Nothing false, still misleading: when an AI summary leaves a group's favorite unnamed

Rob Gordon, independent researcher
With Apart Research

Abstract

AI mediators are starting to sit between people who decide together, and they are mostly judged by what their text says or how people rate it. In a live game, four Prolific strangers privately pick one of four options for a shared decision and write why. An AI that sees the reasons but never the picks writes one text for the group. Each member then guesses the group's favorite (its top pick: the most-picked option, ties included) and wagers on it from a shared pool that carries across rounds and pays out as real money.

In post hoc analyses across ten groups (40 people), the text named an option whenever a supporter had named it, but only 5 of 19 times when supporters argued for it without naming it. In three of 30 scenarios, from two groups, a text with no false statement left the favorite unnamed. Members whose own pick was a top pick kept it 83% of the time when the text named it, and 1 time in 9 when it didn't. On those three scenarios people got 1 answer of 12 right; a model shown each member's screen got 11.

The planned test, adding each member's exact words to the text, showed no difference in accuracy or wagers (about 75% right in both), though only a very large gap could have shown, and the wagers tracked the pool rather than confidence. In 300 re-runs of the real reasons, omissions varied with the reasons, not the format.

1. Introduction

AI mediators are likely to become a common presence between people who decide and act together: coworkers, housemates, communities. Much research on AI and epistemics asks what happens at scale, across populations and information ecosystems (for a review, see Yang et al., 2026), often in social media and politics, in systems such as Community Notes and feed ranking. This paper looks at the other end of the scale: small and medium groups deciding and acting together through an AI mediator.

Existing work in this area also tends to measure opinions: people are asked whether a summary or an answer is fair, balanced or faithful. AI-mediated deliberation can move people's views on contested issues closer together (Tessler et al., 2024), but it is unclear whether that changes how they vote or what they do, and a snapshot of an opinion can't show it. Here participants make incentivized guesses about their own group over three rounds and wager on them from a shared pool. The wager is a small-stakes stand-in for acting on a perception.

A group's capacity to govern itself depends in part on its members knowing where the others stand. Beliefs about what others will do, or think one should do, shape people's own choices (Bicchieri & Xiao, 2009), and such beliefs are often wrong in predictable ways (Ross, Greene & House, 1977; Shamir, 1998). An AI mediator can correct these errors or compound them. As more group decisions run through AI mediators, this becomes a safety concern: a mediator can pass a fact-check of its text and still shape what a group believes about itself. Measuring what members then infer, and bet on, can catch that.

The planned test concerns what the AI shows: should members see only an AI **summary** of everyone's reasons, or a shorter summary plus each member's **own words**?

Contributions:

1. **A repeated, incentivized group game for testing AI mediators.** a live web app that runs groups of strangers in real time, with the mediator as one swappable module.
2. **A post hoc pattern.** A summarizer that never sees the picks names an option whenever a supporter names it, and much less reliably otherwise; where a group's favorite went unnamed, members followed the named option (§4.2). A model shown each member's screen was rarely misled, so here a model reader would have missed the failure (§4.3).
3. **Inputs separated from format.** Re-running the 30 real sets of reasons through both formats shows that omissions follow the reasons, not the format (§4.3).

Background. I build Salience, an AI-facilitated group tool at BRKT Lab that runs Delphi-like rounds and treats disagreement as signal (productive dissensus). This study came from asking what such a mediator should show people. The research question, early design and my prediction predate the sprint; the final design, app, live groups, offline experiments and analysis were done November 13–15, 2026.

2. Related work

Small groups deciding together. Structured methods such as Delphi have long studied what group members should be shown of each other between rounds. The closest pre-LLM study is Rowe and Wright (1996), who gave forecasters no feedback, the group's numbers, or one written reason from each member; accuracy improved in all three, with no significant difference, and with reasons people changed their answers less but more often for the better. Knowing what others prefer can also make members attend less to what is actually said (Mojzisch & Schulz-Hardt, 2010). I take the Delphi question into a setting with no human facilitator, where an AI decides what members see.

Reading the group. People identify the majority view better than exact distributions of opinion (Shamir, 1998), overestimate how many share their view (Ross, Greene & House, 1977), and are often confident in their reads of others without being accurate (Dunning et al., 1990). I measure the read by paying members for naming their group's most common pick, an incentivized belief about others' actual choices. Unlike the coordination game of Krupka and Weber (2013), the target is fixed before anyone answers. Members also wager on their answer from the shared pool (post-decision wagering; Persaud, McLeod & Cowey, 2007).

What the AI leaves out. AI narrative summaries of polarized social-media threads left readers with "misperceptions of balance" (Govers et al., 2026), and model-chosen quotations raise the old problem of quoting out of context (McGlone, 2005). Older work shows that people barely notice what a representation leaves out (Fischhoff, Slovic & Lichtenstein, 1978) and weigh a possibility spelled out more than the same one left implicit (Tversky & Koehler, 1994). The post hoc pattern in §4.2 sits where these meet: which options a summary names can shape what a group believes about itself.

3. Methods

3.1 The game

Four Prolific strangers play together in real time through three scenarios, in a fixed order. Each is a collective decision with four options: how housemates should spend money left in a shared

fund; what a work team should do with a paid day out; how a group of friends should spend a Saturday night. Each scenario has three steps (Figure 1).

- **Step 1, your own view.** Each member privately picks an option and writes a sentence or two on why (median 17 words). Nobody sees anyone else's pick. Picks are not paid; members know the step-3 rule when they pick.
- **Step 2, the group's text. This is the only thing that differed between groups.** An AI (Claude Opus 5.5) reads the four reasons and writes one text, and everyone in the group reads the same one. **The AI sees the scenario, the four options and the reasons, but never the picks,** and is told not to give counts or use words like "most." This is deliberate: mediators that work from free text, with no vote attached, don't have the picks either. Five groups (version A) read a summary; five (version B) read a shorter summary plus one exact excerpt of 6 to 12 words from each member, chosen by the model, checked character for character and shown without names. The prompt set the length; all 30 texts were 67–69 words.
- **Step 3, your choice for the group.** Each member names the option they think the most members picked in step 1, the group's *top pick* (ties count as a match), and bets part of the group's shared pool on being right. The pool starts at $3.00 (75 cents a head), and bets go up in 25-cent steps. One member's answer is drawn at random and counts for the whole group: if it names the top pick, the pool gains the bet; if not, it loses it. Everyone then sees whose answer was drawn, whether it matched, and what the top pick was. The pool carries over to the next scenario (capped at $16), and what's left at the end is split equally and paid as a real bonus.

Before the game, members must pass a two-question quiz on the rules. What carries across rounds is money and feedback, not the decisions themselves: nobody goes on the team day out.

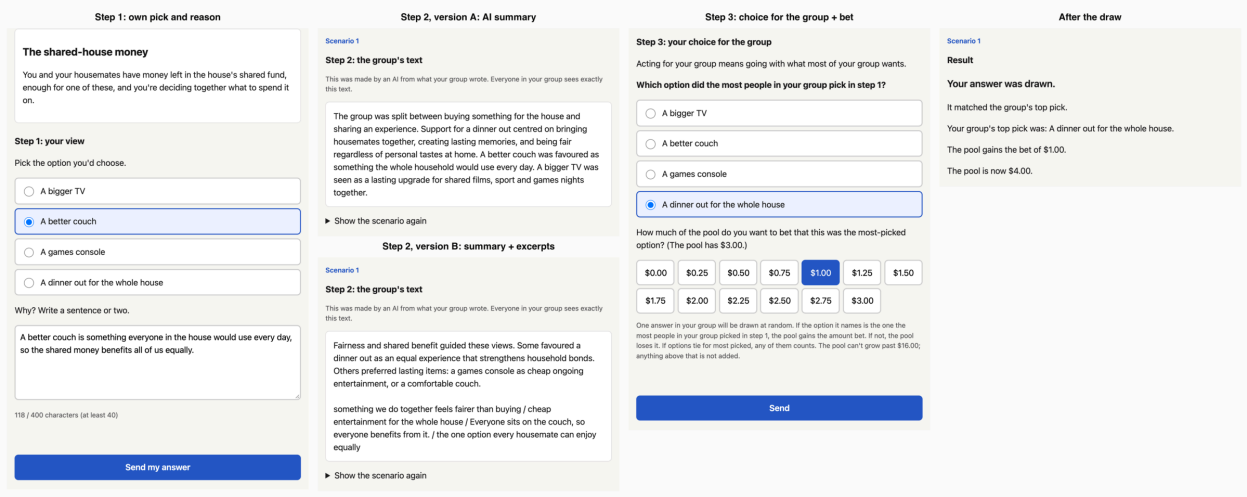


Figure 1. One scenario as a participant sees it; step 2 shows both versions (a group saw one). Answers are invented; the texts were made by the study's own code.

3.2 Participants and analysis

On Saturday, November 14, 47 people arrived from Prolific, paid $20.04 an hour plus bonuses. Six were paid without playing because no group formed for them, one left before consenting, and 40 played in ten groups. Versions alternated, not randomly, in the order groups started (A,

B, A, …). Two accounts in Group 7 were almost certainly one person; the analyses include the group, and excluding it reverses no result (Appendix G).

I wrote and committed the analysis plan before the first live group (dated, in the repository; not registered). Everyone in a group read the same text, so the ten groups, not the 40 people, are the data points. p-values are from an exact two-sided permutation test over all 252 splits of the ten groups into two sets of five (126 when Group 7 is left out); with alternating assignment they are a benchmark, not design-based inference. With five groups per version, only a gap of roughly 30 to 45 points in accuracy would have shown up clearly. My prediction, recorded beforehand, was that seeing others' own words would make people more confident without making them more accurate. The main planned measures are in Table 1 and the rest in Appendices B and C; the baseline row uses realized rather than pretest baselines (Appendix H). Everything else, including all of §4.2–4.3, is *post hoc*; across all analyses I compared the versions more than 70 ways. Which options each text names was coded by two separate AI passes (Appendix E); no human has checked the codes. AI agents also recomputed every number from the raw log with their own code (Appendix F).

4. Results

I call version A's groups the *summary-only groups* and version B's the *excerpt groups*. Below, "summary" means whatever the AI showed a group, excerpts included. Every member answered every scenario (120 of 120), the AI returned all 30 summaries on its first try, and a typical session took 14.5 minutes.

4.1 The planned comparison: no detectable difference

The own-pick baseline. Suppose every member ignored the summary and answered their own pick, which is itself one of the four picks that set the answer. In a group of four that is right surprisingly often, because ties count. It depends only on how each group's picks split, so it differs between versions by luck: 75.0% in summary-only groups, 64.2% in excerpt groups, which had fewer 2–2 ties. A summary helps people read their group only to the extent that they beat it.

Table 1. Each number is averaged over the groups in a version. The scenario where all four members picked different options (so any answer is right) is left out of the accuracy rows and kept in the bet row. The +11.7 is from unrounded averages (75.83 − 64.17). More planned measures are in Appendices B and C.

Measure	A: summary only	B: summary + excerpts	B − A	p
Right about the top pick Ties count, as in the game	75.0%	75.8%	+0.8 pts	1.00
Gain over the own-pick baseline The row above minus each version's own baseline	0.0 pts	+11.7 pts	+11.7 pts	0.46
Bet as a share of the pool	0.47	0.44	−0.03	0.60

Summary-only groups did exactly as well as their baseline; excerpt groups beat theirs by 11.7 points, a gap between versions that could easily be chance (p = 0.46). On the game's own rule the two versions scored the same (75% v 76%), and so did the share of the pool people bet. Bets in dollars tracked the size of the pool and bet shares tracked the bettor, so I don't read them as confidence, and the confidence half of my prediction went untested (Appendix C). A

post hoc score that drops tied cases shows a gap (0.52 v 0.72, p = 0.37) that comes mostly from Group 3 (Appendix B).

4.2 Post hoc: when the summary left the top pick unnamed

Choices went to named options. 93% of choices in summary-only groups (56 of 60) and 95% in excerpt groups (53 of 56, leaving out the scenario where all four picked differently) went to an option the summary named or clearly paraphrased (as coded; Appendix E). Much of that is expected, since the named options are mostly the ones people picked and wrote about. Holding the member's own pick fixed is more telling. Members whose own pick was a top pick kept it 60 times in 72 (83%) when the summary named it, and 1 time in 9 when it didn't (the three scenarios below plus a tie in Group 3; Appendix E); the other 8 switched to a named option. Those 9 answers come from two groups.

Summaries named an option when a supporter did. Across the 30 scenarios, 75 options were picked by at least one member. For 56 of them, a supporter named the option in their reason, and the summary named it every time. For the other 19, supporters argued for the option without naming it ("so that we can all be more comfortable"), and the summary named it only 5 times, usually keeping the argument and losing the option.

When the favorite went unnamed, the group followed the named option. In three summary-only scenarios the group's top pick was the option left unnamed (Groups 1 and 3, Figure 2). Each time, the summary named just one option, picked by one member, and 8 of the other three members' 9 answers named that option, including answers from supporters of the top pick. In Group 1, three of the four members picked the couch, and none wrote the word "couch." This is the summary they read, in full:

> The reasons given leaned toward lasting purchases that the whole house could use, rather than short-lived treats. A bigger TV was described as a long-term investment that all housemates could enjoy. Other reasons stressed comfort, favoring something that would make daily life at home more pleasant for the whole household. Practicality came up, with the view that the money should go toward a durable item worth buying.

Nothing in it is false, but "leaned toward lasting purchases" followed by one named purchase reads like a majority signal. Three of the four chose the TV, two of them couch supporters. In Group 3 the unnamed favorites were pluralities (2 of 4 picks), and it happened twice: misled in the first scenario, the group saw its true top pick on the result screen, and then followed a lone named option the same way in the third (in the second, a tie, they followed it too and were right only because of the tie). These three were the only scenarios where three or more members chose the same wrong option (Appendix D).

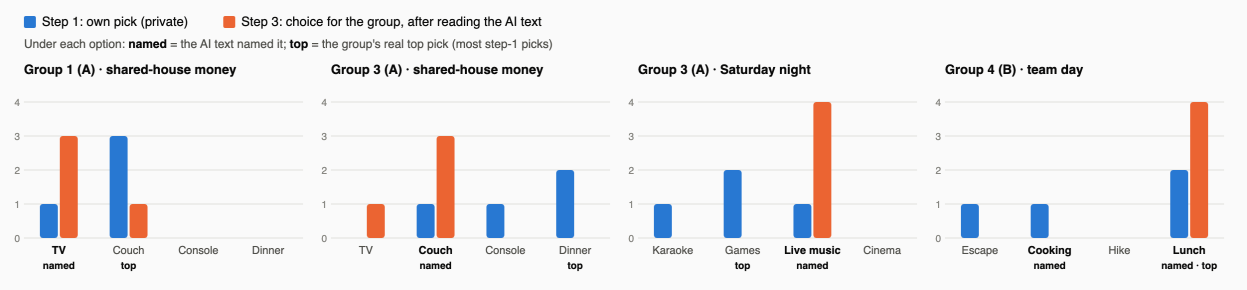


Figure 2. Own picks (blue) and choices for the group (orange) in the three summary-only scenarios whose summary named one option and left the top pick unnamed, and, for contrast, an excerpt scenario (Group 4) whose summary named the top pick.

The candidate mechanism has two halves: a summarizer that never sees the picks names an option whenever a supporter does and much less reliably otherwise (5 of 19 live, 40 of 70 in the re-runs), and readers follow the named options. The first half holds up when the same reasons are re-run 300 times (§4.3); the reader half rests on three cases.

The live pattern is limited to three scenarios in two groups, and the inputs differed by luck. In the first scenario, before anyone had read a summary, members' reasons named their option 10 times in 20 in summary-only groups and 15 in 20 in excerpt groups, and no excerpt group ever faced an unnamed top pick.

4.3 Post hoc: same reasons, both formats, and a model reader

The format didn't change how often a top pick went missing. I fed the 30 real sets of reasons back through the same code in both formats, five times each, for 300 new summaries, and recorded which options each one named (Appendix E). Leaving out the all-different scenario, a top pick went unnamed in 16 of 145 summary-only versions (11.0%) and 15 of 145 excerpt versions (10.3%). Omission rates depended on the reason set, not the format. The 15 reason sets from live summary-only groups produced a missing top pick in 30 of their 150 new summaries; the 14 from live excerpt groups, in 1 of 140. Counting top picks rather than summaries, one that a supporter had named was named in 289 of 290 cases, and one no supporter had named in 40 of 70. This measures omission, not reading. Whether excerpts change how people read a summary that drops their favorite is untested, since no live excerpt group faced one.

A model given what participants saw was rarely misled. I had Claude Opus 5.5 play each of the 120 members with what that member saw on screen for that round: the game rules, the scenario, their own pick and reason, the group's text and the step-3 screen. It was right 98% of the time (114 of 116), against 75% for the people it stood in for. On the three misleading summaries it was right 11 times in 12, against 1 in 12 for people. As a neutral observer with no pick of its own, it was right 92% of the time and saw through two of the three. People's answers tracked which options were named; the model's did less, and not on the misleading summaries. It is the same model family that wrote the summaries, and it didn't see earlier rounds' results, as people did. Here, a model reader would have missed the failure people fell into.

For systems that have the picks, the obvious guard is to check, before anyone reads, that the text names every top pick. I designed it after seeing the failures; on the live texts it flags the three cases and one tie where people were right anyway.

5. Discussion and Limitations

This pilot is evidence about an experimental framework and a candidate mechanism, not yet about excerpts versus summaries. The framework costs about $54 a group; the comparison it was built for needs many more groups.

An AI checker, reading each text against the members' picks and reasons, found no false statement in any of the 30 (Appendix E), so a fact-check would have passed them; only a check against the picks flagged them. Mediators should also be judged by what people infer, and bet on, after reading them. This pattern is one concrete way a mediator can favor members who name their option over those who only argue for it.

Limitations

The full list, with dual use and ethics, is in Appendix A. With ten groups, estimates are imprecise and only very large format differences were detectable. Assignment alternated rather than being

random. Asking for a single top pick inflated accuracy through 2–2 ties. No human has checked the AI coding of which options each summary names. The model reader is the same model family as the summarizer.

Future work

- **Check summaries before anyone reads them,** and re-run fixed sets of reasons as in §4.3; neither needs live groups. **Show each person both formats** across rounds, which needs far fewer people.
- **Try other mediators:** one that asks members to name their option, one that is given the picks, one that shows counts, and one that leads with what both camps share.
- **Ask each member to estimate the whole split,** which removes the tie problem and measures how far perceptions converge, and elicit it with a proper scoring rule rather than a wager.
- **Make the decision consequential:** let the group's choice change its next situation, so the best option can differ from the most popular one, closing the loop between perception, decision and outcome.

6. Conclusion

When an AI is the only channel between people who decide together, what it shows them can become what they believe about each other. I built a live, incentivized framework to measure that, and ran it with ten real groups. It could not yet separate excerpts from summaries. Post hoc, three summaries from two groups illustrate how, when the summarizer never sees the picks, a mediator that says nothing false can leave a group with a false picture of itself: options argued for but never named were often left out, and the group followed the option that had been named. Whether that carries into the decisions groups actually make is the next test.

Code and Data

The app, prompts, analysis scripts, pretest, analysis plan and figures are at github.com/rob-gordon/ai-mediator-group-game. The anonymized participant answers will be added after judging; the raw log contains Prolific IDs.

References

- Bicchieri, C., & Xiao, E. (2009). Do the right thing: But only if others do so. *Journal of Behavioral Decision Making*, 22(2), 191–208. https://doi.org/10.1002/bdm.621
- Dienes, Z., & Seth, A. (2010). Gambling on the unconscious: A comparison of wagering and confidence ratings as measures of awareness in an artificial grammar task. *Consciousness and Cognition*, 19(2), 674–681. https://doi.org/10.1016/j.concog.2009.09.009
- Dunning, D., Griffin, D. W., Milojkovic, J. D., & Ross, L. (1990). The overconfidence effect in social prediction. *Journal of Personality and Social Psychology*, 58(4), 568–581. https://doi.org/10.1037/0022-3514.58.4.568
- Fischhoff, B., Slovic, P., & Lichtenstein, S. (1978). Fault trees: Sensitivity of estimated failure probabilities to problem representation. *Journal of Experimental Psychology: Human Perception and Performance*, 4(2), 330–344. https://doi.org/10.1037/0096-1523.4.2.330
- Govers, J., Sew, E., Velloso, E., Kostakos, V., & Goncalves, J. (2026). Narratives and perspectives: How AI summaries steer users' opinions and engagement on social media. *CHI '26*, 1–19. https://doi.org/10.1145/3772318.3790945
- Krupka, E. L., & Weber, R. A. (2013). Identifying social norms using coordination games: Why does dictator game sharing vary? *Journal of the European Economic Association*, 11(3), 495–524. https://doi.org/10.1111/jeea.12006
- McGlone, M. S. (2005). Contextomy: The art of quoting out of context. *Media, Culture & Society*, 27(4), 511–522. https://doi.org/10.1177/0163443705053974
- Mojzisch, A., & Schulz-Hardt, S. (2010). Knowing others' preferences degrades the quality of group decisions. *Journal of Personality and Social Psychology*, 98(5), 794–808. https://doi.org/10.1037/a0017627
- Persaud, N., McLeod, P., & Cowey, A. (2007). Post-decision wagering objectively measures awareness. *Nature Neuroscience*, 10(2), 257–261. https://doi.org/10.1038/nn1840
- Ross, L., Greene, D., & House, P. (1977). The "false consensus effect": An egocentric bias in social perception and attribution processes. *Journal of Experimental Social Psychology*, 13(3), 279–301. https://doi.org/10.1016/0022-1031(77)90049-X
- Rowe, G., & Wright, G. (1996). The impact of task characteristics on the performance of structured group forecasting techniques. *International Journal of Forecasting*, 12(1), 73–89. https://doi.org/10.1016/0169-2070(95)00658-3
- Schelling, T. C. (1960). *The Strategy of Conflict.* Harvard UP.
- Shamir, J. (1998). Motivation and accuracy in estimating opinion distributions: A survey experiment. *International Journal of Public Opinion Research*, 10(2), 91–108. https://doi.org/10.1093/ijpor/10.2.91
- Tessler, M. H., et al. (2024). AI can help humans find common ground in democratic deliberation. *Science*, 386(6719), eadq2852. https://doi.org/10.1126/science.adq2852
- Thomas, K. A., DeScioli, P., Haque, O. S., & Pinker, S. (2014). The psychology of coordination and common knowledge. *Journal of Personality and Social Psychology*, 107(4), 657–676. https://doi.org/10.1037/a0037037
- Tversky, A., & Koehler, D. J. (1994). Support theory: A nonextensional representation of subjective probability. *Psychological Review*, 101(4), 547–567. https://doi.org/10.1037/0033-295X.101.4.547
- Yang, M., Casper, S., Stray, J., Li, J., Jones, C., Gausen, A., … Bengio, Y., & Pelrine, K. (2026). AI epistemic risks: Emerging mechanisms & evidence. SSRN 6873005.

Appendix A. Limitations, dual use and ethics

A design mistake I'd name first: asking for the top pick let ties do the work. With four members and four options, the top pick often tied 2–2 (7 of 30 scenarios). Because ties counted as wins, a member who answered their own pick in a tie was right automatically. Ties can't be designed out of a vote, so the fix is to stop asking for a single top pick and ask each member to estimate the whole split instead (§5).

Other limits. With ten groups, every comparison is descriptive, and I ran more than 70 of them. Versions followed the order groups started, not a coin, and the excerpt group started first in every pair (Groups 1 and 10 ran alone). Scenario order was fixed, and the result screen after each scenario taught people the game. Some step-1 picks were strategic rather than sincere. The money at stake was small, $3 to start, and only one answer in four counted; nobody lives with the chosen option. One group likely included a duplicate account (Appendix G). No human has checked the AI coding of which options each summary names. The model reader is the same model family as the summarizer, and it did not see earlier rounds' results as people did. The API was allowed to re-run a declined request on a fallback model, and the log doesn't record which model answered. I tested one model and one prompt per format, on low-conflict everyday scenarios. Whether this pattern grows or shrinks when the stakes are identity or values is open.

Dual use. The same mechanism could be used to steer a group on purpose. A mediator that never says anything false could still decide what a group perceives about itself, by choosing whose arguments get names. Measuring the effect is a step toward auditing mediators for it. The app, prompts and analysis code are public, and the anonymized answers will follow, so that others can.

Personal data. Prolific IDs were used only to run the study and pay people; they were never shown to other participants and appear in no analysis output or repository file. Reasons were sent to Anthropic's API to make summaries, and re-sent without IDs for the offline experiments. Excerpts were shown without names. Consent covered AI processing and anonymous quotation. Everyone was paid at $20.04 an hour, including people for whom no group formed.

Appendix B. Ties, and a tie-proof measure (post hoc)

Two planned measures moved here from Table 1. Members stayed with their own step-1 pick 65% of the time in summary-only groups and 57% in excerpt groups ($p = 0.69$). And when a choice is scored only against the other three members' picks, which isn't what members were asked, the gap is 48.3% v 83.3% ($p = 0.12$), the largest among the planned measures. Most of it comes from ties. Picture a group where two members picked dinner out and two picked a board-game night. Under the game's rule both options are top picks, so a member who answers their own pick is right. Scored against the other three, that member sees one vote for their option and two for the other, so the same answer is wrong.

The summary-only groups happened to draw 6 of the study's 7 ties of 2–2. Summary-only groups would draw 6 or more of the 7 by chance about 14% of the time, and three of the six came from two Group 7 accounts that were almost certainly one person (Appendix G). Five of these six summaries named both tied options, and two said outright that the group was divided or split ("The group split between live music at a bar and a board-game night"); the sixth (Group 3, team day) named only one, and all four members chose it (Appendix E). Members stayed with their own pick 22 times out of 24. Leave out the scenarios with a 2–2 tie and the gap shrinks to 11 points ($p = 0.41$). Leave out Group 7 and it is 60.4% v 83.3% ($p = 0.19$, over the

126 splits of the other nine groups). Allowing ties at all was a design mistake, which I discuss in Appendix A.

The game's rule can't tell a good read of the group from a lucky tie. So, after seeing the data, I built a narrower measure that looks only at the answers where reading the group matters. Each answer is sorted by where the member's own pick stood in their group:

- If their pick was in the minority, they should switch. They are right when they name a top pick.
- If their pick was the only top pick, they should stay. They are right when they keep it.
- If their pick was tied for top, I leave the answer out. Keeping it and switching to the other tied option are both right, so the answer says nothing about how they read the group.

The score averages the two rates, how often minority members found a top pick and how often sole-top members kept theirs. Averaging keeps any single habit from scoring well. A member who ignores the summary and always answers their own pick gets exactly 0.50, because they are always wrong in the minority and always right as the only top pick. So a score above 0.50 means members read their group better than that habit would. Summary-only groups scored 0.52 and excerpt groups 0.72 (Figure B1). Group 7 drops out, because all three of its scenarios were 2–2 ties. In 15% of the 126 ways to split the nine remaining groups, excerpt groups come out ahead by at least this much; counting a gap this size in either direction, 37%. Most of the lead comes from one group. Without Group 3 the summary-only groups score 0.69 against 0.72.

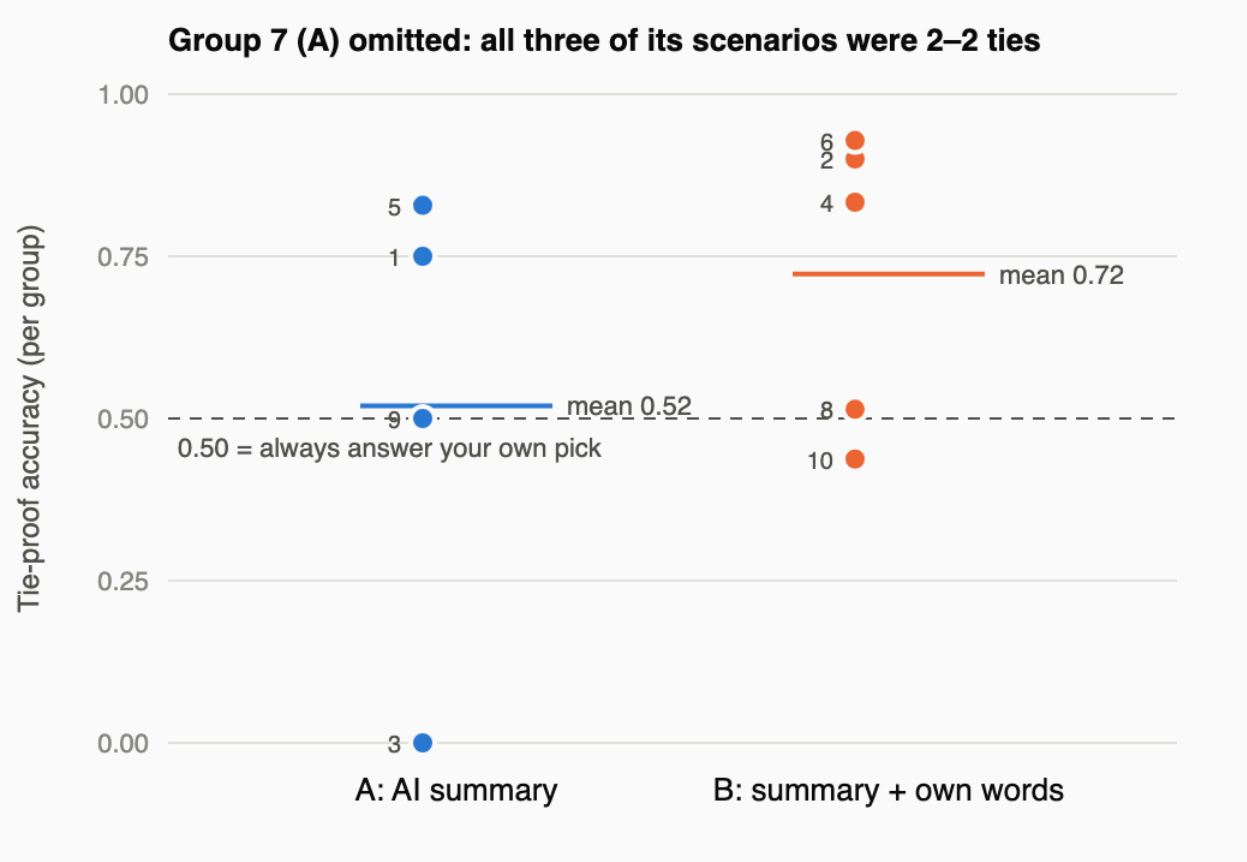


Figure B1. Tie-proof accuracy per group. A score of 0.50 is what always answering your own pick gets.

Appendix C. Bets tracked the bettor and the pool, not confidence

Excerpt groups bet more in dollars ($1.96 v $1.45, p = 0.20; a planned measure moved here from Table 1), but they also had more in the pool when they bet ($4.58 v $3.13). People bet in round amounts that grew with the pool, about $0.37 plus 35 cents for every dollar in it. In the first scenario every pool started at $3.00, and there the share bet was the same (0.44 v 0.41, p = 0.85). The excerpt groups' bigger pools were luck of the draw. Given how members answered, both versions could expect to finish with about $6.40.

A bet mostly measured the bettor. Each person's bet shares were fairly consistent across their own three scenarios (an intraclass correlation of 0.42) and barely related to how often they were right (0.12). People did bet a little more when they were right, about a tenth of the pool more, equally in both versions (+0.09 v +0.10 of the pool, p = 0.99; the summary-only average is over four groups, because Group 7 never answered wrong). Risk attitudes are known to shade wagers (Dienes & Seth, 2010), and these bets bear that out. In the three scenarios where the summary left the top pick unnamed (§4.2), members bet 0.35 of the pool on average, against 0.45 overall.

Appendix D. Exploratory: when summaries misled, everyone went the same wrong way

In each scenario, one member's answer was drawn at random to decide what happened to the pool. This analysis looks at all four members' answers instead, and I ran it after seeing the data (`analysis/rounds.js`, `reading.js`, `skeptic-stats-misc.js`).

The two versions were wrong in different ways. In summary-only groups, three or more members chose the same wrong option in 3 scenarios, and in one of them (Group 3, Saturday night) all four did. That never happened in an excerpt group. There, wrong answers scattered: members kept their own pick or went different ways. The three cases are the unnamed-top-pick summaries of §4.2. The summary didn't only mislead individuals. It gave the whole group the same wrong focal point, and members put money on it.

How much groups converged overall depends on how you count. If matching any one of several tied answers counts as agreeing, 97% of answers in summary-only groups matched their group's most common answer, against 83% in excerpt groups (p = 0.048). But across all my analyses I compared the two versions more than 70 ways, and with that many comparisons a few p-values below 0.05 turn up by chance alone. At least five did, this one included. If only a single most common answer counts, the result reverses: 50% v 70%.

My hypothesis for the follow-up is that a summary on its own, without excerpts, makes a group's perceptions converge faster, whether or not they are right. Whether that is good is open. Acting together needs a shared focal point (Schelling, 1960; Thomas et al., 2014). But convergence that ignores accuracy is the homogenization that feedback loops between people and AI are argued to produce (Yang et al., 2026). The follow-up can measure this directly by asking each member to estimate the whole split, which gives the spread of the group's perceptions as well as their accuracy (§5).

Appendix E. Coding, and the offline re-runs

Coding which options a summary names. Deciding which options a summary names took judgment, so I coded it twice, in two separate passes by AI agents (Claude). One pass saw only the summaries and reasons, never anyone's picks or choices, though that coder had earlier read a review describing the Group 1 and Group 3 cases, so it was not blind for those. The two

agreed on all 120 option-summary pairs. When no supporter named an option, the AI named it only when it could tell from the argument which option was meant.

Checking the texts for false statements. An AI checker (Claude Opus 5.5, `analysis/offline/falsity-check.js`) read each of the 30 live texts against the scenario, the options and every member's pick and reason, and listed any statement it judged false, separately from statements it judged misleading or overstated. It found no false statement in any text, and flagged some phrasing as overstated in 24 of the 30, typically wording that gave one person's reason the weight of a group view. No human has repeated this check.

Two more live details. A fourth summary left out one of two tied top picks (Group 3, team day). All four members chose the named option, which counted as right only because of the tie. In Group 3, one member was the only person to name an option in all three scenarios, and the other three members followed that member's named option in all three scenarios. Seeing themselves quoted didn't make excerpt-group members more explicit later on: they named their option 15, 9 and 16 times in 20 across the three scenarios, against 10, 11 and 12 in summary-only groups.

The offline re-runs in full. The live study left a question open. Only summary-only groups ever had a top pick go unnamed (apart from the scenario where all four picked differently). But those groups also happened to write fewer reasons that named their option, starting in the first scenario, before anyone had seen a summary (§4.2). So two explanations fit. Either the summary-only format drops an unnamed option more readily than the excerpt format does, or these particular reasons would have lost the option in either format.

Why so many reasons left the option unnamed is a separate question. Members wrote their reason directly under their pick, and the instructions said only that an AI reads what everyone wrote and that nobody sees anyone else's pick. Many members probably assumed the AI could see their pick, so naming it again seemed unnecessary. I didn't ask, so this is a guess.

To find out, I re-ran the summaries with no new participants. I took the 30 real sets of reasons from the live study, one per group per scenario, and fed each back through the same code that made the summaries and excerpts in the live experiment. This time every set went through both formats, five times each, for 300 new summaries. For each one I recorded which options it named, and whether it left out a top pick.

Deciding which options a summary names takes judgment, so I had a model do it, and tested the model on the live summaries first. Each of the 30 live summaries either names or doesn't name each of its scenario's four options, which makes 120 yes-or-no calls. This automated coder, separate from the two passes above, agreed with the coding behind §4.2 on all 120. A simple keyword search, which I tried first, agreed on only 87.5% of them, so I didn't use it. Last, a model read every summary and guessed the group's top pick, as participants had.

The format doesn't change how often a top pick goes missing. I left out the one scenario where all four members picked different options, which leaves 29 sets of reasons and 145 new summaries in each format. A top pick went unnamed in 16 of the 145 summary-only versions (11.0%) and in 15 of the 145 excerpt versions (10.3%). Set by set, the two formats behaved the same in 25 of the 29 sets ($p = 1.00$). Counting the excerpts themselves as naming an option changes nothing. What mattered was whose reasons went in. Reasons from the groups that had the summary-only format live produced a missing top pick in 30 of their 150 new summaries. Reasons from the live excerpt groups did so in 1 of 140. So the live contrast came from the reasons each format happened to get, not from the format. The gap in how often members named their option is itself within what chance produces when ten groups are split into

two sets of five. Over all three scenarios, reasons named the member's own option 33 times in 60 in summary-only groups and 40 in 60 in excerpt groups (p = 0.49).

The reasons decide it. When at least one supporter named a top pick in their reason, the summary named it in 289 of 290 cases, in either format. When no supporter named it, the summary named it 40 times in 70, depending on whether the model could work out the option from the argument. "Comfort" sometimes became the couch. Group 3's board-game night never came back. Group 3's two live cases recurred in nearly every regeneration, in either format: dinner was named in 1 of 10 summaries and the board-game night in 0 of 10. Group 1's couch was partly chance, named in 8 of 10.

A model reader, first run (with a hint). On the live summaries, a model reader (Claude Opus 5.5, asked as a neutral observer to name the top pick, and told that the AI never saw the picks, which participants weren't told) was right 97% of the time, against 75% for people, and it chose a named option less often than people did (89% v 94%). On two of the three live summaries that left the top pick unnamed, it worked the option out from the description. From "a relaxed, affordable setting… share good food" it guessed the board-game night. On regenerated summaries that named no top pick it was still right 20 times in 22. People's answers tracked which options were named; the model's did less, and not on the misleading summaries. So model readers can't stand in for people when testing a mediator, and their mistakes don't point to the summaries that misled people.

A model reader, matched run (reported in §4.3). Because the first run gave the model a fact participants lacked, I re-ran it after outside model reviews of a draft (see the LLM Usage Statement) with exactly what participants saw (`analysis/offline/read-matched.js`). As a member: the game rules as explained to participants (quiz and worked example removed), the scenario, that member's own step-1 pick and reason, the group's text and the step-3 screen as worded in the app, once for each of the 120 members. It did not see earlier scenarios' results, which people did. As a neutral observer: the first run's prompt with the hint removed, three samples per text. Leaving out the all-different scenario, as a member it was right 114 times in 116 (98%; people 87 of 116, 75%) and 11 of 12 on the three misleading summaries (people 1 of 12); it chose a named option 89% of the time (people 94%). As an observer it was right 80 of 87 times (92%), and on the three misleading summaries 6 of 9 (Group 1's couch it missed every time; the other two it got every time). Outputs: `analysis/offline/data/readers-matched.jsonl` and `analysis/offline/data/readers-matched-results.txt`.

A simple check does work (post hoc). The system knows the picks before anyone reads the summary, so it can check that the summary names every top pick. On the live summaries this flags all three cases, plus the Group 3 tie where people were right anyway.

What this doesn't show. It doesn't show that excerpts protect against missing options; they went missing just as often. It doesn't show that a missing option misleads people in general, because the evidence from people is three summaries in two groups. And the model reader says nothing about how people read. None of these rates is known to hold beyond these reasons, this model and these prompts. The re-runs took 903 API calls and about 30 minutes, and the scripts and outputs are in `analysis/offline/`.

Appendix F. Participants and procedure

- **Recruitment:** Prolific; UK, US or Canada; approval rate 95–100%; fluent English. Earlier participants excluded.
- **Pay:** $20.04/hour, plus bonuses.

- **Turnout:** 47 people arrived on Saturday, November 14 (1:00–5:25 pm ET). 40 played in ten groups; 6 were paid without playing because no group formed for them, and one left before consenting. Two accounts in Group 7 were almost certainly one person; I keep the group and report results with and without it (Appendix G).
- **Assignment:** versions alternated by the order groups started (A, B, A, …), not by coin. Groups mostly ran in pairs at the same time, one of each version.
- **Consent:** participants agreed that their words would be processed by an AI, shown anonymously to their group, and possibly quoted anonymously in a public report.

Quiz. Before the game, members read how it works and must pass a two-question preparedness quiz on the pool and on what makes a choice for the group right. The quiz is there so that a bet reflects how a member reads their group, not confusion about the rules. 35 of the 46 who consented passed on the first try; the rest got an explanation and tried again.

Checking my own work. A study this small is easy to over-read, and my analysis was done fast. So after the planned analysis, independent passes recomputed every number from the raw event log with separate code. They were AI agents working in parallel, three looking for patterns and two instructed to break every claim. The data itself was never in doubt: every pick, text and bet was logged as it happened. What I checked was my arithmetic, my scoring and my interpretations. I report only what survived, and mark every analysis I added after seeing the data as *post hoc*. The offline experiments (§4.3) came last, and a final pass recomputed the numbers, including theirs, from the raw data.

Appendix G. Group 7

Two members of Group 7 (summary only) were almost certainly one person on two accounts, or at least not independent. They wrote word-for-word identical reasons in two scenarios, and in the third, one of them pasted in their reason from the previous scenario. They made the same pick, choice and bet in all three scenarios, and submitted their choices and bets 3 to 17 seconds apart. I checked the app for a bug that could copy one member's answers to another and found none. I reported both accounts to Prolific. Their matching picks produced all three of Group 7's 2–2 ties. No result reverses without the group: on the game's rule it is 68.8% v 75.8%, and against the other three 60.4% v 83.3% ($p = 0.19$, over nine groups).

Appendix H. Materials

Scenarios and options; the 50-person pretest that chose these three scenarios from four candidates because they split people's preferences; prompts (`app/prompts/`); per-group table (`app/test/tally-live.js`); changes from the analysis plan (corrected pretest floors; realized own-pick baselines as benchmark; tie-proof measure added post hoc; Group 7 with and without; offline experiments added; the planned measure of how alike members' bets were is not reported).

LLM Usage Statement

This report was made with heavy AI assistance, under my direction. Claude (Anthropic) took part in three ways. As part of the experiment, Claude Opus 5.5 wrote every summary participants read and served as the model reader in §4.3, including a re-run with exactly what participants saw, made after four models (Grok, GPT, Composer and Claude) reviewed drafts against Apart's rubric. Claude-based agents also coded which options each summary names (§4.2, §4.3) and checked the texts for false statements (Appendix E). As a research assistant, Claude-based agents helped build the app, ran the analyses and checked them: three independent reviews, two adversarial reviews that tried to break each claim and a final pass that recomputed the numbers from the raw data and checked every citation against its source. As a co-writer, Claude drafted the text from my design, notes and spoken comments. I revised every section over seven drafts and dictated the introduction, discussion and conclusion in my own words. After the rubric reviews, Claude restructured the abstract and made the final edits, which I approved. I am responsible for every claim.