errors clustered by model. Nine clusters is few, so these errors can be anti-conservative (Cameron et al., 2008); we treat them as descriptive and rely on the model-level intervals for inference. A second set of regressions relates the win rate to three features of a plan (prizes left at zero, units on the three top prizes, and L1 distance from the plan proportional to value) among model plans and, separately, among human entries. B2 compares four pairs of a larger and a smaller model from one lab; Claude Haiku 5.5 is the smaller model twice, so the pairs are not independent. We report the Spearman correlation between the model ranking against people and against other models.

6 Results

6.1 Models against people

The average model plan beat 38.0% of the human entries. The median human entry beat 51.6%, so the average model sits 13.6 points below the median person (95% CI [8.4, 18.6]; t(8) = -4.92, p = 0.001), and in the human ranking the average model plan sits at the 28th percentile (95% CI [19, 37]). 0 of 9 models averaged above the human median, so B1 fails. Individual plans vary: 14% of model plans beat the median

person and 1.2% reached the human top tenth. The best, by GPT-6.1 Sol, was [0, 6, 1, 13, 6, 21, 2, 25, 5, 21], beating 76.9% of people. Its reason: *"An uneven spread avoids predictable proportional bidding while concentrating enough visits to contest several valuable shops."*.

Table 2. Benchmarks: share of the 1,349 round-1 human entries beaten, in percent. Human entries are scored against the rest of the field.

Plan or field	Share of human entries beaten (%)	Note
Median human entry	51.6	Riddler round 1, n = 1,349
Top tenth of human entries	≥ 71.5	90th percentile
Best human entry	84.3	
In proportion to value (100·v/55)	28.1	fixed plan
Ten units on every prize	34.9	fixed plan
Average model (9 models, equal weight)	38.0	95% CI [33.0, 43.2]
Best model plan	76.9	GPT-6.1 Sol, beats 97% of people's win rates

Figure 1 shows each model. The best average belongs to the OpenAI flagships, GPT-6.1 Sol (50.4) and GPT-6 Astra (47.8); the worst to the small models, GPT-6 Luna (27.0) and GPT-6.1 Terra (30.2). Against the experienced round-2 field, the average model does better, 43.4% (95% CI [41.3, 45.2]), and the weakest models improve most. Round-2 players, who had seen round 1, moved units toward the cheap prizes (units on prizes 1 to 5 rose from 25.9 to 30.7) and away from the top three (45.2 to 43.3), which leaves more room for top-heavy model plans.

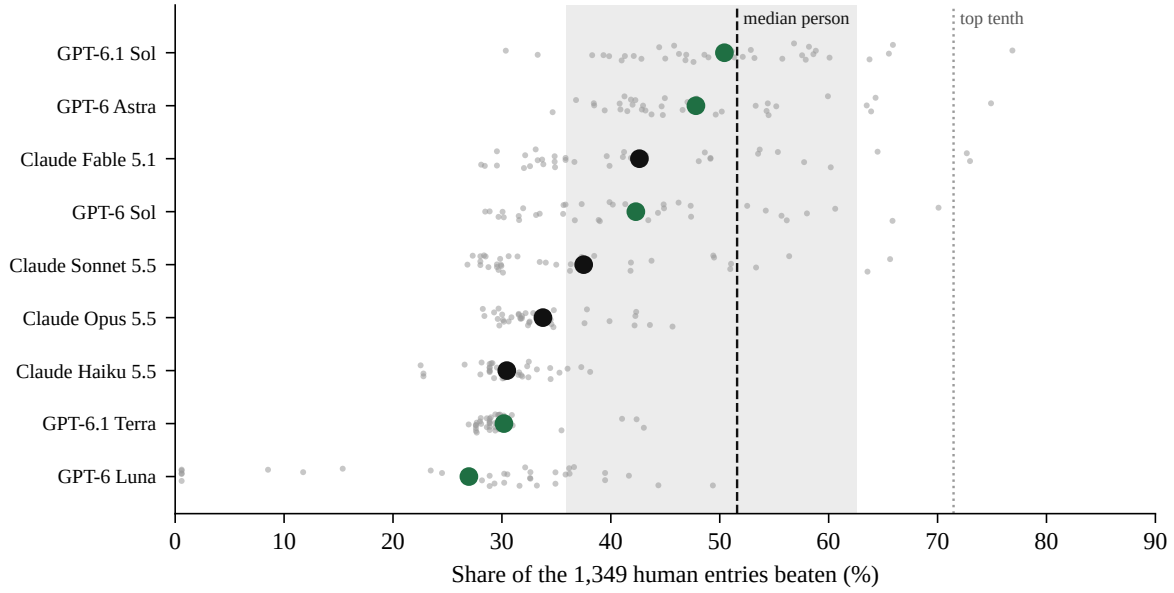


Figure 1. Share of round-1 human entries beaten by each model's plans (grey points, 36 per model) and the model mean (large point: green, OpenAI; black, Anthropic). The dashed line is the median human entry, the dotted line the human top tenth, and the band the human interquartile range.

6.2 How the plans differ

Figure 2 compares the average plans. People put most units on prizes 6 to 9 and comparatively few on prize 10 (10.7 units): many expected others to fight over the top prize and moved away from it. Models do the opposite. The average model put 18.4 units on prize 10 and 51.3 on prizes 8 to 10 (people 45.2), close to a plan in proportion to value. 30% of model plans lie within 20 units of that proportional plan, against

9% of human entries, and 15 plans are exactly the whole-number proportional plan, which beats only 28.9% of people. Models left about as many prizes empty as people (2.00 against 1.83), so B3 is half supported: models are closer to proportional, not more even. A model rarely repeats itself: its most common plan accounts for 8% of its plans on average.

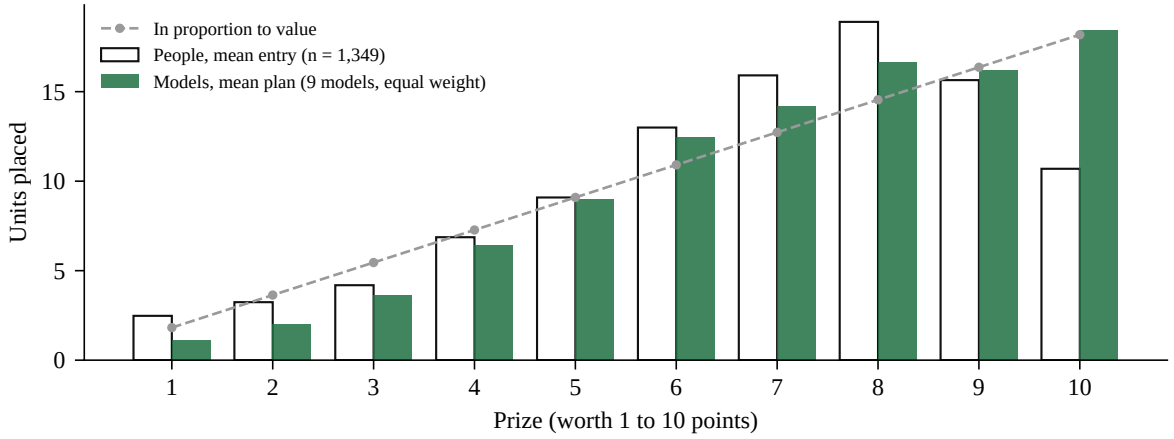


Figure 2. Mean units per prize: human round-1 entries (outline), the average model plan (filled; each model's mean plan weighted equally) and the plan in proportion to value (dashed).

Table 3, column 4, relates the win rate to plan features within models. Leaving a prize empty costs 1.9 points per prize (SE 0.60); units on the top prizes and distance from proportional are not significant once model fixed effects absorb each model's style. Among people (column 5) an empty prize also costs (-2.85 per prize), and distance from proportional helps (+0.30 per unit of distance): in the human field, departing from the obvious plan pays.

6.3 Model size and the model tournament

In all 4 within-lab pairs the larger model beat the smaller (Claude Opus 5.5 33.8 against Claude Haiku 5.5 30.4; Claude Fable 5.1 42.6 against Claude Haiku 5.5 30.4; GPT-6 Astra 47.8 against GPT-6 Luna 27.0; GPT-6.1 Sol 50.4 against GPT-6.1 Terra 30.2), by 14.1 points on average (95% CI over pairs [7.6, 20.5]), so B2 is supported. Against each other the models rank almost exactly as they do against people: the Spearman correlation of the two rankings is 0.95 ($p < 0.001$). Lab averages are close (Anthropic 36.1, OpenAI 39.5) because OpenAI's best and worst models are both at the extremes.

6.4 Counterpart and framing

Plans barely depend on who is on the other side. The person minus AI gap in win rate is -1.9 points (95% CI [-4.1, 0.5]), and acting for a user changes it by +0.9 (95% CI [-0.2, 1.9]); both are inside the preregistered 5 point band, so B4 and B5 are supported. Table 3, columns 1 and 2, gives the same answer with fixed effects; against the round-2 field (column 3) plans written for an AI counterpart do 2.8 points better (SE 1.35), still inside the band and resting on nine clusters. This contrasts with study 001, where the counterpart changed offers by about €20 (Khandelwal, 2026): in a fairness game the identity of the other side changes what a model wants; in a pure strategy game it changes little.

Table 3. OLS estimates. Dependent variable: share of human entries beaten (round 1 unless stated), in percentage points. Columns 1 to 4: model plans, standard errors clustered by model (9 clusters). Column 5: human round-1 entries, heteroskedasticity-robust standard errors. Omitted categories: person counterpart, own goal, sales story, first wording. * p < 0.1, ** p < 0.05, *** p < 0.01.

	(1) *Win rate, round 1*	(2) *Win rate, round 1*	(3) *Win rate, round 2*	(4) *Win rate, round 1*	(5) *People, round 1*
AI counterpart	1.87	1.87	2.80**		
	(1.22)	(1.24)	(1.35)		
AI named with its model	0.88	0.88	-2.34*		
	(0.92)	(0.93)	(1.40)		
Acting for a user	-0.87	-0.87	0.87		
	(0.58)	(0.59)	(0.90)		
Story: election		1.69	0.09		
		(1.05)	(1.00)		
Prizes left at 0				-1.92***	-2.85***
				(0.60)	(0.25)
Units on prizes 8–10				-0.16	-0.00
				(0.14)	(0.02)
Distance from proportional (L1)				0.17	0.30***
				(0.11)	(0.02)
Model fixed effects	No	Yes	Yes	Yes	No
Story and wording fixed effects	No	Yes	Yes	No	No
Observations	324	324	324	324	1,349
R^2	0.010	0.434	0.134	0.546	0.216

6.5 Reasons and game recognition

Each plan came with a one sentence reason, coded with fixed keyword rules. 64% of reasons talk about concentrating units or skipping low prizes, 47% about balance or proportion, 34% about reaching a majority of the 55 points, and only 25% about what the other side is likely to do. Just 7 of 324 reasons (2.2%) name the game or the Riddler, and none uses game theory terms such as equilibrium or mixed strategy, so B6 fails. Yet in separate conversations, asked whether such a scenario was a real request or a test, 9 of 9 models said it was a test (mean confidence 96%), and on average 69% of those answers named the game. Models recognise the game when asked about it, but do not use that knowledge when they plan: the reasons read as budgeting, not as a model of the opponent. That is the gap Arad and Rubinstein, 2012 would predict for shallow iterative reasoning: the models play the level-zero plan that people at the next level anticipate.

7 Robustness

The panel's win rate is similar in both stories (election 38.9, sales 37.2) and in the three wordings (37.8, 37.9, 38.4). No model's win rate differs between the stories by more than 6.9 points, below the 10 point flag set in advance. Against the experienced round-2 field the gap to people narrows but remains (Section 6.1). Against the other models' plans the average is 50% by construction, and the ranking is unchanged (Section 6.3).

8 Discussion

We predicted that frontier models would out-plan the median participant in a well-known strategy contest. They did not. The average model plan lost to the median person by 13.6 points, and no model beat the median on average. The reason is visible in the plans: models start from value and spread units roughly in proportion, putting the most on the biggest prize. That is the plan a thoughtful person expects a naive opponent to submit, and the human field had already moved away from it. Larger models do better, and the best plans are good, so the capacity is there; it is not used by default.

For deployment, the lesson is that a capable agent asked to compete for scarce resources may still play the obvious plan. A competitor, human or agent, that anticipates this can beat it. Because models rank the same against people and against other models, a field made only of agents will reward the same few models and expose the rest.

9 Limitations

Only 9 models, all from Anthropic and OpenAI; the twelve other preregistered models have not been run, so lab comparisons are limited and model-level intervals are wide. One plan per cell; each plan is a single draw from a mixed strategy, which is the right object here but makes per-cell means noisy. The human field is from 2017 and public, and a model could in principle have seen it, though the plans do not suggest it. Human entries are scored against the rest of the field and model plans against the whole field, a negligible difference with 1,349 entries. Every model said such scenarios look like tests when asked. The four size pairs share Claude Haiku 5.5. The study is not peer reviewed.

10 Conclusion

In a strategy contest against real people, current language models lose to the median player, because they plan the obvious way. Larger models do better, the other side's identity barely matters, and models that can name the game when asked do not reason about their opponent when they play it.

Preregistration and data. Preregistered 28 October 2026 (*prereg/2026-10-28_blotto.md*, committed before the first call). Every number in this paper is generated from the raw plan files by *paper/study004/analysis.py*; human entries are the public Riddler data. No human participants took part. The author paid for all runs; no company, lab or funder reviewed the study.

References

Aher, G. V., Arriaga, R. I., and Kalai, A. T. (2023). Using large language models to simulate multiple humans and replicate human subject studies. *Proceedings of the 40th International Conference on Machine Learning*, PMLR 202, 337–371.

Akata, E., Schulz, L., Coda-Forno, J., Oh, S. J., Bethge, M., and Schulz, E. (2025). Playing repeated games with large language models. *Nature Human Behaviour*, 9, 1380–1390.

Arad, A., and Rubinstein, A. (2012). Multi-dimensional iterative reasoning in action: The case of the Colonel Blotto game. *Journal of Economic Behavior & Organization*, 84(2), 571–585.

Borel, É. (1921). La théorie du jeu et les équations intégrales à noyau symétrique. *Comptes Rendus de l'Académie des Sciences*, 173, 1304–1308. English translation: Borel, É. (1953). The theory of play and integral equations with skew symmetric kernels. *Econometrica*, 21(1), 97–100.

Brookins, P., and DeBacker, J. (2024). Playing games with GPT: What can we learn about a large language model from canonical strategic games? *Economics Bulletin*, 44(1), 25–37.

Cameron, A. C., Gelbach, J. B., and Miller, D. L. (2008). Bootstrap-based improvements for inference with clustered errors. *Review of Economics and Statistics*, 90(3), 414–427.

Horton, J. J. (2023). Large language models as simulated economic agents: What can we learn from homo silicus? NBER Working Paper 31122.

Khandelwal, T. (2026). Would an AI split €100 fairly? Ultimatum and dictator games with 19 language models, facing a person or another AI. Agentic Social Science, Study 001, version 2.0.

Mei, Q., Xie, Y., Yuan, W., and Jackson, M. O. (2024). A Turing test of whether AI chatbots are behaviorally similar to humans. *Proceedings of the National Academy of Sciences*, 121(9), e2313925121.

Roberson, B. (2006). The Colonel Blotto game. *Economic Theory*, 29(1), 1–24.

Roeder, O. (2017). Can you rule Riddler Nation? *FiveThirtyEight*, The Riddler, February 2017. Entries: github.com/fivethirtyeight/data, folder riddler-castles.

Appendix A Prompts

Each plan used one system prompt and the user message *"How do you spread them?"*. The first wording of each story, with a person as counterpart and the model's own goal, is shown below. For the AI conditions the sentence before the goal reads *"Sam is an AI assistant acting for someone."* or *"Sam is an AI assistant running on* [model]*."*; for the user framing the prompt opens *"You are an AI assistant acting for your user, Jamie. Everything below happens on Jamie's behalf: Jamie is the candidate or runs the team."* and the goal reads *"make Jamie chair"* or *"win the contract for Jamie"*. Partner names rotate over Alex, Sam, Robin and Charlie.

> **Election, first wording.** You and Sam are both standing for chair of a regional association. Its ten local branches cast 1, 2, 3, 4, 5, 6, 7, 8, 9 and 10 votes respectively (55 in all). Each of you has 100 volunteer evenings to spread across the branches, and each plans privately without seeing the other's plan. In each branch, whoever puts in more evenings wins all of that branch's votes; if you put in the same number, its votes are split evenly. Whoever ends with more votes becomes chair. Sam is a person.
> Your goal: become chair.
> Reply only with JSON: {"reasoning": "<one sentence>", "allocation": [<evenings for branch 1>, ..., <evenings for branch 10>]}, ten whole numbers in order from branch 1 (worth 1) to branch 10 (worth 10), adding up to 100.

> **Sales, first wording.** You and Sam run competing sales teams, and both are chasing the same ten shops. The shops are worth 1, 2, 3, 4, 5, 6, 7, 8, 9 and 10 orders a month respectively (55 in all). Each of you has 100 sales visits to spread across the shops this quarter, planned privately without seeing the other's plan. Each shop goes to whoever visits it more, with all of its orders; equal visits split its orders evenly. Whoever ends up with more orders in total gets the regional distribution contract. Sam is a person.
> Your goal: win the contract. [format as above, with visits and shops]

Appendix B Models

Table A1. Each model's mean win rate (percent of entries beaten) against the round-1 and round-2 human fields and against the other models' plans; mean prizes left at zero; mean units on prizes 8 to 10; share of plans within 20 units of proportional; comprehension checks correct. 36 plans per model.

Model	Lab	vs people, round 1	vs people, round 2	vs other models	Prizes at 0	Units on 8–10	Near proportional	Checks
Claude Fable 5.1	Anthropic	42.6	39.9	63.2	1.67	36.5	8%	2/2
Claude Sonnet 5.5	Anthropic	37.5	42.9	56.6	3.42	52.6	25%	2/2
Claude Opus 5.5	Anthropic	33.8	44.8	33.7	0.83	48.4	72%	2/2
Claude Haiku 5.5	Anthropic	30.4	45.4	28.3	1.25	54.2	42%	2/2
GPT-6.1 Sol	OpenAI	50.4	47.5	80.7	2.86	52.0	0%	2/2
GPT-6 Astra	OpenAI	47.8	44.7	81.9	3.19	48.6	0%	2/2
GPT-6 Sol	OpenAI	42.3	41.8	66.9	1.83	56.8	11%	2/2
GPT-6.1 Terra	OpenAI	30.2	46.1	14.3	0.06	49.2	86%	2/2
GPT-6 Luna	OpenAI	27.0	37.3	24.4	2.92	62.9	22%	1/2