# Virtual Reporter: Automation of News Report and Journalistic Data Collection Using LLMs

**Rafaela Almeida de Souza[1], Luís Flávio Ferreira Monteiro[1], Natan Carvalho de Barros Taques[1], Cristiano Maciel[1], Renan Vinicius Aranha[2], Thiago Meirelles Ventura[1]**

[1]Instituto de Computação – Universidade Federal de Mato Grosso (UFMT)
Cuiabá – MT – Brasil

[2]Escola Politécnica – Universidade de São Paulo
São Paulo – SP – Brasil

rafaela.souza@ifmt.edu.br,luisflaviomont@gmail.com,
natanbtaques@gmail.com, cristiano.maciel@ufmt.br, renanvinicius@usp.br,
thiago@ic.ufmt.br

**Abstract.** *Brazilian Higher Education Institutions (HEIs) receive many incomplete news proposals, increasing screening effort. This article presents the Virtual Reporter, an LLM-based workflow for academic communication that supports missing-field detection, similar-news retrieval, and briefing generation for human review. We evaluate four configurations (fixed-form, direct LLM, structured workflow, and Virtual Reporter + LLM/RAG) on a corpus of 10,189 news records, which we consolidate into 8,681 canonical records. Results on 300 incomplete pitches show that the structured workflow achieved perfect factual precision and no detected hallucinations, while the LLM/RAG variant obtained slightly higher completeness (0.3950 vs. 0.3896) at the cost of a 0.0517 hallucination rate. Factuality is automatically estimated, leaving final editing and publication to journalists.*

## 1. Introduction

Higher Education Institutions (HEIs) in Brazil produce a large volume of information through public notices, events, research and outreach projects, academic performance data, and administrative decisions. Their communication portals are not only publication channels; they also support active transparency and provide institutional, academic, and administrative information for internal and external audiences (Bastos et al., 2025).

This communication flow is intensified by the territorial distribution of the Federal Network. In large states, several campuses and sectors may be administratively linked to one central rectorate, creating practical distance between local information producers and the central communication department. As a result, story suggestions often arrive in fragmented form, without dates, responsible sources, locations, supporting links, or a clear indication of newsworthiness. Since news is produced through the perception, selection, and transformation of events into publishable information (Traquina, 2005), these incomplete demands require journalistic screening before any final text can be written. This challenge also occurs under personnel

constraints in the Brazilian federal administration, including Decree No. 10,185/2019, which abolished vacant positions in several areas and limited new appointments for the specified careers (Brasil, 2019).

The research problem addressed in this paper is: how can an LLM-based workflow support institutional communication teams in screening incomplete story pitches, identifying missing journalistic information, retrieving useful institutional context, and producing an editorial briefing without replacing human editorial judgment?

To address this problem, we propose the Virtual Reporter, a conversational workflow based on Large Language Models (LLMs) for the pre-reporting stage of academic institutional communication. The system receives an initial story pitch, asks follow-up questions, checks whether basic journalistic information is provided, retrieves similar institutional news from a historical corpus, and generates a structured summary for the communication department. Journalists remain responsible for final curation, editing, and publication.

Methodologically, the study is applied and artifact-oriented. It combines corpus construction, software workflow design, weakly supervised classification, information retrieval, and automatic evaluation of briefing generation. The contributions are fourfold: (i) the construction of a corpus of 10,189 raw news records from the Federal Institute of Mato Grosso (IFMT) and 8,681 canonical records; (ii) the design of a human-in-the-loop workflow for screening institutional news demands; (iii) preliminary evaluation of weakly supervised news type classification and similar-news retrieval; and (iv) comparison of fixed-form, direct LLM, structured workflow, and LLM/RAG (Retrieval-Augmented Generation) configurations using metrics related to completeness, factual precision, hallucination, and question coverage. The remainder of the paper presents related work, materials and methods, results, and final considerations.

## 2. Related Work

This section was prepared as a narrative review, following the review typology discussed by Grant and Booth (2009). The objective is not to provide an exhaustive systematic review, but to position the proposed workflow around four conceptual axes: computational news discovery, journalistic support tools, conversational and multi-agent systems, and retrieval-augmented generation.

Research on AI in journalism has explored computational methods for monitoring, filtering, selecting, and organizing information. Diakopoulos (2020) defines Computational News Discovery (CND) as an area of computational journalism concerned with algorithms that direct editorial attention to potentially newsworthy information. This perspective is closely related to the Virtual Reporter because the

proposed system operates before final news production, helping communication teams identify and structure institutional information.

Semi-automated news discovery tools also motivate this work. Diakopoulos, Trielli, and Lee (2021) present Algorithm Tips, a system that helps journalists identify potential stories using computational signals and expert assessment. More recently, Hagar et al. (2025) studied LLMs as initial filters in high-volume information flows and found that they can support lead extraction and general newsworthiness assessment, while still struggling with nuanced editorial judgments. These findings support the decision to keep the journalist as the final reviewer.

A second line of work uses generative agents to produce intermediate materials rather than final articles. Veerbeek and Diakopoulos (2024) propose specialized agents, such as analyst, reporter, and editor, to generate fact sheets for data-driven investigative reporting. The Virtual Reporter follows the same general idea of producing an intermediate document, but differs in context and input: it starts from incomplete institutional story suggestions and produces a briefing for academic communication. The design is also informed by studies on software-based dialogue systems (Motger et al., 2022), reflective multi-agent systems (Mohamed et al., 2025), and orchestrated conversational multi-agent architectures (Silva et al., 2025).

Finally, information retrieval is central because institutional communication depends on previous examples, local vocabulary, and available evidence. RAG combines language models with external memory or indexes, allowing generated outputs to use retrieved documents as support (Lewis et al., 2020). This is important in journalistic contexts because natural language generation may produce unsupported or hallucinatory content (Maynez et al., 2020; Ji et al., 2023). For this reason, the proposed workflow separates confirmed facts from pending information and produces a briefing for human review rather than a publishable news text.

## 3. Materials and Methods

According to Gil (2019), this study can be classified as applied, exploratory, and descriptive research, since it develops a computational artifact for a practical institutional problem and analyzes a real institutional corpus. In software engineering terms, it follows an artifact-oriented design-science logic, in which a system is built and evaluated against explicit problem-oriented criteria (Wieringa, 2014). The approach is qualitative in the problem framing and workflow design, complemented by quantitative experimental indicators for classification, retrieval, and briefing generation.

The study was organized into three stages: construction of a structured corpus of institutional news, design of the Virtual Reporter workflow, and evaluation of briefing generation from incomplete story pitches. The methodological description focuses on the information required to replicate the experiment: data sources, preprocessing decisions, system components, evaluated configurations, and automatic metrics.

## 3.1 Dataset construction and preprocessing

The corpus was collected from four public institutional news environments, kept separate because layouts and URL patterns differed: a historical IP-based portal, legacy subdomains, the current main portal, and campus websites. Certificate verification was disabled only for legacy or IP-based hosts when needed; HTML snapshots, logs, and source identifiers were preserved for auditing.

**Table 1. Source environments included in the final institutional news dataset from IFMT.**

| Source environment | Description | Raw records |
|---|---|---|
| legacy_ip_portal | Historical IP-based portal | 7,601 |
| old_subdomain_portal | Legacy subdomains | 2,247 |
| current_wordpress_portal | Current main news portal of IFMT | 212 |
| campus_current_portal | Campus news archives | 129 |
| **Total** | **Before duplicate filtering** | **10,189** |

For each article, the scraper stored source environment, original and final URLs, title, subtitle or lead when available, publication date, publisher unit or campus, body text, extracted links, attachments, images, and extraction status. The full dataset preserves duplicate publications because republication across portals and campuses is part of institutional communication history. A canonical view was also generated by grouping duplicates through normalized URLs, slugs, title-date signatures, content hashes, and approximate title/body similarity. This view contains 8,681 records and is used when experiments require one representative version of each news item.

Table 2 presents a compact example of the stored schema, showing how article metadata and cleaned textual content are represented in the dataset.

**Table 2. Example of a real record from the institutional news dataset from IFMT.**

| Field | Anonymized example value |
|---|---|
| **global_article_id** | gart_00000047 |
| **source_environment** | legacy_ip_portal |
| **title** | IFMT research transforms texts into comic books using AI |
| **publication_date** | 2025-09-19 |
| **publisher_unit_normalized** | campus_cuiaba |
| **body_text excerpt** | The initiative aims to modernize pedagogical practices [...] |
| **metadata flags** | has_images=true; has_links=true; has_attachments=true; is_canonical=true |

Preprocessing removed navigation menus, footers, sidebars, and repeated portal elements from article bodies. Dates were normalized when possible, and records with weak extraction were flagged rather than discarded. In the final consolidated dataset, all records have titles, 10,164 have body text, 9,075 have normalized publication dates, 6,039 contain links, 1,126 contain attachments, and 1,568 contain images.

### 3.2 Virtual Reporter workflow

The Virtual Reporter was designed as an LLM-based workflow for the pre-reporting stage of institutional communication. Based on an initial story proposal submitted by an institution staff member, the system first classifies the type of request, such as an event, public notice, research project, extension project, administrative information, or academic achievement. Next, an interviewer component asks follow-up questions guided by basic journalistic criteria, as presented by Lage (2008), who identifies the main elements of a news lead that should be developed throughout the paragraphs: What, to whom, when, where, how, why, and for what purpose? The process concludes with the identification of the responsible source and the institutional unit.

After the conversation, an integrity checker identifies missing or uncertain information. A retrieval module searches the institution’s corpus for similar news articles, following the Retrieval-Augmented Generation (RAG) approach (Lewis et al., 2020), so that the generated output can be guided by the institution's vocabulary and prior writing patterns. The final component produces a structured summary containing confirmed facts, pending issues, a suggested headline, a possible highlight, a recommended category, and examples of related news articles. The system does not automatically publish the text; instead, it generates an intermediate document for human editorial review.

In the evaluation, we distinguish between a structured workflow simulation and an LLM-augmented variant. The structured workflow simulation implements the pipeline deterministically, using rule-based type prediction, missing-field detection, adaptive question generation, retrieval, and structured briefing assembly. The Virtual Reporter + LLM/RAG variant uses the same workflow signals and retrieved examples, but delegates briefing generation to a locally deployed LLM. This distinction allows us to separate the effect of workflow structure from the effect of LLM generation.




**Figure 1: Virtual Reporter Institutional Pre-Reporting Workflow Diagram**

### 3.3 Experimental setup and evaluation

The evaluation used 300 incomplete institutional news pitches generated from unique canonical articles held out in the test split. Each article contributed one evaluation case, and each case was processed by all four configurations, yielding 1,200 outputs. For each case, a complete article was converted into an incomplete pitch through controlled removal of selected journalistic fields. The original article was retained only as ground truth for scoring and was not provided to any generation method. Retrieval-based configurations could access only documents from the permitted retrieval index, which excluded the target article. Controlled removal targeted fields such as the responsible actor, location, official source, institutional unit, date information, and participation or access information.

Four configurations were compared. The fixed-form baseline is a deterministic template that uses neither retrieval nor LLM generation. The direct LLM baseline uses a single prompt over the incomplete pitch only, with the local gemma4:e4b model executed through Ollama. The structured workflow simulation implements the pipeline without LLM generation. Finally, the Virtual Reporter + LLM/RAG variant uses the same local model as the direct LLM baseline, but conditions generation on workflow signals and retrieved examples. No configuration received the full original article, expected fields, or original title during generation.

Outputs were scored with automatic metrics aligned with the pre-reporting objective. Completeness measures the number of required journalistic fields that were correctly filled. Factual precision and hallucination rate estimate the proportion of generated claims that are supported or unsupported by the available evidence, namely the pitch and, when applicable, retrieved documents. Question coverage measures whether removed fields are addressed by at least one follow-up question. These automatic metrics do not replace human editorial assessment.

## 4. Data Processing and Results

This section summarizes the data processing and evaluation results. The experiments were implemented as reproducible scripts that read the consolidated dataset, generate weak labels, train classification models, build retrieval indexes, create incomplete test pitches, execute the evaluated configurations, and export tables and reports.

### 4.1 Dataset filtering and labeling

The modeling stage used the canonical dataset with 8,681 records. After filtering records with insufficient textual content, 8,474 articles remained. Because human labels were not available, news types were assigned with weak supervision based on institutional keywords and title/body patterns, a strategy commonly adopted when labeled data are scarce and programmatic rules are used to create noisy training signals (Ratner et al., 2017).

The distribution is highly imbalanced. The largest class is edital_selection, followed by event, while service_announcement has very few examples. For this reason, the classification results should be interpreted as preliminary evidence of textual regularity in the corpus, not as a final human-validated taxonomy.

**Table 3. Distribution of weak labels assigned to the institutional news records.**

| Weak label | Records |
|---|---|
| edital_selection | 5,096 |
| event | 1,505 |
| award_result | 562 |
| research | 491 |
| teaching | 354 |
| extension | 196 |
| institutional_management | 174 |
| other | 83 |
| service_announcement | 13 |

### 4.2 Classification and similar-news retrieval

The news type classification task evaluated whether institutional news could be automatically categorized from textual patterns. The best result was obtained with TF-IDF features (Salton and Buckley, 1988) and a Linear Support Vector Classifier (Cortes and Vapnik, 1995), reaching 0.7657 accuracy, 0.5173 macro-F1, and 0.7650 weighted-F1. The gap between macro-F1 and weighted-F1 reflects the imbalance of the weak-label distribution: frequent classes dominate the corpus and are easier to identify.

The retrieval task evaluated whether incomplete pitch-like queries could recover institutionally similar news. After a sanity-check stage, a stricter evaluation removed the target article from the retrieval index and used incomplete queries derived from held-out articles. In this setting, the goal is not to retrieve the same article, but useful examples with a similar type or institutional unit. The evaluated strategies included BM25 retrieval (Robertson and Zaragoza, 2009) and a hybrid retriever that combines lexical and vector-based signals. Approximate nDCG@10 was used as an information retrieval indicator (Järvelin and Kekäläinen, 2002).

**Table 4. Strict retrieval evaluation results.**

| Method | same_type@5 | same_type_ratio@5 | same_unit@5 | approx. nDCG@10 |
|---|---|---|---|---|
| Hybrid | 0.7220 | 0.4008 | 0.8004 | 0.3520 |

The hybrid retriever obtained the best strict retrieval result. At least one of the top-5 retrieved articles had the same weak news type in 72.20% of the test cases, and at least one retrieved article had the same institutional unit in 80.04% of cases when unit information was available. This suggests that the corpus can support the retrieval of useful institutional examples even when the target article is not present in the index.

### 4.3 Baseline comparison for briefing generation

The briefing-generation experiment compared the four configurations across 300 incomplete institutional news pitches, yielding 1,200 outputs. The direct LLM baseline obtained a completeness score of 0.2890, factual precision of 0.8886, a hallucination rate of 0.1114, and question coverage of 0.3073. The fixed-form baseline and structured workflow simulation both achieved completeness of 0.3896, factual precision of 1.0000, zero detected hallucinations, and full question coverage. The Virtual Reporter + LLM/RAG variant achieved the highest completeness (0.3950) and full question coverage, but factual precision decreased to 0.9483 and the hallucination rate increased to 0.0517.

**Table 5. Automatic evaluation on 300 incomplete institutional news pitches.**

| Config. | Compl. | Fact. Prec. | Halluc. | Q. Cov. | N |
|---|---|---|---|---|---|
| Fixed-form baseline | 0.3896 | 1.0000 | 0.0000 | 1.0000 | 300 |
| Direct LLM baseline | 0.2890 | 0.8886 | 0.1114 | 0.3073 | 300 |
| Structured workflow simulation | 0.3896 | 1.0000 | 0.0000 | 1.0000 | 300 |
| Virtual Reporter + LLM/RAG | 0.3950 | 0.9483 | 0.0517 | 1.0000 | 300 |

For readability, Table 5 abbreviates the evaluated metrics as follows: i) Compl. = completeness; ii) Fact. Prec. = factual precision; iii) Halluc. = hallucination rate; and iv) Q. Cov. = question coverage.

These results suggest that workflow structure provides a strong factuality-completeness trade-off for incomplete story pitches. Compared with the deterministic structured workflow, the LLM/RAG variant increased completeness by 0.0054 (0.3950 vs. 0.3896), but introduced a hallucination rate of 0.0517 and reduced factual precision to 0.9483. The deterministic workflow, therefore, remained the most conservative configuration, while LLM-generated briefings still require human review.

### 4.4 Representative case

To illustrate the workflow beyond aggregate metrics, we selected a representative case involving an incomplete selection-notice pitch. This case is useful because selection notices are frequent in the corpus and because it includes fields commonly missing from evaluation cases, such as the responsible actor, date, access channel, official source, and institutional unit.

The initial pitch states only: “I want to publicize a free course on innovation in the pet sector. Registration is open.” Although it provides the topic and indicates that registration is available, it omits essential information such as the registration period, course modality, responsible unit, registration link, number of seats, target audience, contact person, and official source. The workflow classifies the demand as a course or training-related announcement, while keeping this classification open to editorial revision.

The retrieval module then searches the permitted institutional index for similar previous news about free courses, registrations, extension activities, and public calls. These examples support vocabulary, structure, and category cues, but they are not treated as facts about the current demand. Based on the missing fields, the system generates follow-up questions about the registration period, modality, responsible campus or unit, registration link, number of seats, target audience, and official document or contact.

After the staff member provides additional information, the workflow separates confirmed facts from pending or uncertain details. The final output is a structured briefing for editorial review, including confirmed information, remaining gaps, a suggested headline or category, related examples from the corpus, and a warning that uncertain fields must be checked before publication. A journalist then verifies the sources, edits the text, and decides whether the demand should become a news story, a short note, an agenda item, a service announcement, or remain under review.

This example shows that the RAG component provides contextual support without guessing missing facts. The output is therefore a pre-reporting artifact for human review, not an automatically publishable news story.

## 5. Final Considerations

This work contributes to AI in journalism by focusing on a specific workflow: the pre-reporting stage of academic institutional communication. The Virtual Reporter is not presented as a generic news generator, but as a support tool for collecting, clarifying, retrieving, and organizing information before journalistic writing. This position is important because communication departments in educational institutions

handle distributed demands from campuses, projects, public notices, events, and administrative units.

The study also contributes a structured corpus of institutional news from a selected model institution. The final dataset contains 10,189 raw records and 8,681 canonical records with text, source metadata, publication dates, institutional units, links, attachments, and images. The processing stage generated weak labels, trained a news-type classifier, evaluated retrieval strategies, and compared four briefing-generation configurations. The full system implementation and dataset are available at https://github.com/luisflaviomf/virtual-reporter.

The results indicate that workflow structure and retrieval are relevant to institutional pre-reporting. The structured workflow provided the most conservative balance between completeness and factuality, while the LLM/RAG variant slightly improved completeness at the cost of unsupported claims. The direct LLM baseline was less reliable overall. These findings reinforce the need to keep journalists responsible for final curation, editing, and publication.

This work has limitations related to data scope, weak supervision, and automatic evaluation. Because the corpus comes from a single HEI and public web archives, the results may reflect institutional, regional, temporal, and publication-practice biases. Automated preprocessing, duplicate grouping, weak labeling, and synthetic pitch generation may also influence the results. Finally, factual precision and hallucination were estimated automatically and should be complemented by human evaluation in future work.

Future work should include manual annotation of a subset of the dataset, evaluation with communication professionals, refinement of question generation, and broader testing with additional institutions and LLMs. Less frequent categories remain harder to model and should be revised with manually annotated labels. The current system should therefore be interpreted as preliminary evidence that LLM-assisted workflows can support institutional pre-reporting, not as evidence that automated systems should replace editorial judgment.

## Acknowledgement and use of generative AI

Generative AI tools were used only as auxiliary support for language revision, image creation, and manuscript condensation. The authors reviewed and take responsibility for all content, experiments, analyses, and claims.

## References

Bastos, Bárbara L. E.; GONÇALVES, Estael L.; XAVIER, Cleon; VALLE, Pedro Henrique Dias; ARANHA, Renan V.. Uma página, muitas informações: como as universidades federais lidam com diferentes requisitos legais em seus sites?. In:

SIMPÓSIO BRASILEIRO DE ENGENHARIA DE SOFTWARE (SBES), 39., 2025, Recife/PE. Anais [...]. Porto Alegre: Sociedade Brasileira de Computação, 2025. p. 748-754. ISSN 2833-0633. DOI: https://doi.org/10.5753/sbes.2025.11364.

Brasil. Decreto nº 10.185, de 20 de dezembro de 2019. Extingue cargos efetivos vagos e que vierem a vagar dos quadros de pessoal da administração pública federal e veda a abertura de concurso público e o provimento de vagas adicionais para os cargos que especifica. Brasília, DF, 2019. Available at: https://www2.camara.leg.br/legin/fed/decret/2019/decreto-10185-20-dezembro-2019-789637-publicacaooriginal-159749-pe.html. Accessed on: 31 May 2026.

Cortes, Corinna; VAPNIK, Vladimir. Support-vector networks. Machine Learning, v. 20, p. 273-297, 1995. DOI: 10.1007/BF00994018.

Diakopoulos, Nicholas. Computational News Discovery: Towards Design Considerations for Editorial Orientation Algorithms in Journalism. Digital Journalism, v. 8, n. 7, p. 945-967, 2020. DOI: 10.1080/21670811.2020.1736946.

Diakopoulos, Nicholas; TRIELLI, Daniel; LEE, Grace. Towards Understanding and Supporting Journalistic Practices Using Semi-Automated News Discovery Tools. Proceedings of the ACM on Human-Computer Interaction, v. 5, CSCW2, Article 406, p. 1-30, 2021. DOI: 10.1145/3479550.

Gil, Antonio Carlos. Como elaborar projetos de pesquisa. 6. ed. São Paulo: Atlas, 2019.

Grant, Maria J.; BOOTH, Andrew. A typology of reviews: an analysis of 14 review types and associated methodologies. Health Information & Libraries Journal, v. 26, n. 2, p. 91-108, 2009. DOI: 10.1111/j.1471-1842.2009.00848.x.

Hagar, Nick; SILVER, Ethan; SPENCER, Clare; DIAKOPOULOS, Nicholas. LLM-Assisted News Discovery in High-Volume Information Streams: A Case Study. arXiv preprint arXiv:2509.25491, 2025. Available at: https://arxiv.org/abs/2509.25491. Accessed on: 31 May 2026.

Järvelin, Kalervo; KEKÄLÄINEN, Jaana. Cumulated gain-based evaluation of IR techniques. ACM Transactions on Information Systems, v. 20, n. 4, p. 422-446, 2002. DOI: 10.1145/582415.582418.

Ji, Ziwei; LEE, Nayeon; FRIESKE, Rita; YU, Tiezheng; SU, Dan; XU, Yan; ISHII, Etsuko; BANG, Yejin; CHEN, Delong; DAI, Wenliang; CHAN, Ho Shu; MADOTTO, Andrea; FUNG, Pascale. Survey of Hallucination in Natural Language Generation. ACM Computing Surveys, v. 55, n. 12, Article 248, p. 1-38, 2023. DOI: 10.1145/3571730.

Lage, Nilson. Estrutura da notícia. 6. ed. São Paulo: Ática, 2008.

Lewis, Patrick; PEREZ, Ethan; PIKTUS, Aleksandra; PETRONI, Fabio; KARPUKHIN, Vladimir; GOYAL, Naman; KÜTTLER, Heinrich; LEWIS, Mike; YIH, Wen-tau; ROCKTÄSCHEL, Tim; RIEDEL, Sebastian; KIELA,

Douwe. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In: Advances in Neural Information Processing Systems 33 (NeurIPS 2020), 2020. DOI: doi.org/10.48550/arXiv.2005.11401

Maynez, Joshua; NARAYAN, Shashi; BOHNET, Bernd; MCDONALD, Ryan. On Faithfulness and Factuality in Abstractive Summarization. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, p. 1906-1919, 2020. DOI: 10.18653/v1/2020.acl-main.173.

Mohamed, Aissa Hadj; SANTOS, Frances A.; REIS, Julio Cesar dos. Enhancing LLM Agent Effectiveness via Reflective Multi-Agent System. In: Workshop-Escola de Sistemas de Agentes, seus Ambientes e Aplicações (WESAAC), 19., 2025, Fortaleza/CE. Anais [...]. Porto Alegre: Sociedade Brasileira de Computação, 2025. p. 206-217. DOI: 10.5753/wesaac.2025.37536.

Motger, Quim; FRANCH, Xavier; MARCO, Jordi. Software-Based Dialogue Systems: Survey, Taxonomy, and Challenges. ACM Computing Surveys, v. 55, n. 5, p. 1-42, 2022. DOI: 10.1145/3527450.

Ratner, Alexander; BACH, Stephen H.; EHRENBERG, Henry; FRIES, Jason; WU, Sen; RÉ, Christopher. Snorkel: Rapid Training Data Creation with Weak Supervision. Proceedings of the VLDB Endowment, v. 11, n. 3, p. 269-282, 2017. DOI: 10.14778/3157794.3157797.

Robertson, Stephen; ZARAGOZA, Hugo. The Probabilistic Relevance Framework: BM25 and Beyond. Foundations and Trends in Information Retrieval, v. 3, n. 4, p. 333-389, 2009. DOI: 10.1561/1500000019.

Salton, Gerard; BUCKLEY, Christopher. Term-weighting approaches in automatic text retrieval. Information Processing & Management, v. 24, n. 5, p. 513-523, 1988. DOI: 10.1016/0306-4573(88)90021-0.

Silva, Eryck; SANTOS, Frances A.; THOMPSON, Pedro; REIS, Julio C. dos. LLM-Powered Conversational Multi-Agent Cognitive System for Collaborative Task Solving. In: Workshop-Escola de Sistemas de Agentes, seus Ambientes e Aplicações (WESAAC), 19., 2025, Fortaleza/CE. Anais [...]. Porto Alegre: Sociedade Brasileira de Computação, 2025. p. 59-70. DOI: 10.5753/wesaac.2025.37528.

Traquina, Nelson. Teorias do jornalismo: porque as notícias são como são. 2. ed. Florianópolis: Insular, 2005.

Veerbeek, Joris; Diakopoulos, Nicholas. Using Generative Agents to Create Tip Sheets for Investigative Data Reporting. arXiv preprint arXiv:2409.07286, 2024. DOI: 10.48550/arXiv.2409.07286.

Wieringa, Roel J. Design Science Methodology for Information Systems and Software Engineering. Berlin, Heidelberg: Springer, 2014. DOI: 10.1007/978-3-662-43839-8.