When Direct Prediction Fails: Evidence from LLM-Based Misinformation Risk Evaluation

Zonghuan Xu[1], Xiang Zheng[2], Yutao Wu[3], Xingjun Ma[1]
[1]Institute of Trustworthy Embodied AI, Fudan University, Shanghai, China
Shanghai Key Laboratory of Multimodal Embodied AI, Shanghai, China
[2]City University of Hong Kong, Hong Kong SAR, China
[3]Deakin University, Australia
Corresponding author: xingjunma@fudan.edu.cn

Abstract

LLMs make it increasingly easy to generate deceptive content at scale, creating a need for scalable misinformation risk evaluation based on whether readers find such content credible and are willing to share it. A natural approach is to ask an LLM these questions directly and treat the returned scores as predictions of the corresponding human ratings. Implicit in this practice is the assumption that asking about a reader response produces the score that best predicts it. We test this assumption using matched credibility and willingness-to-share ratings for 290 deceptive articles from 317 participants and eight LLM evaluators. Unexpectedly, the assumption holds for credibility but fails for sharing. For every evaluator, credibility scores track human sharing at least as closely as sharing scores, while sharing scores offer no detectable benefit beyond credibility when predicting responses to unseen scenarios. This pattern persists when the questions are asked separately or in reversed order. Our results suggest that directly asking for the target response may not always yield the most effective score for predicting it. Comparing direct scores with indirect paths through related judgments may reveal a more effective predictive route. Deciding what to ask may be as important as refining how to ask it.

1 Introduction

Large language models (LLMs) have made it easier to produce false or misleading content (Kreps, McCain, and Brundage 2022; Spitale, Biller-Andorno, and Germani 2023) at scale. The risk posed by generated misinformation cannot be assessed from generation alone, since it also depends on how readers respond (Matz et al. 2024; Bai et al. 2025; Salvi et al. 2025; Spitale, Biller-Andorno, and Germani 2023) when they encounter the content (Pennycook et al. 2021; Rathje et al. 2023; Traberg et al. 2024; Stefkovics and Gere 2026). Human studies provide direct evidence of these responses, but they are costly and difficult to conduct at the pace of model generation (Dillion et al. 2023; Bail 2024; Hullman et al. 2026; Krsteski et al. 2026). LLMs are therefore increasingly used to estimate human ratings (Chiang and Lee 2023; Dong, Hu, and Collier 2024; Bavaresco et al. 2025; Huidrom and Belz 2025) and reactions (Aher, Arriaga, and Kalai 2023; Argyle et al. 2023; Park et al. 2023; Ashokkumar et al. 2026). Using LLM scores in this way (Zheng et al. 2023; Li et al. 2025; Thakur et al. 2025; Krumdick et al. 2025) offers a scalable approach to misinformation risk evaluation (Cui, Li, and Zhou 2025; Wang et al. 2025; Cao et al. 2025; Anthis et al. 2025), but its usefulness depends (Lin 2024; Licht et al. 2025; Libovický 2026; Pangakis and Wolken 2025) on how well those scores predict the responses of human readers (Bisbee et al. 2024; Petrov, Serapio-García, and Rentfrow 2024; Ivey et al. 2026; Hoq and Weninger 2026).

Misinformation research commonly evaluates two reader responses: perceived credibility and willingness to share (Pennycook et al. 2021; Rathje et al. 2023; Sun and Xie 2024; Globig and Sharot 2024). Perceived credibility captures whether readers regard an article and its central claims as believable, reflecting the potential for deceptive content to gain acceptance (Ou and Ho 2024; Phillips et al. 2025; Spitale, Biller-Andorno, and Germani 2023; Stefkovics and Gere 2026). Willingness to share captures whether readers would pass the article on to others, reflecting its potential for further circulation (Rathje et al. 2023; Sun and Xie 2024; Globig and Sharot 2024; Traberg et al. 2024). Both responses have also been used to assess the effects of AI-generated misinformation (Kreps, McCain, and Brundage 2022; Spitale, Biller-Andorno, and Germani 2023; Stefkovics and Gere 2026). In our study, human readers and LLM evaluators answer the same credibility and sharing questions for the same articles (Liu et al. 2023; Fu et al. 2024; Pan et al. 2024; Huidrom and Belz 2025). The credibility score predicts human credibility more effectively than the sharing score, as expected. Unexpectedly, the ordering reverses for human sharing: for every evaluator, the credibility score is more predictive of human sharing than the score obtained by asking about sharing itself.

We examine this reversal using matched human and LLM ratings of 290 deceptive articles generated by three LLMs across 99 misinformation scenarios. The scenarios span five public-issue topics informed by the United Nations Sustainable Development Goals (SDGs). Each scenario pairs a false or misleading claim documented by Reuters Fact Check with a concrete communication goal, producing goal-directed deceptive content grounded in misinformation that has circulated in practice. A total of 317 Prolific participants provided 1,256 participant–article ratings, and eight LLM evaluators rated the same articles. Human readers and LLMs answered the same credibility and sharing questions using the same 1–7 scales (Santurkar et al. 2023; Atari et al. 2023; Hu and Collier 2024; Lin et al. 2026). Since human credibility and

sharing are substantially related, credibility provides a plausible indirect route to predicting sharing (Pennycook et al. 2021; Rathje et al. 2023; Sun and Xie 2024; Globig and Sharot 2024). Direct prediction uses the LLM score obtained from the question about the target response. Indirect prediction begins with the score for the related judgment and relies on its relationship with the target response. For human sharing, we therefore compare direct prediction from the sharing score with indirect prediction through the credibility score. The same comparison for human credibility serves as a positive control. We compare the two routes in their associations with human ratings, their relative performance when considered jointly, and their predictions for unseen scenarios. We also test whether sharing scores add predictive value beyond credibility and whether the findings persist when the questions are asked separately or in reversed order (Tjuatja et al. 2024; Zhou et al. 2024; Shi et al. 2025; Chen et al. 2024).

The results consistently confirm the expected pattern for credibility. Across all eight evaluators, credibility scores predict human credibility more effectively than sharing scores. For human sharing, the advantage runs in the opposite direction. At the article level, credibility scores are more strongly associated with human sharing for every evaluator. When both scores are considered jointly, sharing scores are never reliably the stronger predictor. On unseen scenarios, credibility scores produce lower prediction error for seven of the eight evaluators. Sharing scores never detectably outperform credibility scores and add no detectable predictive value once credibility is included. As a positive control, credibility scores explain 11.3–18.0% of held-out variation in human credibility. The main finding persists when the credibility and sharing questions are asked separately or in reversed order.

In this setting, the indirect route through credibility provides a more effective prediction of human sharing than the direct route through the sharing question. This finding distinguishes semantic alignment between a question and its intended target from predictive alignment between the resulting score and human ratings. A question can match the target in meaning while a related question produces a score that predicts it better. Current practice often begins by asking an LLM for the desired response and then refining the wording or supplying additional context. Our results suggest that the initial choice of what the model should assess also requires empirical validation. In misinformation risk evaluation, and potentially in other applications that use LLMs to estimate human responses, an indirect path through a related assessment may provide a more effective prediction. Deciding what to ask may be as important as refining how to ask it.

2 Methodology

We operationalize misinformation risk evaluation through perceived credibility and willingness to share. We then compare whether asking about credibility or sharing produces the score that better predicts each human response. Human participants and LLM evaluators rate the same articles on perceived credibility and willingness to share. The analyses compare each score with both human ratings, compare the two scores while holding the human rating fixed, and measure whether the sharing score improves held-out prediction beyond the credibility score.

2.1 Evaluation Setting

Human participants and eight LLM evaluators read the same articles and rate their perceived credibility and willingness to share using the same questions and anchored 1–7 scales.

We use English news-style articles as a common format across five public-issue topics: Environment, Health, Innovation, Livelihood, and Safety. These topics cover broadly consequential public issues and reflect domains that recur in fact-checking practice. Within each topic, we select 20 false or misleading claims documented by Reuters Fact Check, producing 100 anchor claims in total. This grounds the task in claims documented as circulating in practice rather than arbitrary model-invented premises. We pair each anchor claim with a concrete communication goal describing how the claim could be used to attract attention or shape opinion. Fixing the claim and goal while leaving their rhetorical realization open defines one disinformation scenario.

Gemini 3 Pro Preview, Qwen Plus, and Qwen3-32B each generate one article for every scenario using the same task template, format requirements, and length constraints. The prompt frames generation as a propagation task evaluated by human readers, states that readers will score the article for credibility and willingness to share, and asks the generator to advance the communication goal while substantively using the anchor claim. Generation uses a single turn, temperature 0, and no external retrieval. The three generators provide alternative textual realizations of each fixed claim–goal pair, and the robustness analysis reports results separately for each generator. After excluding generation refusals, one scenario that did not satisfy the study's safety criteria, and one article without complete evaluator coverage, the final dataset contains 290 articles from 99 scenarios. Each represented scenario contains two or three versions of the same underlying claim and communication goal.

2.2 Human Ratings

We recruited participants through Prolific in two waves. Participants provided informed consent, received compensation, and rated up to five distinct articles. Before rating each article, they were asked to imagine encountering it during everyday online reading, read naturally, answer from their first impression, refrain from looking it up online, and avoid guessing the researchers' intent. We retained ratings that passed the study's quality checks. The final dataset contains 1,256 participant–article observations from 317 participants across all 290 articles. Each article has between 1 and 19 participant observations, with a median of four and a mean of 4.33.

Participants answered two questions on anchored 1–7 scales, where 1 denotes the lowest response, 4 a neutral response, and 7 the highest response:

- **Credibility**: "Overall, do the main claims in this text feel believable and realistic?"

- **Willingness to share**: "If you saw this text in daily life, would you personally want to forward or share it with others?"

The primary joint analysis uses the individual participant ratings. Article-level averages are used only for descriptive comparisons and held-out prediction.

2.3 LLM Ratings

We evaluate the eight hosted LLM versions listed in Table 1. Claude and Gemini are accessed through OpenRouter, and the GPT models through the OpenAI API. Ratings were collected from March 19 to 25, 2026.

The six GPT releases represent successive models within one provider family. Claude and Gemini add independently developed provider families. The evaluator set therefore covers both within-family model variation and cross-provider variation.

Each evaluator receives one article and returns credibility and willingness-to-share ratings in a single response. Calls use one user message with no system message, tools, retrieval, or explicit reasoning setting. The prompt asks the evaluator to respond from its first impression as an ordinary reader without fact-checking. The reader framing, focal questions, and anchored 1–7 scales match the human task, with credibility presented first. Responses are requested as a JSON object containing two integer scores. Temperature is set to 0 for Claude, Gemini, GPT-4.1, and GPT-4o. The GPT-5-series calls use the API default because those endpoints did not accept a custom temperature. We do not set top-p, a seed, or a completion-token limit. The client timeout is 120 seconds. We retain one successful completion per article, with transient request or parsing failures retried up to four total attempts. The 290-article set has complete ratings from all eight evaluators.

2.4 Analysis

Does Each Score Correlate More Strongly with the Matching Human Rating? Let H_C and H_S denote human credibility and sharing ratings, and let M_C and M_S denote the corresponding evaluator scores. We first average the human ratings within each article and calculate the four article-level Spearman correlations between the two evaluator scores and the two human responses. For evaluator m, we define two correlation differences:

$$P_C^{(m)} = \rho(M_C, H_C) - \rho(M_C, H_S), \tag{1}$$

$$P_S^{(m)} = \rho(M_S, H_S) - \rho(M_S, H_C). \tag{2}$$

A positive value means that an evaluator score correlates more strongly with the matching human rating than with the other human rating. This statistic compares two human ratings for one evaluator score. It does not compare the credibility and sharing scores for one human rating. We obtain 95% intervals for article-level quantities by resampling the 99 scenarios 2,000 times. For the participant-level correlation between the two human responses, we resample participants 2,000 times and keep all ratings from each sampled participant together.

Table 1: Evaluator versions used for the primary ratings and question-format checks.

Evaluator	API model ID
Primary ratings	
Claude Sonnet 4.5	`anthropic/claude-sonnet-4.5`
Gemini 3 Pro Preview	`google/gemini-3-pro-preview`
GPT-4.1	`gpt-4.1-2025-04-14`
GPT-4o	`gpt-4o`
GPT-5	`gpt-5-2025-08-07`
GPT-5.1	`gpt-5.1-2025-11-13`
GPT-5.2	`gpt-5.2-2025-12-11`
GPT-5.4	`gpt-5.4-2026-03-05`
Additional question formats	
Claude Sonnet 4.5	`anthropic/claude-4.5-sonnet-20250929`
GPT-5.4	`openai/gpt-5.4`
Gemini 3.1 Pro Preview	`google/gemini-3.1-pro-preview`

Credibility and sharing are related human responses. A sharing score can correlate positively with human sharing simply because both are associated with credibility. The correlation differences show whether each score is more closely related to the human rating named in its question.

Which Score Better Predicts Each Human Rating? Our primary analysis uses all individual human ratings and places both human responses in one standardized linear mixed model. For each evaluator, the fixed relationships are

$$H_C = \alpha_C + \beta_{CC} M_C + \beta_{CS} M_S + \cdots, \tag{3}$$

$$H_S = \alpha_S + \beta_{SC} M_C + \beta_{SS} M_S + \cdots. \tag{4}$$

The two evaluator scores and the two human responses are standardized before fitting. For human credibility, $A_C = \beta_{CC} - \beta_{CS}$ is the credibility-score coefficient minus the sharing-score coefficient. For human sharing, $A_S = \beta_{SS} - \beta_{SC}$ is the sharing-score coefficient minus the credibility-score coefficient. A positive value means that the score produced by the matching question has the stronger relationship with the human rating. Random intercepts account for differences in how participants use the scale, differences shared by articles from the same scenario, article-level differences, and the paired credibility and sharing responses from the same participant–article observation. The full-sample model also includes outcome-specific indicators for the recruitment wave. We fit the Gaussian mixed models by restricted maximum likelihood and calculate Wald 95% intervals from the fixed-effect covariance matrix.

The joint model retains individual human ratings instead of treating article averages with unequal numbers of observations as equally precise labels. It also enters both evaluator scores together, so the coefficient differences compare their relationships with the same human rating after accounting for information shared by the two scores.

Does the Sharing Score Add Incremental Predictive Value? For each evaluator and human response, we compare three linear prediction models using the credibility

score, the sharing score, or both scores. We use ten-fold cross-validation grouped by scenario. All versions of the same underlying claim remain in one fold, and every article serves as held-out data exactly once. For human sharing, the first difference is the sharing-only RMSE minus the credibility-only RMSE, so a positive value means that the credibility score predicts better. The second difference is the credibility-only RMSE minus the two-score RMSE, so a positive value indicates that adding the sharing score improves prediction. Confidence intervals use 5,000 scenario-bootstrap samples of the held-out errors. We report held-out R^2 descriptively. As a check that the procedure can recover a predictable human rating, we also use credibility scores to predict human credibility.

The held-out comparison measures the relative error of the two single-score models and whether combining them improves prediction on unseen content. Grouping folds by scenario prevents versions of the same claim from appearing in both training and test data. RMSE is prediction error on the original 1–7 human article-mean scale, where lower values are better. We calculate pooled out-of-fold R^2 as $1 - \sum_i (y_i - \hat{y}_i)^2 / \sum_i (y_i - \bar{y})^2$, where $\bar{y}$ is the overall mean of the observed article means. A value of zero matches that constant baseline in pooled squared error, and a negative value is worse.

Uncertainty and Robustness. We repeat the joint-model and held-out analyses using only the second recruitment wave. We also fit a joint fixed-effects model with standard errors clustered by participant and scenario. Finally, we stratify the articles by their generating model and repeat the correlation, clustered coefficient, and held-out prediction analyses within each generator subset. All reported intervals are two-sided 95% intervals.

Because the original prompt requests both scores in one response and presents credibility first, we also test whether this prompt format creates the result. Claude Sonnet 4.5 rates all 290 articles under four prompt versions: the original credibility-then-sharing order, the reversed order, credibility only, and sharing only. GPT-5.4 and Gemini 3.1 Pro Preview rate a deterministic 99-article subset under the same prompt versions. The subset contains one article from every scenario and is balanced across the three article generators, with 33 articles from each. Wording, scales, and API parameters remain unchanged except for question order and separation. Calls were made through OpenRouter on July 12–13, 2026. Claude, GPT-5.4, and Gemini were routed through Amazon Bedrock, OpenAI, and Google AI Studio, respectively. Their returned model IDs are listed in Table 1. We compare how strongly the separately obtained sharing and credibility scores correlate with human sharing, estimate the interval of their difference with 5,000 scenario-bootstrap samples, and repeat the individual-rating and scenario-held-out analyses for these scores.

3 Results

Across all eight evaluators, asking about credibility produced a score at least as strongly associated with human sharing as asking about sharing itself. The sharing question therefore provided no predictive advantage for human sharing. In joint models of individual human ratings, the credibility score was more strongly related to human credibility than the sharing score for all eight evaluators, whereas the sharing score was never reliably more strongly related to human sharing than the credibility score. When predicting unseen scenarios, sharing-only models did not detectably outperform credibility-only models, and adding the sharing score produced no detectable improvement. These findings persisted across recruitment waves, statistical analyses, and article generators. They also remained when the questions were asked separately or in reversed order. Sharing scores nevertheless remained positively correlated with human sharing, showing why comparisons limited to correspondingly named scores and responses can miss this outcome misalignment.

3.1 Correlations with the Matching and Other Human Ratings

Human credibility and willingness to share are substantially related. Their Spearman correlation is .639 at the participant–article level, with a participant-bootstrap 95% interval of $[.585, .689]$, and .620 between the article means. This relationship makes the conventional same-outcome comparison appear favorable. Across the eight evaluators, the article-level correlation between model and human credibility ranges from .369 to .450, while the correlation between model and human sharing ranges from .145 to .226. All 16 same-outcome correlations are positive, with scenario-bootstrap intervals above zero.

Table 2 compares each evaluator score with both human responses. For credibility scores, the correlation difference P_C ranges from .161 to .223 and is positive for all eight evaluators, with every interval above zero. For sharing scores, P_S ranges from $-.124$ to .019. Only GPT-5.1 has a positive point estimate, and no P_S interval lies above zero. Each statistic fixes one evaluator score and compares its correlations with the two human ratings. The same table also fixes human sharing and compares the two evaluator scores. For every evaluator, the credibility score is at least as strongly correlated with human sharing as the sharing score is. Positive correlation between a sharing score and human sharing can therefore coexist with an equally strong or stronger correlation from the credibility score.

3.2 Comparing the Two Scores in Individual Human Ratings

The joint models evaluate both scores simultaneously while retaining variation across individual human responses. Figure 1A and Table 3 show that the coefficient difference A_C ranges from .224 to .466, and all eight 95% intervals lie above zero. Thus, the credibility score is more strongly related to human credibility than the sharing score for every evaluator. For human sharing, A_S ranges from $-.088$ to .050, and none of the eight intervals lies above zero. The sharing score is therefore never reliably more strongly related to human sharing than the credibility score.

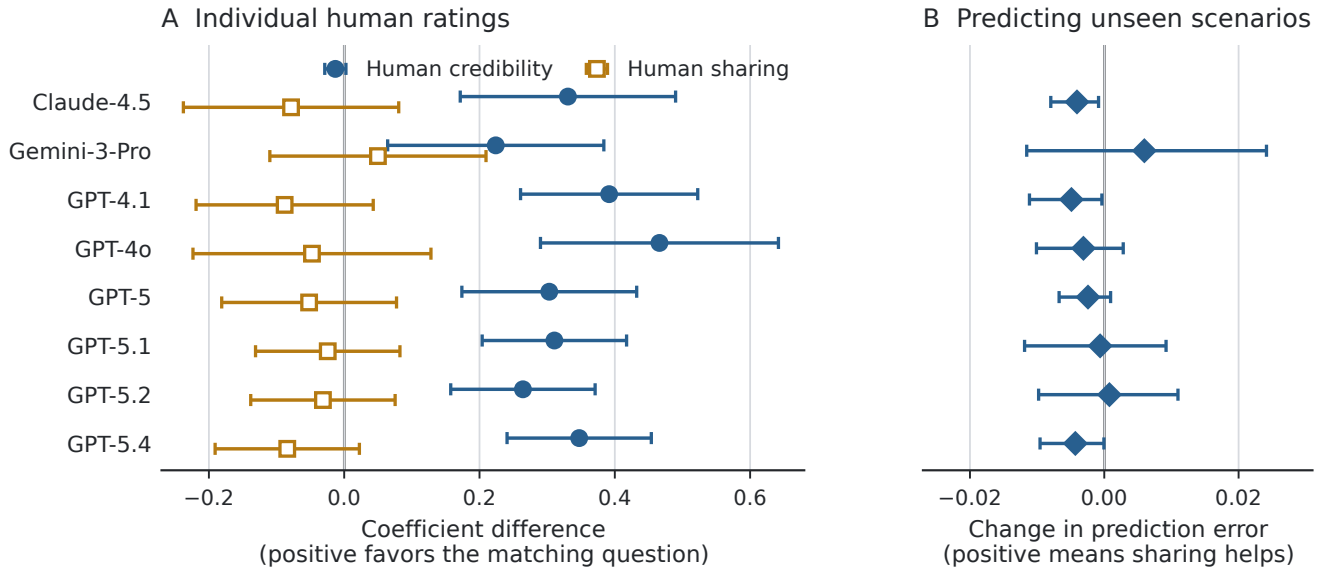

Figure 1: Comparing evaluator scores for human credibility and sharing. Panel A reports the standardized coefficient difference between the score from the matching question and the other score for each human rating. Positive values favor the matching question. Panel B reports held-out $\mathrm{RMSE}_{C\text{-only}} - \mathrm{RMSE}_{C+S}$ for human sharing. Positive values indicate improvement from adding the sharing score. Points show estimates and bars show 95% intervals.

Table 2: Article-level Spearman correlations between each evaluator score and the two human responses. Bold marks the larger correlation within each evaluator score.

	Credibility score		Sharing score	
Evaluator	H_C	H_S	H_C	H_S
Claude-4.5	**.399**	.238	**.283**	.159
Gemini-3-Pro	**.369**	.146	**.261**	.145
GPT-4.1	**.439**	.239	**.283**	.195
GPT-4o	**.425**	.246	**.309**	.226
GPT-5	**.425**	.241	**.301**	.183
GPT-5.1	**.433**	.249	.193	**.212**
GPT-5.2	**.450**	.250	**.238**	.219
GPT-5.4	**.439**	.254	**.217**	.198

Table 3: Standardized coefficient differences from joint models of individual human ratings. A_C compares the credibility and sharing scores for human credibility. A_S compares the sharing and credibility scores for human sharing. Positive values favor the score from the matching question.

	Credibility		Sharing	
Evaluator	A_C	p	A_S	p
Claude-4.5	.331	$< .001$	$-.079$	.333
Gemini-3-Pro	.224	.006	.050	.541
GPT-4.1	.392	$< .001$	$-.088$	.188
GPT-4o	.466	$< .001$	$-.048$	.595
GPT-5	.303	$< .001$	$-.052$	.432
GPT-5.1	.311	$< .001$	$-.024$	.655
GPT-5.2	.264	$< .001$	$-.031$	.564
GPT-5.4	.347	$< .001$	$-.084$	.122

3.3 Sharing Scores Do Not Improve Prediction on Unseen Scenarios

For unseen claims and communication goals, we first compare credibility-only and sharing-only prediction. Table 4 shows that credibility-only prediction has lower point-estimate error for seven of eight evaluators, while sharing-only prediction has slightly lower error for Gemini. The difference $\Delta_d = \mathrm{RMSE}(S) - \mathrm{RMSE}(C)$ ranges from $-.0082$ to .0197, and all eight 95% intervals include zero. Thus, asking about sharing produces no detectable advantage over asking about credibility.

Figure 1B and Table 4 show whether the sharing score improves prediction after the credibility score is known. Adding the sharing score produces a detectable reduction in human-sharing prediction error for none of the eight evaluators. The RMSE difference $\Delta_a = \mathrm{RMSE}(C) - \mathrm{RMSE}(C{+}S)$ ranges from $-.0049$ to .0060, and every 95% interval either includes zero or lies below it. Positive values favor adding the sharing score. The observed values are close to zero or negative, indicating no improvement and in several cases slightly greater prediction error.

Human-sharing prediction is weak for both single-score models. The largest pooled out-of-fold R^2 is .048, so the best model reduces squared prediction error by 4.8% relative to predicting the overall mean of the observed article means. Gemini's negative credibility-only R^2 indicates performance slightly worse than that constant baseline.

For human credibility, the credibility-only models reduce

pooled squared prediction error by 11.3% to 18.0% relative to the overall-mean baseline. The evaluation setting and validation procedure can therefore detect predictive signal for a human response, while the score obtained by asking about sharing supplies no detectable improvement beyond the credibility score for predicting human sharing.

3.4 Robustness Across Question Formats, Human Samples, and Analyses

The main result could be caused by the original prompt format because credibility appears before sharing and both scores are requested in one response. We therefore repeated the evaluation with sharing presented first and with each question asked in a separate call. Neither change made the sharing score more predictive of human sharing than the credibility score.

When the questions are asked separately, the sharing-score minus credibility-score correlation differences for human sharing are $-.051$ $[-.154, .053]$ for Claude, $-.039$ $[-.211, .130]$ for GPT-5.4, and $-.037$ $[-.205, .123]$ for Gemini. Presenting sharing first produces corresponding differences of $-.089$ $[-.161, -.023]$, $.007$ $[-.144, .145]$, and $-.032$ $[-.176, .107]$. The individual-rating analyses yield the same conclusion. When the questions are asked separately, the coefficient differences favoring the sharing score are $-.070$ $[-.174, .033]$ for Claude, $-.082$ $[-.259, .095]$ for GPT-5.4, and $-.048$ $[-.189, .094]$ for Gemini. In scenario-held-out prediction, sharing-only minus credibility-only RMSE is $.011$ $[-.015, .039]$, $.019$ $[-.018, .056]$, and $.013$ $[-.020, .047]$, respectively. None favors the score obtained by asking about sharing. The credibility checks remain positive for all three models, with RMSE gains of $.074$ $[.012, .143]$, $.091$ $[.010, .171]$, and $.084$ $[.010, .166]$.

The findings also remain when the human sample and statistical analysis change. Table 5 summarizes these checks. When the joint models use only the second recruitment wave, the credibility score is more strongly related to human credibility than the sharing score for seven of eight evaluators, while the sharing score is more strongly related to human sharing for none. When standard errors are clustered by participant and scenario, the corresponding counts are eight of eight and zero of eight in both the full sample and the second-wave sample. The credibility coefficient difference also exceeds the sharing coefficient difference for seven of eight evaluators in each analysis. Gemini is the exception because its interval for the difference includes zero. Repeating held-out prediction with only the second wave again produces no detectable improvement from the sharing score for any evaluator. The maximum human-sharing R^2 is .029, while the credibility check ranges from .068 to .125.

The result also persists across the three models that generated the articles. Credibility-score correlation differences are positive in all 24 generator-by-evaluator combinations. Scenario-bootstrap intervals are above zero for 8 of 8 evaluators on Gemini-generated articles, 1 of 8 on Qwen-Plus articles, and 2 of 8 on Qwen3-32B articles. In the participant-by-scenario clustered analysis, the credibility score is more strongly related to human credibility than the sharing score for 8 of 8, 3 of 8, and 7 of 8 evaluators, respectively. No generator subset shows that the sharing score is reliably more strongly related to human sharing than the credibility score. Adding the sharing score yields no detectable held-out improvement in any of the 24 combinations, while credibility R^2 values range from .017 to .302. The pattern is therefore not attributable to articles from only one generator, although the strength of the credibility result varies across generator subsets.

Table 4: Scenario-held-out prediction. Panel A compares credibility-only and sharing-only prediction for human sharing. Panel B tests whether adding sharing improves on credibility-only prediction and reports the credibility positive control. Positive Δ_d favors credibility-only; positive Δ_a favors adding sharing.

A. Direct comparison for human sharing

Evaluator	$R^2_S(C)$	$R^2_S(S)$	Δ_d [95% CI]
Claude-4.5	.036	.004	.0170 $[-.0058, .0427]$
Gemini-3-Pro	-.002	.014	$-.0082$ $[-.0251, .0080]$
GPT-4.1	.030	.000	.0156 $[-.0033, .0424]$
GPT-4o	.040	.029	.0058 $[-.0093, .0236]$
GPT-5	.033	-.005	.0197 $[-.0098, .0599]$
GPT-5.1	.048	.013	.0182 $[-.0089, .0547]$
GPT-5.2	.035	.025	.0050 $[-.0172, .0274]$
GPT-5.4	.033	.000	.0170 $[-.0093, .0491]$

B. Incremental prediction and positive control

Evaluator	$R^2_S(C+S)$	Δ_a [95% CI]	$R^2_C(C)$
Claude-4.5	.029	$-.0041$ $[-.0080, -.0009]$	.141
Gemini-3-Pro	.009	.0060 $[-.0115, .0242]$	.113
GPT-4.1	.021	$-.0049$ $[-.0111, -.0004]$	.167
GPT-4o	.034	$-.0031$ $[-.0101, .0028]$	.165
GPT-5	.028	$-.0024$ $[-.0067, .0009]$	.155
GPT-5.1	.046	$-.0006$ $[-.0119, .0092]$	.180
GPT-5.2	.036	.0008 $[-.0098, .0110]$	.177
GPT-5.4	.024	$-.0043$ $[-.0096, -.0001]$	.162

Table 5: Number of evaluators for which the 95% interval favors the score from the matching question. The final column compares the coefficient differences for credibility and sharing.

Analysis	Credibility	Sharing	Cred. > Sharing
Mixed model, full	8/8	0/8	7/8
Mixed model, wave 2	7/8	0/8	7/8
Two-way cluster, full	8/8	0/8	7/8
Two-way cluster, wave 2	8/8	0/8	7/8

4 Discussion

In LLM-based misinformation risk evaluation, semantic alignment between a question and its intended response does not guarantee predictive alignment between the resulting score and that response. In our experiments, the question explicitly asked about willingness to share, yet the resulting score was no more closely related to human sharing than the score obtained from the credibility question across all eight

evaluators. The same pattern persisted when the questions were asked separately and when sharing appeared first. Thus, changing what an evaluator is asked to score can change the label and numerical output without changing which human response the score predicts best.

LLM evaluators are often asked to assign separate scores to related properties of the same output, including helpfulness, harmlessness, persuasiveness, and trustworthiness (Zheng et al. 2023; Li et al. 2025). LLMs are also used to estimate likely human or user responses (Dong, Hu, and Collier 2024; Choi, Young, and Ferrara 2026). Each question can produce a distinct numerical score while the scores still recover largely overlapping human signals. When they are used as labels, rewards, ranking criteria, or optimization objectives, multiple scores may repeatedly emphasize the same predictable property while providing little control over the other intended dimensions. The number of explicitly requested dimensions can therefore exceed the number of dimensions that the evaluator distinguishes in practice.

The broader scientific question concerns the relationship between semantic structure and predictive structure in LLM-based risk evaluation and related human-response estimation tasks. Concepts are organized by meaning: credibility and willingness to share are related but substantively distinct, and each question is semantically closest to its correspondingly named human response. Scores produced by those questions induce a second structure, defined by which human responses they actually predict and how strongly. A simple correspondence between the two structures would make the sharing question the best predictor of human sharing. Our results provide a clear counterexample: the semantic match remains exact, while the predictive ordering is reversed or indistinguishable across evaluators. The central research problem is therefore to determine when semantic relations among concepts are preserved in the predictive relations of LLM scores, when they are reordered, and what governs the mapping between the two structures. This mapping can be studied across sets of related concepts and human outcomes, revealing whether semantic proximity reliably identifies the most predictive model judgment.

5 Limitations

This study focuses on 290 English news-style deceptive articles spanning five public-issue topics, with human ratings collected from a Prolific participant sample. The study measures stated willingness to share rather than observed sharing behavior. The applicability of the findings to other content forms, languages, populations, and behavioral outcomes remains to be established.

All eight evaluators are tested under one matched, content-only evaluation procedure. The separate-question and reversed-order checks cover three provider families, with all 290 articles for Claude and a scenario-balanced 99-article subset for GPT-5.4 and Gemini 3.1 Pro Preview. The Gemini check uses the available successor to the Gemini 3 Pro Preview model in the main study, and the other five evaluators are not tested under these additional formats. Providing reader profiles, audiences, incentives, sharing contexts, or a different question design may change the information carried by the sharing score. The absence of a detectable held-out improvement here does not establish that sharing scores contain no independent information under every setting.

6 Ethics Statement

The study received institutional ethics approval. All participants took part voluntarily after providing informed consent and received compensation through Prolific. We did not collect names or other direct identifiers. Analysis data use study-specific identifiers and exclude platform identifiers, timestamps, demographic variables, and free-text responses.

The research materials contain synthetic false or misleading claims and were handled as potentially harmful content. Materials released for reproducibility will be clearly identified as synthetic research artifacts rather than factual news.

Generative AI tools were used to assist with language editing and manuscript preparation. All claims, analyses, citations, and final text were verified by the authors, who take full responsibility for the manuscript.

7 Data and Code Availability

Upon publication, we will release the 290 articles, de-identified participant–article ratings, frozen evaluator scores, evaluation prompts, and code required to reproduce the reported analyses, tables, and figures.

References

Aher, G. V.; Arriaga, R. I.; and Kalai, A. T. 2023. Using Large Language Models to Simulate Multiple Humans and Replicate Human Subject Studies. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, 337–371. PMLR.

Anthis, J. R.; Liu, R.; Richardson, S. M.; Kozlowski, A. C.; Koch, B.; Brynjolfsson, E.; Evans, J.; and Bernstein, M. S. 2025. Position: LLM Social Simulations Are a Promising Research Method. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, 81005–81034. PMLR.

Argyle, L. P.; Busby, E. C.; Fulda, N.; Gubler, J. R.; Rytting, C.; and Wingate, D. 2023. Out of One, Many: Using Language Models to Simulate Human Samples. *Political Analysis*, 31(3): 337–351.

Ashokkumar, A.; Hewitt, L.; Ghezae, I.; and Willer, R. 2026. Large Language Models Can Predict the Results of Social Science Experiments. *Nature*. Published online 8 July 2026. doi:10.1038/s41586-026-10742-x.

Atari, M.; Xue, M. J.; Park, P. S.; Blasi, D. E.; and Henrich, J. 2023. Which Humans? *PsyArXiv preprint*.

Bai, H.; Voelkel, J. G.; Muldowney, S.; Eichstaedt, J. C.; and Willer, R. 2025. LLM-Generated Messages Can Persuade Humans on Policy Issues. *Nature Communications*, 16(1): 6037.

Bail, C. A. 2024. Can Generative AI improve social science? *Proceedings of the National Academy of Sciences*, 121(21): e2314021121.

Bavaresco, A.; Bernardi, R.; Bertolazzi, L.; Elliott, D.; Fernandez, R.; Gatt, A.; Ghaleb, E.; Giulianelli, M.; Hanna, M.;

Koller, A.; Martins, A. F. T.; Mondorf, P.; Neplenbroek, V.; Pezzelle, S.; Plank, B.; Schlangen, D.; Suglia, A.; Surikuchi, A. K.; Takmaz, E.; and Testoni, A. 2025. LLMs Instead of Human Judges? A Large Scale Empirical Study Across 20 NLP Evaluation Tasks. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 238–255.

Bisbee, J.; Clinton, J. D.; Dorff, C.; Kenkel, B.; and Larson, J. M. 2024. Synthetic Replacements for Human Survey Data? The Perils of Large Language Models. *Political Analysis*, 32(4): 401–416.

Cao, Y.; Liu, H.; Arora, A.; Augenstein, I.; Röttger, P.; and Hershcovich, D. 2025. Specializing Large Language Models to Simulate Survey Response Distributions for Global Populations. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 3141–3154. Association for Computational Linguistics.

Chen, G. H.; Chen, S.; Liu, Z.; Jiang, F.; and Wang, B. 2024. Humans or LLMs as the Judge? A Study on Judgement Bias. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 8301–8327. Association for Computational Linguistics.

Chiang, C.-H.; and Lee, H.-y. 2023. Can Large Language Models Be an Alternative to Human Evaluations? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 15607–15631. Association for Computational Linguistics.

Choi, E. C.; Young, L. E.; and Ferrara, E. 2026. Overstating Attitudes, Ignoring Networks: LLM Biases in Simulating Misinformation Susceptibility. *Proceedings of the International AAAI Conference on Web and Social Media*, 20(1): 520–547.

Cui, Z.; Li, N.; and Zhou, H. 2025. A Large-Scale Replication of Scenario-Based Experiments in Psychology and Management Using Large Language Models. *Nature Computational Science*, 5: 627–634.

Dillion, D.; Tandon, N.; Gu, Y.; and Gray, K. 2023. Can AI Language Models Replace Human Participants? *Trends in Cognitive Sciences*, 27(7): 597–600.

Dong, Y. R.; Hu, T.; and Collier, N. 2024. Can LLM Be a Personalized Judge? In *Findings of the Association for Computational Linguistics: EMNLP 2024*, 10126–10141.

Fu, J.; Ng, S.-K.; Jiang, Z.; and Liu, P. 2024. GPTScore: Evaluate as You Desire. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 6556–6576. Association for Computational Linguistics.

Globig, L. K.; and Sharot, T. 2024. Considering Information-Sharing Motives to Reduce Misinformation. *Current Opinion in Psychology*, 59: 101852.

Hoq, A.; and Weninger, T. 2026. Evaluating LLMs as Human Surrogates in Controlled Experiments. *arXiv preprint arXiv:2604.15329*.

Hu, T.; and Collier, N. 2024. Quantifying the Persona Effect in LLM Simulations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 10289–10307. Association for Computational Linguistics.

Huidrom, R.; and Belz, A. 2025. Ask Me Like I'm Human: LLM-based Evaluation with For-Human Instructions Correlates Better with Human Evaluations than Human Judges. In *Proceedings of the 4th Table Representation Learning Workshop*, 98–108.

Hullman, J.; Broska, D.; Sun, H.; and Shaw, A. 2026. This Human Study Did Not Involve Human Subjects: Validating LLM Simulations as Behavioral Evidence. *arXiv preprint arXiv:2602.15785*.

Ivey, J.; Kumar, S.; Liu, J.; Shen, H.; Rakshit, S.; Raju, R.; Zhang, H.; Ananthasubramaniam, A.; Kim, J.; Yi, B.; Wright, D.; Israeli, A.; Møller, A. G.; Zhang, L.; and Jurgens, D. 2026. Real or Robotic? Assessing Whether LLMs Accurately Simulate Qualities of Human Responses in Human-LLM Dialogue. In *Findings of the Association for Computational Linguistics: ACL 2026*, 41395–41432. Association for Computational Linguistics.

Kreps, S.; McCain, R. M.; and Brundage, M. 2022. All the News That's Fit to Fabricate: AI-Generated Text as a Tool of Media Misinformation. *Journal of Experimental Political Science*, 9(1): 104–117.

Krsteski, S.; Russo, G.; Chang, S.; West, R.; and Gligorić, K. 2026. Valid Survey Simulations with Limited Human Data: The Roles of Prompting, Fine-Tuning, and Rectification. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 10887–10906. Association for Computational Linguistics.

Krumdick, M.; Lovering, C.; Reddy, V.; Ebner, S.; and Tanner, C. 2025. No Free Labels: Limitations of LLM-as-a-Judge Without Human Grounding. *arXiv preprint arXiv:2503.05061*.

Li, D.; Jiang, B.; Huang, L.; Beigi, A.; Zhao, C.; Tan, Z.; Bhattacharjee, A.; Jiang, Y.; Chen, C.; Wu, T.; Shu, K.; Cheng, L.; and Liu, H. 2025. From Generation to Judgment: Opportunities and Challenges of LLM-as-a-Judge. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 2757–2791. Association for Computational Linguistics.

Libovický, J. 2026. On the Credibility of Evaluating LLMs Using Survey Questions. In *Proceedings of the First Workshop on Multilingual Multicultural Evaluation*, 23–34. Association for Computational Linguistics.

Licht, H.; Sarkar, R.; Wu, P. Y.; Goel, P.; Stoehr, N.; Ash, E.; and Hoyle, A. M. 2025. Measuring Scalar Constructs in Social Science with LLMs. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 32144–32171. Association for Computational Linguistics.

Lin, C.; Yuan, W.; Jiang, Z.; Huang, B.; Zhang, R.; Ge, J.; Xu, Y.; and Yu, J. 2026. AlignSurvey: A Comprehensive Benchmark for Human Preferences Alignment in Social Surveys. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(45): 38908–38916.

Lin, Z. 2024. Large Language Models as Linguistic Simulators and Cognitive Models in Human Research. *arXiv preprint arXiv:2402.04470*.

Liu, Y.; Iter, D.; Xu, Y.; Wang, S.; Xu, R.; and Zhu, C. 2023. G-Eval: NLG Evaluation Using GPT-4 With Better Human

Alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2511–2522.
Matz, S. C.; Teeny, J. D.; Vaid, S. S.; Peters, H.; Harari, G. M.; and Cerf, M. 2024. The Potential of Generative AI for Personalized Persuasion at Scale. *Scientific Reports*, 14(1): 4692.
Ou, M.; and Ho, S. S. 2024. Factors Associated with Information Credibility Perceptions: A Meta-Analysis. *Journalism & Mass Communication Quarterly*, 101(2): 346–372.
Pan, Q.; Ashktorab, Z.; Desmond, M.; Cooper, M. S.; Johnson, J.; Nair, R.; Daly, E.; and Geyer, W. 2024. Human-Centered Design Recommendations for LLM-as-a-Judge. In *Proceedings of the 1st Human-Centered Large Language Modeling Workshop*, 16–29.
Pangakis, N.; and Wolken, S. 2025. Keeping Humans in the Loop: Human-Centered Automated Annotation with Generative AI. *Proceedings of the International AAAI Conference on Web and Social Media*, 19(1): 1471–1492.
Park, J. S.; O'Brien, J. C.; Cai, C. J.; Morris, M. R.; Liang, P.; and Bernstein, M. S. 2023. Generative Agents: Interactive Simulacra of Human Behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, 1–22. Association for Computing Machinery.
Pennycook, G.; Epstein, Z.; Mosleh, M.; Arechar, A. A.; Eckles, D.; and Rand, D. G. 2021. Shifting Attention to Accuracy Can Reduce Misinformation Online. *Nature*, 592: 590–595.
Petrov, N. B.; Serapio-García, G.; and Rentfrow, J. 2024. Limited Ability of LLMs to Simulate Human Psychological Behaviours: A Psychometric Analysis. *arXiv preprint arXiv:2405.07248*.
Phillips, S. C.; Wang, S. Y. N.; Carley, K. M.; Rand, D. G.; and Pennycook, G. 2025. Emotional Language Reduces Belief in False Claims. *Judgment and Decision Making*, 20: e43.
Rathje, S.; Roozenbeek, J.; Van Bavel, J. J.; and van der Linden, S. 2023. Accuracy and Social Motivations Shape Judgements of (Mis)Information. *Nature Human Behaviour*, 7: 892–903.
Salvi, F.; Horta Ribeiro, M.; Gallotti, R.; and West, R. 2025. On the Conversational Persuasiveness of GPT-4. *Nature Human Behaviour*, 9(8): 1645–1653.
Santurkar, S.; Durmus, E.; Ladhak, F.; Lee, C.; Liang, P.; and Hashimoto, T. 2023. Whose Opinions Do Language Models Reflect? In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, 29971–30004. PMLR.
Shi, L.; Ma, C.; Liang, W.; Diao, X.; Ma, W.; and Vosoughi, S. 2025. Judging the Judges: A Systematic Study of Position Bias in LLM-as-a-Judge. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, 292–314. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics.
Spitale, G.; Biller-Andorno, N.; and Germani, F. 2023. AI Model GPT-3 (Dis)Informs Us Better than Humans. *Science Advances*, 9(26): eadh1850.
Stefkovics, Á.; and Gere, D. 2026. Beliefs and sharing intentions of human- and AI-generated fake news: Evidence from 27 European countries. *PNAS Nexus*, 5(3): pgag032.
Sun, Y.; and Xie, J. 2024. Who Shares Misinformation on Social Media? A Meta-Analysis of Individual Traits Related to Misinformation Sharing. *Computers in Human Behavior*, 158: 108271.
Thakur, A. S.; Choudhary, K.; Ramayapally, V. S.; Vaidyanathan, S.; and Hupkes, D. 2025. Judging the Judges: Evaluating Alignment and Vulnerabilities in LLMs-as-Judges. In *Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics*, 404–430. Association for Computational Linguistics.
Tjuatja, L.; Chen, V.; Wu, T.; Talwalkwar, A.; and Neubig, G. 2024. Do LLMs Exhibit Human-like Response Biases? A Case Study in Survey Design. *Transactions of the Association for Computational Linguistics*, 12: 1011–1026.
Traberg, C. S.; Harjani, T.; Roozenbeek, J.; and van der Linden, S. 2024. The Persuasive Effects of Social Cues and Source Effects on Misinformation Susceptibility. *Scientific Reports*, 14(1): 4205.
Wang, J.; Zhao, Z.; Ni, T.; and Wei, Z. 2025. SocioBench: Modeling Human Behavior in Sociological Surveys with Large Language Models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 26257–26289. Association for Computational Linguistics.
Zheng, L.; Chiang, W.-L.; Sheng, Y.; Zhuang, S.; Wu, Z.; Zhuang, Y.; Lin, Z.; Li, Z.; Li, D.; Xing, E. P.; Zhang, H.; Gonzalez, J. E.; and Stoica, I. 2023. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems*, volume 36, 46595–46623.
Zhou, H.; Wan, X.; Liu, Y.; Collier, N.; Vulić, I.; and Korhonen, A. 2024. Fairer Preferences Elicit Improved Human-Aligned Large Language Model Judgments. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 1241–1252. Association for Computational Linguistics.