PNAS RESEARCH ARTICLE | PSYCHOLOGICAL AND COGNITIVE SCIENCES
APPLIED MATHEMATICS OPEN ACCESS

The simulation of judgment in LLMs

Edoardo Loru[a], Jacopo Nudo[b], Niccolò Di Marco[c], Alessandro Santirocchi[d], Roberto Atzeni[d], Matteo Cinelli[b], Vincenzo Cestari[d], Clelia Rossi-Arnaud[d], and Walter Quattrociocchi[b,1]

Affiliations are included on p. 10.

Edited by Susan Fiske, Princeton University, Princeton, NJ; received July 10, 2025; accepted September 18, 2025

Large Language Models (LLMs) are increasingly embedded in evaluative processes, from information filtering to assessing and addressing knowledge gaps through explanation and credibility judgments. This raises the need to examine how such evaluations are built, what assumptions they rely on, and how their strategies diverge from those of humans. We benchmark six LLMs against expert ratings—NewsGuard and Media Bias/Fact Check—and against human judgments collected through a controlled experiment. We use news domains purely as a controlled benchmark for evaluative tasks, focusing on the underlying mechanisms rather than on news classification per se. To enable direct comparison, we implement a structured agentic framework in which both models and nonexpert participants follow the same evaluation procedure: selecting criteria, retrieving content, and producing justifications. Despite output alignment, our findings show consistent differences in the observable criteria guiding model evaluations, suggesting that lexical associations and statistical priors could influence evaluations in ways that differ from contextual reasoning. This reliance is associated with systematic effects: political asymmetries and a tendency to confuse linguistic form with epistemic reliability—a dynamic we term epistemia, the illusion of knowledge that emerges when surface plausibility replaces verification. Indeed, delegating judgment to such systems may affect the heuristics underlying evaluative processes, suggesting a shift from normative reasoning toward pattern-based approximation and raising open questions about the role of LLMs in evaluative processes.

Large Language Models | evaluation and judgment | human–LLM comparison | epistemic alignment | epistemia

Significance

Large Language Models (LLMs) are used in evaluative tasks across domains. Yet, what appears as alignment with human or expert judgments may conceal a deeper shift in how "judgment" itself is operationalized. Using news outlets as a controlled benchmark, we compare six LLMs to expert ratings and human evaluations under an identical, structured framework. While models often match expert outputs, our results suggest that they may rely on lexical associations and statistical priors rather than contextual reasoning or normative criteria. We term this divergence epistemia: the illusion of knowledge emerging when surface plausibility replaces verification. Our findings suggest not only performance asymmetries but also a shift in the heuristics underlying evaluative processes, raising fundamental questions about delegating judgment to LLMs.

Large Language Models (LLMs) are increasingly embedded in workflows involving classification, evaluation, recommendation, and decision support (1). This rapid integration of AI technologies is not just reshaping industries but also presenting a fundamental choice between a path of pure automation and one of human complementation, where technology is designed to augment human spread and mitigate widespread unemployment (2). Beyond assistive tools, they may influence institutional decision-making, where automated outputs support high-stakes judgments. Advances like chain-of-thought prompting (3) and semiautonomous agents (4, 5) mark a broader shift: We are no longer just automating tasks, but embedding evaluative functions into sociotechnical systems (6). This delegation carries significant risks, as AI systems trained on biased historical data may replicate and amplify societal inequalities in critical domains, such as the labor market (7).

As these systems scale, the issue is no longer only whether outputs are correct, but how the very notion of judgment is operationalized once decisions are delegated to statistical models. This raises a key question: What heuristics are encoded when decisions are delegated to LLMs, and how are classifications produced, justified, and interpreted? LLMs can produce outputs similar to those of humans in structured tasks (8), but the similarity concerns results, not the process. What appears as alignment at the output level may conceal a deeper epistemic shift, where normative reasoning is replaced by surface-level approximation.

We analyze how LLMs apply the concepts of reliability and bias, normative categories that influence which content is shown, hidden, or ignored. These classifications shape information exposure, platform moderation, and public trust. Understanding how models handle these constructs is necessary to assess their societal and epistemic impact. To investigate this, we provide a benchmark involving six LLMs, two expert rating systems, and a sample of human evaluators. News outlets are classified by reliability and political bias, allowing us to compare both outputs and the procedures behind them. The news domain serves here only as a controlled testbed, enabling us to isolate and

Author contributions: W.Q. designed research; E.L., J.N., and N.D.M. performed data extraction and analysis; A.S. and R.A. designed and conducted the human experiments; M.C., V.C., C.R.-A., and W.Q. supervised the work; and all authors contributed to writing the manuscript and provided critical feedback.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence may be addressed. Email: walter.quattrociocchi@uniroma1.it.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2518443122/-/DCSupplemental.

Published October 13, 2025.

analyze the mechanisms behind automated evaluations rather than focusing on domain-specific accuracy. This helps show how automated evaluations change not just efficiency but also the criteria and assumptions used in decision-making. Indeed, LLMs often use surface-level text patterns instead of reasoning from evidence (9, 10).

In online spaces where information spreads fast and users are split into echo chambers (11–15), judging source credibility as well as their biases is a core problem (16–18). These judgments shape what people believe and how public debate evolves—driving polarization, spreading misinformation, and eroding trust (19–23). Humans tend to rely on basic standards: Is it accurate, independent, and transparent? (24–26) Expert human evaluators—like NewsGuard and Media Bias/Fact Check (MBFC)—apply clear rules to rate thousands of news sites (27–29). Our goal is not to assess whether LLMs can replace human raters but to use these benchmarks to analyze the heuristics guiding their evaluations and how they operationalize concepts such as credibility and bias. Widely used models such as GPT (30), Gemini (31), and Llama (32)—already employed in numerous classification and fact-checking tasks (33–41)—offer a natural testbed for this analysis. Recent studies have explored whether LLMs replicate human heuristics (4, 42, 43) or reflect ideological biases learned during training (44–46), but most focus on output-level metrics such as accuracy or bias. Few examine how these outputs are produced. We address this gap by analyzing how LLMs generate judgments about reliability and political bias, and how their procedures compare to human evaluation. Six LLMs classify 2,286 domains by reliability and political leaning. Their outputs are compared to expert ratings, and we examine the lexical patterns and explanatory cues they provide to infer the underlying heuristics. We do not assume human-like reasoning. Rather, we empirically analyze how models operationalize evaluative tasks, allowing us to test whether their decisions rely on statistical associations shaped by training and prompting. This leads us to ask whether delegating judgment to LLMs preserves its normative and epistemic meaning or whether it transforms it into what we call epistemia—a condition in which the appearance of coherent and authoritative judgment arises from statistical patterning alone, producing the illusion of knowledge when surface plausibility substitutes for evidence-based reasoning.

Our findings show that model outputs often align with expert ratings of reliability and bias, yet systematic asymmetries emerge across the political spectrum. Moreover, LLMs generate consistent linguistic markers when explaining their evaluations. To test these differences, we implement a structured protocol in which LLMs simulate evaluative behavior—selecting criteria from a predefined set, retrieving content, and producing justifications—while human participants follow the same procedure in a controlled setting. The results show that LLMs and humans prioritize different reliability criteria, consistent with a shift from context—dependent, normative reasoning-understood here as the application of explicit quality standards and contextual reasoning rather than implying perfectly rational agents—toward pattern-based approximation.

Overall, our findings indicate that delegating evaluations to models transforms how reliability and bias are assessed, replacing human judgment with statistical approximation.

Results and Discussion

To investigate how LLMs perform in practice, we begin by analyzing their classification of online news outlets along the dimensions of reliability and political bias. Specifically, we evaluate six state-of-the-art models—Deepseek V3, Gemini 1.5 Flash, GPT-4o mini, Llama 3.1 405B, Llama 4 Maverick, and Mistral Large 2—by comparing their outputs to expert human benchmarks from NewsGuard and MBFC. Beyond assessing alignment with expert judgments, we aim to examine the heuristics and decision patterns these models deploy, providing an approach to investigate the underlying processes that guide model evaluations. To support this analysis, we construct a diverse dataset of 7,715 English-language news outlet domains. The sample spans multiple countries and includes outlets with both national and international reach.

We extract a snapshot of each domain's homepage, removing nonessential elements (e.g., scripts, styling) to isolate relevant textual content such as headlines and descriptions. This preprocessing step ensures that all LLMs are evaluated on the same textual input a human assessor would plausibly consider. The final dataset includes 2,286 active domains successfully classified and output in well-formed JavaScript Object Notation (JSON) documents by all models. A detailed breakdown of the data collection and processing is provided in *Materials and Methods*.

Beyond analyzing the final labels produced by the models, we examine the strategies they use to generate these judgments, thus providing insight into how LLMs encode and operationalize the notion of reliability.

We begin our assessment by querying each model using a zero-shot, closed-book approach, providing no examples or explicit definitions of reliability. This setup constrains models to rely on internal representations acquired during training. Our goal is to analyze, through model outputs as a proxy, the heuristics guiding their classifications and to evaluate where their judgments align with or diverge from structured human evaluations.

To move beyond a simple binary classification (Reliable or Unreliable), we prompt each model to assign a political orientation label to each outlet and justify its assessment by generating explanatory keywords. Using a standardized prompt across all six models, this setup enables direct comparison of their outputs and provides insight into how models construct their reliability judgments and how these compare to expert human evaluations.

We also implement an agentic framework in which LLMs autonomously retrieve news outlet pages and follow a structured evaluation pipeline to assess reliability. This approach allows for a controlled comparison between LLMs and human evaluators when given the same task.

LLMs vs. Expert-Driven Assessments. Fig. 1*A* compares the classifications produced by each model with the reliability ratings assigned by NewsGuard. These ratings are not arbitrary but derived from a structured evaluation protocol based on systematic assessments of editorial standards, transparency, and factual accuracy. By contrast, LLMs make decisions without directly accessing these guidelines, relying instead on internal heuristics formed during training.

All six models accurately identify "Unreliable" sources, consistently flagging domains that NewsGuard associates with low credibility or lack of transparency. In contrast, classifying "Reliable" sources proves more difficult. GPT-4o mini and Llama 4 Maverick, in particular, misclassify 32% and 35% of reliable domains, respectively—substantially more than the other models. This asymmetry may reflect the multifaceted nature of NewsGuard's evaluation criteria (e.g., editorial standards, correction policies, ownership transparency), which may not be fully captured from homepage content alone. We further assess model alignment with expert judgments by comparing

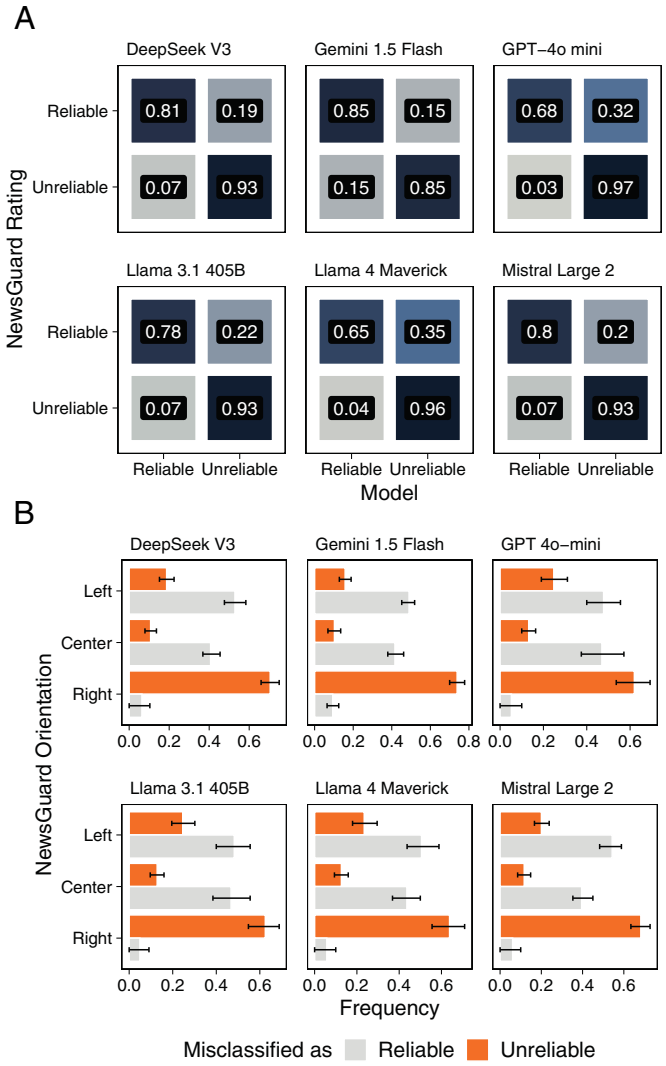

Fig. 1. LLMs' classification against expert human evaluators. (*A*) Each panel compares how domains rated as "Reliable" or "Unreliable" by NewsGuard are classified by each LLM (Deepseek V3, Gemini 1.5 Flash, GPT-4o mini, Llama 3.1 405B, Llama 4 Maverick, Mistral Large 2). All six models accurately identify Unreliable sources, with agreement ranging from 85 to 97% across models. However, Reliable domains show greater classification variability, particularly in Llama 4 Maverick and in GPT 4o-mini, which classify a significant portion (35% and 32%) as "Unreliable." (*B*) We randomly sample 40 domains from each pairing of NewsGuard's political orientation and reliability rating and compute the average misclassification rate across political orientations over 10,000 resamples. The error bars report the first and third quartiles of the resulting frequencies per group. Compared with NewsGuard, LLMs appear to overestimate or underestimate the reliability of news outlets based on their political orientation. In particular, *Right*-leaning news outlets tend to be consistently misclassified by the LLMs as unreliable, whereas the *Center* and *Left*-leaning as reliable.

their classifications against the Credibility ratings assigned by MBFC for a subset of 916 overlapping domains.

Consistent with Fig. 1*A*, LLMs often achieve over 90% accuracy with MBFC ratings at the extremes: Sources labeled Low or High credibility are correctly classified as Unreliable and Reliable, respectively. For Medium credibility sources, however, model performance diverges—both from MBFC and across models. For instance, GPT-4o mini and Llama 4 Maverick classify the majority of these domains as Unreliable (75% and 77%, respectively), while Gemini 1.5 Flash produces a more balanced distribution. This pattern is consistent with the interpretation that LLMs rely primarily on clear-cut textual cues when available, while showing lower accuracy on ambiguous or borderline cases. The confusion matrices for each model are provided in *SI Appendix*, Fig. S1. Although the models lack explicit access to the evaluation procedures of NewsGuard and MBFC and are not provided with their methodological criteria, their outputs are broadly consistent with the credibility assessments made by expert human fact-checkers.

We next examine whether misclassifications by LLMs are uniformly distributed across the political orientation labels assigned by NewsGuard or whether specific orientations are disproportionately affected. To do so, we draw random samples of 40 domains for each combination of NewsGuard's political orientation and reliability labels—the smallest group size in the dataset—and compute the proportion of reliability misclassifications per group. This sampling procedure is repeated 10,000 times to estimate average misclassification frequencies.

As shown in Fig. 1*B*, classification errors are not evenly distributed across the political spectrum. Among domains rated as Reliable by NewsGuard, *Right*-leaning outlets are consistently misclassified as Unreliable more often than *Center* or *Left*-leaning ones, whose reliability tends instead to be overestimated. Importantly, this asymmetry does not indicate that LLMs hold partisan preferences. Rather, consistent with recent work on value alignment and political bias in LLMs (e.g., refs. 41 and 43), it likely reflects correlations in the training data—for instance, the co-occurrence of extremist rhetoric and misinformation-rather than an explicit ideological stance. As a result, models may overgeneralize, conflating legitimate right-leaning journalism with toxic or conspiratorial sources when linguistic markers overlap.

We also evaluate how the political orientation labels assigned by the models compare to those from human annotators. All six LLMs show strong agreement with NewsGuard ratings, as illustrated in *SI Appendix*, Fig. S2, with substantial overlap across the political spectrum. Some discrepancies arise from the finer-grained set of labels used by the models compared to NewsGuard's coarser taxonomy. This alignment is further supported by comparisons with MBFC's "Bias Rating," focusing on strictly political classifications.

Explaining Reliability Ratings with Keywords. We examine the factors driving LLMs' reliability judgments by analyzing three sets of keywords generated by each model for every news outlet, alongside their assigned reliability and political orientation labels. These lexical cues provide indirect evidence about the heuristics models used to approximate credibility in the absence of explicit scoring guidelines. Unlike human evaluators, LLMs rely on implicit heuristics—emerging from patterns in their training data—underscoring the importance of analyzing the associations reflected in their outputs.

For each domain, every LLM generates three distinct sets of keywords: i) classification keywords, reflecting the rationale behind the assigned reliability rating; ii) determinant keywords, extracted directly from the domain's homepage and considered critical for the classification; and iii) summary keywords, broadly capturing the overall content of the homepage. All keywords are converted to lowercase prior to analysis. We impose no constraint on the number of keywords generated, allowing us to observe each model's typical lexical output and assess whether keyword volume varies by reliability label or across models. Constraining output length would risk limiting the models' expressive capacity and reducing the interpretability of their reliability assessments.

To quantify the political leaning of each keyword, we compute the average orientation of the domains in which it appears, using

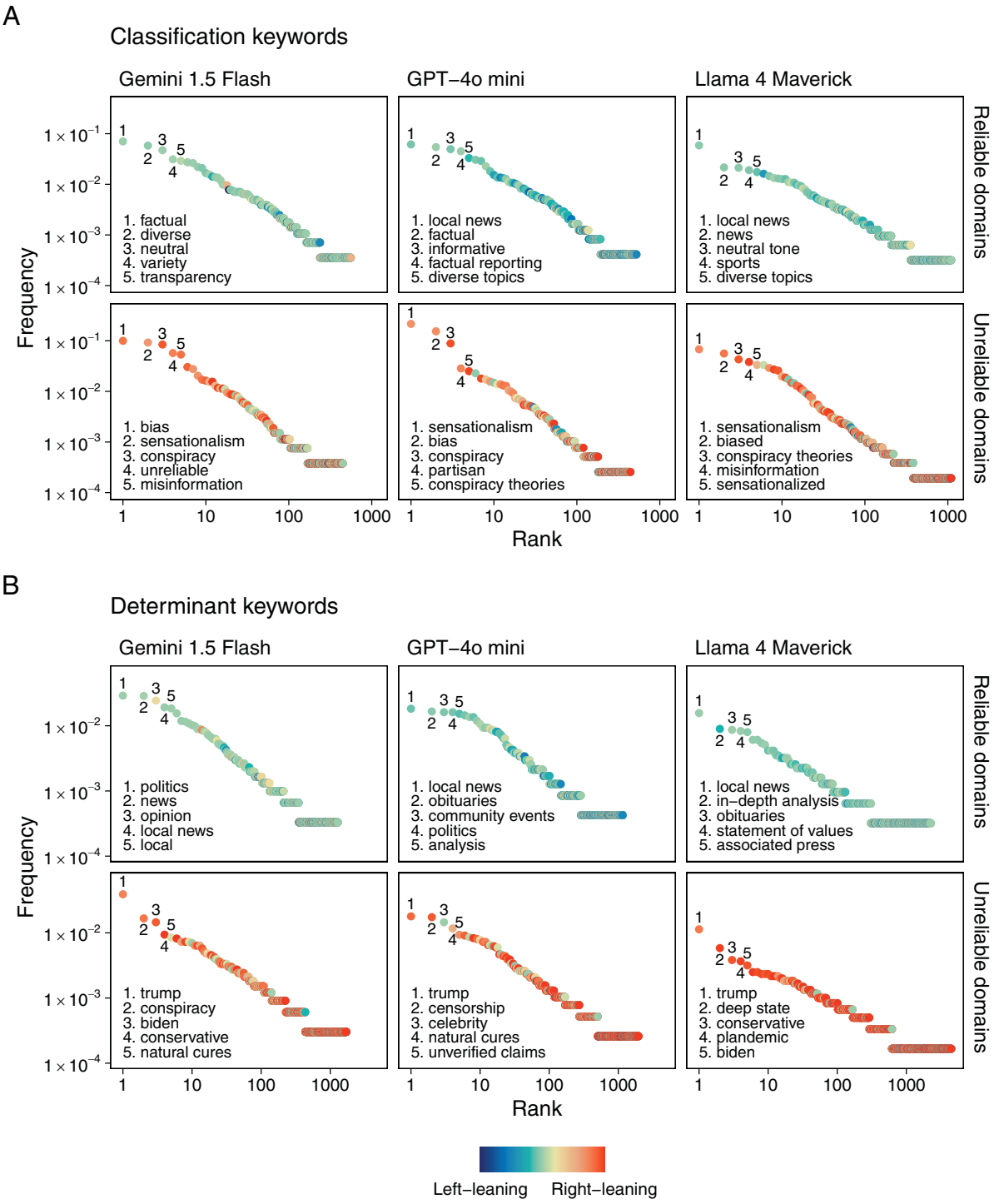

Fig. 2. Rank–frequency distributions of keywords used by each LLM to describe domains. Each panel presents the most frequently used classification (*A*) and determinant (*B*) keywords for Reliable and Unreliable domains. Only the five most common keywords per panel are labeled to enhance readability. The color gradient represents the inferred political orientation of each keyword, ranging from *Left*-leaning to *Right*-leaning, based on the political leaning of the domains they are most frequently associated with. *Right*-leaning keywords appear almost exclusively in descriptions of Unreliable domains, whereas politically neutral or *Left*-leaning keywords are more characteristic of Reliable domains. All distributions exhibit heavy-tailed behavior, as indicated by their roughly linear shape on a log–log scale, where a small set of highly frequent keywords dominate the descriptions, while the majority appear less frequently. This indicates that LLMs produce consistent markers when explaining their reliability evaluations.

the political labels assigned by the models. These categorical labels are mapped onto a numerical scale from −1 (*Left*) to 1 (*Right*), with intermediate values assigned as follows: −0.5 (*Center-Left*), 0 (*Center*), and 0.5 (*Center-Right*).

To analyze keyword usage, we construct rank–frequency distributions separately for each model, keyword type, and reliability label. In these distributions, terms are ordered by frequency of occurrence, with rank 1 assigned to the most frequent.

To streamline the presentation, we focus on three representative models—Gemini, GPT, and Llama 4—with results for the remaining models reported in *SI Appendix*. Fig. 2 shows the rank–frequency distributions of classification and determinant

keywords by model and reliability label. All models exhibit a heavy-tailed pattern, indicating that a common core set of linguistic markers is consistently associated with the classification procedure. This is consistent with natural language corpora, where few words occur frequently while most appear rarely (47).

As shown in Fig. 2, classification keywords capture linguistic markers associated with model classifications in reliability assessment. Reliable domains are frequently linked to terms denoting neutrality, transparency, and factual reporting, suggesting a focus on balanced communication and professional presentation. In contrast, unreliable domains are consistently associated with terms such as "misinformation," "conspiracy," and "bias," consistent with patterns often linked to sensationalism and partisanship. These patterns indicate that LLM classifications exhibit structured linguistic heuristics that partially mirror human evaluative criteria.

Determinant keywords offer further insight into the patterns underlying model classifications. Reliable domains are often associated with references to editorial standards and institutional transparency. GPT-4o mini and Llama 4 notably emphasize "local news," suggesting that community-based reporting appears as a marker of credibility. In contrast, unreliable domains are consistently linked to politically charged terms, with keywords such as "trump," "biden," and "deep state" recurring prominently, suggesting that politicized content is often associated with unreliability.

Additionally, Fig. 2 highlights an asymmetry in the political connotation of keywords: *Right*-leaning terms are more prevalent in descriptions of unreliable sources, whereas neutral or *Left*-leaning terms appear more often in association with reliable domains.

Keywords used to describe both reliable and unreliable domains are shown in Fig. 3, which compares their rank across the two classifications. The farther a keyword lies from the diagonal, the more distinctive it is of either reliable or unreliable domains.

When examining classification and determinant keywords, clear differences emerge between the two reliability groups. Reliable domains are associated with terms such as "local news," "scientific," "diverse," and "evidence-based," while unreliable classifications involve more controversial or politically charged terms, including politician names (e.g., "trump," "biden") and topics such as "genocide" and "vaccines." In contrast, summary keywords—describing the overall content of a domain—show substantial overlap between the two groups. This suggests that reliable and unreliable sources often cover similar topics and that the distinction lies less in what is covered than in how it is presented. We also find that terms with no clear semantic polarity tend to be unevenly distributed across classes, suggesting that model judgments may be associated with framing and usage context rather than meaning alone.

Human and LLM Credibility Assessment. Our previous analysis shows that LLMs often achieve high agreement with expert evaluations from NewsGuard and MBFC. This suggests that, despite lacking access to structured evaluation criteria, their outputs reflect heuristics that approximate human judgments. Yet, a key question remains: What procedures or proxies underlie these evaluations?

A critical observation emerges when models are prompted with only a domain URL, without access to any homepage content (38). We find that, even under these minimal conditions, LLMs generate reliability ratings that broadly align with expert assessments. For instance, Gemini achieves an F1-score of 0.78—only slightly below the 0.86 obtained with full HTML input—while GPT reaches 0.77, compared to 0.79.

This raises a foundational question: Are LLMs engaging in content-specific evaluation or relying primarily on statistical associations learned during training? If a model can classify a domain without analyzing its content, it becomes difficult to disentangle content-based assessment from statistical recall. More broadly, such behavior is consistent with the interpretation that model assessments may be shaped by prior knowledge about the news outlet rather than by content-specific evaluation.

To analyze this issue, we introduce a structured agentic workflow that enables a direct comparison between LLMs and human evaluators. Rather than treating models as black boxes producing binary outputs, we instantiate a multiagent system designed to simulate the procedural steps involved in human evaluation—retrieving, processing, and integrating information before rendering a judgment. An agentic pipeline is well suited to this task, as it provides LLMs with access to modular tools (i.e., external functions implemented in code) and enables the composition of deterministic multistep workflows, where intermediate outputs from one LLM can be passed to the next. To implement this workflow effectively, we focus on Gemini 2.0 Flash, which optimally supports the tooling required for implementing the agentic pipeline. Given the consistency observed across models in prior sections, we expect the insights to generalize.

The protocol, detailed in *Materials and Methods*, unfolds as follows. First, an LLM agent selects five out of six predefined evaluation criteria and ranks them by importance. This selection happens before the LLM is exposed to any content from the news outlet apart from its URL, which is necessary to start the pipeline. However, further experiments we conducted, presented in *SI Appendix*, suggest that exposure to the URL does not appear to influence this selection. Additionally, the criteria are provided to the model in randomized order. Two additional agents then retrieve content: One downloads the homepage, while the other extracts up to two articles deemed informative for assessing credibility, using the URLs of articles found in the homepage. This content is collected using a tool available to agents that downloads webpages in a well-structured Markdown document. The number of retrieved articles is not fixed, allowing the agent to autonomously select zero, one, or two. Subsequently, five additional agents independently evaluate the selected criteria, each using only the materials retrieved in the prior step. Their outputs include both a numeric reliability rating from 1 to 5 and a written assessment. A final agent then aggregates these assessments into a structured summary and produces a binary reliability classification, based upon the other agents' evaluations. This workflow enables a direct procedural comparison between LLMs and humans, both following the same evaluation steps in a controlled setting.

To facilitate this comparison, we replicated the agentic protocol in a human subject experiment. We recruited $N = 50$ participants and tasked them with assessing the reliability of online news outlets using the same structured procedure: selecting five criteria, ranking them, browsing the homepage and up to two articles, and finally producing a binary reliability judgment. Full details on participant recruitment and setup are provided in *Materials and Methods*.

Although a total of 37 Italian-language news outlet domains were evaluated by the recruited human participants, only 27 of these could also be evaluated using our agentic workflow, as reported in *SI Appendix*, Table S2. This limitation was due

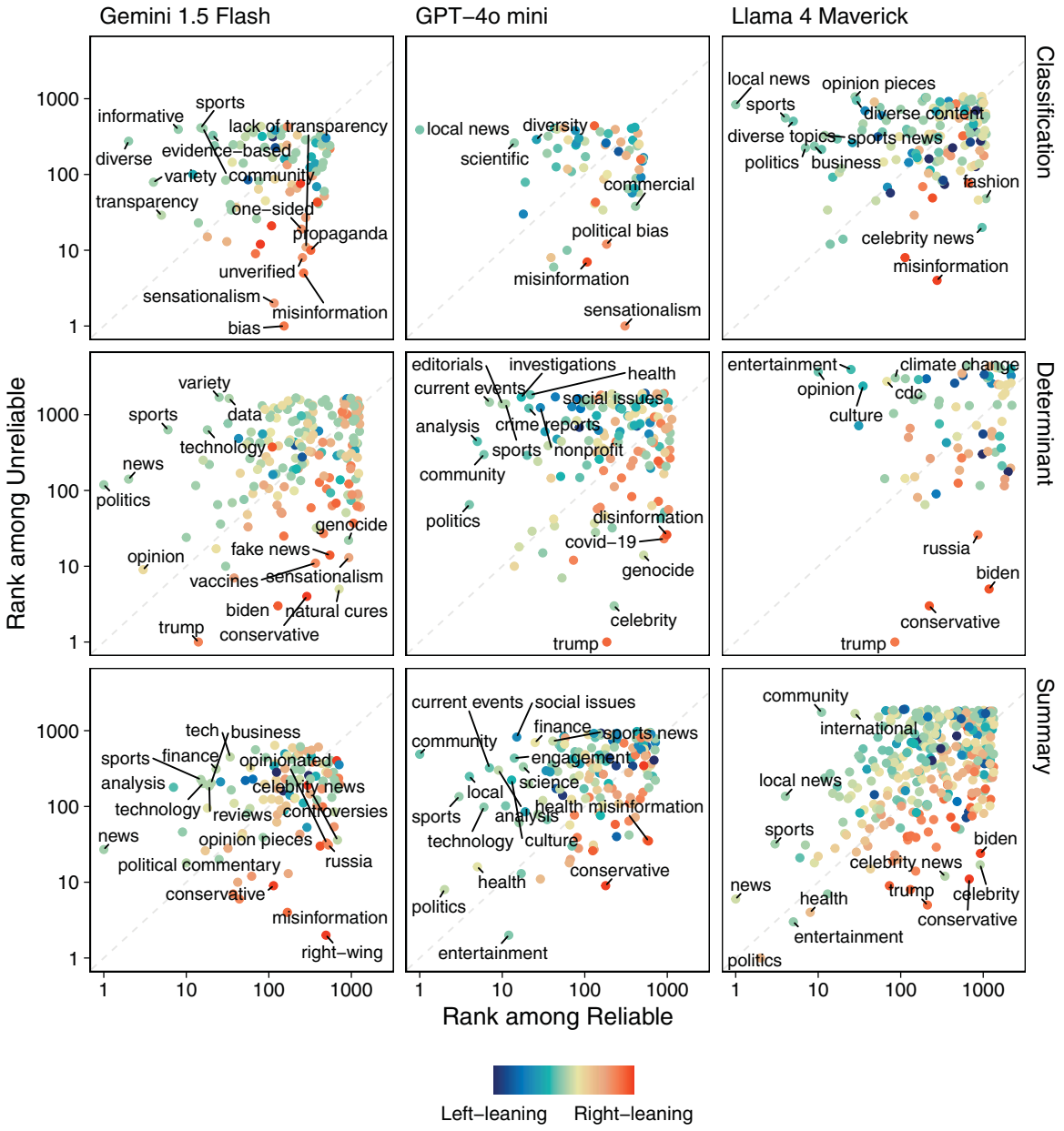

Fig. 3. Keywords' rank among Reliable and Unreliable domains. We label only keywords sufficiently distant from the diagonal, meaning they are predominantly used to describe reliable or unreliable domains rather than being evenly distributed across both classifications. Additionally, we label the top 5 keywords per reliability rating. The color gradient represents the inferred political orientation of each keyword, from *Left*-leaning to *Right*-leaning, based on the domains with which they are most frequently associated. While summary keywords (*Bottom* row) appear with similar frequency in both reliable and unreliable domains, classification and determinant keywords (*Top* and *Middle* rows) exhibit sharper separation. This result suggests that reliable and unreliable sources may cover similar topics but differ in framing tone or contextual emphasis. Notably, keywords related to transparency, objectivity, and credibility are more common among reliable domains. At the same time, sensationalist and politicized terms such as "misinformation," "propaganda," and "bias" are frequently linked to unreliable sources.

to webpage retrieval requests being blocked by some domains. Fig. 4 summarizes the results for this subset of domains, while in *SI Appendix*, Fig. S7 we report the results from human evaluators for the full set. Panel (*A*) compares the reliability ratings produced by LLMs and humans against NewsGuard's classifications. LLMs, even when constrained to a more rigid evaluation pipeline and operating on non-English content, maintain consistency with earlier findings, though with lower accuracy. In contrast, human participants show no meaningful alignment with NewsGuard: Reliable and unreliable domains are classified with roughly equal probability, suggesting that nonexpert evaluators rely on different and less consistent indicators. Panel (*B*) compares LLM and human ratings directly, treating human judgments as a referential baseline. While both groups agree on which outlets are unreliable, major divergence arises for reliable sources: LLMs classify as unreliable almost 80% of the domains rated as reliable by humans.

This asymmetry aligns with prior evidence that individuals are often more prone to reject accurate information than to believe falsehoods. A recent meta-analysis covering over 195,000 participants across 67 studies confirms this tendency—commonly referred to as skepticism bias—and shows that

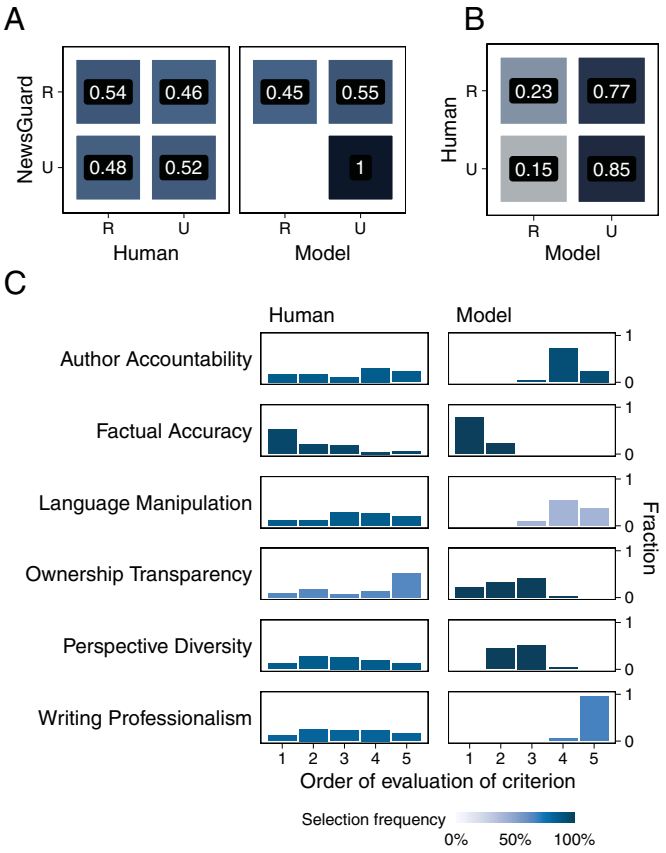

Fig. 4. Reliability evaluations by Gemini-powered LLM agents and nonexpert humans in a controlled experimental setting. (*A*) The two panels compare humans' and agents' reliability ratings against NewsGuard's classifications. Models consistently identify all Unreliable (U) sources and struggle with the Reliable (R). In contrast, humans show little to no alignment with NewsGuard, for both reliable and unreliable domains. (*B*) Confusion matrix of ratings provided by humans and agents, with the human ratings used as the ground truth. The two show strong agreement on unreliable sources, while 77% of sources rated as reliable by humans are considered unreliable by the LLM. (*C*) Distributions of order choices for each criterion by humans (*Left*) and models (*Right*). The human distributions appear more uniform than those of the models, indicating that most criteria are roughly equally likely to appear in any position compared to LLMs.

nonexperts are especially likely to misclassify true information as false (48). Our findings suggest that this bias persists even under structured and controlled evaluation. By directly integrating our experiment with this literature, we show that LLMs reproduce some behavioral regularities observed in humans (e.g., skepticism bias) while diverging sharply in the indicators that guide their judgments.

To explore the basis of this divergence, we examine the evaluation criteria selected by each group. Fig. 4*C* shows the ranking distributions across the six available criteria (Table 1). Both groups consistently prioritize "Factual Accuracy," defined as the extent to which the website's content is "accurate and free from false or misleading information." Nearly all participants—human and LLM—selected this criterion and ranked it first. However, this convergence masks deeper differences in reasoning processes. For human participants, prioritizing accuracy likely reflects deliberative reasoning based on content comprehension and analytical judgment (49). In contrast, for LLMs this emphasis appears to be operationalized via lexical associations and patterns learned during training, as suggested by proxy analyses.

Beyond this shared top criterion, the hierarchies diverge sharply. LLMs consistently rank "Ownership Transparency" among the top three criteria, while humans rarely select it and often place it last. This is consistent with findings that most individuals—unlike professional fact-checkers—rarely engage in "lateral reading" to verify sources or consult external indicators of trustworthiness (50). LLM outputs display patterns consistent with more structured verification behavior, possibly reflecting exposure to factuality benchmarks during training. Human evaluators, by contrast, often rely on surface-level cues and intuitive judgments—especially when cognitive resources are limited (51, 52).

Compared to LLMs, human participants tend to prioritize rhetorical and stylistic cues, such as "Language Manipulation" and "Writing Professionalism." Emotional tone is often interpreted as a sign of persuasive intent rather than factual reliability (26), and fluency of expression is linked to perceived truthfulness through the processing fluency heuristic (53) and the well-documented illusory truth effect, the tendency to perceive repeated information as more likely to be true regardless of its actual accuracy (54). Given that LLMs lack affective or metacognitive grounding, these features appear to carry little epistemic weight in their evaluations, functioning mainly as superficial markers. Taken together, this integration with prior literature suggests that LLMs approximate some dimensions of human evaluation while omitting others. Overall, this agentic comparison enables a more granular understanding of how LLMs appear to operationalize credibility. What emerges is not just a difference in outcomes but a difference in the observable patterns through which evaluation processes are instantiated.

Humans blend intuitive and analytical processes (55); LLM outputs, by contrast, reflect different patterns of evaluation shaped by statistical learning. Recognizing this distinction does not diminish the utility of LLMs in evaluative tasks, but it clarifies the boundaries of their competence. Our findings point

Table 1. Criteria and corresponding questions

Criterion	Question
Author accountability	To what extent does the site provide the names of content authors, along with their biographies or contact information?
Factual accuracy	To what extent do you believe the content presented on the site is accurate and free from false or misleading information?
Language manipulation	To what extent does the site use emotionally charged, exaggerated, or manipulative language?
Ownership transparency	To what extent does the site clearly declare who owns it and who provides funding for it?
Perspective diversity	To what extent does the site present content offering diverse perspectives without ideological or political bias?
Writing professionalism	To what extent does the site adhere to grammatical rules and use a clear, consistent, and professional writing style?

to a shift from content-based deliberation toward plausibility-driven approximations when judgment is delegated to automated systems—a dynamic we term epistemia, where statistical plausibility risks replacing deliberative reasoning with the illusion of knowledge rather than its verification. Future research could link behavioral experiments, cognitive models, and automated evaluation pipelines to map where human and machine heuristics converge and where they fundamentally diverge.

Conclusions

This study examined how LLMs operationalize core evaluative concepts—such as reliability and bias—when tasked with assessing online news outlets, comparing their judgments to expert benchmarks and to human participants following the same protocol. While model outputs often align with expert classifications—especially when flagging unreliable sources—this apparent agreement conceals a deeper divergence in the evaluative mechanisms themselves, raising broader questions about delegating judgment to automated systems in information environments already shaped by infodemics and platform-driven filtering.

Consistent with our findings, LLMs operate through lexical associations, statistical priors, and structural cues, rather than genuine contextual interpretation. This statistical approximation produces systematic asymmetries: *Right*-leaning outlets are disproportionately classified as unreliable. When comparing LLM outputs to expert human judgments, we find further divergence: Models display a pattern reminiscent of the skepticism bias observed in humans—an overrejection of accurate information documented in large-scale studies (56).

By integrating our human experiment with existing work on credibility assessment and cognitive biases (25, 26, 48), we show that LLMs partially replicate behavioral regularities identified in psychology while relying on fundamentally different evaluative mechanisms. This link clarifies that what emerges is not simply an accuracy gap, but a structural shift in how evaluation itself is operationalized when judgment is delegated to automated systems. Both groups prioritize "Factual Accuracy" as the most important criterion, but the underlying mechanisms differ. Human participants likely interpret accuracy through content comprehension and pragmatic reasoning; LLMs, instead, derive it from statistical regularities encoded during training. This distinction becomes clearer in the secondary criteria: LLMs emphasize "Ownership Transparency"—aligning with professional fact—checking protocols—while humans give more weight to stylistic and rhetorical features such as tone and writing fluency. These preferences mirror known heuristics like processing fluency, whereby clarity and emotional neutrality enhance perceived truth.

Such discrepancies underscore a structural difference between intuitive, context-aware human evaluation and the pattern-based, procedural mechanisms of LLMs. As these systems are increasingly embedded in decision-making pipelines—moderation, classification, prioritization—it becomes critical to assess not just whether their outputs appear reasonable, but how their internal procedures operationalize normative categories like reliability and bias. This is especially urgent in an information ecosystem already marked by infodemics, where the oversupply of low-quality or contradictory information erodes trust and amplifies polarization (18). In this context, the rise of what we term epistemia—the illusion of knowledge emerging when plausibility replaces verification—illustrates the risk that statistical approximation could displace deliberative reasoning if adopted uncritically.

While the first wave of research on social media emphasized the volume and velocity of information flows (17, 23), our findings highlight a second phase in which the problem shifts from information overload to the nature of judgment itself, as automated pipelines introduce epistemic opacity into evaluative processes. Our structured agentic framework enables such comparison by aligning inputs, tasks, and justification protocols across humans and LLMs. While our sample size limits generalizability, the controlled designed provides a solid foundation for further research. Future work should examine how this epistemic shift interacts with governance, transparency, and human oversight, especially as automated judgment pipelines expand beyond content moderation into law, policy, and scientific evaluation.

In sum, the apparent alignment between LLMs and expert judgments may mask only superficial output convergence. Delegating evaluative tasks to these systems risks embedding frameworks driven by lexical and statistical associations rather than deliberative reasoning, amplifying existing information pathologies. Addressing this shift requires transparency, human oversight, and potentially new training paradigms that explicitly disentangle factual reliability from ideological or stylistic cues. Hybrid approaches that combine statistical models with explicit reasoning criteria or retrieval-based evidence may offer a promising way forward.

Materials and Methods

Data Collection and Preprocessing. All data were collected by downloading the HTML homepages of domains rated by NewsGuard as "reliable" or "unreliable," using the requests library available on Python (57). These domains have been selected among outlets reported by NewsGuard as English-speaking, based in an English-speaking country (US, GB, CA, AU, NZ), and with a National or International focus. Not all domains could be downloaded, as many were either no longer active at the time of downloading, only accessible from specific regions, or designed in such a way as to render automatic scraping difficult.

The downloaded pages are then filtered to retain only the information relevant to the LLMs to evaluate the reliability of each domain. This process involves removing unnecessary HTML components, such as scripts, styles, and navigation elements, using the beautifulsoup4 library on Python (58). The resulting document almost exclusively includes textual content, particularly related to news or information about the outlet. While domain-specific preprocessing techniques can also be implemented, we opted for a more conservative, universal approach that can easily be applied to any news outlet's homepage. The procedure yields a dataset of 2,302 domains with at least 100 characters of content. Finally, we retain only the 2,286 domains successfully classified and output in well-formed JSON by all six LLMs, with 1,188 rated by NewsGuard as "reliable" and the remaining 1,098 as "unreliable."

Model Prompting. We follow a zero-shot, closed-book approach to prompting for all six LLMs we experiment with, by providing no example classifications and no definitions of "reliability," respectively. Additionally, for all models, we use the system prompt *"You are a service that gives a reliability rating to news outlets."* and set the temperature parameter to 0. Our prompt, which we use for all models, is split into two parts.

The first part gives the model basic instructions regarding the classification task, providing no context to the concept of "reliability" beyond mentioning the terms "reliable" and "unreliable." In the second part, we provide the models with specific instructions concerning their output, specifying both the required content and its structure. Fig. 5 reports the exact prompt used for all models.

Gemini 1.5 Flash, DeepSeek V3, and GPT-4o mini were prompted by performing calls to the official API endpoints made available by Google, DeepSeek, and OpenAI, respectively. Llama 3.1 405B was prompted through SambaNova Cloud, while Llama 4 Maverick and Mistral Large 2 were accessed via OpenRouter, both third-party platforms.

Given the HTML of the homepage of a news outlet, you classify it as either 'reliable' or 'unreliable' based on its content. Focus on the entire homepage, rather than just the main content or story that is presented. Make your judgment solely on the contents of the webpage.

What rating ('reliable', 'unreliable') would you assign to the news outlet with the homepage in the following HTML? I want the output as a JSON with the following attributes:

- 'rating': your overall classification rating ('reliable', 'unreliable')
- 'text': your raw response
- 'orientation': political orientation of the contents of the webpage ('left', 'center-left', 'center', 'center-right', 'right')
- 'classification_keywords': a list of keywords to justify your reasoning for classification
- 'determinant_keywords': a list of keywords from the webpage that were determinant for your classification
- 'summary_keywords': a list of keywords that summarize the contents of the webpage

HTML:
`{scraped HTML}`

Fig. 5. Prompt used for all LLMs when provided the scraped HTML homepage.

Queries sent to DeepSeek, GPT, Llama 3.1, and Mistral were truncated to ensure that they fit within the models' context length (128,000 tokens for all), which is the maximum number of tokens they can process at once. Specifically, the scraped webpages provided to these models were limited to the first 50,000 characters. However, this truncation affected less than 2% of the domains.

Each domain was evaluated individually, as simultaneous classification of multiple inputs may introduce unwanted bias. For example, reliability might be assessed relative to the specific subset of domains provided in the query, rather than based on the model's inherent notion of "reliability."

If the output political orientation label fell outside the specified 5-point scale, we reassigned it to "Center." For all models, this affected none or fewer than 1% of the domains, except for Mistral Large 2, where approximately 4% of the domains received an out-of-scale label.

When evaluating the LLMs' ability to classify news outlets using only their domain names, we slightly altered the prompt in Fig. 5 by substituting the first paragraph with the text "*Given the domain of a news outlet, you classify it as either "reliable" or "unreliable" based on its content.*," and by replacing all other occurrences of "HTML" with "URL."

Agentic Workflow. We implemented the agentic workflow for outlet reliability classification with Google's Agent Development Kit (ADK) (59). ADK is a Python toolkit that allows for developing and orchestrating agentic systems. In particular, for our study, we developed a multiagent system where each subtask in the reliability evaluation procedure is delegated to a dedicated agent. The implementation first relies on what ADK calls a "workflow agent," specifically a "Sequential Agent." This is not an LLM-powered agent, but rather a component in ADK that allows for orchestrating the other agents in a deterministic manner. Specific to our case, this agent is what enables our procedure to follow a well-defined path, as agents are called one after the other. The initial prompt used to start the workflow is *"Select the most appropriate reliability criteria among those provided and evaluate the reliability of {URL}."* We provide all additional prompts used in this workflow in *SI Appendix*.

The first agent is a "code agent," that is, an LLM agent that performs actions by writing code and calling tools written in a programming language (60), Python in our case. Specifically, it is a tool-augmented agent that is tasked with selecting five criteria to evaluate from a list of six, and to rank them from *"most to least important for assessing reliability."* To avoid our chosen order of criteria influencing the model's decision, we do not include the list directly in the prompt. Instead, the agent calls a Python function that returns the criteria in a randomized order.

The next two agents, which are also code agents, are responsible for retrieving all data required for the reliability evaluation. The first one is tasked with retrieving the domain's homepage as a structured Markdown document, by leveraging the markitdown Python library (61). Contrary to the data collection process used for our first analyses, in this case, we are not just interested in the textual content of the page but also in the URLs of the news articles presented on the homepage. The Markdown format is particularly effective at concisely structuring this information in a way that can be easily understood by LLMs. This approach enables the second code agent to analyze the scraped homepage and assess whether to collect up to two articles by scraping their corresponding webpages. All this content is thus stored in the workflow's state to allow all subsequent agents to read it.

Once the criteria are selected and the data collected, the workflow activates all agents tasked with evaluating the criteria, assigning one criterion per agent. These LLM agents are not provided with any coding tool, meaning they function analogously to any other LLM assistant when given a prompt. Further, they produce their assessment independently from each other. Each agent is asked to output a rating from 1 to 5, with higher scores corresponding to higher reliability, a written summary explanation with examples and quotes from the analyzed content, and a binary flag indicating whether the agent also analyzed the downloaded articles for evaluating the criterion. Additionally, we instruct a separate agent to analyze the news outlet's political orientation on a 5-point scale from *Left* to *Right*.

The information produced by all agents is then reviewed by a final LLM agent, which is tasked with assigning an overall binary reliability rating: "reliable" or "unreliable."

We set the temperature to 0 for all LLM agents, except for the one responsible for selecting the evaluation criteria. For this agent, we instead used a temperature of 1 to introduce more variability in the outputs. However, as shown in *SI Appendix*, Figs. S5 and S6, the prioritization of criteria observed in Fig. 4*C* remains largely consistent across the full range of temperatures available for Gemini 2.0 Flash, likely because the task is highly constrained rather than open-ended.

Experimental Design. Here, we provide details about the experimental setting with human participants.

Participants. A total of 50 participants (28 females, 22 males; $\mu_{age} = 28.4$, $\sigma_{age} = 9.6$) took part in the in-person experiment sessions at the Department of Psychology of Sapienza University of Rome. All participants were recruited online through advertisements distributed via social media platforms (e.g., LinkedIn, Facebook) and by snowball sampling. Eligibility was limited to native Italian-speaking adults (aged $\geq$18) with normal or corrected vision. All human participants were nonexperts with no prior training in credibility assessment, ensuring that judgments reflected naïve, uncoached heuristics rather than professional fact-checking protocols.

Materials and procedure. The procedure was carried out in a controlled laboratory setting and administered to participants using a Lenovo laptop with a 15.6-inch screen. The experiment was conducted through the Google Chrome web browser to ensure compatibility and correct display of the stimuli. Before taking part in the testing phase, participants provided informed consent.

The testing phase consisted of two parts: a criteria selection task and an evaluation task.

In the first part, participants were given a set of six criteria in question form (Table 1) for assessing the reliability of news domains.

Participants were asked to identify five criteria they deemed the most relevant for evaluating news credibility and to rank them from most (1) to least (5) relevant. They were asked to exclude one criterion that they deemed nonrelevant for the assessment, in order to encourage critical prioritization and to avoid uniform ratings across all criteria. In the second part, participants were

shown six real and publicly accessible Italian-language news outlets, one at a time. The order in which the websites were presented was randomized across participants. Each participant was instructed to freely navigate the websites and to read up to two full articles to deepen their evaluation. Subsequently, they responded to each question corresponding to the criteria selected in the first phase on a 5-point Likert scale (1 = not at all, 5 = completely) for each website. Additionally, participants provided a binary judgment (Yes/No) concerning the overall reliability of the website, answering the question *"Is this website reliable?"* A time limit was imposed for the evaluation of each website to standardize the procedure across participants. The full testing session lasted approximately 15 to 20 min. Instructions and interface were given in Italian. All stimuli were tested before the experimental procedure to ensure usability.

Ethics Statement. The study protocol was approved by the Ethics Committee for Transdisciplinary Research of Sapienza University of Rome (Protocol ID: 339/2025).

Data, Materials, and Software Availability. The list of domains used for the analyses has been deposited in OSF (https://osf.io/d8npc/) (62).

ACKNOWLEDGMENTS. We thank Luciano Floridi, Mattia Samory, Giovanni Pezzulo, and the Hypnotoad for precious insights during the development of the work. This work was supported by the IRIS Infodemic Coalition (UK government, grant no. SCH-00001-3391); by SEcurity and RIghts in the CyberSpace (SERICS, code PE00000014) under the Piano Nazionale di Ripresa e Resilienza of the Ministero Università e Ricerca, funded by the European Union–NextGenerationEU; by Comunicare il rischio nelle emergenze per la Sanità Pubblica from the Italian Ministry of Health under the Centro Controllo Malattie 2022 program; by the National Operational Program "Ricerca e Innovazione" 2014–2020; and by Progetti di Rilevante Interesse Nazionale 2022 "Multimedia Understanding meets Social Media Analysis" (MUSMA; CUP G53D23002930006), funded by the European Union–NextGenerationEU, Mission 4 Component 2 Investment 1.1.

Author affiliations: [a]Department of Computer, Control and Management Engineering, Sapienza University of Rome, Rome 00185, Italy; [b]Department of Computer Science, Sapienza University of Rome, Rome 00161, Italy; [c]Department of Legal, Social, and Educational Sciences, Tuscia University, Viterbo 01100, Italy; and [d]Department of Psychology, Sapienza University of Rome, Rome 00185, Italy

1. M. Binz *et al*., How should the advancement of large language models affect the practice of science?. *Proc. Natl. Acad. Sci. U.S.A.* **122**, e2401227121 (2025).
2. M. Özer, M. Perc, "Human complementation must aid automation to mitigate unemployment effects due to AI technologies in the labor market" in *Istanbul Bilgi University* (Istanbul Bilgi University, 2024).
3. Y. Gao, D. Lee, G. Burtch, S. Fazelpour, Take caution in using LLMs as human surrogates. *Proc. Natl. Acad. Sci. U.S.A.* **122**, e2501660122 (2025).
4. N. Yax, H. Anlló, S. Palminteri, Studying and improving reasoning in humans and machines. *Commun. Psychol.* **2**, 51 (2024).
5. L. Wang *et al*., A survey on large language model based autonomous agents. *Front. Comput. Sci.* **18**, 186345 (2024).
6. H. Torkamaan *et al*., Challenges and future directions for integration of large language models into socio-technical systems. *Behav. Inf. Technol.* **57**, 1–20 (2024).
7. M. Özer, M. Perc, H. Suna, Artificial intelligence bias and the amplification of inequalities in the labor market. *J. Econ. Cult. Soc.* **69**, 159–168 (2024).
8. M. Binz *et al*., A foundation model to predict and capture human cognition. *Nature* **644**, 1–8 (2025).
9. Y. Qu, J. Wang, Performance and biases of large language models in public opinion simulation. *Humanit. Soc. Sci. Commun.* **11**, 1–13 (2024).
10. R. Stureborg, D. Alikaniotis, Y. Suhara, Large language models are inconsistent and biased evaluators. arXiv [Preprint] (2024). http://arxiv.org/abs/2405.01724 (Accessed 1 October 2025).
11. M. Cinelli, G. De Francisci Morales, A. Galeazzi, W. Quattrociocchi, M. Starnini, The echo chamber effect on social media. *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2023301118 (2021).
12. A. E. Holton, M. Coddington, S. C. Lewis, H. G. De Zuniga, Reciprocity and the news: The role of personal and social media reciprocity in news creation and consumption. *Int. J. Commun.* **9**, 22 (2015).
13. M. L. Khan, Social media engagement: What motivates user participation and consumption on Youtube? *Comput. Hum. Behav.* **66**, 236–247 (2017).
14. M. Avalle *et al*., Persistent interaction patterns across social media platforms and over time. *Nature* **628**, 582–589 (2024).
15. E. Kubin, C. Von Sikorski, The role of (social) media in political polarization: A systematic review. *Ann. Int. Commun. Assoc.* **45**, 188–206 (2021).
16. C. Budak, B. Nyhan, D. M. Rothschild, E. Thorson, D. J. Watts, Misunderstanding the harms of online misinformation. *Nature* **630**, 45–53 (2024).
17. D. M. Lazer *et al*., The science of fake news. *Science* **359**, 1094–1096 (2018).
18. M. Del Vicario *et al*., The spreading of misinformation online. *Proc. Natl. Acad. Sci. U.S.A.* **113**, 554–559 (2016).
19. K. Gallup, Indicators of news media trust (2018). https://knightfoundation.org/reports/indicators-of-news-media-trust/. Accessed 1 October 2025.
20. N. Newman, R. Fletcher, D. Levy, R. Nielsen, The Reuters institute digital news report(2018). https://ssrn.com/abstract=3245355. Accessed 1 October 2025.
21. C. A. Bail *et al*., Exposure to opposing views on social media can increase political polarization. *Proc. Natl. Acad. Sci. U.S.A.* **115**, 9216–9221 (2018).
22. M. Falkenberg *et al*., Growing polarization around climate change on social media. *Nat. Clim. Change* **12**, 1114–1121 (2022).
23. M. Cinelli *et al*., The COVID-19 social media infodemic. *Sci. Rep.* **10**, 1–10 (2020).
24. M. J. Metzger, A. J. Flanagin, "Psychological approaches to credibility assessment online" in *The Handbook Psychology Communication Technology* (Wiley, 2015), pp. 445–466.
25. S. Y. Rieh, Credibility and cognitive authority of information. *Encycl. Libr. Inf. Sci.* **1**, 1337–1344 (2010).
26. M. J. Metzger, A. J. Flanagin, R. B. Medders, Social and heuristic approaches to credibility evaluation online. *J. Commun.* **60**, 413–439 (2010).
27. J. Lühring, H. Metzler, R. Lazzaroni, A. Shetty, J. Lasser, Best practices for source-based research on misinformation and news trustworthiness using Newsguard. *J. Quant. Descr. Digit. Media* **5**, e003 (2025).
28. P. Nakov, H. T. Sencar, J. An, H. Kwak, "A survey on predicting the factuality and the bias of news media" in *Findings of the Association for Computational Linguistics ACL* (Association for Computational Linguistics, 2024), pp. 15947–15962.
29. NewsGuard Technologies, Rating process and criteria. https://www.newsguardtech.com/ratings/rating%20-process-%20criteria/. Accessed 26 November 2024.
30. OpenAI, "Gpt-4o" (Tech. Rep., 2024).
31. G Team *et al*., Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. arXiv [Preprint] (2024). http://arxiv.org/abs/2403.05530 (Accessed 1 October 2025).
32. M AI, Llama 3.1: Advancements in open-weight LLMs (2024) https://ai.meta.com/blog/meta-llama-3-1/. Accessed 1 October 2025.
33. P. Törnberg, Large language models outperform expert coders and supervised classifiers at annotating political social media messages. *Soc. Sci. Comput. Rev.* e08944393241286471 (2024).
34. F. Gilardi, M. Alizadeh, M. Kubli, Chatgpt outperforms crowd workers for text-annotation tasks. *Proc. Natl. Acad. Sci. U.S.A.* **120**, e2305016120 (2023).
35. P. Y. Wu, J. Nagler, J. A. Tucker, S. Messing, Large language models can be used to estimate the latent positions of politicians. arXiv [Preprint] (2023). http://arxiv.org/abs/2303.12057 (Accessed 1 October 2025).
36. C. H. Chiang, Hy. Lee, "Can large language models be an alternative to human evaluations?" in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics* (Association for Computational Linguistics, 2023).
37. J. O. Krugmann, J. Hartmann, Sentiment analysis in the age of generative AI. *Cust. Needs Solut.* **11**, 3 (2024).
38. K. C. Yang, F. Menczer, "Accuracy and political bias of news source credibility ratings by large language models" in *Proceedings of the 17th ACM Web Science Conference* (ACM, 2025).
39. E. Hoes, S. Altay, J. Bermeo, Leveraging chatGPT for efficient fact-checking. PsyArXiv [Preprint] (2023). https://europepmc.org/article/PPR/PPR640906# (Accessed 1 October 2025).
40. D. Quelle, A. Bovet, The perils and promises of fact-checking with large language models. *Front. Artif. Intell.* **7**, 1341697 (2024).
41. R. Hernandes, G. Corsi, LLMs left, right, and center: Assessing GPT's capabilities to label political bias from web domains. arXiv [Preprint] (2024). http://arxiv.org/abs/2407.14344 (Accessed 1 October 2025).
42. T. Hu *et al*., Generative language models exhibit social identity biases. *Nat. Comput. Sci.* **5**, 1–11 (2024).
43. F. Y. Motoki, V. P. Neto, V. Rangel, Assessing political bias and value misalignment in generative artificial intelligence. *J. Econ. Behav. Organ.* **234**, 106904 (2025).
44. J. W. Strachan *et al*., Testing theory of mind in large language models and humans. *Nat. Hum. Behav.* **8**, 1–11 (2024).
45. E. Coppolillo, G. Manco, L. M. Aiello, Unmasking conversational bias in AI multiagent systems. arXiv [Preprint] (2025). http://arxiv.org/abs/2501.14844 (Accessed 1 October 2025).
46. M. Safdari *et al*., Personality traits in large language models. arXiv [Preprint] (2023). http://arxiv.org/abs/2307.00184 (Accessed 1 October 2025).
47. N. Di Marco *et al*., Patterns of linguistic simplification on social media platforms over time. *Proc. Natl. Acad. Sci. U.S.A.* **121**, e2412105121 (2024).
48. J. Pfänder, S. Altay, Spotting false news and doubting true news: A systematic review and meta-analysis of news judgements. *Nat. Hum. Behav.* **9**, 1–12 (2025).
49. G. Pennycook, D. G. Rand, Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition* **188**, 39–50 (2019).
50. S. Wineburg, S. McGrew, Lateral reading and the nature of expertise: Reading less and learning more when evaluating digital information. *Teach. Coll. Rec.* **121**, 1–40 (2019).
51. S. Fiske, *Social Cognition: Selected Works of Susan Fiske* (Routledge, 2018).
52. A. Tversky, D. Kahneman, Judgment under uncertainty: Heuristics and biases: Biases in judgments reveal some heuristics of thinking under uncertainty. *Science* **185**, 1124–1131 (1974).
53. R. Reber, N. Schwarz, Effects of perceptual fluency on judgments of truth. *Conscious. Cogn.* **8**, 338–342 (1999).
54. L. K. Fazio, N. M. Brashier, B. K. Payne, E. J. Marsh, Knowledge does not protect against illusory truth. *J. Exp. Psychol. Gen.* **144**, 993 (2015).
55. D. Kahneman, Maps of bounded rationality: Psychology for behavioral economics. *Am. Econ. Rev.* **93**, 1449–1475 (2003).
56. C. S. Taber, M. Lodge, Motivated skepticism in the evaluation of political beliefs. *Am. J. Polit. Sci.* **50**, 755–769 (2006).

57. GitHub - psf/requests: A simple, yet elegant, HTTP library – github.com (2024), https://github.com/google/adk-python. Accessed 1 October 2025.
58. L. Richardson, Beautiful Soup: We called him Tortoise because he taught us –crummy.com (2024), https://www.crummy.com/software/BeautifulSoup/. Accessed 1 October 2025.
59. GitHub - google/adk-python: An open-source, code-first Python toolkit for building, evaluating, and deploying sophisticated AI agents with flexibility and control – github.com (2025), https://github.com/google/adk-python. Accessed 1 October 2025.
60. X. Wang *et al*., "Executable code actions elicit better LLM agents" in *Proceedings of the 41st International Conference on Machine Learning, Proceedings of Machine Learning Research*, R. Salakhutdinov *et al*, Eds. (PMLR, 2024), vol. 235, pp. 50208–50232.
61. GitHub - microsoft/markitdown: Python tool for converting files and office documents to Markdown – github.com (2025), https://github.com/microsoft/markitdown. Accessed 1 October 2025.
62. E. Loru *et al*., The simulation of judgment in LLMs. Open Science Framework. https://doi.org/10.17605/OSF.IO/D8NPC. Deposited 18 September 2025.