Social-RAG: A Retrieval-Augmented Generation Pipeline for Computational Social Science Research on Telegram

Leonardo Nascimento[1], **Eric Brasil**[2], **Arthur Lima**[1]
Gabriel Andrade[1], **Ricardo Sodré Andrade**[3], **Tarssio Barreto**[4]
[1] Federal University of Bahia, Salvador, Bahia, Brazil
[2] UNILAB, São Francisco do Conde, Bahia, Brazil
[3] National Archives, NE Regional Office, Salvador, Bahia, Brazil
[4] Ministry of Health, Brasília, DF, Brazil
{leofn3,profericbrasil,arthurlimareserva,gabriel.tarssioesa}@gmail.com
rsandrade@ufba.br

Abstract

Digital trace data expand empirical opportunities in the social sciences but intensify the challenge of scale: many corpora are now too large and fast-moving to read exhaustively without losing interpretive rigor. We present Social-RAG, a modular Retrieval-Augmented Generation (RAG) architecture for scalable qualitative inquiry that preserves evidence traceability, auditability, and researcher control. We apply it to public Telegram messages organized into two thematic subsets—vaccine discourse and debates on Brazil's Lei Rouanet cultural funding policy—and detail core design choices: "one post = one chunk" indexing, embedding-based semantic retrieval, an Adaptive-K cutoff for context selection, MMR re-ranking for diversity, and structured analytical instructions that constrain generation to retrieved evidence. We evaluate the system with hermeneutic and factual question blocks and compare three models (local open-weight, cloud open-weight, and commercial closed) using an LLM-as-judge protocol with qualitative criteria. Across both corpora, the larger models perform robustly on narrative and factual tasks, while the local model remains useful for exploratory narrative synthesis but is less reliable for strict factual extraction and attribution. We close with implications, limitations, and directions for improving scalability and extensibility.

1 Introduction

Digital trace data (Howison et al., 2011) and digital methods (Jungherr, 2015; Omena, 2019; Rieder and Röhle, 2017; Rogers, 2013) have introduced new empirical possibilities — and methodological tensions — into social science research (Amaturo and Aragona, 2019; Carrigan, 2014; Conte et al., 2012; Lupton, 2015; Marres, 2017; Nascimento, 2016). The massive, continuous production of data, the processes of datafication of social behavior (van Dijck, 2014; Lomborg et al., 2020; Sadowski, 2019; Southerton, 2020), and the subsequent reuse of these traces in socio-anthropological research (Salganik, 2018; Rogers, 2013) have posed renewed methodological challenges across the humanities. On the one hand, the abundance of digital traces expands the possibilities for empirical inquiry into human behavior. On the other hand, it creates a structural mismatch: the volume, velocity, and heterogeneity of the data produced often overwhelm the analytical capacity of traditional social-scientific approaches (Abbott, 2000). The scale of available data challenges established methods in the humanities, making manual pattern identification difficult. As a result, the abundance of data coexists with the risk of methodological paralysis or, worse, the uncritical adoption of tools that claim to tame the complexity of digital data traces.

These trends require re-evaluating not only data collection and processing techniques but also the epistemological framework mobilized in social research using digital data. The challenges are multiple and interconnected: they involve the critical evaluation of digital sources (Gebru et al., 2021) and their algorithmic pre-construction (boyd and Crawford, 2012; TacticalTechVideos, 2014); the articulation — rather than opposition — between qualitative and quantitative approaches; epistemological vigilance against the allure of objectivity and the power of visual evidence (Rieder and Röhle, 2017); and the fundamental distinction between data elicited by the researcher and data collected from pre-existing records (Salganik, 2018). These challenges are multifaceted, encompassing not only the need for enhanced digital literacy among researchers and sufficient computational capacity for large-scale laboratory work, but also issues related to publicity, accessibility, and representativeness.

Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026)
Vol. 2, pages 255–265
April 13-16, 2026

This article [1] presents the development of Social-RAG, a Retrieval-Augmented Generation (RAG) architecture for analyzing digital trace data from Telegram groups and channels. We detail the key design decisions, technical choices, and methodological limits of this approach in humanities and social science research, situating it within broader debates on RAG as an epistemic resource for large-scale digital data analysis. We take a technical and methodological approach, describing the implemented pipeline and its practical effects on textual analysis, while avoiding black-box treatment of AI components in the research process (Schwandt, 2022).

The article has five sections: (1) a brief overview of RAG components, variants, and key debates; (2) Social-RAG and its fit to our research needs and data; (3) implementation methods (Telegram data collection, vectorization/indexing, Adaptive-K retrieval, system instructions, and the Streamlit interface); (4) evaluation (hermeneutic and factual experiments, criteria, cross-model results, limitations, and future work); and (5) a conclusion summarizing the main contributions.

2 Retrieval-Augmented Generation (RAG): definition and operating structure

The core principle of Retrieval-Augmented Generation (RAG) architectures is to retrieve relevant documents in response to a query and incorporate them into the processing context of large language models (LLMs) (Lee et al., 2025; Lewis et al., 2021). In general, a RAG system comprises two distinct components: a retrieval module, responsible for identifying and selecting pertinent passages from an external knowledge base, and a generative module, which produces answers conditioned on both the original query and the retrieved material (Zhou et al., 2023). This functional separation allows the model to rely less on knowledge internalized in its parameters during training, and instead to draw on information that is up-to-date, specialized, or situated at inference time. By grounding text generation in retrieved evidence, RAG systems tend to reduce hallucinations (fluent outputs that are nevertheless factually incorrect), increasing the traceability and verifiability of responses (Filippova, 2020; Gao et al., 2023; Maynez et al., 2020; Singh et al., 2025).

A basic RAG pipeline has three steps: index documents as embedded chunks in a vector store, retrieve semantically similar passages for an embedded query, then prompt the LLM to synthesize an answer grounded in that evidence. Output quality depends on choices such as chunking, embeddings/normalization, metadata filtering, re-ranking, and citation rules, which make the claim–source link explicit.

RAG's core value is factual grounding: it reduces hallucinations and obsolescence by forcing generation to rely on a bounded, verifiable corpus rather than opaque parametric recall. This makes outputs auditable — claims can be traced to specific passages, preserving verification practices akin to footnotes — and shifts the LLM from a "stochastic parrot" (Bender et al., 2021) toward a more controllable research instrument under the researcher's interpretive authority.

3 Implementing a Retrieval-Augmented Generation architecture: Social-RAG

RAG architectures vary widely in complexity, modularity, and methodological sophistication (Gangavarapu et al., 2025; Oche et al., 2025), and recent work has introduced iterative retrieval, knowledge structures, and tighter evidence controls to improve both retrieval and generation (Brontes et al., 2025); yet most systems still start from a common baseline that remains a useful reference point. This baseline—often called naive RAG—follows a simple Retrieve–Read paradigm (Gao et al., 2023), with single-pass retrieval (no query chaining, reassessment, or dedicated filtering) and generation without explicit strategies for interpreting or reconciling retrieved material, making it prone to weakly relevant, redundant, or contradictory passages and offering limited support for integrating multiple sources or perspectives. More broadly, RAG quality hinges on design choices across its two core axes, retrieval and generation (Zhang and Zhang, 2025): on the retrieval side, naive approaches lack systematic denoising and reasoning mechanisms to articulate lines of evidence (Cheng et al., 2025), while on the generation side, weak controls over evidence use hinder handling ambiguity, interpretive conflict, and source prioritization—central concerns in humanities and social-science research (Babbie, 2013).

[1] Esse trabalho é uma versão do artigo submetido para a revista Journal of Computational Social Sciences e publicado como preprint em https://doi.org/10.31235/osf.io/wmc2q_v1

To overcome naive RAG limits, recent approaches make the pipeline more modular and dynamic, adapting indexing, retrieval, re-ranking, and generation to task goals: Self-RAG trains the model to signal when more evidence is needed and to trigger new searches (Asai et al., 2023), Adaptive-RAG retrieves only when the current context is judged insufficient (Lee et al., 2025), other methods monitor context cutoffs (Xie et al., 2025) or learn retrieval/reordering policies via reinforcement learning (Dynamic-RAG) (Sun et al., 2025), and Agentic RAG adds explicit cycles of planning, querying, evaluation, and synthesis (Singh et al., 2025). Beyond performance gains, these designs operationalize procedures familiar to social-science inquiry — gathering relevant evidence, integrating multiple sources, and producing traceable, verifiable syntheses — so, given the demands of misinformation research and the scale, heterogeneity, and context dependence of Telegram data, we propose Social-RAG as a task-oriented architecture that balances methodological control, evidence traceability, and operational feasibility.

Social-RAG builds on core ideas from prior RAG proposals while introducing design choices tailored to humanities and social-science research needs. Its development responds to the scale and velocity of digital-platform data, which increasingly outpace traditional close-reading workflows, especially in domains such as vaccine misinformation and political extremism where rhetorical variation and platform dynamics demand methods that operate at scale without sacrificing traceability or interpretive control (Nascimento et al., 2023; Cesarino et al., 2025; Scheren et al., 2024). These corpora are also temporally volatile: messages are produced, circulated, and recontextualized in continuous flows with bursts driven by political and public health events, requiring tools that support exploratory, hypothesis-driven interpretation within short windows and sometimes near real time. We therefore conceived Social-RAG as a technical mediation layer that helps researchers navigate large corpora, surface and organize relevant evidence, and sustain auditable analytical practices.

4 Methods

Social-RAG was designed to align computational choices with corpus and common analytical workflows in the humanities and social sciences. We use a modular pipeline — vectorization, indexing, retrieval, re-ranking, and generation — so that each component can be inspected and adjusted. This approach allowed us to balance technical sophistication, methodological traceability, and operational feasibility, ensuring that the system functions as a research support instrument rather than as an opaque layer of interpretive mediation.

Social-RAG is implemented in Python using a reproducible, modular stack that integrates local and cloud components. LangChain orchestrates retrieval, context assembly, and model calls; Ollama runs open-weight models locally; and the OpenAI API is used when proprietary models are required. We use ChromaDB for persistent vector storage and embedding-based search, pandas for data ingestion and preprocessing, FastAPI as the backend service layer, and a lightweight Streamlit frontend for iterative querying and qualitative inspection, with integrated logs.

Over eight months, we followed a human-in-the-loop development cycle (Afzal et al., 2024), holding regular meetings with specialists and iteratively revising the full pipeline (dataset curation, chunking, embeddings, retrieval, and generation). Repeated team testing across datasets and query scenarios drove refinements to chunking, embedding choice, Adaptive-K retrieval, prompts and test questions, and the Streamlit interface, informed by five years of lab experience with Telegram data to assess retrieval relevance and the evidentiary quality of generated answers.

4.1 Ethics, privacy, and data governance

Following guidance in computational social science and platform research (Salganik, 2018), we treat public Telegram accessibility as insufficient to waive obligations around privacy, contextual integrity, potential harm, and researcher safety. We therefore collect only from publicly accessible groups/channels, exclude private 1:1 messages, secret chats, and closed groups, and adopt a read-only ("lurker") posture (Ferguson, 2017; Barratt and Maddox, 2016). We minimize re-identification risk by removing direct identifiers, processing user/chat references via platform IDs and/or cryptographic hashing, retaining only fields required for the analyses, and avoiding re-identification or cross-platform triangulation, mindful of "surveillance-as-method" tensions (Topinka et al., 2021). Because releasing a large, searchable corpus can amplify harm, we do not publish the full dataset (Nascimento et al.,

2023; Cesarino et al., 2025).[2] Instead, we share reproducible artifacts (evaluation materials, aggregated source data, and plotting scripts) and small anonymized samples (1,000 messages per theme) to support transparency without enabling harmful redistribution (Nascimento et al., 2023; Cesarino et al., 2025).[3]

4.2 Telegram data-collection pipeline

The data analyzed in this study comes from a computational infrastructure that automates the collection and storage of messages from public Telegram groups and channels.[4] We perform extraction via the official API (MTProto), accessed through Python libraries, and stream messages in real time for indexing in an Elasticsearch cluster together with their metadata (e.g., authorship, timestamp, content type, and forwards). After indexing, we apply transformation and enrichment routines to support querying, visualization, and analysis, with exploration via Kibana. Continuous collection is crucial because administrators or users themselves often delete content in groups and channels; nevertheless, messages already captured are preserved within the infrastructure (Ferguson, 2017).

To support analysis and experimentation, we extract thematic subsets from Elasticsearch with a script that automates querying, filtering, and export to standardized CSV/JSON, including normalization and deduplication to reduce noise from reposts and improve retrieval precision in the RAG pipeline. We evaluate the system on two contrasting subsets: vaccines (large, heterogeneous) and Lei Rouanet (smaller, more concentrated). The vaccine subset was built with wildcard queries (e.g., *vacin*, *vaccin*) over messages since 2022, yielding 116,284 unique messages after deduplication; it spans explicitly anti-vaccine posts and broader discussions of immunization, adverse effects, pharmaceutical companies, and public health policy, capturing how misinformation intersects with institutional distrust and conspiratorial narratives. The Lei Rouanet subset used orthographic variants (e.g., *rouane*, *ruane*) over messages since 2017, yielding 3,284 unique messages after deduplication, and is dominated by cultural-funding debates and "culture war" framings targeting artists and institutions.[5]

4.3 Vectorization, embeddings, and metadata

In Social-RAG, we implement a strict "**one message = one chunk**" policy: each Telegram post is indexed and retrieved as a single unit, reflecting the corpus's short, self-contained, and highly contextual discourse structure. Splitting messages risks semantic loss, while aggregating them introduces noise by mixing voices, topics, and temporalities; keeping posts intact preserves each retrieved item as an identifiable discursive act that can be cited, verified, and contextualized (Lee et al., 2025). Concretely, we embed each message with `text-embedding-3-large` (OpenAI, 2024), enabling semantic retrieval beyond keyword overlap—crucial for misinformation settings characterized by rhetorical variation, irony, abbreviations, and indirect strategies—while strengthening traceability through message-level sources with associated metadata, consistent with qualitative research practices in the humanities and social sciences (Hatch, 2010; Krippendorff, 2004).

Beyond vector representations, Social-RAG incorporates the metadata associated with each chunk. This additional layer of information is essential for analytical control, source traceability, and flexible retrieval. For each message, we store the following metadata fields: `message_id`, `chat_id`, `strict_date`, `type`, and `chat_title`.

4.4 Efficient retrieval and Adaptive-K context selection

Social-RAG uses HNSW (Hierarchical Navigable Small World) to perform low-latency nearest-neighbor search over message embeddings (Malkov and Yashunin, 2016). HNSW indexes vectors as a multi-layer graph that enables fast navigational search—coarse jumps in sparse upper layers followed by finer search in denser layers—avoiding exhaustive comparisons and supporting near-real-time exploratory querying (Malkov and Yashunin, 2016). For context construction, we avoid a fixed top-k, which is brittle across question types (Mengmeng et al., 2024): small K can miss evidence, while large K increases noise, cost, and latency

[2] Access to the full datasets is handled under controlled conditions for legitimate scholarly purposes, evaluated case-by-case and subject to commitments that prohibit re-identification and redistribution.

[3] Available after anonymous review.

[4] Messages are collected from a set of thousands of public, open groups and channels associated with far-right networks in Brazil, in real time, continuously, since 2021.

[5] Lei Rouanet (Law No. 8,313/1991) created PRONAC and federal cultural-support mechanisms, mainly via tax incentives that allow individuals and firms to allocate part of their tax liability to sponsor approved cultural projects; it also includes the National Culture Fund and other modalities. For discussion of its logic and limits, see (Balbino and Venâncio, 2020).

(Sun et al., 2025; Taguchi et al., 2025). Instead, we adopt Adaptive-K (Taguchi et al., 2025), selecting K from the similarity-score distribution in a single pass:

$$K = \underset{1 \le i < n}{\operatorname{argmax}} \left(s_{i+1} - s_i \right),$$

i.e., the sharpest drop in similarity after ranking candidates ($s_1 \ge \cdots \ge s_n$). In practice, we start from a broad candidate pool, add a small buffer ($B = 5$), restrict the cutoff search to the top 90% to avoid tail artifacts, and clamp K to $[10, 100]$ based on internal tests and corpus redundancy. Finally, we apply Maximal Marginal Relevance (MMR) to the selected set to balance relevance and diversity and reduce near-duplicate reposts (Carbonell and Goldstein, 1998).

4.5 System instructions

We developed a system prompt to guide the model's analytical behavior during Social-RAG's generation stage (the full version is available in the supplementary materials). The prompt casts the model as an analyst with interdisciplinary training in the humanities and social sciences and instructs it to respond exclusively based on retrieved passages, avoiding external knowledge and unsupported extrapolations. We adopt a "thread-of-thought" structure (Zhou et al., 2023), with explicit stages for interpreting the question, planning, selecting evidence, conducting critical analysis, and synthesis, to promote controlled, evidence-oriented outputs. We pass the corpus theme (e.g., vaccination or Lei Rouanet) as a parameter to adapt the analytical framing without changing the prompt structure. This combination strengthens Social-RAG's methodological coherence, aligns retrieval and generation with human-in-the-loop evaluation, and supports reproducibility by making the system's operational rules public.

4.6 Streamlit and the graphical interface: features and resources

Social-RAG is delivered through a Streamlit web interface (Streamlit Inc., 2021), allowing social science researchers to use the system without running scripts. Hosted on a dedicated server, the interface structures the workflow and exposes key controls, including selection of the language model, theme, and pre-vectorized corpus (e.g., Vaccine or Lei Rouanet). These choices set the analytical context and automatically load the corresponding thematic system prompt and parameters. The app supports iterative, chat-style querying with persistent history and provides access to dataset reports, the active system prompt, and a downloadable Markdown conversation log for documentation and auditing.

4.7 Models available in Social-RAG

We evaluate Social-RAG with three deliberately distinct LLM profiles to test how openness/licensing, deployment mode (local vs. cloud), and model capacity affect performance on narrative and factual tasks. Model A (`gemma3:12b`) runs locally via Ollama (Ollama, 2024a), supporting low-cost, reproducible experimentation under hardware constraints.[6] Model B (`gpt-oss:120b-cloud`) is an open-weight, large-scale cloud model accessed via Ollama (Ollama, 2024b), providing greater synthesis capacity without local infrastructure limits. Model C (`gpt-5-mini`) is a commercial closed model (OpenAI, 2025) used as a robust reference.[7]

Together, these models span common research conditions—local/control, open/scalable, and closed/optimized—enabling comparison of how cost, governance, infrastructure, and generation quality interact within the pipeline.

Unlike benchmark studies that primarily compare alternative RAG strategies (Gangavarapu et al., 2025; Zheng et al., 2023), our evaluation tests Social-RAG against the methodological and theoretical choices that guided its design. We assess performance under the corpus's analytical constraints using a human-in-the-loop procedure (Afzal et al., 2024) complemented by an LLM-as-judge protocol (Zheng et al., 2023).

Our tests use two complementary question blocks. Hermeneutic (narrative) questions evaluate whether the system can identify and synthesize recurring discursive patterns—framings, metaphors, moral judgments, and political associations—grounded in what the messages actually contain, rather than producing a single "correct" factual answer; this is central because misinformation and polarization often stabilize through narrative coherence. Factual questions evaluate precise evidence retrieval and faithful answering (enti-

[6] Ollama is a tool that allows users to download, manage, and run large language models (LLMs) through a command-line interface and a local API, facilitating both on-premise execution and, when applicable, the use of cloud backends within the same execution ecosystem. For more information, see: `https://docs.ollama.com`

[7] Through a partnership with OpenAI. OpenAI does not publicly disclose parameter counts for models in the GPT-5 family, including `gpt-5-mini`.

ties, numbers, dates, links, explicit claims), including negation and absence of information—crucial for detecting hallucinations, unwarranted extrapolations, and attribution errors in misinformation-heavy corpora.

Separating narrative from factual tasks targets distinct capabilities that automated RAG evaluations often aggregate (Zheng et al., 2023), despite their different epistemological and technical demands. Consistent with human-in-the-loop evaluation (Afzal et al., 2024), we do not rely on a closed gold standard; instead, we interpret outputs against discursive patterns established in the literature and empirical research. Applying the same protocol to two thematically distinct datasets (vaccines and Lei Rouanet) further enables comparison under variation in scale, redundancy, and the stabilization of ideological framings.

4.8 Criteria and methods for evaluating responses

We evaluated Social-RAG with task-specific qualitative criteria summarized in Table 1. For hermeneutic questions, we focused on thematic accuracy, analytical adequacy, evidence precision, synthesis, and political sensitivity; for factual questions, we prioritized verifiable extraction (question–evidence correspondence, coverage, entity extraction, sensitivity to negation/absence, and accurate recovery of numbers and links). Across both blocks, we also scored clarity/organization and the absence of hallucinations, especially corpus-unsupported gap-filling. Methodologically, this protocol approximates AI-assisted grounded theory (Charmaz, 2006; Corbin and Strauss, 2008), combining inductive pattern identification with computational retrieval and synthesis while preserving researcher interpretive control.

For comparative evaluation, we adopted an LLM-as-judge protocol to minimize order effects and label bias and to assess judge stability (Zheng et al., 2023). We conducted three blinded rounds in which the three models were rotated across positions A/B/C (each model appearing once per position), while judges saw only Response 1/2/3 with randomized order; the hidden mappings were logged, and the same two LLM judges (GPT-5 and Gemini Pro) scored all outputs under identical instructions. We replicated the same blinded procedure with a human judge (one author) using the same criteria, prompts, and response sets; Table 2 reports the blinding schedule. Across two themes (vaccine, Rouanet) and two question types (hermeneutic, factual), this yielded 12 LLM judgment tables (3 rounds × 2 themes × 2 types) plus one human table, totaling 13.

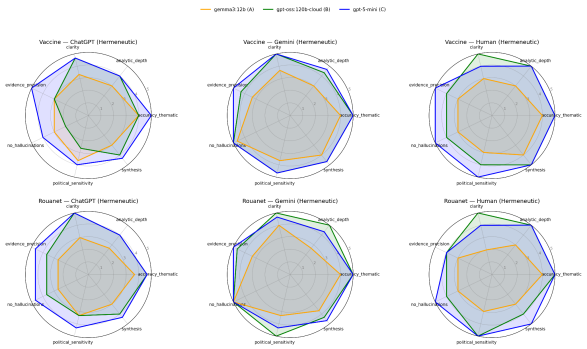

Figure 1: Hermeneutic Evaluation (by Judge): Vaccine and Rouanet — Criterion-Level Radar Profiles.

4.9 Results

All score sheets from the three blinded rounds were merged into a single long-format dataset (`judgments_long_full.csv`) with columns `theme`, `category`, `judge`, `blind`, `response_id`, `model_label`, `model_real`, `criterion`, `score`. From this file we computed criterion-level and overall mean scores by theme (vaccine/Rouanet), task type (hermeneutic/factual), model, and judge, plus judge-aggregated criterion profiles. Results are summarized with per-judge radar plots, a heatmap of overall mean variation, and a consolidated four-panel radar that aggregates criterion patterns across judges for each theme × question type.

Across both themes, hermeneutic radar profiles (Figure 1) show a stable ordering: Model C (`gpt-5-mini`) scores highest, followed by Model B (`gpt-oss:120b-cloud`), with Model A (`gemma3:12b`) trailing. This ranking is consistent across the two LLM judges and the blinded human evaluation, and is most pronounced on criteria tied to interpretive rigor (analytical depth, synthesis, clarity, and disciplined use of retrieved evidence), where Model A more often produces thinner or more generic summaries.

In the factual block (Figure 2), the radar plots show a clearer gap between the local model (A) and the larger models (B and C). Models B and C sustain consistently high scores for literal precision, coverage, entity extraction, and sensitivity to negations/absences, indicating stronger reliability when answers must be tightly anchored in explicit corpus

Question type	Criterion	Brief description
Hermeneutic	Thematic accuracy	Correspondence between the answer and the main discursive framings present in the corpus
	Analytical depth	Ability to articulate ideological, moral, and political dimensions in a non-superficial way
	Evidence precision	Appropriate and consistent use of retrieved passages to support the interpretation
	Synthesis capacity	Coherent organization of multiple discursive fragments into an intelligible explanation
	Political sensitivity	Attention to polarization contexts, avoiding undue simplifications or neutralizations
	Clarity and organization	Clear, comprehensible, and well-structured textual organization
	Absence of hallucinations	Does not introduce information, interpretations, or patterns not supported by the corpus
Factual	Literal precision	Direct correspondence between the question and retrieved textual passages
	Factual coverage	Ability to identify multiple relevant occurrences when present
	Entity extraction	Correct identification of proper names, institutions, artists, and political actors
	Sensitivity to negations	Explicit recognition of negations, absences, or contradictions in the corpus
	Number and percentage extraction	Correct retrieval of values, amounts, dates, and percentages
	URLs and links	Accurate identification and retrieval of cited links and domains
	Clarity and organization	Clear and structured presentation of factual information
	Absence of hallucinations	No fabrication of data, numbers, or sources not present in the corpus

Table 1: Evaluation criteria used in Social-RAG tests.

Blind round	Response 1	Response 2	Response 3	Judges
Blind 1	gemma3-12b	gptoss-120b	gpt5-mini	GPT-5; Gemini Pro
Blind 2	gptoss-120b	gpt5-mini	gemma3-12b	GPT-5; Gemini Pro
Blind 3	gpt5-mini	gemma3-12b	gptoss-120b	GPT-5; Gemini Pro

Table 2: Blinded evaluation schedule: model-to-response mapping across the three rounds.

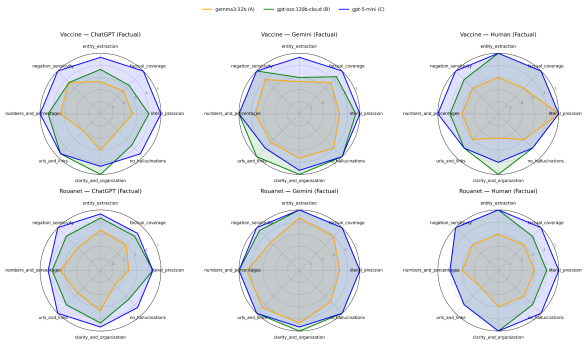

Figure 2: Factual Evaluation (by Judge): Vaccine and Rouanet — Criterion-Level Radar Profiles.

content. Model A remains usable in many cases but more often drops on precision- and evidence-dependent criteria (especially literal precision, coverage, and negation sensitivity), where small errors materially affect interpretation.

To test judge stability under the same blinding, we computed overall mean scores for each judge × subset × model (averaging across items and criteria). The heatmap (Figure 3) shows two consistent patterns: model ranking is stable across judges and subsets (C highest, B close behind, A lowest), and although judges differ in absolute severity (Gemini

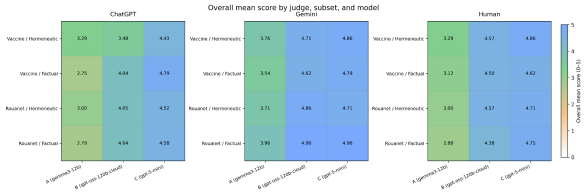

Figure 3: Overall Mean Score Variation: Judge × Subset × Model.

Pro scores higher on average than ChatGPT/GPT-5 and the human judge), the direction of differences is preserved, indicating robust comparative judgments.

Across subsets, model gaps are slightly larger in the vaccine corpus than in Rouanet, most clearly for hermeneutic questions. This likely reflects vaccines' greater volume and narrative heterogeneity, which accentuates differences in organization, inference control, and evidential discipline. In Rouanet, where framings are more stabilized and repetitive, performance remains differentiated but converges somewhat, suggesting that discursive redundancy can narrow the advantage of higher-capacity models for some interpretive tasks.

To summarize criterion-level patterns, we report consolidated radar plots by theme and question type, aggregating across judges with equal judge weighting (Figure 4). We first compute criterion means within each judge (pooling the three blinded rounds for ChatGPT/GPT-5 and Gemini Pro), then average across judges so the LLM

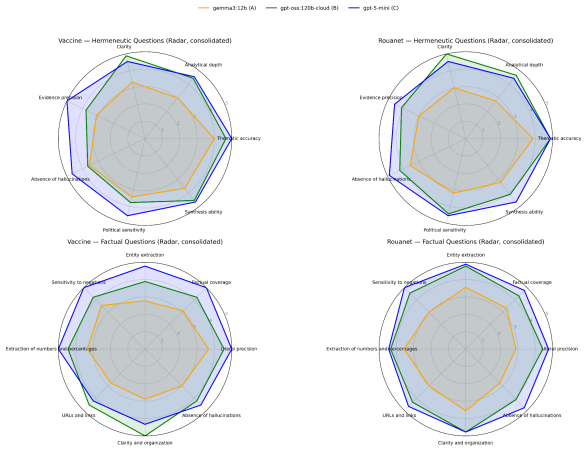

Figure 4: Consolidated Criterion Profiles (Equal Judge Weighting): Vaccine/Rouanet × Hermeneutic/Factual.

judges are not implicitly over-weighted relative to the single human evaluation. The consolidated profiles confirm the main trade-offs: in hermeneutic tasks, Models B and C maintain broad, strong profiles, while Model A is more uneven—often thematically aligned but weaker in depth, synthesis, and evidence precision; in factual tasks, Models B and C cluster at the top across most criteria, whereas Model A drops more on extraction- and verification-dependent criteria. Overall, Social-RAG yields stable comparisons across corpora and task types, and the results show that model capacity interacts with task demands: smaller local models can support exploratory interpretation, but the reliability gap widens when tasks require literal recovery, negation handling, and traceable evidence.

4.10 Limitations and future work

Social-RAG is not designed for exact counting or classical descriptive statistics (e.g., term/frequency counts, link totals), which are better handled by regex and traditional NLP pipelines. Instead, it retrieves semantically relevant messages and uses LLMs to synthesize and interpret them under explicit analytical instructions; its outputs are therefore evidence-oriented samples rather than exhaustive measurements. Social-RAG should be read as a qualitative, hermeneutic aid for exploring discursive patterns and narrative framings—supporting iterative, reflexive analysis—rather than a metric-producing system.

Future work focuses on modular evolution and scalability. Because components (embedding, indexing, retrieval, re-ranking, generation, interface) are replaceable, the system can track rapid changes in models and methods, but scaling to larger corpora and higher query concurrency will depend on compute/storage capacity and API costs. We also plan to add knowledge-graph modules to connect entities, actors, and recurring claims across messages, enabling graph-informed retrieval and more structured analysis.

5 Conclusions

This paper presented the implementation and evaluation of Social-RAG, a Retrieval-Augmented Generation architecture tailored to humanities and social-science analysis. Our starting point is that the scale and velocity of digital trace data require more than "computational power": they demand pipelines that preserve evidential traceability, interpretive control, and critical verification.

Social-RAG operationalizes this through design choices matched to Telegram-style corpora — one-post-per-chunk indexing, Adaptive-K context selection, MMR diversification, and structured analytical instructions — and we show, across two thematic datasets (vaccines and Lei Rouanet), that the system behaves consistently on both hermeneutic and factual tasks, supporting narrative synthesis and evidence recovery. Comparative experiments indicate a clear trade-off: larger models (B and C) are more reliable across both task types when evidential discipline is enforced, while the smaller local model (A) remains useful for exploratory interpretation but is less dependable for strict factual extraction, negation handling, and precise attribution. Finally, by documenting prompts, parameters, and design decisions, we make the pipeline auditable and reproducible, enabling inspection and adaptation to other corpora and research constraints.

Social-RAG neither replaces critical reading nor automates interpretation; it functions as a mediation layer that expands researchers' exploratory capacity when corpora exceeds the reach of exhaustive reading. Looking ahead, we highlight three development priorities: improving system scalability, conducting systematic evaluation across additional thematic domains, and integrating knowledge graphs to enrich contextualization. Ultimately, this work contributes to the development of computational infrastructures that serve social research rather than black-box systems that substitute for it.

Conflict of Interest

On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

Andrew Abbott. 2000. Reflections on the future of sociology. *Contemporary Sociology*, 29(2):296.

Asad Afzal, Ashwin Kowsik, Reza Fani, and Florian Matthes. 2024. Towards optimizing and evaluating a retrieval augmented QA chatbot using LLMs with human in the loop. *Preprint*, arXiv:2407.05925.

Enrica Amaturo and Biagio Aragona. 2019. Per un'epistemologia del digitale: Note sull'uso di big data e computazione nella ricerca sociale. *Quaderni di Sociologia*, 81.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2023. Self-RAG: Self-reflective retrieval augmented generation.

Earl R. Babbie. 2013. *The Practice of Social Research*, 13 edition. Wadsworth, Cengage Learning.

Gustavo Matias Soares Balbino and Renato Pinto Venâncio. 2020. Políticas culturais e arquivos públicos: o caso da Lei Rouanet. *ÁGORA: Arquivologia em debate*, 30(60):57–74.

Monica J. Barratt and Alexia Maddox. 2016. Active engagement with stigmatised communities through digital ethnography. *Qualitative Research*, 16(6):701–719.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, pages 610–623.

danah boyd and Kate Crawford. 2012. Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society*, 15(5):662–679.

Flavius Brontes, Janus Genesis, Zephyrine Noa, and Stavros Nymphodoros. 2025. Learning to retrieve, generate, and compress: A unified view of efficient RAG.

Jaime Carbonell and Jade Goldstein. 1998. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 335–336.

Mark Carrigan. 2014. An agenda for digital sociology.

Letícia Cesarino, Leonardo Nascimento, and Priscila Fonseca. 2025. Democracy "inside out": On far-right refracted publics in Brazil. In Zizi Papacharissi, editor, *The Routledge Companion to Digital Media and Democracy*, page 490. Routledge.

Kathy Charmaz. 2006. *Constructing Grounded Theory*. Sage Publications.

Ming Cheng, Yang Luo, Jianguo Ouyang, and 1 others. 2025. A survey on knowledge-oriented retrieval-augmented generation. *Preprint*, arXiv:2503.10677.

Rosaria Conte, Nigel Gilbert, Guido Bonelli, and 1 others. 2012. Manifesto of computational social science. *The European Physical Journal Special Topics*, 214(1):325–346.

Juliet Corbin and Anselm Strauss. 2008. *Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory*, 3 edition. SAGE Publications.

Ross-Helen Ferguson. 2017. Offline 'stranger' and online lurker: Methods for an ethnography of illicit transactions on the darknet. *Qualitative Research*, 17(6):683–698.

Katja Filippova. 2020. Controlled hallucinations: Learning to generate faithfully from noisy data. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 864–870.

Rohith Gangavarapu, Abhinav R. A. Srinivasan, and Venkatesh Moparthi. 2025. Evaluating accuracy in large language models: Benchmarking corrective RAG vs. naive retrieval augmented generation approach. In *2025 IEEE International Conference on AI and Data Analytics (ICAD)*, pages 1–7.

Luyu Gao, Xueguang Ma, Jimmy Lin, and Jamie Callan. 2023. Precise zero-shot dense retrieval without relevance labels. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1762–1777.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, and 1 others. 2021. Datasheets for datasets. *Preprint*, arXiv:1803.09010.

J. Amos Hatch. 2010. *Doing Qualitative Research in Education Settings*. SUNY Press.

James Howison, Andrea Wiggins, and Kevin Crowston. 2011. Validity issues in the use of social network analysis with digital trace data. *Journal of the Association for Information Systems*, 12(12).

Andreas Jungherr. 2015. *Analyzing Political Communication with Digital Trace Data: The Role of Twitter Messages in Social Science Research*. Springer.

Klaus H. Krippendorff. 2004. *Content Analysis: An Introduction to Its Methodology*, 2 edition. Sage Publications.

Jong Hyuk Lee, Ghulam Ali, and Jeong-In Hwang. 2025. A retrieval-augmented generation system for accurate and contextual historical analysis. *Computer Animation and Virtual Worlds*, 36(4):e70048.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, and 1 others. 2021. Retrieval-augmented generation for knowledge-intensive NLP tasks. *Preprint*, arXiv:2005.11401.

Stine Lomborg, Lina Dencik, and Hallvard Moe. 2020. Methods for datafication, datafication of methods: Introduction to the special issue. *European Journal of Communication*.

Deborah Lupton. 2015. *Digital Sociology*. Routledge.

Yury A. Malkov and Dmitry A. Yashunin. 2016. Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. *Preprint*, arXiv:1603.09320.

Noortje Marres. 2017. *Digital Sociology: The Reinvention of Social Research*. Polity Press.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919.

Song Mengmeng, Liu Zhibin, Wang Qingwei, Huang Man, and Xu Feiyang. 2024. An effective retrieval method to improve RAG performance. In *2024 7th International Conference on Data Science and Information Technology (DSIT)*, pages 1–5.

Leonardo F. Nascimento, Taciana Barreto, Leticia Cesarino, Vânia Mussa, and Priscila Fonseca. 2023. Públicos refratados: Grupos de extrema-direita brasileiros na plataforma Telegram. *Internet & Sociedade*.

Leonardo Fernandes Nascimento. 2016. A sociologia digital: Um desafio para o século XXI. *Sociologias*, 18:216–241.

Akhere Josephine Oche, Adegboyega G. Folashade, Tirthankar Ghosal, and Arindam Biswas. 2025. A systematic review of key retrieval-augmented generation (RAG) systems: Progress, gaps, and future directions. *Preprint*, arXiv:2507.18910.

Ollama. 2024a. Gemma3.

Ollama. 2024b. gpt-oss.

Janna Joceli Omena. 2019. *Métodos Digitais: Teoria-Prática-Crítica*. ICNOVA.

OpenAI. 2024. text-embedding-3-large model.

OpenAI. 2025. GPT-5 mini model.

Bernhard Rieder and Theo Röhle. 2017. Digital methods. In Mirko Tobias Schäfer and Karin van Es, editors, *The Datafied Society*, pages 109–124. Amsterdam University Press.

Richard Rogers. 2013. *Digital Methods*. MIT Press.

Jathan Sadowski. 2019. When data is capital: Datafication, accumulation, and extraction. *Big Data & Society*, 6(1).

Matthew J. Salganik. 2018. *Bit by Bit: Social Research in the Digital Age*. Princeton University Press.

Marcos L. Scheren, Vinícius S. Rodrigues, Guilherme D. López Zamora, and 1 others. 2024. Métodos mistos para a antropologia digital: Um relato de experiência sobre a análise de grupos bolsonaristas na plataforma Telegram. *Horizontes Antropológicos*, 30:e680407.

Silke Schwandt. 2022. Opening the black box of interpretation: Digital history practices as models of knowledge. *History and Theory*, 61(4):77–85.

Aditi Singh, Abul Ehtesham, Saket Kumar, and Tala Talaei Khoei. 2025. Agentic retrieval-augmented generation: A survey on agentic RAG. *Preprint*, arXiv:2501.09136.

Clare Southerton. 2020. Datafication. In Laurie A. Schintler and Connie L. McNeely, editors, *Encyclopedia of Big Data*, pages 1–4. Springer International Publishing.

Streamlit Inc. 2021. Streamlit — a faster way to build and share data apps.

Jiajie Sun, Xin Zhong, Shirui Zhou, and Jiawei Han. 2025. DynamicRAG: Leveraging outputs of large language model as feedback for dynamic reranking in retrieval-augmented generation. *Preprint*, arXiv:2505.07233.

TacticalTechVideos. 2014. Smari McCarthy, making data speak.

Chihiro Taguchi, Saku Maekawa, and Nikita Bhutani. 2025. Efficient context selection for long-context QA: No tuning, no iteration, just adaptive-k. *Preprint*, arXiv:2506.08479.

Robert Topinka, Alan Finlayson, and Camille Osborne-Carey. 2021. The trap of tracking: Digital methods, surveillance, and the far right. *Surveillance & Society*, 19(3):384–388.

José van Dijck. 2014. Datafication, dataism and dataveillance: Big Data between scientific paradigm and ideology. *Surveillance & Society*, 12(2):197–208.

Ruijie Xie, Junxiong Wang, Paul Rosu, and 1 others. 2025. Language models (mostly) know when to stop reading. *Preprint*, arXiv:2502.01025.

Wei Zhang and Jing Zhang. 2025. Hallucination mitigation for retrieval-augmented large language models: A review. *Mathematics*, 13(5).

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, and 1 others. 2023. Judging LLM-as-a-judge with MT-Bench and chatbot arena. *Preprint*, arXiv:2306.05685.

Yucheng Zhou, Xiubo Geng, Tao Shen, and 1 others. 2023. Thread of thought unraveling chaotic contexts. *Preprint*, arXiv:2311.08734.