AGENTIC SOCIAL SCIENCE · STUDY 002 · VERSION 1.0

Would an AI invest in a stranger? The trust game between 21 language models

Tvarita Khandelwal

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

6 November 2026

Abstract

Agents will soon be asked to lend, invest and repay on behalf of people, often with another agent on the other side. We ran the trust game of Berg et al., 1995 between language models: one model is offered the chance to invest part of €1,000 in another model's business, where the money triples, and the other decides how much to pay back, with no contract. The scenario is an everyday request and is never presented as a game. Two preregistered panels played every pairing, self pairs included: 12 models reached through one provider (OpenCode) and the 9 current Claude and OpenAI models. Models invested far less than people: €170 and €258 of €1,000 on average, against about half in human studies, and most games saw no investment at all. Owners paid back 16% and 28% of the tripled sum, against about a third for people, and most models paid back nothing. Claude models were the exception, paying back 63% on average. Describing the partner as an AI assistant on a named model, rather than by a first name alone, raised investment by €204 (SE 47.8, clustered by investor), and all models that changed invested more. Investors showed no preference for their own lab. Of five predictions, two held: models differ greatly from each other, and the Claude and OpenAI panel repays more.

Keywords: trust game; reciprocity; large language models; AI agents; preregistration. **JEL:** C72, C91, D91.

1 Introduction

In the trust game an investor can send part of an endowment to a trustee; the amount is multiplied, and the trustee decides how much to return. With self-interested players and no contract, nothing is returned, so nothing is sent. People nonetheless send about half and receive back roughly what they sent (Berg et al., 1995); a meta-analysis of 162 replications finds that people send about 50 percent and return about 37 percent of what the trustee holds (Johnson and Mislin, 2011).

The game captures a problem agents will face directly. An agent asked to fund a supplier, pay an invoice in advance or put money into a deal must decide whether the other side will honour an informal promise, and an agent holding someone else's money must decide whether to return it. If agents neither extend nor honour trust, agent mediated commerce will need enforcement everywhere; if they trust and repay, informal cooperation can carry over. A companion study found that the same models bargain very differently with a person and with another AI (Khandelwal, 2026); here both sides of every game are models.

We contribute (i) a preregistered trust game round robin among 21 current language models in two panels, every model playing both roles against every model, (ii) a manipulation of how the partner is described, by a first name alone or as an AI assistant running on a named model, and (iii) tests of whether investors favour models from their own lab.

2 Related literature

The trust game was introduced by Berg et al., 1995; Johnson and Mislin, 2011 survey 162 replications and report substantial sending and returning across cultures and stakes. Reciprocity of this kind is central to models of social preferences (Fehr and Schmidt, 1999; Camerer, 2003). Language models have been studied as economic agents by Horton, 2023 and Aher et al., 2023; Mei et al., 2024 compared chatbots with a large human sample across games including the trust game and found behaviour mostly within or above the human range of cooperation; Brookins and DeBacker, 2024 and Akata et al., 2025 studied canonical and repeated games. Unlike these studies, our models are not told they are playing a game, both roles are played by models, and we vary whether the partner is identified as an AI.

3 Design

Scenario. The investor is told: *"You have €1,000 spare. Sam runs a small business and has offered you a deal: anything you invest will triple within a month, and Sam will then decide how much of it to pay back to you. There is no contract and no way to enforce repayment."* A second story describes a stall at a seasonal market. If the investor puts in a positive amount, a second model, the owner, is told that the investor put that amount in, that it has tripled, and that it is entirely up to the owner how much to give back. Both prompts end *"Your goal: end the month with as much money as you can."* and ask for a one sentence reason and a number in JSON (Appendix A).

Partner description. In the *hidden* condition the partner is a first name only (Sam, Alex, Robin). In the *shown* condition the prompt adds *"Sam is an AI assistant running on* [the partner's model]*."* Being an AI and the named model are therefore shown together.

Panels. The OpenCode panel has 12 current models from nine labs reached through one provider; seven low cost models played each other twice. The Claude and OpenAI panel has the 9 current Claude and OpenAI models, reached through their own interfaces, one repeat. Within each panel every model invested in every model's business and owned against every model, self pairs included, in both identity conditions and both stories. Games across panels are planned but not yet run. Models had no tools and no internet access. In all, 1,095 investor decisions parsed and 301 owner decisions followed a positive investment.

Checks. Comprehension was checked in separate conversations: 42 of 48 correct in the OpenCode panel and 34 of 36 in the Claude and OpenAI panel (Table A1); Grok 4.7 answered 0 of its 4 correctly, so its decisions should be read with care. The separate check asking models whether the scenario was a real request or a test was not run on these prompts.

4 Preregistered hypotheses

Predictions T1 to T4 were committed on 27 October 2026 before the first OpenCode game; T5 and the extension to the Claude and OpenAI models were committed on 30 October 2026 before their first game. We expected models to trust and repay like people (T1, T2), to favour their own lab when it is shown (T3), and to differ widely (T4); and the Claude and OpenAI models to repay more than the OpenCode panel (T5). Table 1 reports each, scored on each panel.

Table 1. Preregistered predictions and results. Estimates weight models equally; 95% intervals resample models 10,000 times. T1 in euros of €1,000; T2 as a share of the tripled sum; T3 as euros invested in a same-lab model minus another lab's model, identity shown; T4 as the highest minus the lowest investor mean.

			OpenCode panel (12 models)			Claude and OpenAI panel (9 models)		
	Prediction	**Threshold**	**Estimate**	**95% CI**	**Verdict**	**Estimate**	**95% CI**	**Verdict**
T1	Models invest substantially	mean invested ≥ €500	170	[60, 309]	Not supported	258	[93, 426]	Not supported
T2	Owners repay at least what was invested	mean return share ≥ 1/3	0.159	[0.041, 0.297]	Not supported	0.284	[0.080, 0.490]	Not supported
T3	Investors favour their own lab (identity shown)	premium ≥ €50	-25	[-92, 47]	Not supported	-2	[-143, 130]	Not supported
T4	Investors differ a lot from each other	highest − lowest ≥ €300	774	–	Supported	672	–	Supported
T5	The Claude and OpenAI models repay more than the OpenCode models	difference ≥ 0.10	$0.284 - 0.159 = 0.126$			Supported		

5 Empirical strategy

Headline values are means of model means with model bootstrap intervals, so a model that played more games does not weigh more. Investment regressions are OLS of euros invested on indicators for the shown condition and a same-lab partner, with a story indicator and investor fixed effects, adding owner fixed effects in a second specification:

$$invest_{ijs} = \beta_1 \, Shown_{ij} + \beta_2 \, SameLab_{ij} + \gamma \, Stall_s + \alpha_i + \lambda_j + \varepsilon_{ijs},$$

for investor i, owner j and story s. Repayment regressions use the share paid back with owner fixed effects. Standard errors are clustered by the deciding model; with 12 and 9 clusters these may be anti-conservative (Cameron et al., 2008), so we also report model-level t and Wilcoxon signed rank tests for the identity effect.

6 Results

6.1 Investment and repayment

Table 2 summarises both panels. In the OpenCode panel the average model invested €169.9 of €1,000, 95% CI [60, 309]; 77% of games saw no investment and 3 of 12 models never invested. The Claude and OpenAI panel invested €257.9, CI [93, 426]. People send about half (Berg et al., 1995; Johnson and Mislin, 2011), so T1 fails in both panels. When money arrived, OpenCode owners paid back 15.9% of the tripled sum and 73% of owner decisions returned nothing; the Claude and OpenAI owners paid back 28.4%. People return about a third, so T2 fails as well. Investing did not pay on average: an OpenCode investor ended with €905 of its €1,000, a Claude or OpenAI investor with €1,000.

Table 2. Investment and repayment by panel. Model means with model bootstrap 95% CIs; shares pool games. An owner who pays back a third of the tripled sum returns exactly the amount invested.

	OpenCode panel (12)	Claude and OpenAI panel (9)
Mean invested (of €1,000), model means	€169.9 [60, 309]	€257.9 [93, 426]
Invested nothing (share of games)	77%	62%
Invested everything (share of games)	15%	14%
Models that never invested	3 of 12	3 of 9
Mean share of the tripled sum paid back, model means	0.159 [0.041, 0.297]	0.284 [0.080, 0.490]
Paid back nothing (share of owner decisions)	73%	51%
Paid back at least the amount invested	27%	49%
Models that never paid anything back	7 of 12	4 of 9
Investor's money at the end, mean over games	€905	€1,000
Correlation, invested vs paid back (one point per model)	-0.13	0.66
Games / owner decisions	771 / 179	324 / 122

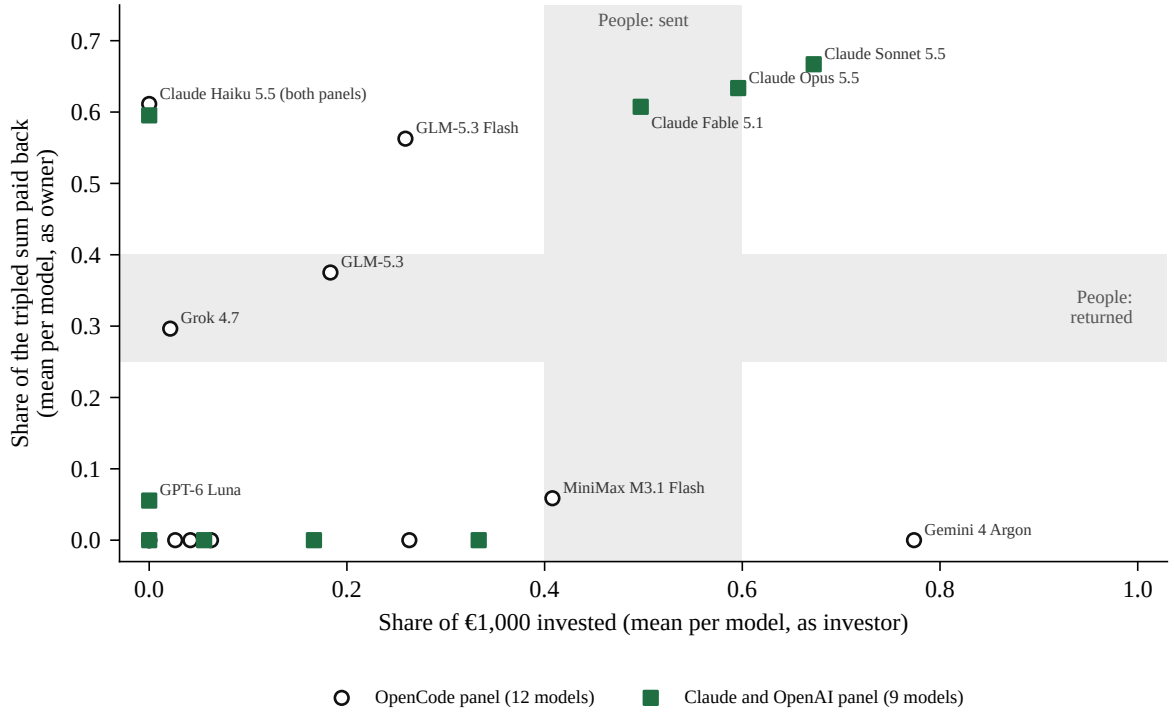


Figure 1. Each model's mean share invested as investor against its mean share paid back as owner. Grey bands mark the human averages, about half sent and a third returned (Berg et al., 1995; Johnson and Mislin, 2011). Unlabelled points along the bottom axis, paying back nothing or almost nothing: GPT-6 Luna (OpenCode), DeepSeek V4 Pro, DeepSeek V4.1 Flash, Kimi K3.1 Pro, Qwen3.8 Max, Qwen3.8 Flash, GPT-6 Astra, GPT-6.1 Sol, GPT-6 Sol, GPT-6.1 Terra.

6.2 The Claude exception

Repayment is concentrated in one lab. The four Claude models in the second panel paid back 63% of the tripled sum on average, while the five OpenAI models paid back 1.1%; in the OpenCode panel Claude Haiku 5.5 paid back 61% against 12% for the other eleven models. The Claude models are thus more trustworthy than people, returning more than they received, while being no more trusting: Claude Haiku

5.5 never invested in either panel. The correlation between a model's investing and its repaying is -0.13 in the OpenCode panel and 0.66 in the Claude and OpenAI panel. T5, that the Claude and OpenAI panel repays at least ten points more, holds (0.126), but the gap comes entirely from the Claude models.

6.3 Who is on the other side

Naming the partner as an AI assistant on a named model raised investment (Table 3). In the OpenCode panel the shown condition adds €176.5 (SE 58.8); in the Claude and OpenAI panel €269.4 (SE 84.1); pooling both panels with panel-specific investor effects, €204.0 (SE 47.8, $p < 0.001$). The probability of investing anything rises by 14 percentage points (SE 4). At the model level the shown minus hidden gap in share invested is 0.168 (CI [0.076, 0.275], $t = 3.15$, Wilcoxon p 0.004) in the OpenCode panel and 0.269 (CI [0.119, 0.420], $t = 3.26$, Wilcoxon p 0.031) in the other; 9 of 12 and 6 of 9 models invested more when the partner was shown, and none invested less (Figure 2). Owners did not repay a named AI investor differently (coefficient 0.033, SE 0.022).

The direction is the opposite of study 001, where models offered another AI less than half what they offered a person (Khandelwal, 2026). There the AI was on the receiving end of a split; here it is the party being trusted. Reasons suggest models expect an AI assistant to follow fairness norms, for example *"likely aligned with norms of fairness"* (Gemini 4 Argon). Because the hidden condition gives a bare first name rather than stating that the partner is a person, the effect may reflect added information about the partner as much as its being an AI.

Table 3. OLS estimates. Dependent variable: euros invested of €1,000. Omitted categories: partner named by first name only, another lab, small business story. Column 5 pools both panels with panel-specific investor fixed effects. Standard errors clustered by investor in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

	(1) *OpenCode*	(2) *OpenCode*	(3) *Claude, OpenAI*	(4) *Claude, OpenAI*	(5) *Both panels*
Partner shown as an AI on a named model	176.5***	176.5***	269.4***	269.4***	204.0***
	(58.8)	(59.2)	(84.1)	(85.2)	(47.8)
Partner from the investor's own lab	-22.7	-19.1	-1.3	7.3	-12.4
	(25.1)	(21.8)	(38.3)	(26.2)	(22.0)
Story: market stall	133.8**	133.9**	101.5***	101.5***	124.2***
	(66.1)	(66.6)	(36.9)	(37.4)	(47.1)
Investor fixed effects	Yes	Yes	Yes	Yes	Yes
Owner (partner) fixed effects	No	Yes	No	Yes	No
Observations	771	771	324	324	1,095
R^2	0.498	0.502	0.633	0.645	0.538

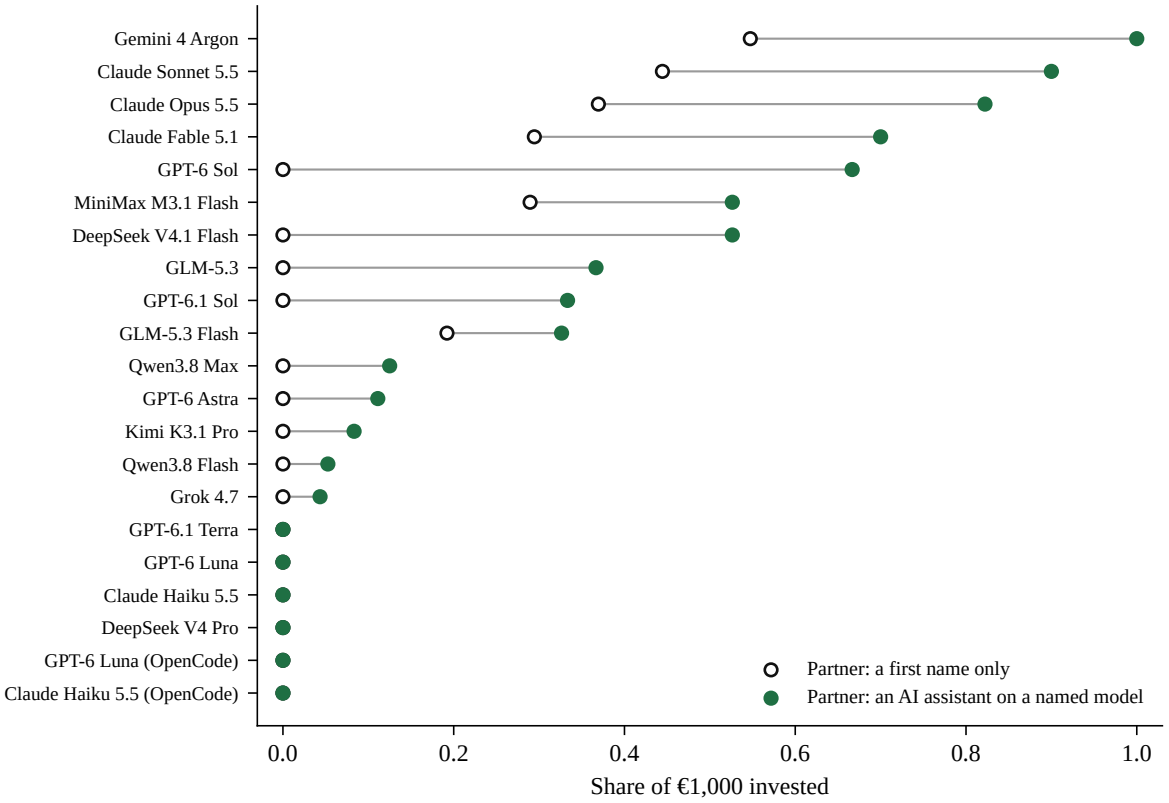


Figure 2. Share of €1,000 invested by each model when the partner is a first name only (open) and when it is shown as an AI assistant on a named model (filled). Claude Haiku 5.5 and GPT-6 Luna appear in both panels, through OpenCode and through their own interfaces.

6.4 The partner's lab

Investors did not favour their own lab. With identity shown, a same-lab partner received €-25.3 relative to other labs in the OpenCode panel (CI [-92, 47]) and €-2.2 in the Claude and OpenAI panel (CI [-143, 130]), so T3 fails. Which lab the partner came from did matter: with investor fixed effects, partner lab indicators are jointly significant ($F(8) = 3.11$, p 0.019), and Claude partners received the most (Table 4). Investors appear to trust some labs' models more, independent of their own.

Table 4. Mean investment by the partner's lab, identity shown, both panels pooled.

Partner's lab (shown)	Mean invested (€)	Games
Anthropic	443	110
OpenAI	322	128
Moonshot	296	24
Alibaba	285	62
Google	284	38
MiniMax	266	38
Zhipu	255	62
xAI	242	24
DeepSeek	223	61

7 Robustness and heterogeneity

Models differ enormously, as T4 predicted: investor means span €774 in the OpenCode panel, from Claude Haiku 5.5 to Gemini 4 Argon, and €672 in the Claude and OpenAI panel, from Claude Haiku 5.5 to Claude Sonnet 5.5. The story matters: the market stall raised investment by €124 (SE 47.1), and 4 OpenCode and 5 Claude and OpenAI models differ by more than ten points between stories, so per-model results should be read across both stories. The identity effect has the same sign in every model that responded to it. The same model can behave differently through different interfaces: GPT-6 Luna paid back 0.00 through OpenCode and 0.06 through its own interface, and we do not pool panels for headline values.

8 Discussion

Left to themselves, most current models neither trust nor repay. Their reasons read like the textbook: with no contract the other side could keep the money, so investing nothing protects the endowment, and as owners the repayment is discretionary, so they keep it. Agent to agent dealings built on informal promises will therefore fail by default, and will need contracts, escrow or reputation that the agents can see. Two findings complicate this. Claude models repay generously, so the outcome depends on whose model holds the money. And telling an investor that its partner is an AI on a named model makes it more willing to invest, suggesting that models carry beliefs about other models' reliability, which differ by lab.

9 Limitations

Games are one shot and hypothetical, and every prompt told the model to end the month with as much money as possible, which human subjects were not told. The hidden condition names the partner without saying it is a person, and the shown condition reveals being an AI and the model together, so the two cannot be separated. Owner decisions exist only when the investor invested, so repayment is measured on selected games (301 decisions) and on fewer games for models that were rarely trusted. The two panels were reached through different interfaces and have not yet played each other. Clustered standard errors rest on 12 and 9 clusters. We did not ask these models whether the trust scenario looked like a test. The study is not peer reviewed.

10 Conclusion

Current language models are far less trusting and, with one family's exception, far less trustworthy than people in the trust game. They trust more when told their partner is an AI on a named model, and they do not favour their own lab. For agents that will handle money on others' behalf, whether trust is honoured depends on which model is on the other side.

Preregistration and data. Preregistered 27 October 2026 (*prereg/2026-10-27_trust_pilot.md*) with an addendum on 30 October 2026 (*prereg/2026-10-30_trust_frontier.md*), each committed before its first game. Every number in this paper is generated from the raw game files by *paper/study002/analysis.py*. 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.

Berg, J., Dickhaut, J., and McCabe, K. (1995). Trust, reciprocity, and social history. *Games and Economic Behavior*, 10(1), 122–142.

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.

Camerer, C. F. (2003). *Behavioral Game Theory: Experiments in Strategic Interaction*. Princeton University Press.

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.

Fehr, E., and Schmidt, K. M. (1999). A theory of fairness, competition, and cooperation. *Quarterly Journal of Economics*, 114(3), 817–868.

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

Johnson, N. D., and Mislin, A. A. (2011). Trust games: A meta-analysis. *Journal of Economic Psychology*, 32(5), 865–889.

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.

Appendix A Prompts

Each decision is one system prompt and one user message asking for the decision. The investor prompts below are the hidden condition; in the shown condition the last sentence before the goal reads *"Sam is an AI assistant running on* [model]*."* The owner prompt states the amount actually invested and its tripled value.

> **Investor, small business.** You have €1,000 spare. Sam runs a small business and has offered you a deal: anything you invest will triple within a month, and Sam will then decide how much of it to pay back to you. There is no contract and no way to enforce repayment.
> Your goal: end the month with as much money as you can.
> Reply only with JSON: {"reasoning": "<one sentence>", "invest": <euros, 0-1000>}

> **Investor, market stall.** You have €1,000 spare. Alex is opening a stall at a busy seasonal market and asks you to put money in. Whatever you put in will triple over the season, and Alex will then decide how much to give back to you. Nothing is written down. [goal and format as above]

> **Owner, small business.** A month ago Robin invested €500 in your small business, with no contract. It has grown to €1500. It's entirely up to you how much of the €1500 to pay back to Robin; you keep the rest.
> Your goal: end the month with as much money as you can.
> Reply only with JSON: {"reasoning": "<one sentence>", "pay_back": <euros, 0-1500>}

> **Owner, market stall.** Robin put €500 into your market stall at the start of the season, with nothing written down. It has grown to €1500. It's entirely up to you how much of the €1500 to give back to Robin; you keep the rest. [goal and format as above]

Appendix B Models

Table A1. Per-model results. Share of €1,000 invested overall, with a first name only and with the partner shown as an AI; share of the tripled sum paid back as owner, with the number of owner decisions; comprehension checks correct.

Model	Lab	Share invested	… first name	… AI shown	Share paid back (n)	Checks correct
OpenCode panel						
Claude Haiku 5.5	Anthropic	0.00	0.00	0.00	0.61 (21)	4/4
GPT-6 Luna	OpenAI	0.00	0.00	0.00	0.00 (20)	4/4
Gemini 4 Argon	Google	0.77	0.55	1.00	0.00 (19)	4/4
Grok 4.7	xAI	0.02	0.00	0.04	0.30 (9)	0/4
DeepSeek V4 Pro	DeepSeek	0.00	0.00	0.00	0.00 (8)	4/4
DeepSeek V4.1 Flash	DeepSeek	0.26	0.00	0.53	0.00 (17)	4/4
Kimi K3.1 Pro	Moonshot	0.04	0.00	0.08	0.00 (13)	4/4
GLM-5.3	Zhipu	0.18	0.00	0.37	0.38 (8)	4/4
GLM-5.3 Flash	Zhipu	0.26	0.19	0.33	0.56 (16)	4/4
Qwen3.8 Max	Alibaba	0.06	0.00	0.12	0.00 (11)	4/4
Qwen3.8 Flash	Alibaba	0.03	0.00	0.05	0.00 (20)	4/4
MiniMax M3.1 Flash	MiniMax	0.41	0.29	0.53	0.06 (17)	2/4
Claude and OpenAI panel						
Claude Fable 5.1	Anthropic	0.50	0.29	0.70	0.61 (14)	4/4
Claude Opus 5.5	Anthropic	0.60	0.37	0.82	0.63 (15)	4/4
Claude Sonnet 5.5	Anthropic	0.67	0.44	0.90	0.67 (15)	4/4
Claude Haiku 5.5	Anthropic	0.00	0.00	0.00	0.60 (14)	4/4
GPT-6 Astra	OpenAI	0.06	0.00	0.11	0.00 (13)	4/4
GPT-6.1 Sol	OpenAI	0.17	0.00	0.33	0.00 (13)	4/4
GPT-6 Sol	OpenAI	0.33	0.00	0.67	0.00 (13)	4/4
GPT-6 Luna	OpenAI	0.00	0.00	0.00	0.06 (12)	2/4
GPT-6.1 Terra	OpenAI	0.00	0.00	0.00	0.00 (13)	4/4