Collective problem decomposition drives the wisdom of deliberative crowds

Federico Barrera-Lermarchand[1,2,3], Victoria Lescano-Charreau[2,4], Julieta Ruiz[1], Nuria Cáceres[1], Facundo Carrillo[3,5], Mariano Sigman[1], & Joaquin Navajas[1,2]

[1] Laboratorio de Neurociencia, Escuela de Negocios, Universidad Torcuato Di Tella, Buenos Aires, Argentina

[2] Comisión Nacional de Investigaciones Científicas y Técnicas (CONICET), Buenos Aires, Argentina

[3] Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires, Argentina

[4] Laboratorio Interdisciplinario del Tiempo y la Experiencia (LITERA), Universidad de San Andrés, Buenos Aires, Argentina

[5] Instituto de Investigación en Ciencias de la Computación (ICC), CONICET; Universidad de Buenos Aires; Argentina.

Corresponding author: Joaquin Navajas (joaquin.navajas@utdt.edu)

Abstract

Understanding when and why social interaction improves human judgment is a central question in the behavioural sciences. We examine whether accuracy increases when groups break problems into parts and generate approximate solutions, a process we call Collective Fermi Estimation. We tested this idea across three studies by analysing over 1,000 online group deliberations in text chatrooms. Study 1 (N=500) shows that greater use of problem decomposition in group discussions predicts lower error. Study 2 (N=240) provides causal evidence: instructing groups to apply problem decomposition leads to higher accuracy than combining initial estimates. Study 3 (N=160) shows the advantage arises when applied collectively rather than by individuals working alone. We also introduce a scalable natural language method to detect problem decomposition in deliberation text and predict collective accuracy. These findings identify problem decomposition as a key mechanism behind the wisdom of deliberative crowds and provide tools to detect and promote it.

1. Introduction

The aggregation of many lay estimates often outperforms individual expert judgments (De Condorcet, 1785; Galton, 1907). This phenomenon, popularly known as the "wisdom of crowds" (Surowiecki, 2005), has been applied to solve diverse problems, from predicting financial markets (Ray, 2006) and forecasting geopolitical events (Mellers et al., 2014a) to improving medical diagnoses (Kurvers et al., 2016) and even decoding the smell of molecules (Keller et al., 2017). Given the ubiquity of this effect and its vast applications to fields as varied as policymaking (Epp, 2017), medicine (Krockow et al., 2020), and finance (Lee et al., 2022), understanding how and when collective judgements succeed or fail, and how to improve them, has become central questions in the social and behavioral sciences (Prelec et al., 2017; Navajas et al., 2018; Kameda et al., 2022).

Previous research has investigated the impact of social influence on the wisdom of crowds, yielding mixed results. While some studies suggest that it can lead to herding and reduce accuracy (Raafat et al., 2009; Lorenz et al., 2011; Madirolas & de Polavieja, 2015), others have shown that social interaction can improve collective estimates (Gürçay et al., 2015; Mellers et al., 2014b; Bahrami et al., 2010; Juni & Eckstein, 2015). For example, it has been found that averaging the consensus estimates reached by small groups can significantly outperform the wisdom of large, independent crowds (Navajas et al., 2018). However, even though this result has been replicated across various contexts (Dezecache et al., 2022; Espina Mairal et al., 2024; Pescetelli et al., 2021), the underlying procedural mechanisms that determine when group consensuses improve or undermine collective accuracy remain poorly understood.

Based on theoretical arguments (Landemore & Page, 2015), one could, in principle, hypothesize that groups can follow two fundamentally different approaches to reach a

collective estimate. To illustrate the contrast between them, consider a group of four individuals estimating the height of the Eiffel Tower—first independently, then through group consensus. Suppose their initial estimates are 100, 200, 500, and 800 meters. The average, 400 meters, overestimates the correct value by 76 meters.

One approach to consensus involves combining the initial private signals. This suggests that group decisions result from aggregating the individual values and confidences, for example, by taking the mean, the median, a confidence-weighted mean, or a robust average, among other procedures. For example, in the case of the Eiffel Tower, the group might discard the most extreme values and average the two middle estimates, arriving at 350 meters—just 26 meters off the correct height. While different aggregation rules could be applied, this strategy assumes that the final estimate is derived directly from the original individual judgments. Therefore, if groups follow this strategy, they must interchange their initial estimates during the deliberation procedure.

A completely different approach involves using the deliberation process to generate a new estimate through reasoning. In this case, the group might discuss structural features of the Eiffel Tower, reasoning that it resembles a 100-story building, with each floor being approximately 3.5 meters tall. This may lead them again to an estimate of 350 meters. This method relies on breaking the problem into simpler components, making rough approximations, and performing basic calculations. Unlike the previous approach, it does not require sharing initial estimates, though individuals may still do so in practice. Because this strategy extends a well-established problem-solving technique for back-of-the-envelope calculations (attributed to Enrico Fermi) to a group setting (Anderson & Sherman, 2010.; Weinstein & Adam, 2009; Nityananda, 2014), we refer to it as the "Collective Fermi Estimation" strategy.

The original study showing that consensus estimates outperform the wisdom of crowds suggests that groups must be doing something more than simply averaging their initial estimates (Navajas et al., 2018). While that work also ruled out the idea that all groups uniformly followed other aggregation procedures, it remained unclear whether groups employed a mix of rules or an entirely different procedure, not evaluated in prior work.

This raises the key research question of this work: What strategies do groups actually follow to reach consensus, and how do these strategies impact collective accuracy? To investigate this, we conducted a series of behavioral studies in online chatrooms, replicating the original finding (Navajas et al., 2018) while analyzing the text produced during group deliberations. Based on prior observations suggesting that collective reasoning enhances performance beyond individual reasoning (Sperber & Mercier, 2012; Moshman & and Geil, 1998; Trouche et al., 2014), we hypothesized that groups engaging in the Collective Fermi Estimation strategy would achieve lower error.

Our findings support this hypothesis: groups that more strongly implemented the Collective Fermi Estimation strategy, as assessed by human ratings or through natural language processing, produced more accurate estimates than those that relied on it less (Study 1). In a follow-up experiment (Study 2), we found that instructing groups to apply the Collective Fermi Estimation strategy causally improves performance. Finally, we show that aggregating estimates from individuals applying the Fermi technique in isolation is worse than doing the same with individuals who engaged in the strategy collectively (Study 3), suggesting that the key advantage of this procedure comes not from individual reasoning but from its application at the group level.

2. Results

Study 1: Replication of Navajas et al. (2018) in chatrooms

500 participants (288 female, mean age 25.4 yr, s.d. 7.7 yr), organized into 125 groups of four individuals, participated in Study 1. Participants were incentivized to provide accurate answers (for details, see Methods). The experiment consisted of three stages (Fig. 1A). During the first stage (stage i1), participants were asked individually a series of eight general-knowledge estimation questions (Table S1). In the second stage (stage c), they were divided into groups of four, and they were asked a random subset of four of the previous questions. They were instructed to discuss these questions in a virtual chatroom, and asked to try to reach consensus. All conversations were recorded. In the last stage (stage i2), which followed exactly the same procedure as the first stage, participants could provide new estimates and confidence values.

Despite methodological differences with the original study, particularly with respect to group size (four vs. five) and modality (face-to-face vs. chatroom conversations), we replicated the observation that averaging collective estimates led to lower estimation error than averaging initial independent values (Figure 1B, see Methods for details). We observed this pattern for all values displayed in the x-axis of Figure 1B: For example, the estimation error of aggregating 8 randomly-chosen consensus estimates was substantially lower than the error obtained by averaging 32 randomly-chosen individual values (Mann–Whitney U test: $z=19.3$, $p=2\times10^{-83}$; effect size: Cohen's $d = 6.22$). This finding implies that collective estimates were more accurate than what we would have observed if groups had reached consensus by computing the mean of their initial values (Fig. 1C, normalized error of within-group averages: 1.18 ± 0.05, normalized error of consensus estimates: 0.72 ± 0.04, Wilcoxon signed-rank test: $z=7.3$, $p=3\times10^{-13}$; effect size: Cohen's $d = 0.95$).

Lastly, comparing participants' estimates from the first and third stages, separating questions that were discussed in groups from those that were not, reveals a markedly greater reduction in error for the discussed questions (Fig. 1D). Although participants showed a modest improvement in accuracy for undiscussed questions, the improvement was substantially larger for the discussed questions (generalized linear mixed-effects model, see Methods for details: β = -0.48 [-0.53, -0.43], SE = 0.03, t(2787) = -17.9, p = 5×10^{-68}, R^2 = 0.148).

In principle, one might assume that group discussions were dominated by individuals with higher accuracy or confidence, and that this alone could account for the improved accuracy of consensus estimates compared to simple averages. However, the data indicate that contributions within groups were approximately uniform. When we rank participants by individual error (Fig. S1A) or by confidence (Fig. S1B), the proportion of interventions made by each of the four group members is statistically indistinguishable from an equal share of 25% (two one-sided tests for equivalence, $p < 3\times10^{-5}$ for all comparisons; see Table S2 for details). This suggests that the accuracy gains associated to group discussions were not simply driven by a few dominant individuals, but rather emerged from broadly distributed contributions within the group.

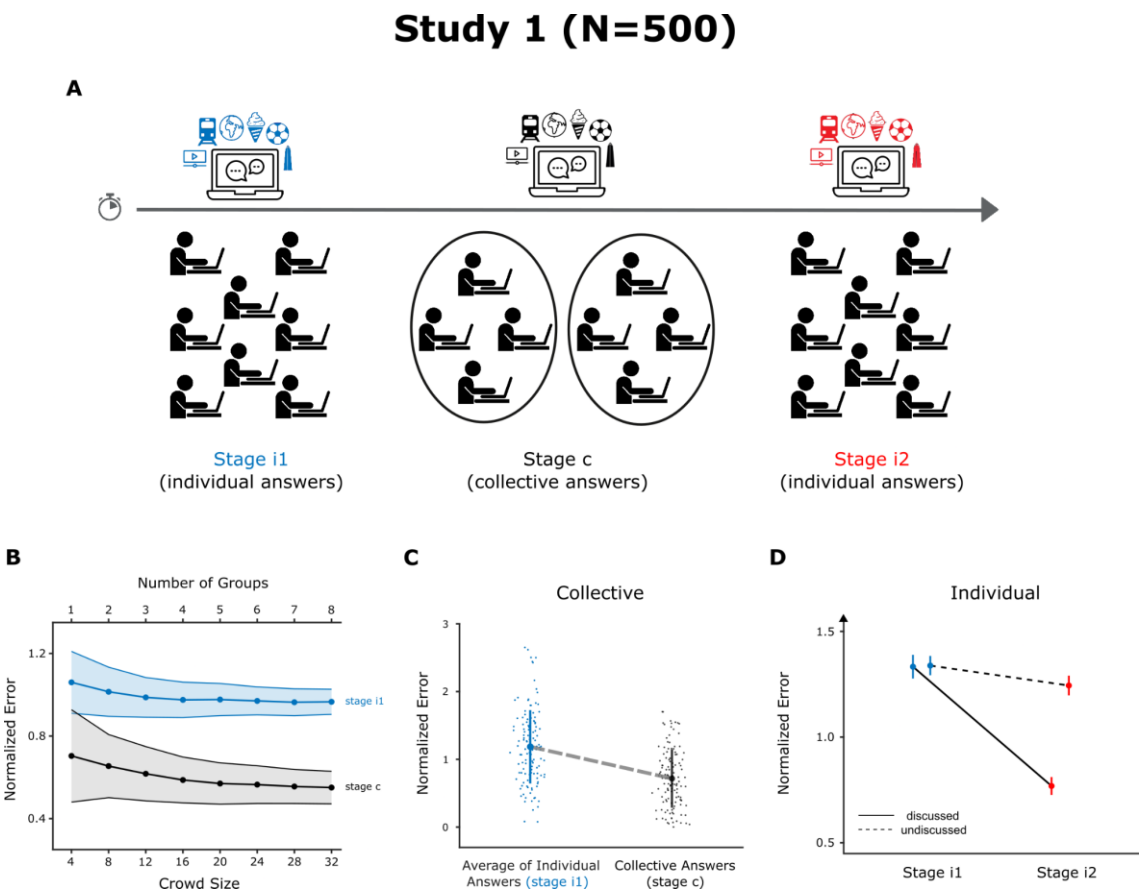

Fig. 1. Empirical results for Experiment 1. (A) The experiment had three stages. On Stage i1 participants answered eight general-knowledge questions individually. On Stage c, participants were divided into groups of four, and were asked half of the questions from the previous stage. Stage i3 followed the same procedure as Stage i1. **(B)** Normalized error of the average of *n* individual answers (blue line for stage i1), and normalized error of the average of $m = n/4$ collective estimates (black line, stage c). The error bars correspond to the standard deviation of the means. **(C)** Average normalized error and distributions of normalized errors of the individual answers (blue), and the collective answers (black). The error bars correspond to the standard deviation of the means. **(D)** Average normalized error for the individual stages, for both discussed and undiscussed questions. The error bars correspond to the 95% confidence intervals.

Collective Fermi Estimation is Associated with Lower Error

We then turned to the central research question of this study: What strategy do groups use to reach consensus? Do they share and combine their initial estimates, or do they recompute the value through problem decomposition? To investigate this, we trained ten

individuals (8 female; mean age = 22.8 years, s.d. = 5.4) to rate each conversation along two dimensions using 11-point Likert scales (see Methods for rating instructions). The first rating assessed the extent to which group members shared and combined their initial judgments, a strategy we call "Sharing Numbers" (or simply "Numbers"). The second rating captured the extent to which participants generated new estimates through rough approximations and basic calculations, a strategy we refer to as "Collective Fermi Estimation" (or simply "Fermi"). Figure 2A illustrates two real examples: one conversation in which participants combined their initial estimates to reach consensus (left panel), and another in which they relied on problem decomposition (right panel). In these examples, the first conversation would be expected to receive a high "Numbers" rating and a low "Fermi" rating, while the second would show the opposite pattern.

We observed moderate-to-strong pairwise Spearman correlations between raters, ranging from .35 to .76 (Fig. 2B; inter-rater correlation for "Numbers": 0.518 ± 0.002, $p < 6\times10^{-20}$; for "Fermi": 0.661 ± 0.001, $p < 3\times10^{-53}$), indicating a high level of consistency across them. Given this reliability, we used the average rating across raters as a summary measure of how strongly each strategy was employed in each conversation. These averaged ratings revealed a strong negative correlation between "Numbers" and "Fermi" scores (Fig. 2C; Pearson $r = -0.92$, $p = 6\times10^{-176}$), suggesting that the two strategies were typically used in a mutually exclusive manner.

We conducted two control analyses to ensure that the patterns we observed were not artifacts of the instructions given to raters. First, we found that the mean "Numbers" rating was positively correlated with the actual proportion of messages in which participants shared a number matching one of the initial individual estimates (Pearson $r = 0.32$, $p = 3.6\times10^{-11}$), and that this behavioral measure also correlated negatively with "Fermi" ratings (Pearson $r = -0.30$, $p = 3\times10^{-10}$). Second, we recruited a separate group of six

raters (5 female; mean age = 25.8 years, s.d. = 5.0) to re-code the conversations using less specific instructions for the "Fermi" dimension. Instead of asking them to identify rough approximations or basic calculations, we asked them to rate the extent to which participants shared reasons to justify the group's consensus value. The mean ratings obtained under this broader instruction strongly correlated with the original "Fermi" ratings (Pearson $r = 0.84$, $p = 2\times10^{-115}$), and all key findings presented in Fig. 2 remained robust under this alternative operationalization of the Collective Fermi Estimation strategy (Fig. S2).

We then examined whether and how the application of each strategy was associated with lower or higher estimation error. We found that conversations with higher "Numbers" ratings produced consensus estimates with higher estimation error (Fig. 2D, Pearson correlation: $r = 0.16$, $p = 0.004$) and that conversations scoring more highly on the "Fermi" strategy led to collective estimates which were less erroneous (Fig. 2E, Pearson correlation: $r = -0.20$, $p = 2 \times 10^{-4}$).

One might assume that groups with higher "Fermi" scores were more accurate at the collective stage simply because they were already more accurate at the individual stage. However, our data rule out this explanation. Conversations in which the Fermi score exceeded the Numbers score did not originate from more accurate individuals than those where the Numbers score was higher (Fig. 2F; Cohen's $d = 0.018$; Bayes Factor $BF_{01} = 8.5$, strong evidence for the null hypothesis). The difference in accuracy between these groups only emerged after deliberation (Fig. 2F; Cohen's $d = 0.37$; Bayes Factor $BF_{01} = 0.085$, strong evidence for the alternative hypothesis).

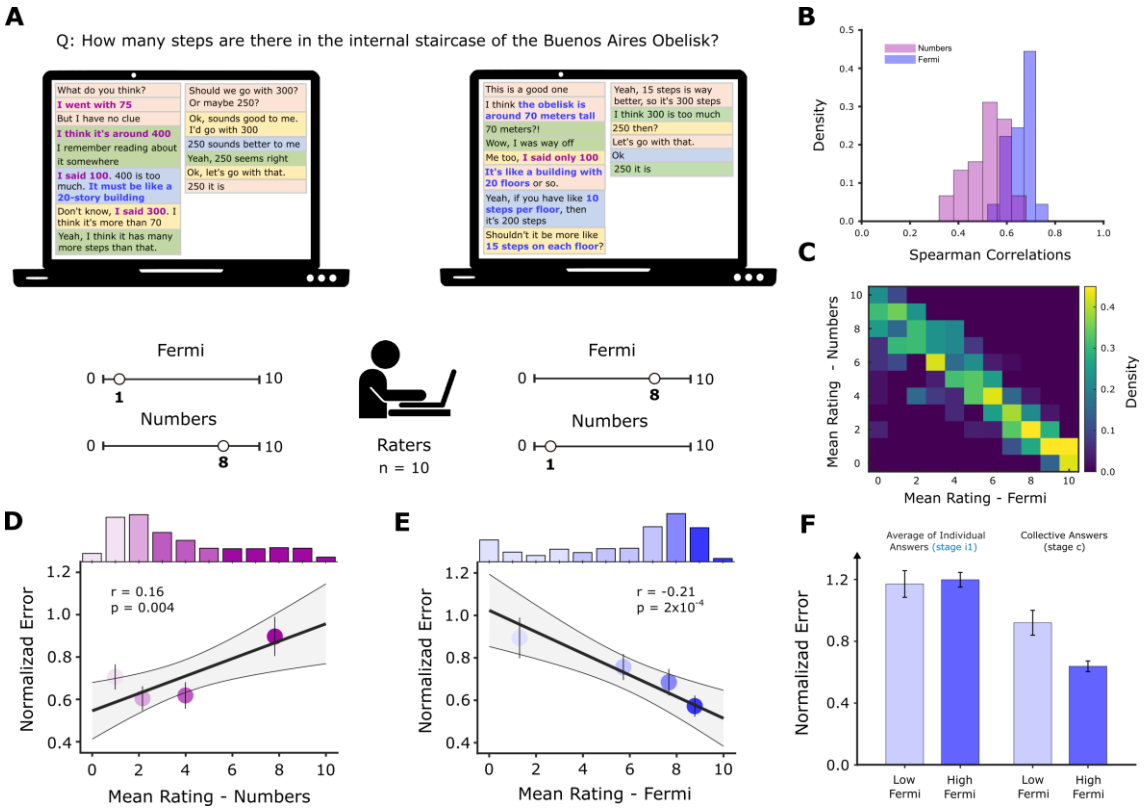

Fig. 2. Analysis of conversations. (A) Examples of conversations rated as high in either the "Numbers" strategy (sharing individual estimates, left panel) or the "Fermi" strategy (collective approximation, right panel). Ten raters classified each conversation along these dimensions. **(B)** Distribution of pairwise Spearman correlations between raters for each strategy. **(C)** Density plot of mean "Numbers" and "Fermi" ratings, showing a strong negative correlation between strategies. **(D–E)** Normalized error as a function of average strategy ratings. Dots represent quartiles in the distribution of mean ratings, which is displayed above each plot. Lines depict best linear fits and shades represent 95% confidence intervals. **(F)** Normalized error before and after deliberation for conversations categorized as "low Fermi" (or "high Numbers") or "high Fermi" (or "low Numbers") based on which strategy received the higher average rating.

Identifying Collective Fermi Estimation Through Automated Language Analysis

Because identifying groups that implemented the Collective Fermi Estimation strategy relies on human ratings, the process remains costly, time-consuming, and difficult to scale. To address this, we developed a computational method to automatically identify

conversations likely to have employed this strategy. The underlying intuition is simple: If groups use the deliberation stage to generate a collective estimate, they are likely to use words that are semantically related to those in the estimation question.

For example, consider a group tasked with estimating the number of steps in the internal staircase of the Buenos Aires Obelisk (Fig. 3A). A group using the Collective Fermi Estimation strategy might compare the Obelisk to a 20-story building, estimate that a typical staircase has around 15 steps per floor, and conclude that the Obelisk should have approximately 250 steps, a solution which is reasonably close to the actual value of 206. Such a conversation would naturally include words that either appear in the question itself (e.g., "staircase," "steps") or are semantically related (e.g., "building," "floor").

To quantify this intuition, we used pre-trained word embeddings (Gutiérrez & Keith, 2019), specifically FastText embeddings trained in Spanish (Bojanowski et al., 2017). Word embeddings (vector representations of words) capture semantic relationships, such that semantically related words are represented by similar vectors. We computed the similarity between all words in the question and all words in the corresponding conversation, excluding stop words and short words. This yielded a distribution of similarity values, from which we extracted the mean to create a "Similarity" index (right panel in Fig. 3A).

We found that this Similarity index was positively correlated with "Fermi" ratings (Pearson $r = 0.21$, $p = 1\times10^{-5}$) and negatively correlated with "Numbers" ratings (Pearson $r = -0.16$, $p = 9\times10^{-4}$). Moreover, the Similarity index was negatively associated with estimation error (Fig. 3B; Pearson $r = -0.15$, $p = 0.007$), and this effect was significantly mediated by "Fermi" ratings (Fig. 3C; total effect = -0.11 ± 0.03, $p = 0.006$; direct effect = -0.08 ± 0.04, $p = 0.02$; indirect effect = -0.03 ± 0.01, $p = 0.009$; see Methods for details).

To assess whether simpler alternatives could reproduce the findings obtained with the Similarity index, we also examined two baseline methods: the total number of words in each conversation, and the number of words from the corresponding question (excluding stop words and short words) that appeared in the conversation. The former could reflect the complexity of the group's reasoning process, while the latter offers a crude approximation of semantic overlap. However, neither measure was significantly associated with normalized group error (Pearson correlations: $r = –0.076$, $p = 0.17$; and $r = –0.082$, $p = 0.14$, respectively). These results suggest that simply counting words or matching lexical items from the question is insufficient to capture the underlying reasoning dynamics. Instead, the semantic similarity captured by word embeddings appears necessary to detect when groups engage in Collective Fermi Estimation.

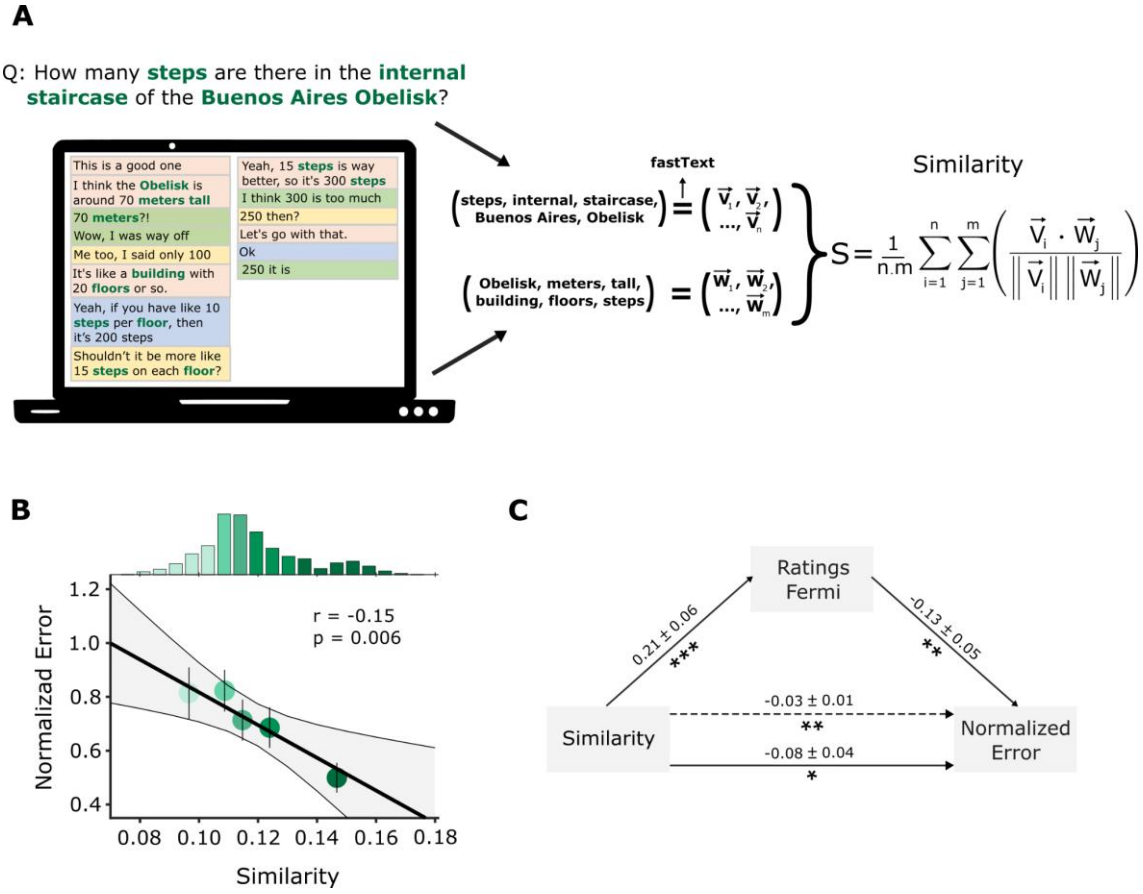

Fig. 3. Analysis of text. (A) For each conversation, we computed the similarity between all words in the conversation and all words in the question (excluding stop words and short words). Bags of words were created for both the question and the conversation, and FastText was used to obtain word embeddings. We then calculated the average similarity between all vectors from the two sets. (B) Normalized error as a function of similarity. Quintiles of the error distribution are shown, along with the best linear fit, 95% confidence intervals, and the Pearson correlation coefficient (r) with its corresponding p-value. Bars above the graph indicate the full distribution of similarity values. (C) Mediation analysis testing whether the relationship between similarity and normalized error is mediated by Fermi ratings.

The results presented so far provide evidence for an association between engaging in problem decomposition and producing collective estimates with higher estimation accuracy. To study if the implementation of this strategy, as opposed to just combining initial estimates, causally drives the observed reduction in collective error, we performed an experiment.

Study 2: Collective Fermi Estimation causally drives collective accuracy

240 participants (148 female, mean age 30.6 yr, s.d. 9.8 yr), organized in 60 groups of four individuals, participated in Study 2. Participants had monetary incentives to estimate these variables as accurately as possible (for details, see Methods). The protocol, hypotheses, and analysis plans were pre-registered at https://aspredicted.org/TL2_Q1Q.

The experiment (Fig. 4A) followed the same procedure as the previous study, with one key difference: after providing their initial estimates and before beginning the deliberation stage, participants watched a short instructional video on how to reach consensus (see Methods for full instructions). Half of the groups were randomly assigned to the "Fermi" condition, where they were instructed to perform rough approximations, break the problem into smaller estimation steps, and recalculate the final answer. The other half were assigned to the "Numbers" condition, where they were instructed to share their initial estimates and combine them in any way they liked.

As predicted, we observed that groups in the "Fermi" condition produced collective estimates with lower error than those in the "Numbers" condition (Fig. 4B; Mann–Whitney U test: $z = 2.1$, $p = 0.04$; Cohen's $d = 0.26$). To confirm that this difference reflected the implementation of different strategies, we trained ten naïve raters (4 female; mean age = 18.2 years, s.d. = 0.4), who had not participated in Study 1 and were blind to condition assignments, to evaluate the extent to which each conversation reflected the "Fermi" or "Numbers" strategy. As expected, conversations in the "Numbers" condition

received higher “Numbers” ratings than those in the “Fermi” condition (Fig. 4C; Mann–Whitney U test: $z = 6.8$, $p = 1\times10^{-11}$; Cohen’s $d = 1.19$). Conversely, conversations in the “Fermi” condition received higher “Fermi” ratings than those in the “Numbers” condition (Fig. 4D; Mann–Whitney U test: $z = 6.3$, $p = 4\times10^{-10}$; Cohen’s $d = 1.06$).

Similarly, we found that instructing groups to share numbers led to a higher proportion of interventions containing numbers from the individual stage (Fig. 4E; Mann–Whitney U test: $z = 5.0$, $p = 6\times10^{-7}$; Cohen’s $d = 0.70$). Finally, the Similarity index (previously correlated with “Fermi” ratings in Study 1) was also higher for groups in the “Fermi” condition than in the “Numbers” condition (Fig. 4F; Mann–Whitney U test: $z = 2.5$, $p = 0.01$; Cohen’s $d = 0.34$). Together, these results provide evidence that the instructions effectively modulated the strategies participants used to reach consensus.

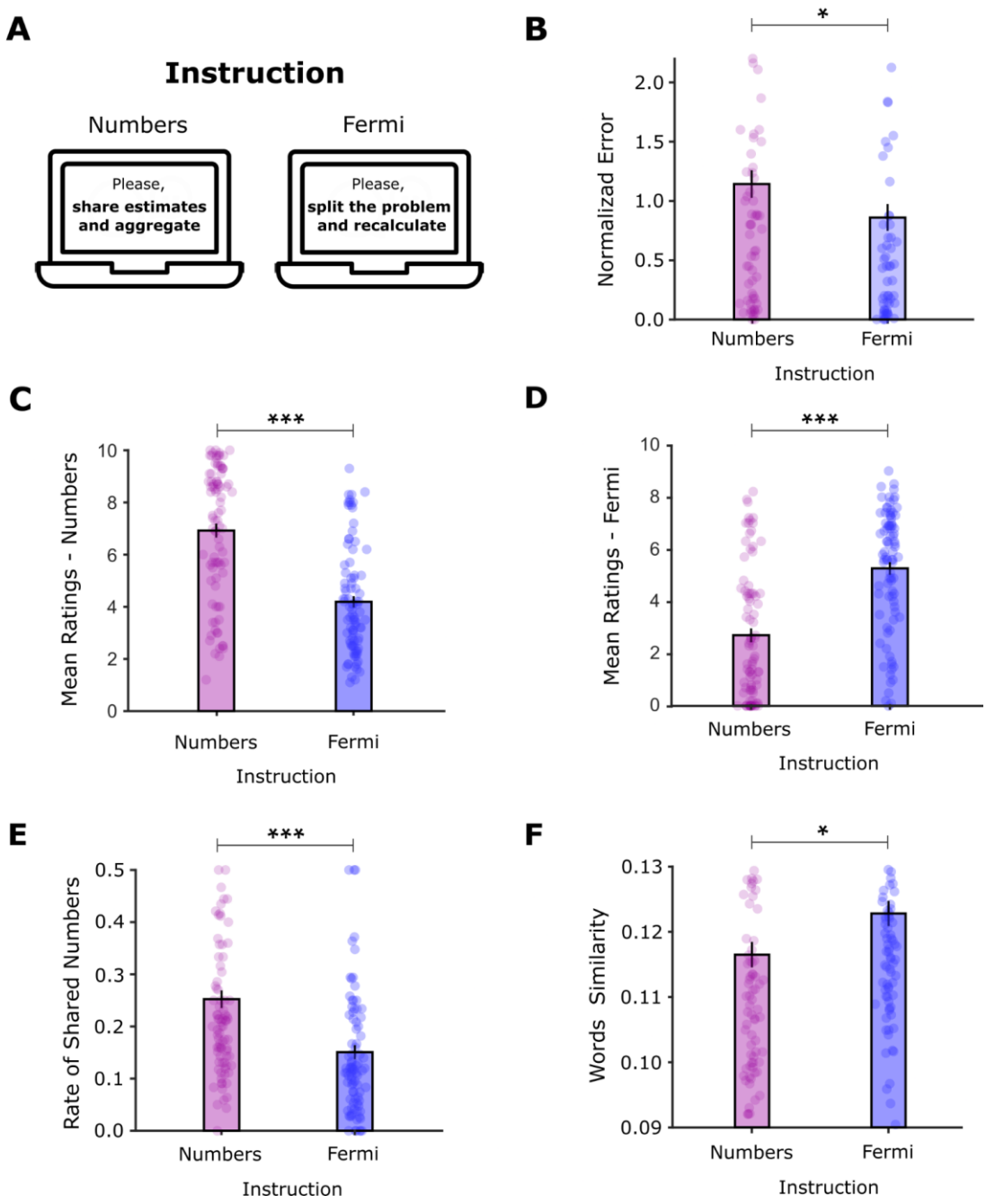

Fig. 4. Results of Experiment 2. (A) The experiment had the same procedure as Study 1, except that before the collective deliberation stage, participants received an instruction either to share numbers or to perform approximations and recompute the value. (B) Normalized error by instruction condition. (C) "Numbers" ratings provided by human raters by condition. (D) "Fermi" ratings provided by human raters by condition. (E) Proportion of shared numbers from the individual stage present in discussions, by condition. (F) Similarity index by instruction condition. In all cases, bars show the mean value, error bars depict s.e.m. and dots show data from individuals. Asterisks denote statistical significance (*: p<.05, **: p<.01, ***: p<.001).

While this experiment offers new insights into the drivers of collective accuracy and supports the idea that groups using the Fermi method achieve better estimates, another explanation is possible. The observed reduction in error in the "Fermi" condition may not be due to collective problem decomposition itself, but rather to individual participants applying the Fermi method within the group, without necessarily engaging in a group-level reasoning process (Landemore & Mercier, 2010). In other words, the strategy might improve accuracy simply by being used, whether individually or collectively. To test this possibility, we conducted another experiment designed to disentangle individual and collective contributions to the effectiveness of the Fermi method.

Study 3: Fermi Estimation is more accurate in groups than alone

N=160 participants (95 female, mean age 29.8 yr, s.d. 10.9 yr) participated in Study 3 (pre-registered at https://aspredicted.org/SZ5_MNP). A random half of the participants were divided into 20 groups of 4 individuals, and the remaining 80 performed the study individually. For half of the participants, the experiment followed exactly the same procedure as the "Fermi" condition in the previous study: Participants received instructions to apply the Fermi method and deliberated in groups to reach a collective estimate (Fig. 5A). The key difference in this experiment was the addition of an individual condition, which applied to the other half of participants, who worked alone rather than in groups. These participants were instructed to apply the Fermi method individually and to write their reasoning and calculations in a chat window, without interacting with others (Fig. 5B). All written responses across both conditions were recorded (see Methods for details).

This design allowed us to directly compare the effectiveness of applying the Fermi method individually versus in groups as a strategy for improving accuracy. To do so, we measured the reduction in error between the initial individual estimates (stage i1) and the

final individual estimates (stage i2) for both discussed and undiscussed questions. Consistent with Study 1 (Fig. 1D), we replicated the finding that group deliberation improves individual accuracy at stage i2 (Fig. 5C). Specifically, while participants showed a modest improvement for undiscussed questions, the improvement was substantially greater for discussed questions (generalized linear mixed-effects model, see Methods for details, β = -0.45 [-0.59, -0.31], SE = 0.07, t(469) = -6.36, p = 5×10^{-10}, R^2 = 0.10).

We then examined the individual deliberation condition (Fig. 5D), where participants worked alone but were instructed to apply the Fermi method. We observed a similar pattern, with a modest improvement for undiscussed questions and a larger improvement for discussed questions. However, the magnitude of this improvement was smaller than the one observed in the group deliberation condition. This difference was statistically significant (Mann–Whitney U test: z = 6.3, p = 3×10^{-10}; Cohen's d = 0.58), as illustrated by the comparison between Fig. 5C and Fig. 5D (generalized linear mixed-effects model, see Methods for details, β = -0.17 [-0.28, -0.07], SE = 0.05, t(469) = -3.28, p = 0.001, R^2 = 0.02). Together, these results suggest that while the Fermi method improves accuracy when applied individually, group deliberation provides an additional benefit beyond individual reasoning.

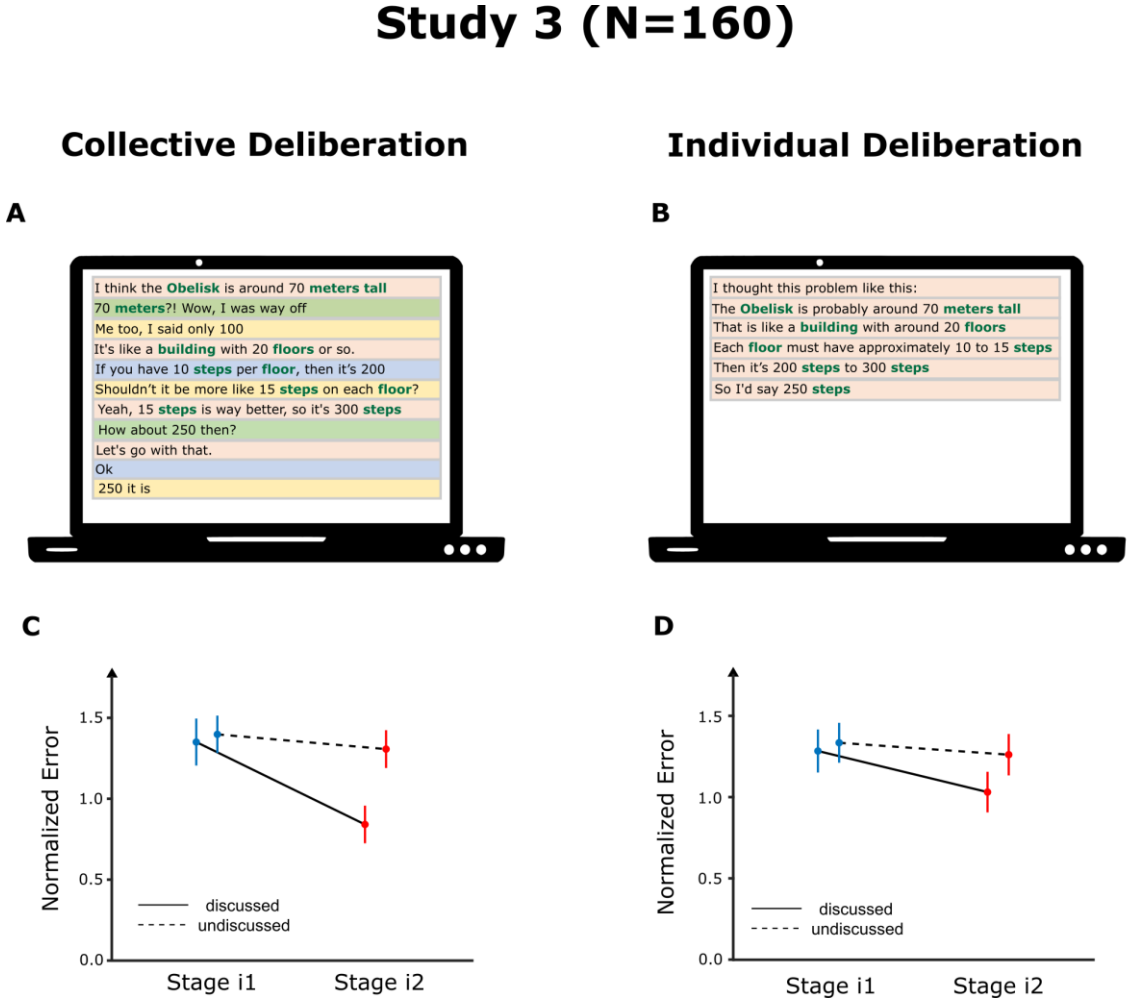

Fig. 5. Empirical results for Experiment 3. The procedure was the same as that of the previous experiment, with two exceptions: **(A)** half of the participants were divided into groups and received the original Fermi instruction, applying the Fermi method as a group. **(B)** the remaining half proceeded individually, and received a modified Fermi instruction prompting them to also apply the Fermi method individually (writing down their procedure). **(C)** Average normalized error for the individual stages, for both discussed and undiscussed questions, for participants that deliberated collectively. The error bars correspond to the 95% confidence intervals. **(D)** Average normalized error for the individual stages for participants that deliberated individually. We followed the same procedure as for the previous panel.

Discussion

Across three studies, we provide converging evidence that small groups can reach more accurate estimates than large, independent crowds when they engage in problem decomposition, a strategy we term "Collective Fermi Estimation". This approach involves making rough calculations and computing an approximate solution, rather than merely aggregating initial signals. In Study 1, we found that groups exhibiting higher use of this strategy, as determined by trained human raters, produced significantly more accurate estimates. Study 2 established a causal link: When groups were explicitly instructed to use the Collective Fermi Estimation strategy, their performance improved relative to groups prompted to aggregate their individual estimates. Finally, in Study 3, we demonstrated that the advantage of this strategy lies in its collective application: Aggregating the estimates of individuals who each applied the strategy independently yielded poorer performance than when the reasoning occurred through group deliberation (Landemore & Mercier, 2010).

One key contribution of this work is the development of a computational method to detect whether a group employed the Collective Fermi strategy based on the content of its conversation. Using word embeddings to measure the semantic similarity between words in the deliberation and those in the original estimation question, we constructed an index that correlates with human Fermi ratings and, critically, predicts group accuracy. This method outperformed simpler proxies such as word count or lexical overlap, suggesting that capturing the semantic structure of a group's reasoning is essential for detecting meaningful deliberation. These results have practical implications as they provide a new scalable method to extract information from crowdsourced estimates under social influence.

These findings also contribute to the broader literature on collective intelligence and the wisdom of crowds by helping to clarify when and how social interaction can improve collective accuracy. Prior work has shown that deliberation can either impair or enhance group performance, depending on the context and the nature of the interaction (Lorenz et al., 2011; Bahrami et al., 2010; Becker et al., 2017; Navajas et al., 2018). Here, we provide evidence in support of the hypothesis that deliberation grounded in reasoning (rather than information exchange) is a key mechanism driving superior group performance. However, several limitations remain. As with much of the crowd wisdom literature, it is unclear whether our findings extend to domains where correct answers are categorical rather than continuous (for a notable exception, see Becker et al., (2022)). Moreover, although our study examined specifically problem decomposition as a technique, it remains possible that other reasoning heuristics could yield similar benefits. Future research should examine whether the gains observed here are specific to the Collective Fermi Estimation strategy or whether they generalize to other structured reasoning processes.

Beyond hypothesis testing, we believe our findings carry broader implications for the design of deliberative environments. The wisdom of crowds has long been interpreted as a model for the epistemic benefits of democratic judgment (Galton, 1907; Landemore, 2012; Surowiecki, 2005). Within this framework, a central theoretical distinction has been drawn between "voting mechanisms," which aggregate individual preferences, and "deliberative mechanisms," which emphasize shared reasoning and mutual justification (Landemore & Mercier, 2010). While deliberative democracy rests on the premise that genuine deliberation can improve collective decisions over information sharing (Landemore, 2017), empirical demonstrations of this principle have remained limited. Our findings help fill this gap by showing not only that groups naturally engage in

collective reasoning under certain conditions, but also that they can be prompted to adopt such a mindset, yielding measurable epistemic benefits.

Overall, this work shows that collective reasoning, and in particular reasoning by approximation, underlies enhanced collective accuracy during deliberation, and offers new tools to detect and foster such processes, contributing to basic understanding of human group decision-making and informing practical applications aimed at improving crowdsourcing strategies.

Methods

Participants

Participants were informed that their participation was completely voluntary, and that they could withdraw their participation at any time. All data were completely anonymous. Overall, we collected data from 900 individuals, all residents of Argentina, across three samples: Study 1 involved 500 participants (288 female, mean age 25.4 yr, s.d. 7.7 yr), in Study 2 we recruited 240 participants (148 female, mean age 30.6 yr, s.d. 9.8 yr), and Study 3 involved 160 participants (95 female, mean age 29.8 yr, s.d. 10.9 yr).

In all cases, participants were entered in a draw for 5 prizes of 100 USD each. This procedure led to a participation fee with an expected value of roughly 6 USD per hour of experiment, which is a fair rate based on the guidelines of subject participation in Buenos Aires, Argentina. All protocols were approved by the ethics committee at Universidad Abierta Interamericana (Buenos Aires, Argentina), protocol 0-1108.

Study 1

The study consisted in three stages. In the first stage, participants independently responded to a series of eight numerical estimation questions within a 5-minute time limit. Alongside their answers, they indicated their level of confidence using a slider bar ranging from low (0) to high (100). Moving on to the second stage, participants communicated

via chat to try to reach a consensus on values for four of the eight questions from the first stage. Participants were instructed to try to reach consensus with a 5-minute time limit. If consensus was achieved earlier, participants had to wait until the 5 minutes elapsed before moving on to the next question. During this stage, participants were not asked to report their confidence levels. The third stage of the study mirrored the first stage, wherein participants individually tackled the same numerical estimation questions within a time limit of five minutes.

Study 2

Study 2 was identical to Study 1, except that they were given instructions on how to reach consensus. In addition to the previous procedure, participants were allocated to one of two conditions. In the *Numbers Condition*, they were explicitly instructed to "share their initial numerical estimates". In the *Fermi Condition* they were explicitly instructed to "divide the problem into smaller estimation problems and perform all relevant computations to reach their final answer". All hypotheses and data analysis plans were preregistered at https://aspredicted.org/TL2_Q1Q.

Study 3

In Study 3, all participants received the same instructions as in the Fermi Condition from the previous study. However, a random half were assigned to groups of four, while the other half worked individually. Those in the individual condition were asked to apply the Fermi Method on their own, documenting their procedure in the same chatroom used by the groups, though they remained alone for the entire experiment. This study was preregistered at https://aspredicted.org/SZ5_MNP.

Estimation questions

The eight questions were presented in random order in the first stage. A random half of them were selected for stage 2. For the third stage, the questions were presented in the same order as the first stage. The complete set of questions and correct answers are available at Supplementary Table 1.

Supervised metrics

For this approach, we trained ten raters (8 female; mean age = 22.8 years, s.d. = 5.4), who were naïve to the hypotheses of the study, on identifying conversations that used the "Fermi" and "Numbers" strategies. Then, they individually read each conversation and determined how much of the Fermi method was actually used on a scale from 0 (completely absent) to 10 (completely present). They were also asked to rate how much the groups implemented the "Numbers" strategy, consisting on exchanging their answers and combining their initial numerical estimates to obtain a final group answer. The order of presentation of the explanation of the two strategies and the conversations to be rated were randomized across raters. The detailed instructions provided to the raters, including the alternative definition of the "Fermi" method, are available in Supplementary Note 1.

Unsupervised metric of problem decomposition

We computed a "Fermi" score using automatized natural language processing (Manning, & Schutze, 1999; Carrillo et al., 2015). To this aim, we leverage pre-trained word embeddings, specifically FastText trained in Spanish (Bojanowski et al., 2017). Word embeddings are vector representations of words in a way that semantically related words have similar vectors. We employed FastText to measure the semantic similarity between the question asked and the entire conversation. Specifically, our method calculates the similarity between all words in the question and all words in the conversation, with the

exclusion of stop words and small words. This process generates a distribution of similarity values:

$$S = \frac{1}{n.m}\sum_{i=1}^{n}\sum_{j=1}^{m}\left(\frac{\vec{V_i}.\vec{W_j}}{\|\vec{V_i}\|\|\vec{W_j}\|}\right) \quad [1]$$

Non-parametric normalization.

The distributions of estimates for each question were centered on different values. To normalize these distributions, we used a non-parametric approach, originally developed in the outlier detection literature, and used in previous studies examining the wisdom of crowds (Barrera-Lemarchand et al., 2024; Navajas et al., 2018). This procedure involves subtracting the median value of the distribution and then dividing by the median absolute deviance, leading to a non-parametric z-score value. The rationale for normalizing our data was twofold. First, we used this procedure to reject outliers in the distribution of responses. Following previous studies, we discarded all responses that deviated from the median by more than 2.5 times the median absolute deviance. The second purpose of normalization was to average our results across different questions. This helps the visualization of our data, but our main findings can be replicated without any normalization (Table S3).

Mixed-effects models.

To compare participants' estimates from the first and third stages in Study 1 (Fig. 1D), we used a generalized linear mixed-effects model, with random intercepts for each question, and fixed effects for each group and question. The same type of generalized linear mixed-effects model was used in Study 3 (Fig. 5C and Fig. 5D).

Code and data availability

All data and codes supporting our analyses are available at the Open Science Framework (https://osf.io/bru7k/?view_only=20b5fd8f315b4cb083186aeba294b632).

References

Anderson, P. M., & Sherman, C. A. (n.d.). Applying the fermi estimation technique to business problems. *Journal of Applied Business & Economics*, *10(5)*.

Bahrami, B., Olsen, K., Latham, P. E., Roepstorff, A., Rees, G., & Frith, C. D. (2010). Optimally Interacting Minds. *Science*, *329*(5995), 1081–1085. https://doi.org/10.1126/science.1185718

Barrera-Lemarchand, F., Balenzuela, P., Bahrami, B., Deroy, O., & Navajas, J. (2024). Promoting Erroneous Divergent Opinions Increases the Wisdom of Crowds. *Psychological Science*, *35*(8), 872–886. https://doi.org/10.1177/09567976241252138

Becker, J. A., Guilbeault, D., & Smith, E. B. (2022). The Crowd Classification Problem: Social Dynamics of Binary-Choice Accuracy. *Management Science*, *68*(5), 3949–3965. https://doi.org/10.1287/mnsc.2021.4127

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, *5*, 135–146.

Carrillo, F., Cecchi, G. A., Sigman, M., & Fernandez Slezak, D. (2015). Fast distributed dynamics of semantic networks via social media. *Computational intelligence and neuroscience*, 2015(1), 712835.

Dezecache, G., Dockendorff, M., Ferreiro, D. N., Deroy, O., & Bahrami, B. (2022). Democratic forecast: Small groups predict the future better than individuals and crowds. *Journal of Experimental Psychology: Applied*, *28*(3), 525–537. https://doi.org/10.1037/xap0000424

De Condorcet, M. M. J. A. N. (1785). *Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix de Caritat*.

Epp, D. A. (2017). Public policy and the wisdom of crowds. *Cognitive Systems Research*, *43*, 53–61. https://doi.org/10.1016/j.cogsys.2017.01.002

Espina Mairal, S., Bustos, F., Solovey, G., & Navajas, J. (2024). Interactive crowdsourcing to fact-check politicians. *Journal of Experimental Psychology: Applied*, *30*(1), 3–15. https://doi.org/10.1037/xap0000492

Galton, F. (1949). Vox Populi (1907) Nature, n. 1949, vol. 75, pp. 450-451 (traduzione di Romolo Giovanni Capuano\copyright). *Nature*, *75*, 450–451.

Gürçay, B., Mellers, B. A., & Baron, J. (2015). The Power of Social Influence on Estimation Accuracy. *Journal of Behavioral Decision Making*, *28*(3), 250–261. https://doi.org/10.1002/bdm.1843

Gutiérrez, L., & Keith, B. (2019). A Systematic Literature Review on Word Embeddings. In J. Mejia, M. Muñoz, Á. Rocha, A. Peña, & M. Pérez-Cisneros (Eds.), *Trends and Applications in Software Engineering* (Vol. 865, pp. 132–141). Springer International Publishing. https://doi.org/10.1007/978-3-030-01171-0_12

Juni, M. Z., & Eckstein, M. P. (2015). Flexible human collective wisdom. *Journal of Experimental Psychology: Human Perception and Performance*, *41*(6), 1588.

Kameda, T., Toyokawa, W., & Tindale, R. S. (2022). Information aggregation and collective intelligence beyond the wisdom of crowds. *Nature Reviews Psychology*, *1*(6), 345–357.

Keller, A., Gerkin, R. C., Guan, Y., Dhurandhar, A., Turu, G., Szalai, B., Mainland, J. D., Ihara, Y., Yu, C. W., Wolfinger, R., Vens, C., Schietgat, L., De Grave, K., Norel, R., DREAM Olfaction Prediction Consortium, Stolovitzky, G., Cecchi, G. A., Vosshall, L. B., & Meyer, P. (2017). Predicting human olfactory perception from

chemical features of odor molecules. *Science*, *355*(6327), 820–826. https://doi.org/10.1126/science.aal2014

Krockow, E. M., Kurvers, R. H. J. M., Herzog, S. M., Kämmer, J. E., Hamilton, R. A., Thilly, N., Macheda, G., & Pulcini, C. (2020). Harnessing the wisdom of crowds can improve guideline compliance of antibiotic prescribers and support antimicrobial stewardship. *Scientific Reports*, *10*(1), 18782. https://doi.org/10.1038/s41598-020-75063-z

Kurvers, R. H. J. M., Herzog, S. M., Hertwig, R., Krause, J., Carney, P. A., Bogart, A., Argenziano, G., Zalaudek, I., & Wolf, M. (2016). Boosting medical diagnostics by pooling independent judgments. *Proceedings of the National Academy of Sciences*, *113*(31), 8777–8782. https://doi.org/10.1073/pnas.1601827113

Landemore, H. E. (2017). Deliberative Democracy as Open, Not (Just) Representative Democracy. *Daedalus*, *146*(3), 51–63. https://doi.org/10.1162/DAED_a_00446

Landemore, H. E. (2012). Why the Many Are Smarter than the Few and Why It Matters. *Journal of Deliberative Democracy*, *8*(1), Article 1. https://doi.org/10.16997/jdd.129

Landemore, H. E., & Mercier, H. (2010). *'Talking it Out': Deliberation with Others Versus Deliberation Within* (SSRN Scholarly Paper No. 1660695). Social Science Research Network. https://doi.org/10.2139/ssrn.1660695

Landemore, H., & Page, S. E. (2015). Deliberation and disagreement: Problem solving, prediction, and positive dissensus. *Politics, Philosophy & Economics*, *14*(3), 229–254. https://doi.org/10.1177/1470594X14544284

Lee, J., Li, T., & Shin, D. (2022). The Wisdom of Crowds in FinTech: Evidence from Initial Coin Offerings. *The Review of Corporate Finance Studies*, *11*(1), 1–46. https://doi.org/10.1093/rcfs/cfab014

Lorenz, J., Rauhut, H., Schweitzer, F., & Helbing, D. (2011). How social influence can undermine the wisdom of crowd effect. *Proceedings of the National Academy of Sciences*, *108*(22), 9020–9025. https://doi.org/10.1073/pnas.1008636108

Madirolas, G., & de Polavieja, G. G. (2015). Improving collective estimations using resistance to social influence. *PLoS Computational Biology*, *11*(11), e1004594.

Manning, C., & Schutze, H. (1999). *Foundations of statistical natural language processing*. MIT press.

Mellers, B., Ungar, L., Baron, J., Ramos, J., Gurcay, B., Fincher, K., Scott, S. E., Moore, D., Atanasov, P., Swift, S. A., Murray, T., Stone, E., & Tetlock, P. E. (2014a). Psychological Strategies for Winning a Geopolitical Forecasting Tournament. *Psychological Science*, *25*(5), 1106–1115. https://doi.org/10.1177/0956797614524255

Mellers, B., Ungar, L., Baron, J., Ramos, J., Gurcay, B., Fincher, K., Scott, S. E., Moore, D., Atanasov, P., Swift, S. A., Murray, T., Stone, E., & Tetlock, P. E. (2014b). Psychological Strategies for Winning a Geopolitical Forecasting Tournament. *Psychological Science*, *25*(5), 1106–1115. https://doi.org/10.1177/0956797614524255

Moshman, D., & and Geil, M. (1998). Collaborative Reasoning: Evidence for Collective Rationality. *Thinking & Reasoning*, *4*(3), 231–248. https://doi.org/10.1080/135467898394148

Navajas, J., Niella, T., Garbulsky, G., Bahrami, B., & Sigman, M. (2018). Aggregated knowledge from a small number of debates outperforms the wisdom of large crowds. *Nature Human Behaviour*, *2*(2), Article 2. https://doi.org/10.1038/s41562-017-0273-4

Nityananda, R. (2014). Fermi and the art of estimation. *Resonance*, *19*(1), 73–81.

Pescetelli, N., Rutherford, A., & Rahwan, I. (2021). Modularity and composite diversity affect the collective gathering of information online. *Nature Communications*, *12*(1), 3195. https://doi.org/10.1038/s41467-021-23424-1

Prelec, D., Seung, H. S., & McCoy, J. (2017). A solution to the single-question crowd wisdom problem. *Nature*, *541*(7638), 532–535.

Raafat, R. M., Chater, N., & Frith, C. (2009). Herding in humans. *Trends in Cognitive Sciences*, *13*(10), 420–428.

Ray, R. (2006). Prediction Markets and the Financial "Wisdom of Crowds." *Journal of Behavioral Finance*, *7*(1), 2–4. https://doi.org/10.1207/s15427579jpfm0701_1

Sperber, D. & Mercier, H. (2012). Reasoning as a social competence. En H. Landemore & J. Elster (Eds.). In Hélène Landemore & Jon Elster (Eds.), *Collective wisdom: Principles and mechanisms* (pp. 368–392). Cambridge University Press.

Surowiecki, J. (2005). *The wisdom of crowds*. Anchor.

Trouche, E., Sander, E., & Mercier, H. (2014). Arguments, more than confidence, explain the good performance of reasoning groups. *Journal of Experimental Psychology: General*, *143*(5), 1958–1971. https://doi.org/10.1037/a0037099

Weinstein, L., & Adam, J. (2009). *Guesstimation: Solving the World's Problems on the Back of a Cocktail Napkin*. Princeton University Press. https://doi.org/10.1515/9781400824441

Supplementary Information

Question	Correct Answer
How many goals were scored in total at the 2014 FIFA World Cup played in Brazil?	171
How many steps are there on the internal staircase of The Obelisk of Buenos Aires?	206
How many people (in millions) live in South America?	373
What is the total length (in meters) of line A of the Buenos Aires underground network?	10800
Until 2021, in how many Copa Libertadores finals has an Argentine team participated in?	37
How many kilograms of ice cream does a person in Argentina eat per year?	7
How many episodes are there in the American television series "Friends"?	236
How many countries have territory in the southern hemisphere, excluding the Antarctic continent?	48

Table S1: Questions tested in all experiments and their correct answers

Ind.	Accuracy			Confidence		
	TOST (d_1 = 20%, d_2 = 30%)		Bayes Factor	TOST (d_1 = 20%, d_2 = 30%)		Bayes Factor
1	p_1 ($x < d_1$)	$p < 10^{-200}$	15.0	p_1 ($x < d_1$)	$4x10^{-5}$	0.005
	p_2 ($x > d_2$)	10^{-8}		p_2 ($x > d_2$)	10^{-32}	
	C.I.	[0,0.022]		C.I.	[-0.036,-0.016]	
2	p_1 ($x < d_1$)	10^{-16}	52.7	p_1 ($x < d_1$)	10^{-11}	18.5
	p_2 ($x > d_2$)	$2x10^{-14}$		p_2 ($x > d_2$)	$3x10^{-21}$	
	C.I.	[-0.008,0.013]		C.I.	[-0.019,0.001]	
3	p_1 ($x < d_1$)	$2x10^{-13}$	48.7	p_1 ($x < d_1$)	$p < 10^{-200}$	3.90
	p_2 ($x > d_2$)	$8x10^{-17}$		p_2 ($x > d_2$)	$4x10^{-9}$	
	C.I.	[-0.014,0.007]		C.I.	[0.004,0.024]	
4	p_1 ($x < d_1$)	$6x10^{-12}$	12.2	p_1 ($x < d_1$)	$p < 10^{-200}$	0.34
	p_2 ($x > d_2$)	$2x10^{-23}$		p_2 ($x > d_2$)	$5x10^{-6}$	
	C.I.	[-0.020,0.001]		C.I.	[0.010,0.032]	

Table S2: Results from two one-sided tests (TOST) of equivalence and Bayes factor analyses for the proportion of interventions of each individual (1-4), ordered by accuracy (first columns) and confidence (last columns). All TOSTs indicate that all proportions of interventions are within the 20-30% interval.

	β	SE	t	p-val	95% C.I.
Intercept	0.20	0.02	8.00	$2x10^{-14}$	[0.15 ; 0.24]
Fermi Ratings	-0.01	0.003	-2.98	0.003	[-0.02 ; -0.003]
DF	323				
Log-likelihood	107				
AIC	-206				
BIC	-191				

Table S3: Replication of the main result of Study 1 without normalization. Generalized linear mixed-effect model of Collective Error, with the Fermi Ratings as a fixed effect, and random intercepts for each question.

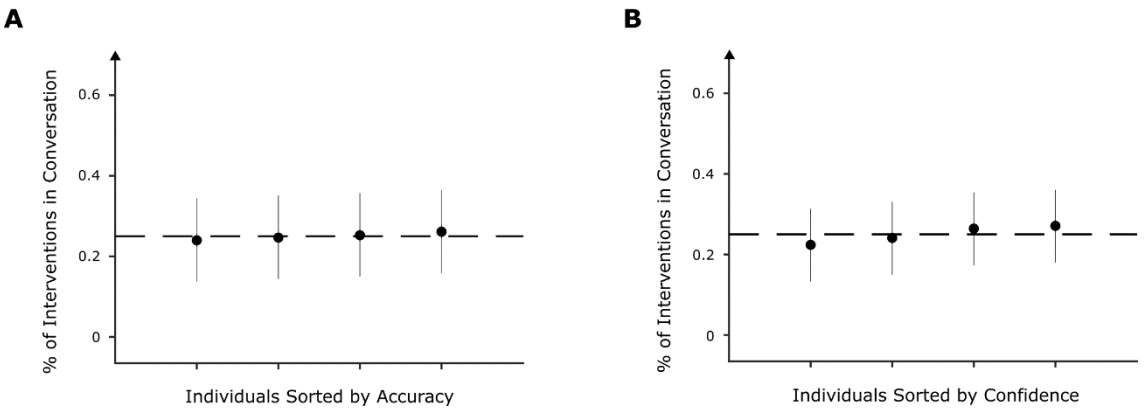

Fig. S1. Accuracy and confidence for each individual for Experiment 1. (A) Percentage of interventions (messages) of each individual for each conversation. Individuals are sorted by accuracy. The full lines represent the standard deviation of the mean, and the dotted line represents 25% **(B)** Percentage of interventions (messages) of each individual for each conversation. Individuals are sorted by confidence. The full lines represent the standard deviation of the mean, and the dotted line represents 25%.

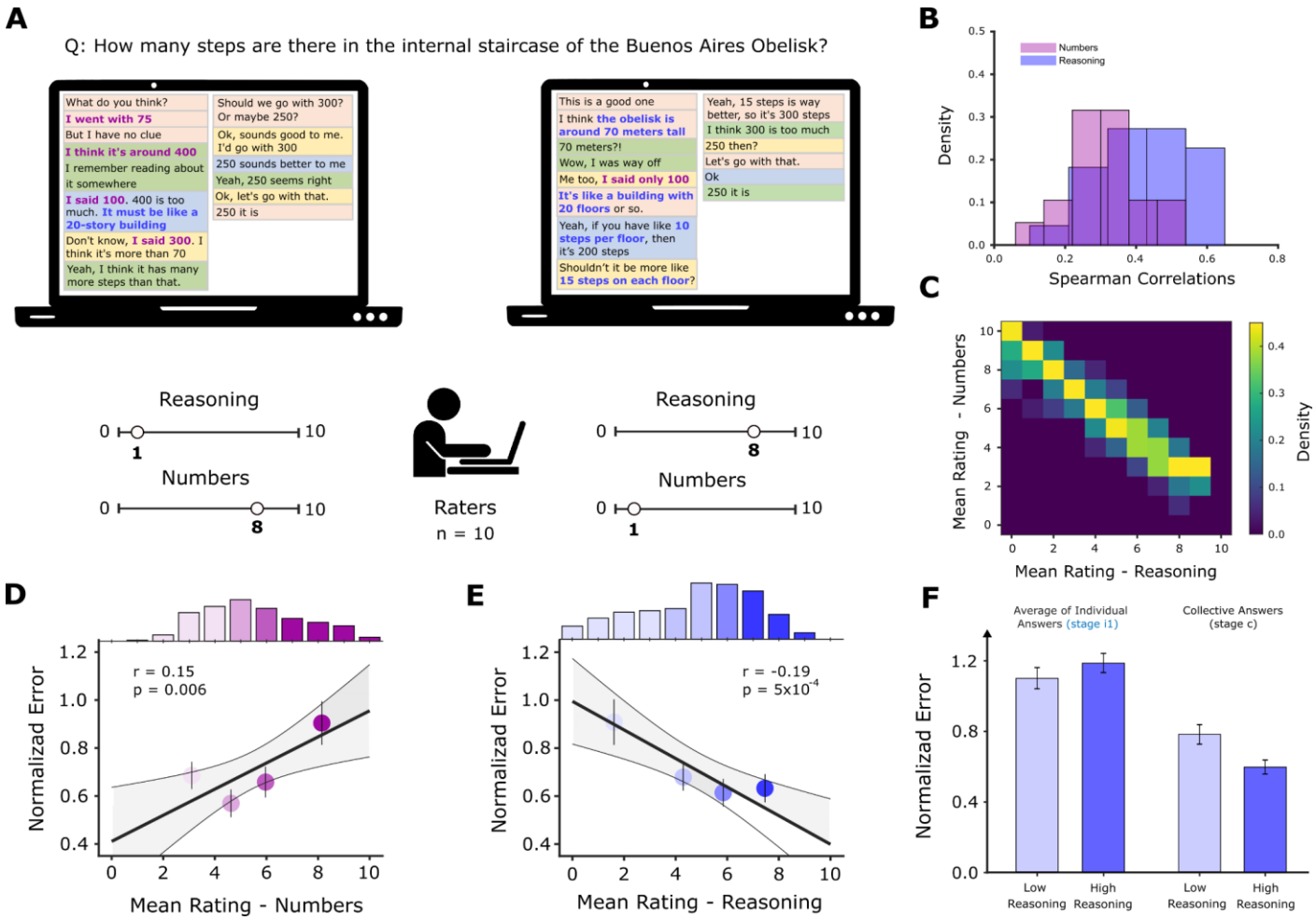

Fig. S2. Analysis of Conversations. Replication of Figure 2 under a different way of coding conversations, (A) Example of conversations that lead to opposing ratings of the strategies used in them. 6 raters were trained on how to classify conversations in terms of whether participants pooled their previous individual answers ("Numbers strategy") or reached a new collective answer by employing a reasoning procedure ("Reasoning strategy"). **(B)** Distribution of correlations between the values provided by each rater with each other, for each strategy. The correlation of the values provided by different raters is generally high, for both strategies. **(C)** Density of ratings for the Reasoning and Numbers Strategies. Ratings for both strategies were strongly negatively correlated: whenever raters on average assigned a high value for the Reasoning strategy in a conversation, they also tended to assign on average a very low value for the Numbers strategy to that conversation. **(D)** Normalized error as a function of the "Numbers strategy" ratings (averaged across raters). We plot the quartiles of the distribution of normalized errors, along with the best linear fit of the data, and the 95% confidence intervals. The bars above the main graph show the actual distribution of normalized errors. We also show the Pearson correlation coefficient, and the corresponding p-value **(E)** Normalized error as a function of the

"Reasoning strategy" ratings (averaged across raters). We plot the quartiles of the distribution of normalized errors, along with the best linear fit of the data, and the 95% confidence intervals. The bars above the main graph show the actual distribution of normalized errors. We also show the Pearson correlation coefficient, and the corresponding p-value. **(F)** Variation of Normalized Error between groups and averages of the individual answers. We categorize a group for each conversation as either "low Reasoning" (when the Reasoning rating was lower than the Numbers rating) or "high Reasoning" (when the Reasoning rating was higher than the Numbers rating).

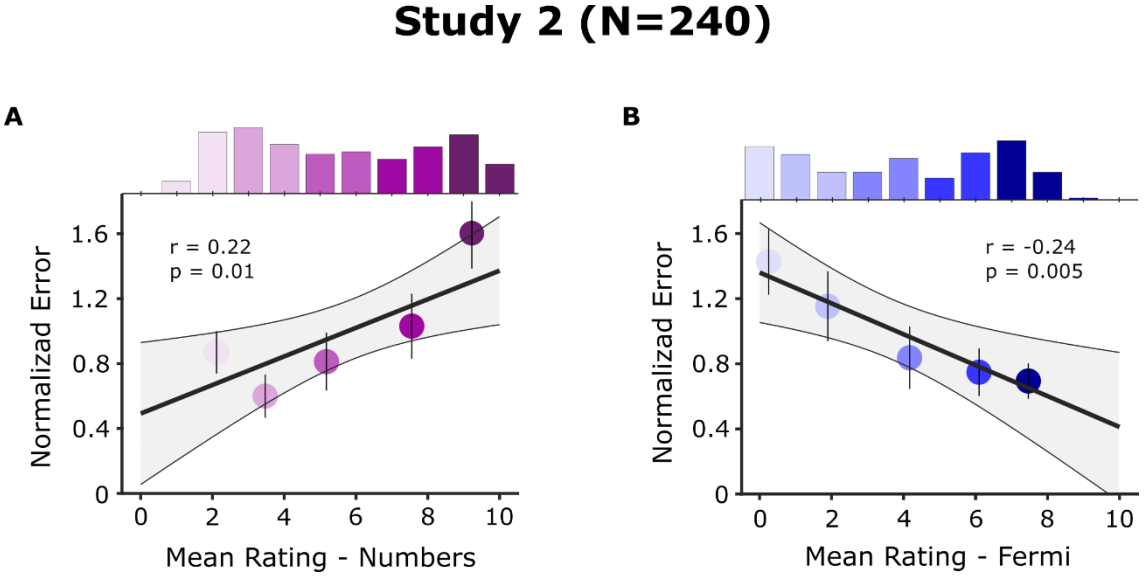

Fig. S3. Normalized error as a function of human ratings. Replication of Figs 2D and 2E (Study 1) in Study 2. (A) Normalized error as a function of the "Numbers strategy" ratings (averaged across raters). We plot the quintiles of the distribution of normalized errors, along with the best linear fit of the data, and the 95% confidence intervals. The bars above the main graph show the actual distribution of normalized errors. We also show the Pearson correlation coefficient, and the corresponding p-value **(E)** Normalized error as a function of the "Fermi strategy" ratings (averaged across raters). We plot the quintiles of the distribution of normalized errors, along with the best linear fit of the data, and the 95% confidence intervals. The bars above the main graph show the actual distribution of normalized errors. We also show the Pearson correlation coefficient, and the corresponding p-value

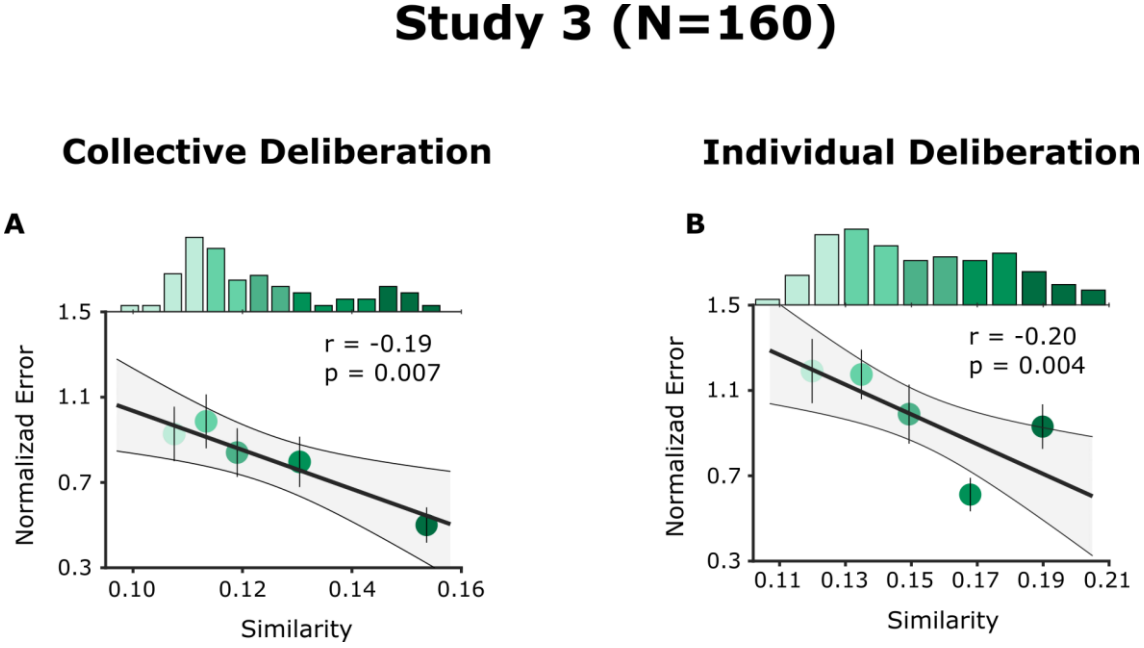

Fig. S4. Normalized error as a function of similarity. Replication of Fig 3B (Study 1) in Study 3. (A) Normalized error as a function of similarity for participants that deliberated collectively. We plot the quintiles of the distribution of normalized errors, along with the best linear fit of the data, and the 95% confidence intervals. The bars above the main graph show the actual distribution of normalized errors. We also show the Pearson correlation coefficient and the corresponding p-value. **(B)** Normalized error as a function of similarity for participants that deliberated individually. We followed the same procedure as for the previous panel.

Supplementary Note 1

The instructions provided to the raters were the following:
Your task is to read conversations between groups of four individuals discussing factual topics (e.g., “What is the height of the Eiffel Tower?”). Each group was instructed to “try to reach consensus” on these topics. There are eight different questions in total.

All participants first completed an individual stage, where they provided answers to the eight questions on their own. Afterward, they were divided into groups and asked to discuss 4 of the 8 questions, with five minutes allotted per question (they could see their own previous answers). Most groups did reach consensus, but this was not required.

As you read each conversation, you will be asked to rate, on a scale from 0 (completely absent) to 10 (completely present), the extent to which two possible strategies for reaching consensus were used.

The first strategy involves combining the estimates from the initial individual stage. This could mean calculating an average, weighting responses by confidence, discarding some answers in favor of others, or any other approach that makes use of their original estimates. We call this the “Numbers Strategy.”

The second strategy involves breaking down the problem into smaller parts, estimating those quantities, and then combining them to arrive at a final answer. For example, for the question “What is the height of the Eiffel Tower?”, one might assume the Tower is similar to a 100-story building (first estimate), and if each story is 3 meters tall (second estimate), then the Tower would be about 300 meters high (which in this case happens to be correct, though this is not always so). We call this the “Fermi Strategy.”

Thus, for each conversation, you will be asked to answer two questions:

a) Did they reach consensus by combining their initial estimates from the first stage? (Numbers Strategy) / 0 = completely absent, 10 = completely present

b) Did they reach consensus by breaking the problem into smaller parts? (Fermi Strategy) / 0 = completely absent, 10 = completely present

Note that the ratings for a) and b) do not need to sum to 10. A group may have used both strategies, or neither. Also, be sure to use the full scale rather than only the extremes (0 and 10).