

Decoding AI Judgment: How LLMs Assess News Credibility and Bias

Edoardo Loru¹, Jacopo Nudo², Niccolò Di Marco²,
Matteo Cinelli², Walter Quattrociocchi^{2*}

¹Department of Computer, Control and Management Engineering,
Sapienza University of Rome, Viale Ariosto 25, Rome, 00185.

²Department of Computer Science, Sapienza University of Rome, Viale
Regina Elena 295, Rome, 00161.

*Corresponding author(s). E-mail(s): walter.quattrociocchi@uniroma1.it;

Abstract

Large Language Models (LLMs) are increasingly used to assess news credibility, yet little is known about how they make these judgments. While prior research has examined political bias in LLM outputs or their potential for automated fact-checking, their internal evaluation processes remain largely unexamined. Understanding how LLMs assess credibility provides insights into AI behavior and how credibility is structured and applied in large-scale language models.

This study benchmarks the reliability and political classifications of state-of-the-art LLMs—Gemini 1.5 Flash (Google), GPT-4o mini (OpenAI), and LLaMA 3.1 (Meta)—against structured, expert-driven rating systems such as NewsGuard and Media Bias Fact Check. Beyond assessing classification performance, we analyze the linguistic markers that shape LLM decisions, identifying which words and concepts drive their evaluations. We uncover patterns in how LLMs associate credibility with specific linguistic features by examining keyword frequency, contextual determinants, and rank distributions.

Beyond static classification, we introduce a framework in which LLMs refine their credibility assessments by retrieving external information, querying other models, and adapting their responses. This allows us to investigate whether their assessments reflect structured reasoning or rely primarily on prior learned associations.

1 Introduction

In a digital environment where information is constantly produced and consumed [1, 2] and users interact and discuss [3, 4], assessing the credibility of sources is a key challenge [5–7]. The way reliability is determined influences public trust [8, 9], shapes social and political discussions [4, 10, 11], and affects decision-making in critical areas like public health [12, 13]. While human evaluators rely on structured criteria to assess credibility [14–16], the rise of Large Language Models raises new questions about how these systems process, interpret, and replicate such judgments.

News rating agencies like NewsGuard and Media Bias Fact Check (MBFC) provide structured, expert-driven assessments of news reliability. These assessments are based on rigorous evaluation criteria, such as factual accuracy, transparency, and editorial independence, and are developed through years of systematic work by human evaluators [17]. These benchmarks serve as operational gold standards in media assessment, widely used by researchers, platforms, and policymakers [18]. However, their reliance on human expertise makes them costly and time-consuming [19, 20].

On the other hand, LLMs, including GPT-4o (OpenAI) [21], Gemini 1.5 Flash (Google) [22], and LLaMA 3.1 (Meta) [23], have demonstrated advanced capabilities in tasks such as text classification [24–27], sentiment analysis [28], and fact-checking [29–32]. Moreover, recent research has increasingly focused on how human heuristics and biases manifest in artificial intelligence models [33–35]. Beyond surface-level bias detection, studies are also investigating whether LLMs encode psychological traits and value orientations, shedding light on the broader implications of their training data and decision-making processes [36–38].

This raises a fundamental question: to what extent do LLMs replicate, diverge from, or even reveal new dimensions of these structured human evaluations? Indeed, little is known about how these models internally process information and build their evaluations. To what extent do LLMs reflect human-driven evaluations’ biases, priorities, and heuristics? How do their decision-making processes differ from or align with those of human experts?

This study examines how Large Language Models (LLMs) make decisions when evaluating the reliability and political orientation of a sample of 2,302 news outlets. Instead of merely assessing their alignment with expert evaluations, we focus on how these models build them. We address the underlying patterns shaping their reasoning by analyzing the linguistic markers, heuristics, and contextual cues that factor into their classifications. Through a systematic comparison with structured human evaluations (i.e., NewsGuard and MBFC), we explore whether LLMs rely on similar principles or develop distinct strategies for credibility assessment.

We also investigate how LLMs behave within an agentic workflow in which models can refine their assessments by retrieving additional information, querying external sources, or interacting with other AI systems. This approach allows us to examine whether LLMs can self-correct, reinforce biases, or adapt their reasoning when faced with new inputs. By integrating these dynamics, we move beyond mere classification and lay the groundwork for a structured human-AI comparison to provide deeper insights into the cognitive mechanisms underlying credibility judgments.

2 Results and Discussion

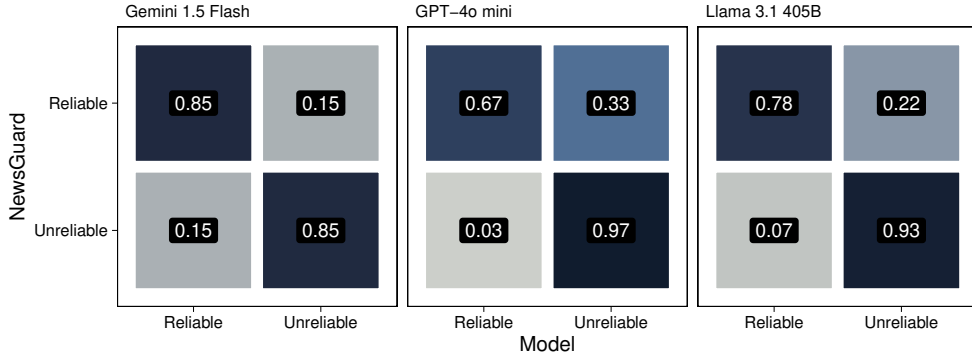
We investigate how three state-of-the-art LLMs—Gemini 1.5 Flash, GPT-4o mini, and LLaMA 3.1 405B—encode and apply credibility assessments by comparing their outputs to expert human benchmarks from NewsGuard and MBFC. Rather than merely measuring classification accuracy, we aim to address the underlying processes guiding their evaluations. To ensure a diverse and representative dataset, we select 7,715 English-speaking news domains, evenly split between those labeled as Reliable and Unreliable by NewsGuard. These sources span multiple countries and include outlets with national or international focus. We retrieve a snapshot of each domain’s homepage, filtering out nonessential elements (e.g., scripts, styling) to isolate relevant textual components, such as news headlines and descriptions. This pre-processing step ensures that all LLMs are evaluated based on the same contextual information a human assessor might use. The final dataset consists of 2,302 active domains with sufficient content for classification. By analyzing not only the final classification labels assigned by the LLMs but also the process behind their assessments, we aim to provide deeper insights into how these models encode the notion of reliability. A detailed breakdown of the data collection and processing is provided in Methods.

We begin our assessment by querying each model using a zero-shot, closed-book approach, meaning no prior examples or explicit definitions of reliability are provided. This ensures that the models, without further context, solely rely on their internalized knowledge and learned heuristics to classify news outlets. By doing so, we aim to investigate these models’ interpretative framework and assess how their reliability assessments mirror or diverge from structured human evaluations. To this end, beyond a simple binary classification (Reliable or Unreliable), we prompt the models to assign a political orientation label to each news outlet and to justify their assessment by generating explanatory keywords. This additional layer of analysis allows us to explore how LLMs construct reliability assessments, whether their justifications align with human evaluators, and whether emerging discrepancies can be observed in their decision-making. Further, as detailed in Methods, we use the same prompt for all three LLMs to allow for a direct comparison between models. Finally, we introduce an agentic framework where LLMs refine their credibility assessments by retrieving external information, interacting with other models, and adapting their responses. This approach allows us to examine whether LLMs apply structured reasoning beyond their internalized priors and sets up the conditions for a direct comparison between human and AI-driven evaluation strategies.

2.1 LLMs vs. Expert-Driven Assessments

In Fig. 1A, we illustrate how the classifications of each model compare with the reliability ratings assigned by NewsGuard. It is important to note that NewsGuard’s ratings are not arbitrary judgments but the result of a structured, operationalized evaluation framework, developed through rigorous, systematic assessments of news outlets. At the same time, LLMs operate without explicit knowledge of these guidelines, meaning their decisions emerge from their internal processes, rather than from strict adherence to predefined criteria.

A



B

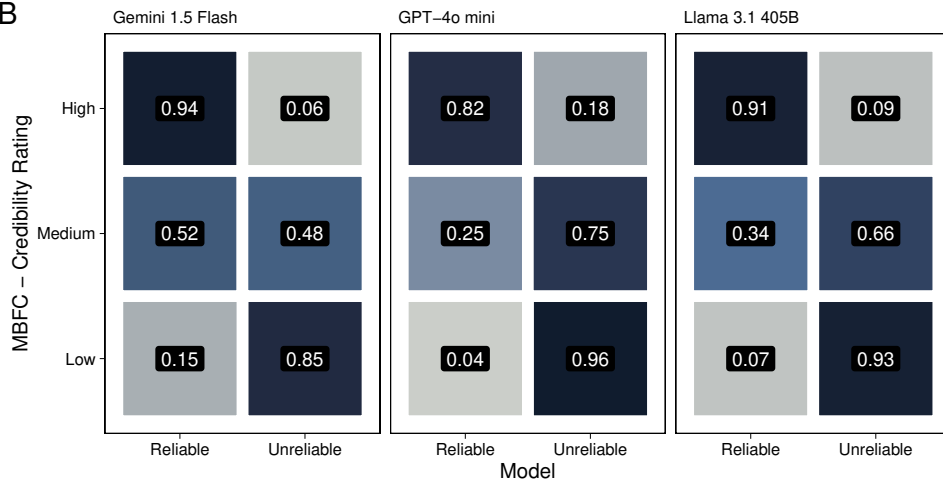


Fig. 1 LLMs’ classification against expert human evaluators. (A) Each panel compares how domains rated as “Reliable” or “Unreliable” by NewsGuard are classified by each LLM (Gemini 1.5 Flash, GPT-4o mini, Llama 3.1 405B). All three models accurately identify “Unreliable” sources, with agreement ranging from 85 to 97% across models. However, “Reliable” domains show greater classification variability, particularly in GPT-4o mini, which classifies a significant portion (33%) as “Unreliable”. (B) Each panel shows how MBFC’s “Credibility” ratings (High, Medium, Low) align with LLM classifications. The models strongly agree on both high- and low-credibility domains, classifying them correctly over 90% of the time. However, “Medium” credibility sources exhibit greater inconsistency across models, with GPT-4o mini and Llama 3.1 tending to classify them as “Unreliable” (75% and 66%, respectively), while Gemini remains more balanced (52%-48%). This further suggests that LLMs are particularly sensitive to sources with lower credibility signals but struggle with intermediate cases.

All three models accurately identify “Unreliable” sources, consistently flagging domains that NewsGuard marks for lack of credibility or transparency. Conversely,

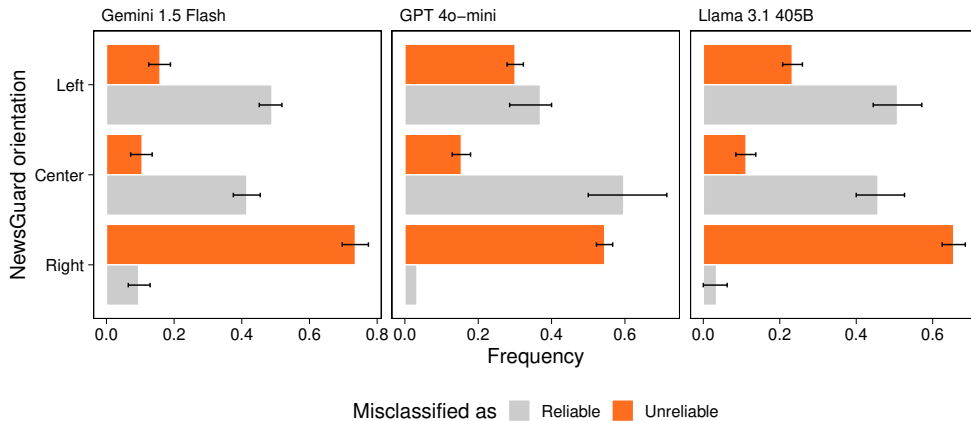


Fig. 2 LLMs’ reliability rating misclassification across political orientation. We randomly sample 40 domains from each pairing of NewsGuard’s political orientation and reliability rating, and estimate the average frequency over 10,000 resamples of reliability misclassification for each. The error bars report the first and third quartile 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.

classifying “Reliable” sources appears to be more challenging for all three LLMs, with GPT-4o mini in particular showing a higher misclassification rate (33%) than the rest. This discrepancy may reflect that NewsGuard’s methodology incorporates multiple dimensions of evaluation, such as editorial standards, correction policies, and transparency, which may not be directly inferable from the homepage’s content alone.

To further assess how the models’ ratings match against expert human evaluators, Fig. 1B shows the alignment with the “Credibility” ratings from Media Bias Fact Check (MBFC), a well-established service that categorizes sources using a formalized three-tier framework based on factual accuracy, bias, and traffic/longevity [39]. Among the 977 domains overlapping with MBFC’s dataset, LLMs exhibit strong agreement for sources with a “Low” or “High” credibility rating, classifying over 90% of them as “Unreliable” and “Reliable”, respectively. However, for “Medium” credibility sources, the models show differences both compared to MBFC and among themselves: GPT-4o mini and LLaMA 3.1 classify most of these sources as “Unreliable” (75% and 66% of them, respectively), whereas Gemini 1.5 Flash remains more balanced. This suggests that LLMs may rely on clear-cut textual cues associated with highly credible or noncredible sources.

Although these models do not have explicit access to the rating process of NewsGuard and MBFC, nor are they provided their methodological criteria in the prompt, their responses suggest that they possess distinct but systematic heuristics that generally approximate human-defined credibility standards.

In light of this, we now investigate the cases of models’ ratings disagreeing with human evaluators. Notably, we analyze whether these classification errors are distributed evenly across the political orientation labels assigned by NewsGuard or only characterize some. To this end, we consider a random sample of domains for each of NewsGuard’s orientation and reliability labels and calculate the percentage of the domains whose reliability rating is misclassified by the LLM. In particular, we focus on a random sample of 40 domains per NewsGuard’s political orientation and rating, as it is the least populated among these groups. Then, we repeat this sampling procedure 10,000 times to obtain average misclassification frequencies.

The results in Fig. 2 show that classification errors are not uniformly distributed across the political spectrum. In particular, focusing on domains rated as “reliable” by NewsGuard, we observe that Right-leaning domains are classified by all models as “unreliable” substantially more often than the Center and Left-leaning, whose reliability appears to be overestimated with respect to NewsGuard.

Finally, beyond assessing the models’ performance in reliability classification, we measure how the political orientation labels they assign to news outlets compare against human evaluators. All three LLMs show strong agreement with human annotations, as seen in Supplementary Fig. S1. Comparing the political labels assigned by the models to those assigned by NewsGuard, we find a substantial overlap across the political spectrum for all three models. However, some differences can be observed due to NewsGuard employing fewer orientation labels than the models. This alignment is further confirmed by comparing the models’ political orientation assessments with the “Bias Rating” from MBFC, focusing specifically on strictly political labels.

2.2 Explaining Reliability Ratings with Keywords

We now investigate the main factors driving LLMs’ reliability ratings and how they relate to the content of a news outlet’s homepage. To achieve this, we analyze three distinct sets of keywords generated by the models for each outlet, alongside their reliability ratings and political orientation. By examining what keywords are used and how they relate to reliability and political orientation, we aim to gain further insights into the mechanisms these models employ to reach their reliability evaluation. Unlike human evaluators, LLMs do not explicitly follow predefined scoring guidelines, so exploring the patterns they exhibit when assigning reliability ratings is essential.

For all domains, each LLM is tasked to provide three types of keywords. The first set of keywords, referred to as “classification keywords”, reflects the model’s rationale behind its classification and summarizes its rating. The second set, “determinant keywords”, comprises terms extracted directly from the domain’s homepage that were critical for the model’s reliability judgment. The final set, “summary keywords”, includes terms that broadly summarize the contents of the domain’s homepage. Before analysis, we convert all keywords to lowercase. Importantly, we do not give the models any constraints on the number of keywords to output. By omitting this constraint, we can observe the typical number of keywords each model associates with a given input and examine whether this number varies between “reliable” and “unreliable” sources or across different models. Furthermore, imposing such a limit may hinder the explainability of each model’s reliability ratings by reducing their expressive power.

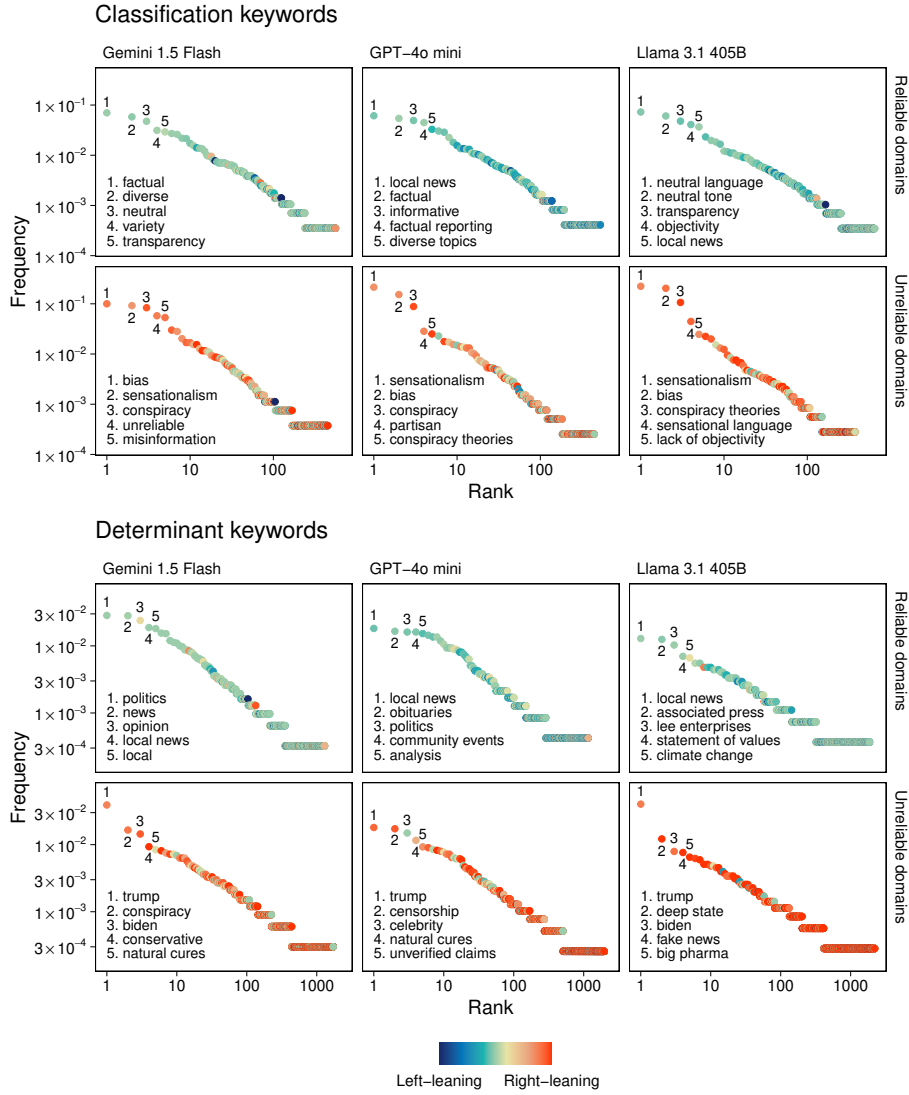


Fig. 3 Rank-frequency distributions of keywords used by each LLM to describe domains. Each panel presents the most frequently used classification and determinant keywords for Reliable and Unreliable domains. Only the five most common keywords per panel are labeled to enhance readability. The color gradient 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.

Using the political orientation label assigned by the models to each domain, we infer the political leaning of each keyword as the average political leaning of the domains it is associated with. To achieve this, we transform the political orientation labels into numerical values ranging from -1 to 1 , assigning -1 to Left, -0.5 to Center-Left, 0 to Center, 0.5 to Center-Right, and 1 to Right.

We construct separate rank-frequency distributions for each model, keyword type, and reliability rating to analyze the models’ keyword usage. A rank-frequency distribution calculates how often an element appears in a sample relative to its rank, where elements are ordered from most to least frequent. These distributions frequently exhibit a heavy-tailed behavior, characterized by a few elements dominating in frequency and the majority appearing rarely. This pattern, commonly observed in natural language studies, reflects the typical number of occurrences of words in a corpus of documents, where a few high-frequency terms account for the bulk of occurrences, and many others are used infrequently.

Fig. 3 displays the rank-frequency distributions of “classification” and “determinant” keywords obtained per model and reliability rating, revealing a consistent heavy-tailed behavior across all models and keyword types. This suggests that LLMs may rely on a core set of linguistic markers to evaluate reliability.

As shown in Fig. 3, classification keywords highlight key markers of a model’s reliability assessments. Reliable domains are frequently associated with terms denoting neutrality, transparency, and factual reporting. Llama also focuses on “neutral language” and “objectivity”, reinforcing the importance of tone and professional framing in its assessments. Conversely, unreliable domains are consistently associated with terms relating to sensationalism, bias, or conspiracy theories. Words like “misinformation”, “conspiracy”, and “partisan” frequently appear, reflecting the models’ alignment with human evaluative criteria for detecting unreliable sources. These findings indicate that LLMs develop structured linguistic heuristics, mirroring some aspects of human reliability evaluation.

Analyzing determinant keywords reveals further insights into the mechanisms driving the models’ classification. Reliable domains are frequently linked to editorial practices and institutional transparency. Notably, GPT-4o mini and Llama 3.1 emphasize “local news” as a relevant descriptor for reliability, suggesting that community-based reporting is perceived as an indicator of credibility. Unreliable domains, in contrast, are strongly associated with politically charged and controversial terms. Words such as “trump”, “biden”, “deep state”, and “fake news” dominate the descriptions of unreliable sources, indicating that highly politicized content correlates with lower reliability classifications.

Additionally, Fig. 3 shows that right-leaning terms appear more frequently in descriptions of unreliable sources, while neutral or left-leaning terms are more common in reliable sources. However, the presence of political keywords alone does not determine reliability. For example, “politics” frequently appears in descriptions of reliable and unreliable sources, suggesting that it is not the topic itself but how it is framed and presented that influences model assessments.

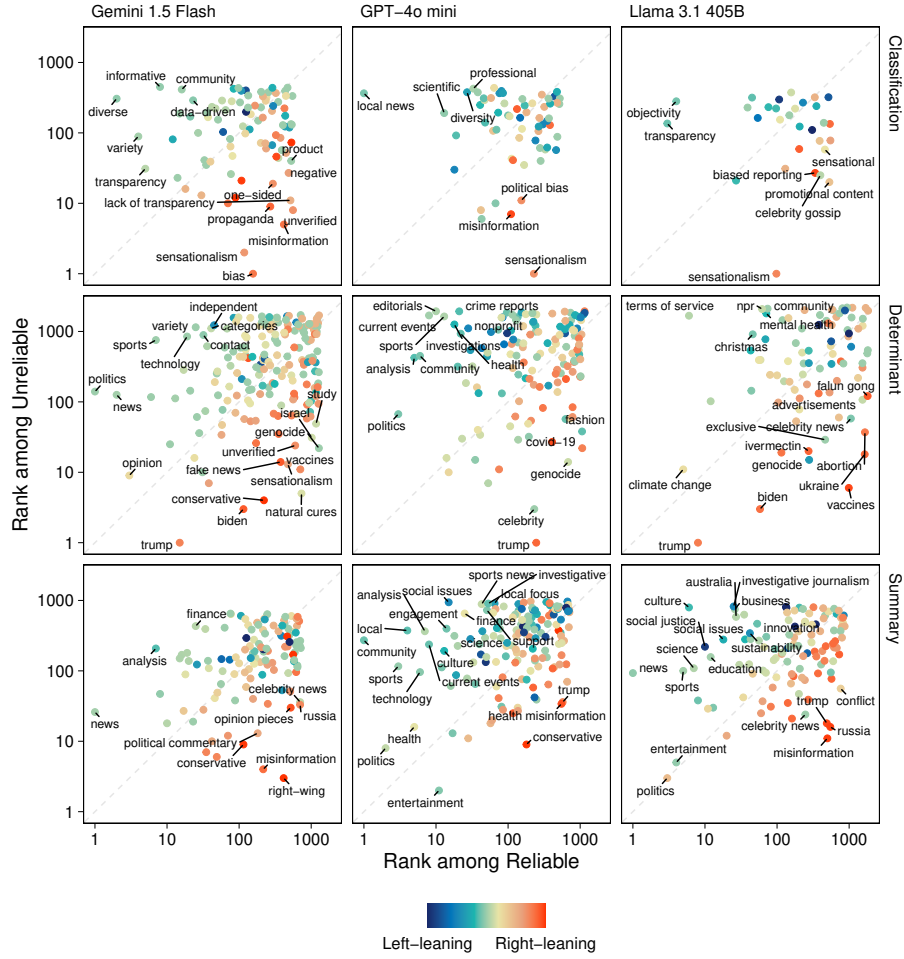


Fig. 4 Keywords’ rank among “reliable” and “unreliable” domains. We label only keywords sufficiently distant from the diagonal, meaning they are predominantly used to describe reliable or unreliable domains rather than being evenly distributed across both classifications. Additionally, we label the top 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.

Keywords used to describe both “reliable” and “unreliable” domains are presented in Fig. 4, which compares their ranks across the two classifications. In this visualization, the further a keyword is from the diagonal, the more characteristic it is of one of the two ratings.

When examining “classification” and “determinant” keywords, we observe that the difference in keyword usage between the two groups is apparent. Reliable classifications produce terms such as “local news”, “scientific”, “diverse” and “data-driven”. In contrast, when explaining unreliable classifications, the models utilize more controversial or politically charged terms, including politician names (e.g., “trump”, “biden”) as well as topics such as “genocide” and “vaccines”. On the other hand, “summary” keywords, which broadly describe the content of a domain’s homepage, tend to show substantial overlap between reliable and unreliable news outlets. This suggests that both types of news outlets cover similar general themes and reinforce the idea that the difference does not necessarily lie in the topics discussed but rather how those topics are framed and communicated. Additionally, some terms that do not inherently indicate reliability or unreliability appear consistently associated with one category over the other, hinting at underlying stylistic or contextual differences that influence model evaluations.

These findings suggest that LLMs do not simply categorize news outlets based on explicit criteria but instead rely on an implicit understanding of reliability, possibly shaped by their training data. Their assessments also appear to be guided by linguistic framing, recurring stylistic patterns, and contextual signals, rather than just the presence of specific factual claims. This raises important questions about how LLMs internalize and apply credibility heuristics and whether they construct their evaluative frameworks based on patterns observed in human discourse.

2.3 Agentic Framework for Investigating LLM Decision-Making

Our analysis shows that LLMs often produce reliability ratings that closely align with expert evaluations from NewsGuard and MBFC. This suggests that these models have developed internal heuristics that approximate human assessments, despite not having explicit access to structured evaluation criteria. However, a critical question remains: how do LLMs actually reach these conclusions?

A key observation emerges when we prompt the models with nothing more than the URL of a domain, without any extracted content from its homepage [29]. Even in this minimal setting, the models generate reliability ratings that broadly align with those assigned by human evaluators. For instance, Gemini achieves an F1-score of 0.78—slightly lower than the 0.85 obtained with the HTML homepage—and GPT an F1-score of 0.77, instead of 0.79. This raises an important issue: are LLMs actively analyzing information, or are they simply recalling prior associations learned during training? If models can classify a news source without even seeing its content, it suggests that their evaluations may be shaped more by pre-existing knowledge than by real-time evaluation. This makes it difficult to determine whether their classifications are based on an actual assessment of the content or just on patterns they have already internalized.

To address this, we introduce a structured agentic workflow designed to probe how LLMs interpret and classify news outlets, which we test with Gemini and GPT on a sample of 71 news outlet domains. Rather than treating these models as black boxes that output a binary reliability label, we create an agent that can actively gather and analyze information before providing a final assessment. We equip this agent with three

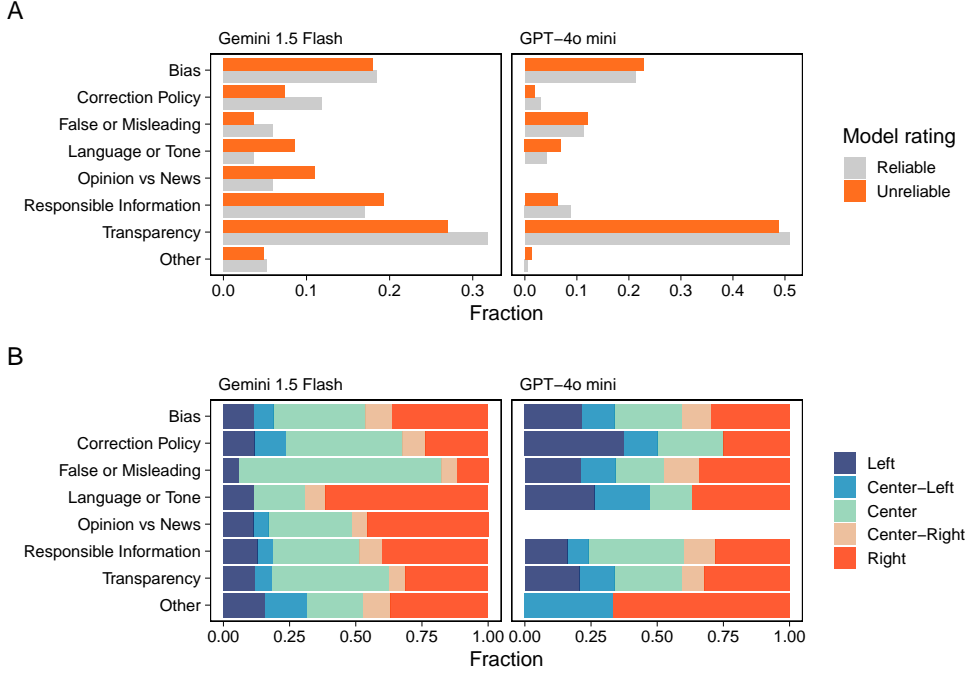


Fig. 5 LLMs’ reliability criteria against reliability and political orientation classification. (A) Frequency of each reliability criterion among news outlets rated as reliable and unreliable by the LLMs. (B) Proportion of domains of political orientation, as classified by the LLMs, for each criterion. We note that Opinion vs News is missing from GPT’s panels as no criteria provided by the model could be associated with it.

main tools: the ability to scrape webpage content directly from a domain’s homepage or subpages found within, perform web searches to retrieve additional information, and prompt another LLM for text analysis. By observing how the agent uses these tools and in what order, we can gain insights into what information LLMs prioritize, how they refine their assessments, and whether their decision-making process resembles human reasoning. This workflow shifts the focus from whether LLMs can classify news sources to how they reach those classifications, by examining them as evaluators that are capable of following multi-step processes to reach conclusions.

In the prompt, we ask the agent to produce a detailed and structured report of the actions it performs. These actions can be summarized as a series of steps, each characterized by several attributes: the criterion the model is assessing, the specific news articles analyzed (if any), the prompt used to query a second LLM for content analysis, and details of any web search performed during that step. Tracking the agent’s steps toward its final classification allows us to distinguish between different strategies the models might employ. For instance, if an agent frequently consults external sources or analyzes specific articles, this suggests it is actively seeking verification rather than relying purely on prior knowledge. By studying these behavioral

patterns, we can better investigate the mechanisms LLMs use to evaluate reliability and understand whether they exhibit adaptive capabilities.

We start by manually inspecting the reliability criteria that the models provided and thus evaluated. Overall, we observe that the same criteria are consistently assessed, although titled with different wording. Therefore, we manually annotate each so they fit into eight more generic criteria: Bias, Correction Policy, False or Misleading, Language or Tone, Opinion vs News, Responsible Information, Transparency, and Other. Then, we investigate how their use relates to the final reliability rating and political orientation label output by the model.

Figure 5A shows how commonly each criterion is employed to evaluate domains classified as reliable/unreliable. The resulting frequencies show that most criteria are generally evaluated with the same frequency for both reliable and unreliable domains, by both models. Some exceptions include ‘Correction Policy’, which both models more often employ for reliable news outlets, and ‘Language or Tone’, which is more common among the unreliable. Overall, Gemini shows greater variability in the choice of criteria compared to GPT. Among these, ‘Transparency’ is the most commonly evaluated criterion by both models, particularly by GPT.

In Fig. 5B, we focus on the relationship between reliability criteria and political orientation. In this analysis, the differences between models and among criteria are more apparent. For instance, while ‘Language or Tone’ is predominantly used by Gemini to evaluate right-leaning domains, GPT employs it more uniformly across the political spectrum. To a lesser extent, similar considerations apply to other criteria such as ‘Fake or Misleading’, which Gemini uses substantially more often for center-leaning domains, unlike GPT, and ‘Correction Policy’.

Concerning how often the models actively seek additional information, we note slight differences between the two agents. In our experiments, Gemini focused on specific articles rather than just the whole homepage for 21% of the domains, while GPT for the 33%. Conversely, the models used with similar frequency the search tool to retrieve further information from the Web, with Gemini employing it for 23% of the domains and GPT for the 20%.

Even if on a small sample, our results show that this approach offers a clearer picture of LLMs’ decision-making process, providing a foundation for future studies on explainability, adaptability, and the extent to which these models replicate human-like reasoning. This framework also opens the door to direct comparisons between LLMs and human decision-making in the same task. By designing experiments where human participants are given similar tools—search engines, document retrieval, and rating prompts—we can analyze how human evaluators approach credibility assessments. Comparing their behavior to that of the agent helps us understand the similarities and differences in how humans and LLMs prioritize and process information. Do humans rely more on search results and external verification, while LLMs default to internalized knowledge? Do both exhibit patterns of confirmation bias, favoring information that aligns with prior beliefs? By structuring the problem in a way that allows side-by-side analysis, it is possible to go beyond simple accuracy comparisons and begin to explore the deeper question of how LLMs and humans make complex decisions in uncertain environments.

3 Conclusions

This study investigates how Large Language Models (LLMs) evaluate news outlet reliability, comparing their judgments to structured human benchmarks provided by NewsGuard and Media Bias Fact Check (MBFC). While prior research has often treated LLMs as potential tools for automating reliability assessments, our findings suggest a broader question: how do these models construct their reasoning, and how does it compare to human evaluative frameworks?

Our results reveal a strong alignment between LLM classifications and human expert ratings, particularly in identifying “unreliable” sources. The models consistently flag domains associated with conspiracy theories, sensationalism, and bias, echoing key criteria used in expert evaluations. However, their classification of “reliable” sources is less consistent, revealing differences in how they interpret credibility when contextual signals may be limited. Interestingly, when analyzing how errors in reliability classification are distributed across the political spectrum, we find that right-leaning news outlets tend to be consistently misclassified as “unreliable”, while the center and the right-leaning as “reliable”. These results raise critical questions about whether LLMs inherit biases from training corpora, how these biases interact with structured evaluative frameworks, and whether their reasoning patterns reflect genuine assessment or learned associations. This is further corroborated by the models producing similar ratings even when prompted only with domain URLs, rather than the scraped domain homepage.

By analyzing keyword usage via their rank-frequency distributions, we further explore how LLMs operationalize reliability. Our findings indicate that all models consistently use certain terms to explain their ratings, as shown by the characteristic heavy-tailed behavior of the distributions. Overall, we find that keywords referring to local news, factual reporting, or neutral language are typically associated with “reliable” domains. Conversely, “unreliable” domains are often characterized by terms relating to sensationalism, controversies, or bias, which reflect commonly used markers employed by human evaluators to identify low credibility sources. Additionally, our results show that keywords that summarize the contents of the webpage are often common to both reliable and unreliable news outlets, pointing toward the role of tone and framing in the models’ reliability evaluations.

Moving beyond simple classification, we introduce an agentic workflow to investigate how LLMs structure their evaluation procedure when given tools to actively seek information. By equipping an AI agent with a webpage scraper, a search engine, and the possibility to query a LLM for content analysis, we gain a more granular view of how these models reach their conclusions. Analyzing whether the criteria the models decide to evaluate change with the final reliability rating, we find that ‘Transparency’ and ‘Bias’ emerge as the more commonly evaluated criteria for both reliable and unreliable domains, while ‘Language or Tone’ or ‘Correction Policy’ are not as employed. Overall, for both Gemini and GPT we observe no substantial differences between reliable and unreliable news outlets in terms of what criteria are prioritized. Conversely, discrepancies emerge when exploring the relationship between the criteria and the final political orientation label. In this case, certain criteria such as ‘Language

or Tone’ and ‘False or Misleading’ for Gemini, and ‘Correction Policy’ for GPT, are more often employed for specific orientations.

Future research should expand this framework by incorporating direct human comparisons, examining how real-world evaluators navigate the same task, and testing whether LLM-based agents can develop more autonomous, context-aware decision strategies.

Ultimately, this study reframes LLMs not merely as automated credibility classifiers but as windows into the cognitive structures underlying both human and machine reasoning. Their evaluations do not just reflect computational heuristics; they offer insight into the challenges of operationalizing credibility in an information ecosystem shaped by competing narratives, institutional frameworks, and algorithmic decision-making. By unpacking their reasoning processes, we move closer to understanding the extent to which LLMs simulate structured evaluation, whether they can adapt to new decision-making environments, and how they compare to human cognitive strategies in complex judgment tasks.

4 Methods

4.1 Data collection and pre-processing

All data was collected by downloading the HTML homepages of domains rated by NewsGuard as “reliable” or “unreliable”, using the `requests` library available on Python. 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 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. The resulting document almost exclusively includes textual content, particularly related to news or information about the outlet. While domain-specific pre-processing 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 final dataset employed includes a total of 2,302 domains with at least 100 characters of content, with 1,196 rated by NewsGuard as “reliable” and the remaining 1,106 as “unreliable”.

4.2 Model prompting

We follow a zero-shot, closed-book approach to prompting for all three LLMs we experiment with, by providing no example classifications and no definitions of “reliability”, respectively. 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

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

instructions concerning their output, specifying both the required content and its structure. Figure 6 reports the exact prompt used for all models.

Gemini 1.5 Flash and GPT-4o mini were prompted by performing calls to the official API endpoints made available by Google and OpenAI, respectively. LLaMA 3.1 405B was prompted by requests to SambaNova Cloud, a third-party online service that allows fast inference with LLMs. However, since LLaMA’s weights are available for download, local inference is also a viable option.

Queries sent to GPT and Llama were truncated to ensure they fit within the models’ context length (128,000 tokens for both), which is the maximum number of tokens they can process at once. Specifically, the scraped webpages provided to GPT and Llama 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”.

When evaluating the LLMs’ ability to classify news outlets using only their domain names, we slightly altered the prompt in Fig. 6 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’.

Given the URL of a news outlet, you classify it as either 'reliable' or 'unreliable' based on its content. Focus on the entire homepage. Don't use any prior knowledge you may have about the news outlet.

This is what you are allowed to do:

- scrape web pages using the `webpage_scraper_tool`
- scrape specific articles using the `webpage_scraper_tool`
- use the `web_search` tool if you fail to scrape a page
- use the `web_search` tool to find information not contained in the scraped pages
- use the `llm_language_analysis` tool to analyze content requiring NLP techniques
- use the `llm_language_analysis` tool to analyze the search results

This is what you are forbidden to do:

- use the `web_search` tool to find human reliability ratings or opinions
- simulate content, URLs, or search results
- use examples
- use prior knowledge you have about the news outlet
- analyze the website's structural elements, such as its layout or navigation

This is what you must do:

1. decide how many and which criteria of 'news outlet reliability' you must evaluate and save them in an ordered list
2. then, evaluate them one by one and in order with the tools you are provided, which are the most sophisticated available
3. after each step ask yourself: can I use the `webpage_scraper_tool` or `web_search` tool to improve my assessment?

Rely on the `webpage_scraper_tool` and `web_search` tool as fallback. Avoid using placeholders or simulated examples.

It is a step-by-step process. Use the information gathered at each step to help you with the next.

Fig. 7 First part of the prompt used for the LLM agent, where instructions about the task are provided.

4.3 Agentic workflow

We implemented the agentic workflow for outlet reliability classification with `smolagents`, a library for Python developed by Hugging Face. In particular, we implement a so-called Code Agent, which is an agent that performs actions via code writing [40]. By allowing an agent to write its actions in code and providing it with a set of tools that can be utilized via code, we obtain a model that is capable of designing

At the end, your output must be a JSON with the following attributes

- 'webpage_url'
- 'rating_report': a report containing a summary of your final assessment
- 'reliability_criteria_evaluated': ordered list of reliability criteria evaluated
- 'rating': your final reliability rating ('reliable' / 'unreliable')
- 'orientation': political orientation of the news outlet ('left', 'center-left', 'center', 'center-right', 'right')
- 'additional_comment': if you have any additional comments that are related to the rating (such as your inability to perform certain steps) then put them here
- 'steps': a list of all steps you made, each reported as a JSON with the following attributes
 - 'step_number'
 - 'reliability_criterion': which criterion from your list are you evaluating in this step
 - 'step_scope': what is the step aimed at
 - 'step_reason': why have you decided to make this step
 - 'step_outcome': explain in natural language the results obtained at this step
 - 'analyzed_article': if at this step you analyze a specific article or list of articles, report the list of URLs here
 - 'llm_prompt' (question asked to the LLM, without the content you have asked to analyze)
 - 'search_results_analyzed': if during this step you have analyzed search results, then provide them here as a list, each reported as JSON with attributes 'search_query', 'webpage_urls' (list), 'reason_for_checking_search_result', 'llm_summary_of_results'

Based on all instructions I provided you, scrape and then give me a reliability rating of the news outlet at this URL: {URL of the news outlet's homepage}

You have no time constraints. Take as long as you need. You have all the sophisticated tools and information you need.

Fig. 8 Second part of the prompt used for the LLM agent, where instructions about the output format and the URL of the news outlet to classify are provided.

and executing a workflow, in our case aimed at news outlet reliability classification. In particular, we provide the agent with these three tools:

- `webpage_scraper_tool`, which is a function that downloads a webpage into a Markdown-formatted document using the `markitdown` Python library
- `llm_language_analysis`, which is a function that prompts a LLM (the same model behind the agent)

- `web_search`, which is a tool built in `smolagents` that retrieves web search results using the DuckDuckGo API, which we limit to the first 20 entries

The exact prompt we provide the agent is reported in Fig. 7 and 8, whereas we leave unchanged the default system instructions implemented in `smolagents`. Finally, we limit the agent’s workflow to a maximum of 20 steps, to prevent infinite loops or an excessive number of calls to LLMs.

Acknowledgements. SERICS (PE00000014) under the NRRP MUR program funded by the European Union - NextGenerationEU, project CRESO from the Italian Ministry of Health under the program CCM 2022, PON project “Ricerca e Innovazione” 2014-2020, and PRIN Project MUSMA for Italian Ministry of University and Research (MUR) through the PRIN 2022.

References

- [1] Holton, A.E., Coddington, M., Lewis, S.C., De Zuniga, H.G.: Reciprocity and the news: The role of personal and social media reciprocity in news creation and consumption. *International journal of communication* **9**, 22 (2015)
- [2] Khan, M.L.: Social media engagement: What motivates user participation and consumption on youtube? *Computers in human behavior* **66**, 236–247 (2017)
- [3] Avalle, M., Di Marco, N., Etta, G., Sangiorgio, E., Alipour, S., Bonetti, A., Alvisi, L., Scala, A., Baronchelli, A., Cinelli, M., *et al.*: Persistent interaction patterns across social media platforms and over time. *Nature* **628**(8008), 582–589 (2024)
- [4] Kubin, E., Von Sikorski, C.: The role of (social) media in political polarization: a systematic review. *Annals of the International Communication Association* **45**(3), 188–206 (2021)
- [5] Budak, C., Nyhan, B., Rothschild, D.M., Thorson, E., Watts, D.J.: Misunderstanding the harms of online misinformation. *Nature* **630**(8015), 45–53 (2024)
- [6] Lazer, D.M., Baum, M.A., Benkler, Y., Berinsky, A.J., Greenhill, K.M., Menczer, F., Metzger, M.J., Nyhan, B., Pennycook, G., Rothschild, D., *et al.*: The science of fake news. *Science* **359**(6380), 1094–1096 (2018)
- [7] Del Vicario, M., Bessi, A., Zollo, F., Petroni, F., Scala, A., Caldarelli, G., Stanley, H.E., Quattrociocchi, W.: The spreading of misinformation online. *Proceedings of the national academy of Sciences* **113**(3), 554–559 (2016)
- [8] Gallup, K.: Indicators of news media trust. John S. and James L. Knight Foundation Miami (2018)
- [9] Newman, N., Fletcher, R., Levy, D., Nielsen, R.: The Reuters Institute Digital News Report 2018

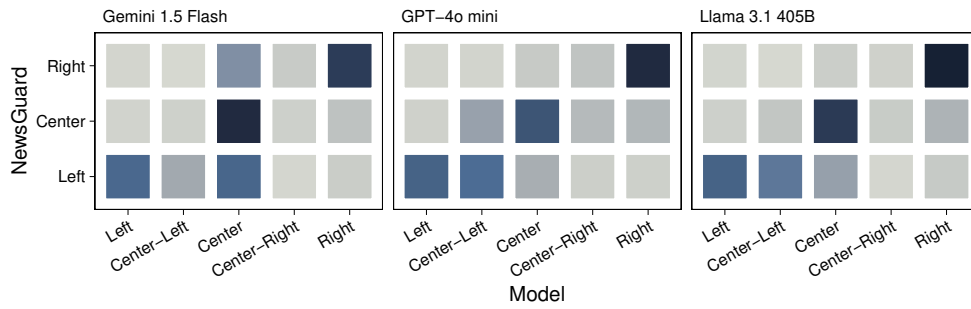
- [10] Bail, C.A., Argyle, L.P., Brown, T.W., Bumpus, J.P., Chen, H., Hunzaker, M.F., Lee, J., Mann, M., Merhout, F., Volfovsky, A.: Exposure to opposing views on social media can increase political polarization. *Proceedings of the National Academy of Sciences* **115**(37), 9216–9221 (2018)
- [11] Falkenberg, M., Galeazzi, A., Torricelli, M., Di Marco, N., Larosa, F., Sas, M., Mekacher, A., Pearce, W., Zollo, F., Quattrocioni, W., *et al.*: Growing polarization around climate change on social media. *Nature Climate Change* **12**(12), 1114–1121 (2022)
- [12] Cinelli, M., Quattrocioni, W., Galeazzi, A., Valensise, C.M., Brugnoli, E., Schmidt, A.L., Zola, P., Zollo, F., Scala, A.: The covid-19 social media infodemic. *Scientific reports* **10**(1), 1–10 (2020)
- [13] Kim, L., Fast, S.M., Markuzon, N.: Incorporating media data into a model of infectious disease transmission. *PloS one* **14**(2), 0197646 (2019)
- [14] Metzger, M.J., Flanagin, A.J.: Psychological approaches to credibility assessment online. *The handbook of the psychology of communication technology*, 445–466 (2015)
- [15] Rieh, S.Y.: Credibility and cognitive authority of information. *Encyclopedia of library and information sciences* **1**(1), 1337–1344 (2010)
- [16] Metzger, M.J., Flanagin, A.J., Medders, R.B.: Social and heuristic approaches to credibility evaluation online. *Journal of communication* **60**(3), 413–439 (2010)
- [17] NewsGuard Technologies: Rating Process and Criteria. <https://www.newsguardtech.com/ratings/rating%20-process-%20criteria/>. Accessed: 2024-11-26
- [18] Lühring, J., Metzler, H., Lazzaroni, R., Shetty, A., Lasser, J.: Best practices for source-based research on misinformation and news trustworthiness using newsguard. *Journal of Quantitative Description: Digital Media* **5** (2025)
- [19] Herrero-Beaumont, E.: Emerging transparency systems for news governance to protect media independence and credibility in the digital infosphere. *Communication Law and Policy* **27**(3-4), 220–249 (2022)
- [20] Aslett, K., Guess, A.M., Bonneau, R., Nagler, J., Tucker, J.A.: News credibility labels have limited average effects on news diet quality and fail to reduce misperceptions. *Science advances* **8**(18), 3844 (2022)
- [21] OpenAI: GPT-4o Technical Report (2024). <https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/>
- [22] Team, G., Georgiev, P., Lei, V.I., Burnell, R., Bai, L., Gulati, A., Tanzer, G.,

- Vincent, D., Pan, Z., Wang, S., et al.: Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. arXiv preprint arXiv:2403.05530 (2024)
- [23] AI, M.: LLaMA 3.1: Advancements in Open-Weight LLMs (2024). <https://ai.meta.com/blog/meta-llama-3-1/>
- [24] Törnberg, P.: Large language models outperform expert coders and supervised classifiers at annotating political social media messages. *Social Science Computer Review*, 08944393241286471 (2024)
- [25] Gilardi, F., Alizadeh, M., Kubli, M.: Chatgpt outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences* **120**(30), 2305016120 (2023)
- [26] Wu, P.Y., Nagler, J., Tucker, J.A., Messing, S.: Large language models can be used to estimate the latent positions of politicians. arXiv preprint arXiv:2303.12057 (2023)
- [27] Chiang, C.-H., Lee, H.-y.: Can large language models be an alternative to human evaluations? arXiv preprint arXiv:2305.01937 (2023)
- [28] Krugmann, J.O., Hartmann, J.: Sentiment analysis in the age of generative ai. *Customer Needs and Solutions* **11**(1), 3 (2024)
- [29] Yang, K.-C., Menczer, F.: Large language models can rate news outlet credibility. arXiv preprint arXiv:2304.00228 (2023)
- [30] Hoes, E., Altay, S., Bermeo, J.: Leveraging chatgpt for efficient fact-checking. *PsyArXiv*. April **3** (2023)
- [31] Quelle, D., Bovet, A.: The perils and promises of fact-checking with large language models. *Frontiers in Artificial Intelligence* **7**, 1341697 (2024)
- [32] Hernandez, R., Corsi, G.: Llms left, right, and center: Assessing gpt’s capabilities to label political bias from web domains. arXiv preprint arXiv:2407.14344 (2024)
- [33] Hu, T., Kyrychenko, Y., Rathje, S., Collier, N., Linden, S., Roozenbeek, J.: Generative language models exhibit social identity biases. *Nature Computational Science*, 1–11 (2024)
- [34] Yax, N., Anlló, H., Palminteri, S.: Studying and improving reasoning in humans and machines. *Communications Psychology* **2**(1), 51 (2024)
- [35] Motoki, F.Y.S., Neto, V.P., Rodrigues, V.: Assessing political bias and value misalignment in generative artificial intelligence (2024)
- [36] Strachan, J.W., Albergo, D., Borghini, G., Pansardi, O., Scaliti, E., Gupta, S.,

- Saxena, K., Rufo, A., Panzeri, S., Manzi, G., et al.: Testing theory of mind in large language models and humans. *Nature Human Behaviour*, 1–11 (2024)
- [37] Coppolillo, E., Manco, G., Aiello, L.M.: Unmasking conversational bias in ai multiagent systems. arXiv preprint arXiv:2501.14844 (2025)
- [38] Safdari, M., Serapio-García, G., Crepy, C., Fitz, S., Romero, P., Sun, L., Abdulhai, M., Faust, A., Matarić, M.: Personality traits in large language models. arXiv preprint arXiv:2307.00184 (2023)
- [39] Methodology - Media Bias Fact Check. <https://mediabiasfactcheck.com/methodology/>
- [40] Wang, X., Chen, Y., Yuan, L., Zhang, Y., Li, Y., Peng, H., Ji, H.: Executable code actions elicit better llm agents. arXiv preprint arXiv:2402.01030 (2024)

Supplementary information

A



B

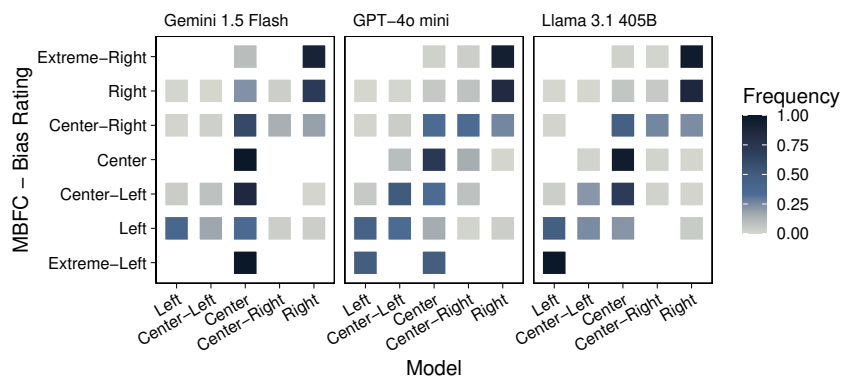


Fig. S1 LLMs' classification against expert human evaluators. All of the three models show to be able to correctly predict the political leaning of the news outlet they are analyzing. (A) Comparing the answers of the models with the NewsGuard labels, the accuracy is high, with a few error on some bias news outlet, classified as center. (B) Using as ground truth the labels of MBFC the accuracy is the same, with a higher value of accordance using LLaMA 3.1 405B.