Media and Journalism Research Center

Evaluating Free AI Chatbots on Civic and Human-Rights Topics:

A 15-Country Expert Study

Written by Marius Dragomir

Research team: Charles Autheman, Jorge Bossio, Luis Miguel Carriedo, Swati Dey, Nurma Fitrianingrum, Rafael da Silva Paes Henriques, Hager Hesham, Silviya Irshad, Guillermo Mastrini, Akira Oikawa, Yotaro Okamoto, Edward Pittman, Ruth Reis, Norina Solomon, Martin Vaz-Álvarez, Andreina Zuñiga

Feb 2026

About

MJRC

MEDIA AND JOURNALISM RESEARCH CENTER

The Media and Journalism Research Center (MJRC) is an independent research and policy center that examines how media, technology, and information governance shape public life and democratic accountability. One of its flagship projects is mapping the behavior of AI companies and chatbots (how they are owned, financed, and governed) to understand how they are changing the media and communication field, including new informational inequalities and patterns of visibility and invisibility across regions and topics. Through cross-country projects on AI systems, newsrooms, and human-rights reporting, MJRC seeks to help journalists and civil-society actors adopt AI responsibly and in ways that strengthen independent reporting and public knowledge.

MARIUS DRAGOMIR

Dr. Marius Dragomir is an award-winning media scholar and journalist, serving as Director of the Media and Journalism Research Center (MJRC). He is also a researcher at the University of Santiago de Compostela (Spain) and has served as a visiting professor at Central European University (CEU) in Vienna, where he has taught journalism, research design, advocacy, and policy engagement.

Dr. Dragomir has designed and led numerous regional and global comparative research initiatives. These include the Media Influence Matrix, examining power relations and undue influence in news media across nearly 60 countries; the Global Media Finances Map, a comprehensive database of major media companies worldwide focused on ownership and funding transparency; the State Media Monitor, which tracks developments in state-controlled media across more than 170 countries; and the Media Capture Monitoring Report, a collaborative project with the International Press Institute. Prior to founding MJRC, Dr. Dragomir directed the Center for Media, Data and Society (CMDS) at CEU and managed the global research and policy portfolio for the Program on Independent Journalism at the Open Society Foundations in London. He was lead editor of the Mapping Digital Media project (covering 56 countries) and principal writer/editor of Television Across Europe, a landmark comparative study of broadcast policies and independence in 20 European states. He has published extensively for international organizations including UNESCO, OSCE, the Council of Europe, the Open Society Foundations, and the World Bank. His research focuses on media capture, ownership transparency, public service media governance, state influence mechanisms, and the evolving relationship between media systems and democratic accountability.

Table of contents

Introduction

Large language models (LLMs) increasingly shape how journalists, researchers, and the general public access information on high-stakes topics such as elections, human rights, identity politics, historical memory, and public services. People around the world now turn to free AI assistants to understand how their electoral systems work, what protections exist for vulnerable groups, or how public services like healthcare are organized.

Most existing evaluations of LLMs, however, focus on single-country settings, synthetic benchmarks, or narrow political spectra rather than cross-country expert assessments of real, free-text answers to sensitive societal questions. They rarely examine how public-facing systems describe the same contested topics across multiple languages and political regimes, nor do they systematically measure differences across countries and models in dimensions that matter for journalism and civic life, such as neutrality, factual accuracy, balance, and tone.

This study by the Media and Journalism Research Center (MJRC) addresses that gap by examining how eight widely used chatbots, namely ChatGPT, Gemini, Perplexity, Claude, Copilot, DeepSeek, Meta AI, and Grok, describe politically and socially sensitive topics in 15 countries that span diverse political regimes, cultural contexts, and levels of press freedom. We focus on the basic, free versions of these assistants, the tools that billions of users worldwide rely on daily to understand complex, contested realities. Using human expert-annotated scores on Neutrality, Factuality, Balance, and Tone, we assess strengths and weaknesses across four broad issue clusters: democratic institutions and power, rights and identity, social policy and welfare, and historical memory and corruption.

We ask three guiding questions. First, how do free AI assistants perform on high-stakes civic and human-rights topics across diverse countries and languages when evaluated by domain human experts? Second, how do their strengths and weaknesses vary across issue types, from present-day service systems to contested histories of violence and corruption? Third, how do differences across bots and countries map onto existing information inequalities, and what does this imply for journalism, human-rights advocacy, and public knowledge more broadly?

Methodology

Country and topic selection

MJRC mapped 15 countries across different regions, political regimes, and human-rights environments: Argentina, Brazil, Canada, Colombia, Egypt, France, India, Indonesia, Japan, Mexico, Peru, Romania, Saudi Arabia, Spain, and the United Kingdom (UK). For each country, MJRC identified two or three topics of high societal relevance and political sensitivity, such as elections, LGBTQ+ rights, caste, secularism, immigration, or historical memory.

This process yielded 30 distinct country–topic pairs, each anchored in a short expert baseline and a set of chatbot outputs for evaluation:

- Argentina: military dictatorship; "memory, truth, and justice."
- Brazil: Amazon/Cerrado/Pantanal (environmental degradation and policy); equity and social recognition.
- Canada: Indigenous sovereignty and land rights; Quebec French-language laws.
- Colombia: corruption; human rights and violence.
- Egypt: elections and power dynamics; LGBTQ+ rights.
- France: secularism; political extremism.
- India: bulldozer justice; caste system.
- Indonesia: Jokowi clinging to power; West Papua.
- Japan: constitutional reform; female emperor.
- Mexico: reform of the judiciary; sexual diversity.
- Peru: ethnic and cultural diversity; Peruvian historical memory.
- Romania: healthcare system; drug use.
- Saudi Arabia: elections and power dynamics; governance and politics.
- Spain: hate crimes; political corruption.
- United Kingdom: National Health Service (NHS); immigration.

Taken together, these 30 country–topic pairs span democratic institutions and power, rights and identity, social policy and welfare, and historical memory and corruption, reflecting our focus on media, governance, and human-rights issues.

Expert roles and evaluation design

The design deliberately separates content production from evaluation through a two-expert structure.

Set 1 consists of country and topic experts (authors), typically local or thematic specialists, who drafted or journalistically verified a concise, one-paragraph baseline text ("expert text") for each country–topic pair. These baselines summarize the issue in a style suitable for research or institutional use and serve as reference descriptions, not as competing texts to be scored.

Set 2 consists of evaluation experts (annotators), a different group of experts with regional and/or thematic competence who did not participate in drafting the baselines. Using the baselines for orientation and context, Set 2 evaluated how chatbots described the same topics and scored only the chatbot outputs. For internal calibration, some evaluators also scored the baselines, but this study reports and analyzes only the chatbot scores.

This two-expert design reduces self-evaluation bias and aligns with emerging best practices in human expert-annotated LLM evaluation.

Prompting and model selection

For each country–topic pair, evaluators posed a set of short, neutral prompts to each model, designed to elicit brief descriptive paragraphs. Prompts were issued in English and in relevant local languages (e.g., Arabic for Egypt, Spanish for Mexico and Peru, Portuguese for Brazil, French for France and Quebec, Hindi for India, Indonesian for Indonesia, etc.). Across all models, topics, and languages, evaluators issued on the order of hundreds of prompts per topic, generating a rich corpus of outputs for scoring and comparison.

The eight chatbots tested were ChatGPT, Gemini, Perplexity, Claude, Copilot, DeepSeek, Meta AI, and Grok. All models were queried in their free, publicly accessible versions during 2025, without ideological personas or custom system prompts, to reflect typical user behavior.

Scoring rubric and dimensions

Evaluators scored each chatbot paragraph on four dimensions using a 1–5 scale:

- Neutrality (1–5): the degree to which the text avoids normative, partisan, or advocacy language and instead uses descriptive, analytical phrasing.
- Factuality (1–5): the accuracy and precision of key facts, figures, dates, actors, and causal relationships, as judged against human expert knowledge and the public record.
- Balance (1–5): the inclusion and fair treatment of relevant sides, actors, or dimensions, and the avoidance of one-sidedness or motivated omission.
- Tone (1–5): the appropriateness of style for professional, institutional, or comparative research use (for example, restrained and analytic versus polemical or emotive).

For each chatbot paragraph, evaluators also computed an Overall score as the arithmetic mean of the four dimensions. Alongside numeric scores, they provided short comparative notes explaining where a chatbot converged or diverged from the baseline (for example, "more abstract, fewer dates and actors"; "richer on legal mechanisms, less on local political context"; "maintains high neutrality but omits concrete scandals").

Limitations

This study has several limitations that qualify its findings.

Time window and model churn. The evaluations capture model behavior during specific months in 2025 and specific model versions. LLMs and their safety layers are continuously updated, often without public change logs, and the default "basic" versions exposed to free users may change abruptly as providers roll out new models, filters, or retrieval systems. The scores reported here are therefore a snapshot, not a fixed property of any given brand.

Prompting style and interaction scope. Although the project does not rely on a single prompt per topic, but on hundreds of prompts per topic across all bots and languages, the interaction style remains relatively constrained. Evaluators use short, neutral, informational prompts designed to elicit brief descriptions and repeat them in English and local languages. The study thus illuminates how chatbots respond to simple, good-faith information requests in multiple languages, but not how they behave under adversarial prompting, extended chat histories, or highly contextualized user personas.

Subjectivity in human expert scoring and topic selection. Despite the two-sets-of-experts design, scoring inevitably reflects human judgment. A different evaluation panel might rate Neutrality or Balance differently, particularly on normatively loaded issues such as secularism, political extremism, caste, West Papua, or historical memory. Topics were purposely selected to match our focus on media, governance, and human rights, rather than randomly sampled from the space of all possible queries. The study therefore maps chatbot performance on a strategic subset of sensitive issues, not an exhaustive catalog of all LLM behavior.

Model and provider coverage. The study covers eight major general-purpose chatbots in their free, public configurations, but does not assess paid tiers, specialized retrieval-augmented tools, or locally developed models deployed in some newsrooms or civic-tech projects. Results thus speak to what ordinary users see when they open a default assistant, not to the full diversity of AI tools in professional settings.

We address these limitations in several ways. To temper the effects of model churn over time, we concentrated data collection within clearly defined months in 2025, documented version identifiers where possible, and applied the same protocol to all systems in that window so that differences are primarily cross-model and cross-topic, not artifacts of staggered sampling. To reduce prompt-design bias, we used large prompt sets per topic, varied them across languages and minor phrasings, and aggregated scores at the level of country–topic pair rather than drawing conclusions from single prompts or isolated answers. To mitigate subjectivity in human expert scoring, we separated authorship and evaluation into two human expert sets, used a shared rubric with detailed anchor descriptions for each score from 1 to 5, and focused our analysis on robust patterns that recur across countries, topics, and bots, rather than on marginal score differences on any one item. Finally, while we restricted this phase to free, public assistants, we explicitly interpreted the findings as a baseline against which ongoing and future MJRC work on paid and enterprise-grade systems can be compared, allowing us to see where access to more capable models could narrow, or where it might widen, the informational gap between well-resourced users and the wider public.

Findings

Overall patterns across countries and bots

Across the 15 countries, chatbots score consistently high on Neutrality and Tone. Country averages typically range between 3.75 and 4.69 for Neutrality and between 4.00 and 4.56 for Tone. The United Kingdom, Spain, and Japan sit at the top end (UK: Neutrality 4.69, Tone 4.56; Spain: 4.38, 4.38; Japan: 4.12, 4.12), while Argentina and Indonesia remain slightly lower but still within a high band (Argentina: 3.75, 4.00; Indonesia: 3.75, 4.00). For a typical user, this means that on most topics and in most countries, free chatbots tend to respond in an institutional, descriptive style rather than in polemical or overtly partisan language.

Factuality and Balance show more variation. Brazil, Egypt, Canada, and Japan achieve the strongest country-level Factuality scores (4.38, 4.38, 4.25, and 4.25, respectively), while Mexico, Romania, and Colombia are somewhat lower (3.75–3.81). Balance is highest in Spain (3.94) and Brazil (4.12), and lowest in Peru (3.38), France (3.50), and India (3.50), suggesting that descriptions of some Global South and identity-sensitive topics tend to under-represent certain perspectives or dimensions. At the bot level, Grok and Gemini stand out with the highest Overall averages (4.59 and 4.35), followed by ChatGPT and Perplexity (4.10 and 4.03), while Meta AI and DeepSeek trail (3.63 and 3.78).

Taken together, the corpus suggests a recurring pattern. When asked baseline descriptive prompts, sometimes in English, sometimes in local languages, on high-stakes civic topics, free public chatbots generally provide neutral, coherent overviews that are factually solid at a structural level but may lack the finer empirical detail (dates, percentages, named actors, recent crises) that human experts view as central to current debates. Users who rely on these systems gain access to an adequate picture of institutions and rights frameworks, but not always to the full depth of evidence and contestation that characterizes high-quality investigative or human-rights reporting.

Country-level variation

Countries differ in how well they are served by chatbots overall. The UK, Spain, Brazil, Canada, and Egypt all have Overall averages above 4.10, while Indonesia, Argentina, and Mexico sit at or just below 3.88. This roughly half-point difference on a five-point scale indicates that human experts consistently judged responses in some countries as only “good enough,” while others came closer to “very good.”

Table 1. Average chatbot scores by country and dimension (all bots pooled)

Country	Neutrality	Factuality	Balance	Tone	Overall
Argentina	3.75	4.06	3.69	4.00	3.88
Brazil	3.94	4.38	4.12	4.00	4.11
Canada	4.19	4.25	3.81	4.19	4.11
Colombia	4.19	3.81	3.81	4.25	4.02
Egypt	3.94	4.38	3.88	4.25	4.11
France	4.25	3.94	3.50	4.19	3.97
India	4.31	3.88	3.50	4.38	4.02
Indonesia	3.75	4.12	3.56	4.00	3.86
Japan	4.12	4.25	3.88	4.12	4.09
Mexico	3.94	3.75	3.56	4.25	3.88
Peru	4.12	4.00	3.38	4.19	3.92
Romania	4.31	3.75	3.81	4.31	4.05
Saudi Arabia	3.94	4.00	3.62	4.19	3.94
Spain	4.38	4.06	3.94	4.38	4.19
UK	4.69	4.19	3.75	4.56	4.30

Source: MJRC chatbot evaluation dataset; means across all bots and topics per country • Created with Datawrapper

These differences align with the availability and visibility of country-specific information in global corpora. Contexts with extensive English-language coverage, strong institutional documentation, and high international salience tend to see higher average scores than those where high-quality material is more localized or fragmented.

Issue-type clustering

Grouping topics into four clusters (democratic institutions and power; rights and identity; social policy and welfare; and historical memory and corruption) shows that social policy and welfare topics (healthcare, NHS, drug use, equity) attract the highest combined scores (Overall ≈4.13), while historical memory and corruption issues sit slightly lower (Overall ≈3.98). This suggests that models are more comfortable describing present-day service systems and policy trade-offs than navigating competing narratives about past violence, transitional justice, or elite wrongdoing.

For democratic institutions and power (elections, extremism, governance, constitutional reform), cluster averages remain high but show moderate drops on Balance, reflecting the difficulty of compressing polarized multiparty landscapes into short neutral paragraphs, especially when recent scandals or legitimacy crises loom large. Rights and identity topics (LGBTQ+ rights, caste, sexual diversity, hate crimes, Indigenous sovereignty) show similar profiles: strong Neutrality and Tone, good Factuality on well-documented abuses, but uneven coverage of less visible or more locally contested issues.

In the historical memory and corruption cluster, the main weakness is Balance rather than Neutrality or Tone, indicating that bots under-represent contested perspectives, actors, and unresolved disputes even when they capture the broad narrative arc.

Table 2. Per-bot average scores across all countries and topics

Bot	Neutrality	Factuality	Balance	Tone	Overall
ChatGPT	4.33	3.70	4.03	4.33	4.10
Claude	3.63	4.07	3.70	4.10	3.88
Copilot	4.03	3.70	3.63	4.10	3.87
DeepSeek	3.93	3.67	3.57	3.97	3.78
Gemini	4.37	4.63	3.83	4.57	4.35
Grok	4.67	4.87	4.10	4.73	4.59
Meta AI	3.77	3.80	3.23	3.73	3.63
Perplexity	4.23	4.00	3.67	4.20	4.03

Source: Aggregated from chatbot scores across all 15 countries • Created with Datawrapper

Table 3. Topic-type clusters: approximate average scores across all countries

Topic cluster	Neutrality	Factuality	Balance	Tone	Overall
Democratic institutions & power	~4.10	~4.12	~3.80	~4.25	~4.07
Rights & identity	~4.15	~4.10	~3.78	~4.30	~4.08
Social policy & welfare	~4.25	~4.05	~3.85	~4.35	~4.13
Historical memory & corruption	~4.05	~4.00	~3.70	~4.15	~3.98

Source: Clustered from chatbot scores by pooling relevant topics (elections, extremism, governance, constitutional reform; LGBTQ+ rights, caste, sexual diversity, hate crimes, indigenous rights; healthcare, NHS, drug use, equity; dictatorship, memory, corruption). • Created with Datawrapper

Topic- and country-level patterns

Looking inside individual country–topic pairs helps clarify how these aggregate scores arise and what kind of content users actually see when they query chatbots on high-impact civic and human-rights issues. Three patterns are particularly salient: asymmetries between structurally described institutions and politically charged dynamics; differences between well-documented and under-documented human-rights issues; and uneven treatment of historical memory and transitional justice.

Institutions versus power

On “elections and power dynamics” in Egypt and Saudi Arabia, most chatbots produce descriptions that match the high Factuality scores observed at country level, but they tend to lean toward structural characterizations of authoritarian systems rather than concrete illustrations of repression.

In Egypt, for example, several models describe tightly controlled multiparty elections, the dominance of the presidency, and the central role of the military and security apparatus, often noting the 2019 constitutional amendments that extended presidential terms. Yet they vary in how clearly they articulate the mechanisms that make elections non-competitive, such as candidate disqualification, media capture, harassment of opposition organizers, or the imprisonment of specific challengers, details that feature prominently in the expert baselines and human annotations.

A similar pattern appears in Saudi Arabia, where models competently outline the formal architecture of monarchical governance and consultative assemblies but are more cautious in describing how power, patronage, and coercion operate in practice. For users, this translates into an understanding of Saudi institutions that is conceptually accurate but thinner on the concrete mechanisms of control and contestation.

France shows a different variant of the same tension. On secularism and political extremism, Neutrality and Tone scores are high, but Balance is lower than in many other countries. Outputs often describe laïcité as a constitutional principle, mention headscarf debates or "tensions around visible religious symbols," and refer to the rise of far-right parties or "polarized politics." However, they frequently omit specific legislative and judicial milestones, namely the 2004 law on religious symbols in schools, the 2010 face-covering ban, and the 2021 "separatism" law, alongside landmark court decisions that shape lived experience and public argument. In the extremism topic, several models center far-right actors and narratives, with less attention to far-left mobilization, Islamist networks, or state overreach, mirroring the relatively low French Balance score despite overall high Neutrality.

Human-rights visibility and data density: Egypt, Canada, India, Indonesia

Human-rights topics illustrate how performance depends not just on the severity of abuses, but on how well they are documented and integrated into the training ecosystem. LGBTQ+ rights in Egypt are a telling case. Several chatbots provide highly detailed accounts of de facto criminalization through debauchery and morality laws, police raids and online entrapment, forced medical examinations, lack of legal recognition, and social stigma. This aligns with the high Factuality and Tone scores for that topic and for Egypt overall. Here, decades of human-rights documentation, international reporting, and global advocacy have created a dense, well-indexed set of sources from which models can generalize.

By contrast, some other rights issues show more uneven coverage. On Indigenous sovereignty and land rights in Canada, models correctly identify treaties, residential-school legacies, and resource conflicts, but vary in how concretely they refer to recent court cases, specific First Nations, or ongoing land-back campaigns. India's caste system and "bulldozer justice" topics reveal a similar pattern: chatbots typically acknowledge caste-based discrimination, violence, and political mobilization, and describe demolition campaigns as a contested law-and-order tactic, but differ in how explicitly they name parties, chief ministers, affected communities, or recent emblematic incidents.

In Indonesia's West Papua topic, descriptions often emphasize "long-standing tensions" and "conflicts over self-determination" without fully capturing the density of allegations about security-force abuses, resource exploitation, or information blackouts that appear in local and specialist reporting.

In all these cases, outputs are recognizably about the right topic, but their Balance scores reflect how much nuance and actor diversity they manage to retain within a short paragraph. The more fragmented or locally embedded the documentation, the more likely bots are to flatten contestation into generic language about "tensions," "debates," or "complex legacies."

Social policy and welfare: strong structure, selective detail

Social-policy topics (healthcare systems, the NHS, drug use, and equity and social recognition) are where chatbots most consistently align with expert expectations on structure while still revealing important gaps in empirical depth. In Romania, models describe a hospital-centered public system with low spending by EU standards, staff shortages due to emigration, urban–rural inequalities, informal payments, and ongoing attempts to modernize infrastructure and digital systems. These descriptions match Romania's high Neutrality and Tone scores and decent Factuality, and they are recognizable to anyone familiar with EU-level health comparisons.

Yet the annotations show that the models largely omit the sequence of hospital fires, nosocomial-infection scandals, patients being asked to procure their own supplies, or high-profile corruption cases that dominate domestic coverage and public distrust. In other words, chatbots provide reliable scaffolding about how the system is supposed to work but underplay the scandals and failures that make it politically salient.

In the UK, chatbots almost uniformly excel on the NHS. They state its founding in 1948, tax-based financing, "free at the point of use" principle, the division into four national systems, and common pressures like aging populations, funding constraints, and workforce shortages. This performance helps explain why the UK sits at the top of the country ranking for Overall scores. However, when asked about immigration, models are more cautious about specific numbers and controversies: many mention points-based systems, net-migration trends, and measures such as electronic travel authorization or stricter language requirements, but fewer provide detail on the now-scrapped Rwanda plan, its cost relative to outcomes, or the divergence between public perceptions and the actual composition of migrant flows. Social policy thus seems to be an area where the models offer solid scaffolding but sometimes underplay the political and moral conflicts that make these issues salient.

Historical memory, transitional justice and corruption: Argentina, Peru, Colombia, Spain

Historical-memory and corruption topics are among the most demanding for chatbots because they require integrating legal, historical, and moral perspectives that remain contested even among human experts.

In Argentina's military dictatorship and "memory, truth, and justice" topics, most models correctly sketch the 1976–1983 regime, state terror, desaparecidos, democratic transition, and truth-commission processes, which underpins the relatively high Argentine Factuality average. But they differ in how carefully they handle contested victim numbers, the evolution of amnesty and trial policies, or the ways in which current political actors invoke or minimize the dictatorship in present-day disputes. Some outputs lean into a museum-style narrative of "never again," while others stay at a very general level that flattens differences between victims, perpetrators, and bystanders.

Peru's "ethnic and cultural diversity" and "historical memory" topics raise similar issues. Chatbots tend to highlight the multi-ethnic composition of the country, the marginalization of Indigenous communities, and the Truth and Reconciliation Commission's account of internal armed conflict but may give less attention to ongoing disputes over how that history is taught, memorialized, or denied. In Colombia's "corruption" and "human rights and violence" topics, models generally mention armed conflict, paramilitaries, state actors, and peace processes, but vary in how they connect corruption to violence and how they distribute responsibility across institutions and armed groups. Spain's "political corruption" and "hate crimes" topics are often narrated through high-level descriptions of party-funding scandals, institutional reforms, and rising reports of hate incidents, again with relatively strong structural accuracy and slightly weaker treatment of specific emblematic cases that shape domestic debate.

Across these historical-memory and corruption topics, the cluster scores in Table 3 (Overall ≈3.98, with Balance at the lower end of all four clusters) resonate with the qualitative evidence. Chatbots can usually situate a country within a broad narrative of dictatorship, conflict, or corruption, but they struggle to simultaneously name enough actors, events, and contesting perspectives to satisfy human experts who work in those fields daily.

Bots, baselines and regional asymmetries

The topic-level analysis confirms that bot-level differences are not merely statistical artifacts. Grok and Gemini stand out not just numerically but also in the richness and apparent recency of their outputs on several complex topics: they tend to integrate more up-to-date electoral and migration figures, provide clearer descriptions of evolving coalitions in Japan, and give more detailed accounts of LGBTQ+ persecution in Egypt or drug-use trends in Romania. Meta AI and DeepSeek, by comparison, are more likely to produce shorter or more generic paragraphs that score lower on Balance and, occasionally, on Factuality.

However, all eight systems show a similar geography of strengths and weaknesses: high performance in information-dense, globally salient contexts and more cautious, thinner answers on issues and regions that sit at the margins of the dominant training corpus. This pattern echoes broader concerns about how generative AI may reinforce existing asymmetries in news coverage, documentation, and visibility, particularly between Global North and Global South contexts.

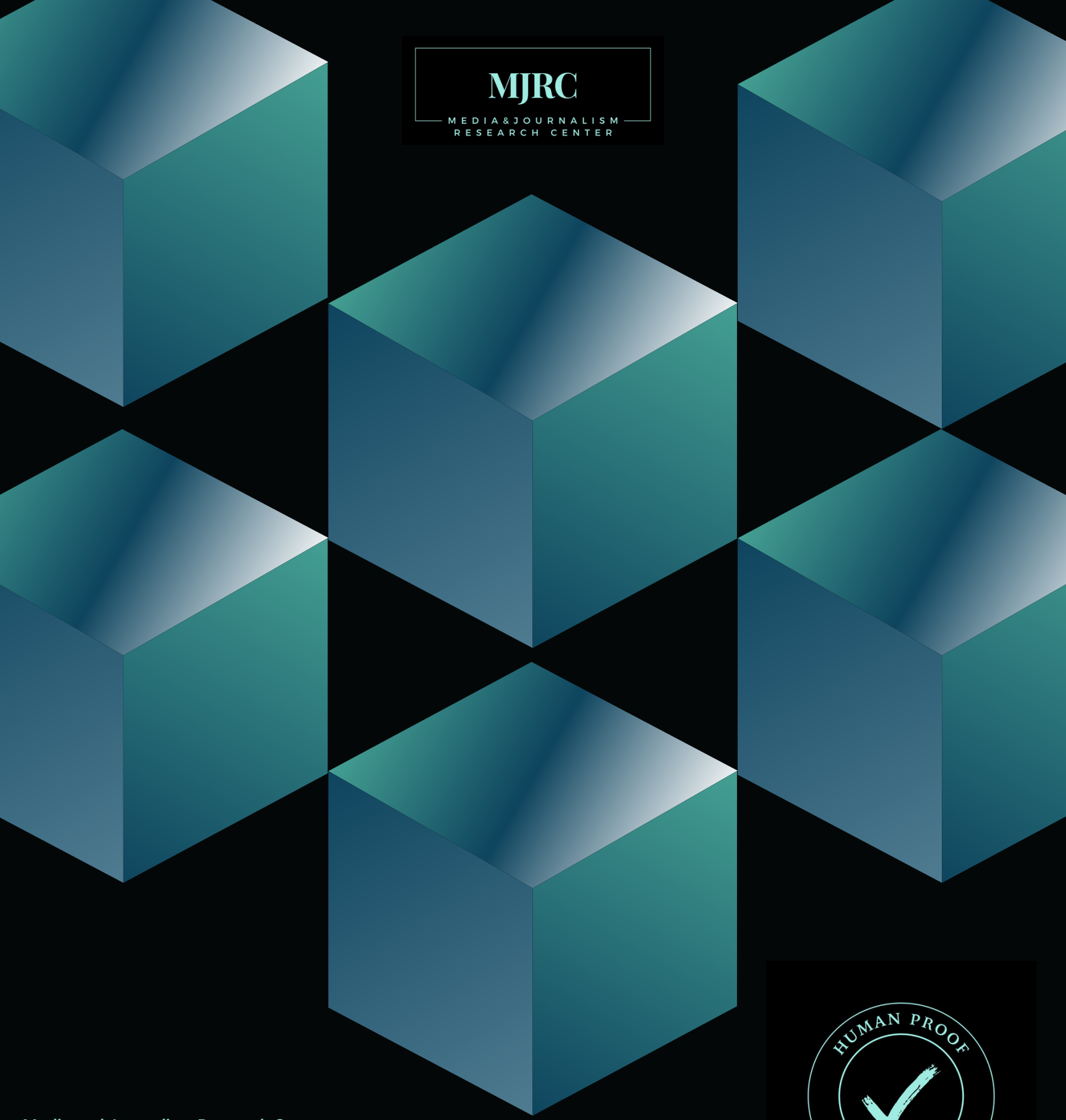
Conclusions: Implications for journalism and society

The results collectively trace a picture of how public-facing chatbots now participate in shaping civic knowledge. For a journalist or editor, these systems function as competent but shallow research assistants. When asked (in English or local languages) about the NHS, Japanese constitutional reform, or Romanian healthcare, the major models deliver reasonably accurate, neutral summaries of institutions, legal frameworks, and long-term challenges. They are particularly good at compressing complex systems into a paragraph that can help a generalist reporter orient themselves on a new beat. However, reporters looking for the heat of current politics, the precise vote shares in an Egyptian election, the cost and legal defeats of the Rwanda scheme, the names and dates of hospital fires, or the exact contours of French headscarf laws, will not find them reliably in these outputs. The models excel at structure, not at the granular evidence that validates strong journalism.

In human-rights work, the picture is similarly mixed. On some highly documented abuses, such as the persecution of LGBTQ+ people in Egypt, several models synthesize legal mechanisms, police practices, and social stigma with impressive clarity, in a neutral tone that can be reused as a primer or briefing note. On others, such as West Papua, bulldozer justice and caste in India, or historical memory in Peru, the scores and annotations suggest a tendency to flatten contestation into vague formulations about "tensions," "debates," or "complex legacies," without giving full weight to victims, contested numbers, or ongoing impunity. For rights advocates, this means that chatbots can help explain the basic facts of an issue, but they cannot be relied on to capture its political charge or the full spectrum of affected voices.

From a societal perspective, the fact that free chatbots reach Overall averages near or above 4.0 in most countries is not trivial. In environments where textbooks are outdated, independent media are constrained, or access to specialist literature is limited, these tools already offer millions of people a reasonably accurate, non-polemical account of how their electoral systems work, what their health services do, or how certain rights are formally defined. At the same time, the subtle skew visible in the data (that the UK, Spain, Canada, and Japan are better served than Indonesia, Argentina, or Peru, and that Balance is lowest on historical memory and contested identity topics) suggests that this new layer of "AI-mediated public knowledge" risks reproducing and amplifying existing information inequalities.

The narrative emerging from the corpus is therefore double-edged. On one hand, the basic versions of ChatGPT, Gemini, Perplexity, Claude, Copilot, DeepSeek, Meta AI, and Grok already deliver a floor of civic information that is higher, more neutral, and more globally accessible than many human-authored sources. On the other, they do so in a way that is more comfortable with institutional description than with power, more precise on laws than on those who break or bend them, and more at ease with Euro-Atlantic evidence than with the harder-to-find documentation of Global South struggles. For journalists, advocates, and policymakers, the task ahead is not to treat these systems as oracles, but to understand their characteristic blind spots and to make deliberate choices about when to lean on them, and when to insist on deeper, locally grounded, human inquiry.



Media and Journalism Research Center

Legal address
Tartu mnt 67/1-13b, 10115,
Tallinn, Harju Maakond, Estonia

Postal address
6 South Molton St, London,
W1K 5QF, United Kingdom

Academic affiliation
Universidade de Santiago de Compostela (USC)
Colexio de San Xerome, Praza do Obradoiro s/n,
CP 15782 de Santiago de Compostela.

Contact
www.journalismresearch.org
mjrc@journalismresearch.org

Artificial Intelligence (AI) Disclosure Statement
AI tools were used in this research project for data collection and preprocessing purposes. Perplexity was used to correct grammar during the editing phase. No AI tool was used in drafting or shaping the analytical content of this study.

doi: 10.5281/zenodo.18620123

File integrity (SHA-256):
A31F35979602924DA4573148D9B8ECC661A01AFA551E828EE06C21F94C01D9B3