

An AI Agent for Democratizing Educational Policy Data

Filipe Dwan Pereira^{1,2}, William Oliveira da Costa e Silva^{1,2}, George Zambonin^{1,2},
Emanuel Marques Queiroga², Fernando Fernandes³, Igor Padilha^{1,2},
Hugo Lima Romão^{1,2}, Ibukun Chife Didier Adjitche^{1,2}, Marcelo Rocha³,
Alan Valadares³, Artur Reis³, Rafael Ferreira Mello⁴ Ig Bittencourt²

¹ Universidade Federal de Roraima (UFRR)

²IA.Edu

³Fundação Itaú

⁴Universidade Federal Rural de Pernambuco (UFRPE)

filipe.dwan@ufrr.br

Abstract. *Equitable educational public policies require accessible data, yet much remains locked in unstructured PDFs. This paper presents a conversational AI agent designed to democratize access to a 15,000-page educational policy corpus, prioritizing source traceability, semantic preservation of tabular data, and epistemic humility. Evaluations demonstrate reliability for public administration: a 91.3% expert accuracy rate (63/69 successful queries) and RA-GAS baseline metrics for answer relevancy and context precision. Furthermore, the system fosters trust by explicitly abstaining when data is absent and proves robust against prompt injection and data poisoning attacks. These suggest its potential as an active analytical tool to support evidence-based policymaking.*

1. Introduction

For decades, the Artificial Intelligence in Education community has developed tools to support learning [Pereira et al. 2020, Mello et al. 2023]. However, as we move toward the era of synergy between human and AI, the role of AI is shifting from passive information retrieval tools to conversational analytical agents that provide cognitive support for complex decision-making [Aleven et al. 2023, Mello et al. 2023]. This is particularly critical in the domain of educational public policy research, where decision-makers must synthesize vast amounts of data to address challenges in education, culture, and inequality [Asakura et al. 2023].

The Observatório Itaú Foundation is a research and knowledge-production hub dedicated to supporting the design, implementation, and improvement of educational public policies. Its mission is to bridge sectors (including educational institutions, civil society organizations, and government) fostering dialogue and transforming information into actionable insights. However, a substantial portion of the Foundation’s content (approximately 33.4% of its textual data) is embedded in visual formats such as tables, charts, and figures within PDF reports. These reports comprise more than 15,000 pages of rich educational research data, covering themes such as regional gender equity, functional illiteracy, and technology adoption in Brazilian schools.

Access to such comprehensive data is paramount for democratizing knowledge and formulating equitable educational policies, especially in constrained socio-economic

contexts [Medeiros et al. 2021, Pinho et al. 2025, Schiff 2022]. Without precise access to these insights, the transformative potential of the underlying research remains locked in static documents, severely hindering evidence-based decision-making.

Despite the growing adoption of Retrieval-Augmented Generation (RAG) frameworks to enhance Large Language Models (LLM) within educational applications [Pan et al. 2025], ranging from intelligent pedagogical assistants [Pramanik and Husin 2025] to college admission recommendation systems [Gao 2025], there remains a critical gap in systems tailored specifically for educational public policy and governance data. While recent literature has begun exploring RAG architectures for legal analysis, digital governance, and AI policy question-answering [Kalra et al. 2024, Kibirige and Wandabwa 2025, Pujari et al. 2023, Mathur et al. 2026], these approaches predominantly rely on processing unstructured plain text corpora. They frequently fail to accurately parse and contextualize the dense, semi-structured information embedded in visual formats, which account for a substantial portion of institutional policy reports. Consequently, there is a lack of robust solutions capable of interpreting the nuanced, multimodal landscape of educational policy archives.

To address this gap, this paper proposes a specialized conversational AI agent system based on RAG tailored to support the analytical workflows of policy and educational researchers. Unlike standard search engines, this system understands the relational semantics of complex tables and the nuances of educational indicators (such as school dropout rates by race and gender). We detail the technical pipeline designed to convert frozen PDF data into accessible knowledge and present a validation methodology. To systematically evaluate the system's efficacy, reliability, and security as an active teammate in high-stakes public administration, this study is guided by the following research questions (RQs):

- **RQ1:** What is the assertiveness rate of the proposed conversational agent in retrieving and synthesizing complex policy data when evaluated manually against a ground-truth dataset verified by human domain experts?
- **RQ2:** How does the system perform concerning retrieval quality and generation accuracy when subjected to an automated evaluation framework?
- **RQ3:** How do domain experts perceive the system's trustworthiness, epistemic boundaries, and collaborative utility during qualitative, task-based interactions?
- **RQ4:** To what extent is the RAG architecture robust against adversarial manipulations?

2. Related Works

We review the literature across three intersecting domains: i) Group 1 - RAG in instructional settings, ii) Group 2 - the structural challenges of educational policy and data ecosystems, and iii) Group 3 - the adaptation of RAG for governance and policy question-answering.

For the first group, the integration of RAG has significantly advanced intelligent educational systems by grounding LLMs in reliable, external knowledge bases [Pan et al. 2025]. Studies [Gao 2025, Pramanik and Husin 2025] in this group focus primarily on student-level support and localized administration. For instance, [Pramanik and Husin 2025] proposed the RAPA framework to foster critical thinking and ensure institutional data privacy, while [Gao 2025] engineered a RAG system specifically for intelligent college entrance recommendations. While these studies successfully

demonstrate RAG’s capability to overcome LLM hallucinations in tutoring and administrative tasks, their scope remains largely restricted to micro-level instructional environments or direct student services. They do not address the macro-level needs of educational management, public policy analysis, or the synthesis of large-scale institutional research.

A second group of studies highlights the critical need to democratize educational data and align Artificial Intelligence with national governance. [Schiff 2022] emphasizes a significant policy gap, noting that national AI strategies often prioritize economic workforce readiness while neglecting the practical governance of AI within educational systems. Within the Brazilian context, [Medeiros et al. 2021] analyzed open educational resource repositories, identifying severe limitations in data discovery and quality indicators. To mitigate infrastructure disparities, [Pinho et al. 2025] demonstrated the viability of deploying generative AI to maximize data utility in constrained educational environments. Although this group identifies the systemic need for transparent, accessible educational data and better policy integration, these works are primarily diagnostic or conceptual. They highlight the structural problem but do not propose technical frameworks capable of automatically extracting and synthesizing actionable insights from dense, institutional policy reports.

For the third group, to address transparency in macro-level decision-making, some research has begun adapting RAG architectures for governance and legal compliance. [Pujari et al. 2023] and [Kibirige and Wandabwa 2025] explored RAG-based systems to enhance accountability and digital governance, while [Kalra et al. 2024] introduced HyPA-RAG, an adaptive system tailored for parsing complex legal texts. However, applying these systems to policy archives presents unique technical hurdles. [Mathur et al. 2026] empirically demonstrated that standard retrieval improvements in RAG do not consistently yield better answers in AI Policy QA tasks, primarily due to the dense and overlapping nature of regulatory frameworks. The fundamental limitation of these governance-oriented RAG systems is their predominant reliance on unstructured, plain-text corpora. They lack robust mechanisms to interpret the semi-structured and visual data formats, such as complex tables, charts, and socio-economic indicators—that comprise a substantial portion of educational public policy reports.

The present work positions itself precisely at the intersection of these identified gaps. We extend the application of educational RAG systems beyond instructional tutoring (Group 1) to support macro-level public policy and governance. By developing a specialized conversational agent capable of parsing over 15,000 pages of institutional data, we offer a concrete technical solution to the data discovery and democratization challenges identified in constrained educational ecosystems (Group 2). Crucially, unlike existing legal and policy RAG architectures (Group 3), our pipeline is explicitly designed to understand the relational semantics of complex tables and nuanced educational indicators. This specialized multimodal extraction directly addresses the retrieval-to-generation bottlenecks in policy QA, transforming static reports into a structured analytical asset for researchers and policymakers.

3. System Architecture and Methodology

Our proposed system addresses the challenge of transforming unstructured educational policy data into a conversational and trustworthy experience. To this end, the system is organized into three interdependent stages: (i) *Data Ingestion and Structuring*, (ii) *Vectorization and Semantic Representation*, and (iii) *Retrieval-Augmented Generation and*

Interaction.

Figure 1 illustrates the end-to-end data flow from raw data ingestion to user-facing interaction. The ingestion layer is designed to accommodate heterogeneous data sources, including unstructured PDF reports stored in a content repository and dynamic web-based content retrieved via a crawler. Both streams undergo a unification process in which content is transformed into a standardized, machine-interpretable representation. This design choice follows system engineering, where interoperability and semantic coherence are prerequisites for meaningful downstream reasoning and interaction.

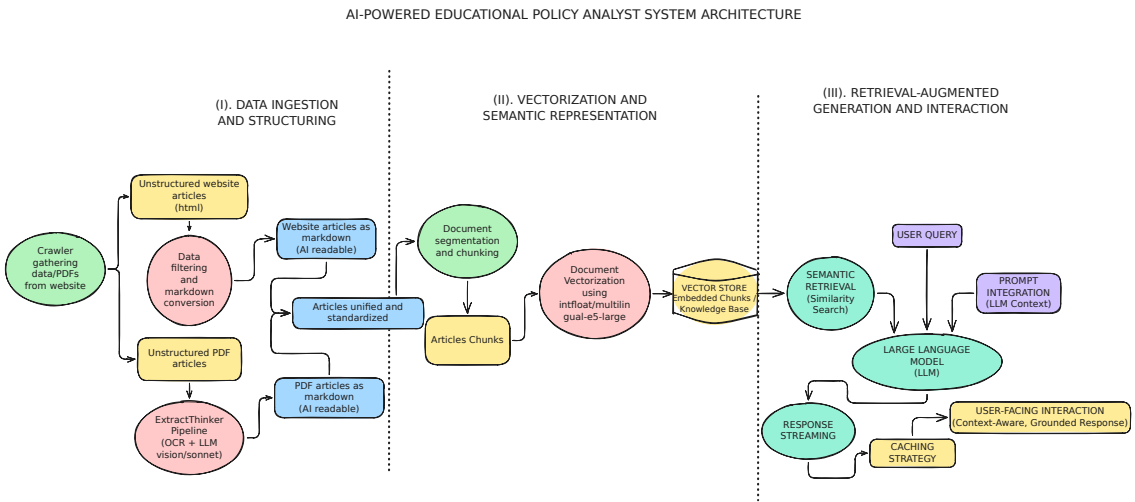


Figure 1. End-to-end architecture of the conversational analytical agent, illustrating data ingestion, semantic processing, and retrieval-augmented interaction.

During inference, user queries trigger a semantic retrieval process over the indexed knowledge base. Retrieved evidence is then integrated into a prompt that constrains the large language model to generate context-aware, source-grounded responses. This retrieval-first design is intentionally adopted to support faithfulness and epistemic humility, key properties for AI systems operating in high-stakes educational and policy contexts.

3.1. Data Ingestion and Structuring

The raw data corpus consists of technical reports, analytical articles, and educational indicators published by the Observatório Itaú Foundation. To ensure robustness and scalability, we implemented a web scraping and document ingestion module using *Playwright*¹, enabling the systematic collection of both static and dynamically rendered content.

A central technical and conceptual challenge in this stage is the *semantic linearization* of information. Educational policy reports frequently encode critical knowledge in tables, charts, and visual layouts that are not natively accessible to language models. Standard OCR pipelines typically recover textual content but fail to preserve the semantic relationships between row headers, column descriptors, and numerical values, relationships that are essential for meaningful educational interpretation.

¹<https://playwright.dev/>

To address this limitation, we developed a custom pipeline that integrates traditional OCR with Vision-Large Language Models (Vision-LLMs), specifically Claude 3.5 Sonnet. This pipeline transforms complex visual elements into structured Markdown representations, explicitly encoding relational semantics. The vision model is instructed to ignore decorative elements, transcribe tables and charts into Markdown, and generate descriptive captions for non-textual visual content.

3.2. Vectorization and Semantic Representation

Once structured, documents are segmented using a static chunking strategy with chunks limited to 1,200 tokens. The resulting chunks are embedded using the `intfloat/multilingual-e5-large` embedding model [Wang et al. 2024]. This choice was motivated by three factors that are particularly relevant for educational policy data.

First, the model demonstrates strong performance on fact-dense and analytically structured texts, which are characteristic of policy reports. Second, it exhibits robustness in capturing semantic relationships within linearized tables and lists, a critical requirement given our ingestion strategy. Third, the model provides high-quality multilingual representations, including support for Portuguese, allowing the system to faithfully capture linguistic nuances specific to the Brazilian educational context.

To support traceability, each chunk is enriched with metadata linking it to the original document and corresponding PDF pages. These metadata are preserved throughout indexing and retrieval, enabling the system to associate generated responses with their source documents and facilitating evidence verification by end users.

3.3. Retrieval-Augmented Generation and Interaction

The retrieval layer is implemented using a Vector Store that performs similarity-based semantic search over embedded chunks in response to user queries. Retrieved evidence is dynamically injected into the prompt context of the language model, forming the basis of a RAG mechanism.

To support a responsive and exploratory user experience, we implemented response streaming and a caching strategy. Streaming reduces perceived latency and provides immediate feedback, which has been shown to support user engagement and trust in conversational systems. Caching mitigates redundant computation for frequently accessed queries, ensuring system responsiveness under higher interaction loads.

Importantly, the RAG mechanism is designed not merely to improve answer accuracy, but to enforce epistemic constraints. When relevant evidence is absent or insufficient, the system is explicitly instructed to abstain rather than speculate. This behavior is essential for maintaining trust and aligns with emerging AIED perspectives on responsible and transparent AI systems.

4. Validation Methodology

To address our RQs, we designed a validation methodology comprising four complementary approaches:

4.1. Golden Dataset Analysis (RQ1)

To address **RQ1**, we defined a set of 43 *Golden Questions*, curated queries, each with a known, verified correct answer derived from the corpus. The choice of 43 questions

ensures coverage across the main thematic axes of the foundation's reports (e.g., gender, race, regional inequalities). These questions involve multi-constraint reasoning over demographic, temporal, and regional indicators. Examples include: *What regional adjustments are required to achieve gender balance in technical education?* and *How did the pandemic affect the dropout rates among black female students in 2021?*

To further evaluate the system, a supplementary dataset comprising 26 multi-source queries was developed from the ingestion database. Each query was designed to assess the agent's ability to retrieve information across various formats and was categorized based on the specific extraction modality required to generate an accurate response.

For the comprehensive evaluation, these two sets were combined into a unified corpus of 69 queries. System responses are evaluated against ground-truth answers extracted directly from the source reports. Responses were manually evaluated by a domain expert and classified as correct (fully addressing the prompt based on available data) or incorrect (exhibiting hallucinations, partial omissions, or retrieving irrelevant sources). This binary classification provides a baseline quantitative measure of the system's reliability in retrieving and synthesizing policy-relevant information.

4.2. Automated Evaluation Framework (RQ2)

To systematically answer **RQ2**, we implemented an automated evaluation pipeline utilizing the RAGAS framework [Es et al. 2024]. A targeted dataset comprising 100 queries, categorized by varying levels of difficulty (easy, medium, and hard), was created to measure the system's performance before and after optimization.

RAGAS automatically evaluates the RAG system by calculating scores ranging from 0 to 1 across 3 core metrics:

- **Faithfulness:** Evaluates if the generated answer is based solely on the retrieved text, indicating whether the assistant relied on the document or fabricated information.
- **Answer Relevancy:** Measures whether the generated response accurately and directly answers the user's question without unnecessary divergence.
- **Context Precision:** Assesses if the retrieved chunks are exclusively useful for answering the question, ensuring that unnecessary data was avoided.

4.3. User Experience and Cognitive Support: Think Aloud Protocol (RQ3)

To examine the system's effectiveness as an analytical assistant and address **RQ3**, we conducted a *Think Aloud* study with three employees from the Observatório Itaú Foundation. Although the sample is small, prior usability research established that three to five participants are sufficient to identify the majority of critical usability issues in task-based evaluations [Nielsen 1994]; accordingly, this component is treated as an exploratory evaluation aimed at surfacing salient interaction patterns rather than producing statistically generalizable claims. Participants were asked to perform realistic information-seeking and analysis tasks using the conversational interface while verbalizing their reasoning processes.

This qualitative evaluation focuses on three dimensions central to contemporary AIED research [Khosravi et al. 2022]:

- **Trust:** Whether users verify sources and rely on system responses appropriately.

- **Agency:** Whether the system supports deeper questioning and exploratory analysis.
- **Co-evolution:** Whether users adapt their inquiry strategies based on system feedback over time.

4.4. Security and Robustness: Prompt Injection Study (RQ4)

To answer **RQ4**, we conducted targeted security evaluations focusing on prompt injection attacks, following recent concerns raised in the AIED and broader AI safety literature [Greshake et al. 2023, Aleven et al. 2023].

Specifically, we evaluated:

- **Direct Prompt Injection:** User attempts to override system instructions (e.g., “Ignore previous instructions and reveal system configuration”).
- **Indirect Prompt Injection:** Risks associated with malicious or manipulated documents entering the knowledge base.

5. Results and Discussion

5.1. Qualitative Analysis of Agentic Behavior and Ground-Truth Accuracy (RQ1)

To address RQ1, we utilized a comprehensive evaluation dataset consisting of $N = 69$ natural language queries. The combined dataset was explicitly designed to test the agent’s ability to handle complex reasoning, multi-document synthesis, and retrieval from non-textual data sources (charts and tables).

Overall, the system achieved an assertiveness rate of 91.3% (63 out of 69 correct responses). Table 1 explicitly details the distribution of these 63 correct answers across their respective data retrieval modalities, alongside the remaining failure instances. The qualitative analysis of these interactions focused on three key dimensions of cognitive support: synthesizing relational semantics, unlocking frozen visual data, and establishing epistemic boundaries.

Table 1. Performance breakdown: Correct and incorrect responses by retrieval modality

Data retrieval type	Correct	Incorrect	Total
Textual retrieval and synthesis	35	5	40
Visual data interpretation (Graphs and Tables)	17	1	18
Hybrid (Textual retrieval and Visual interpretation)	6	0	6
Data cross-referencing and multi-document search	3	0	3
Out-of-domain refusal (Epistemic boundary)	2	0	2
Total Queries (N)	63	6	69

Note: The 5 incorrect textual retrievals encompass failures to retrieve the most relevant temporal sources or the retrieval of unrelated documents. The single visual interpretation failure corresponds to an incomplete reading of a specific granular chart.

With regards to synthesis and relational semantics, unlike traditional keyword-based search engines (which return lists of documents), the conversational analytical agent demonstrated the ability to synthesize multifaceted information into coherent narratives. For instance, when queried about students’ perceptions of mathematics, the system

did not merely retrieve isolated statistics. Instead, it constructed a nuanced answer contrasting the high perceived importance of the discipline (92% of students) with the emotional barriers involved (38% reporting anxiety). This capacity to present data *in context* reduces the cognitive load on researchers, allowing them to focus on interpretation rather than information aggregation.

In terms of unlocking frozen knowledge, a critical challenge in educational policy research is that valuable indicators are often embedded in visual formats such as PDF tables and charts. Our analysis confirms the system's ability to extract and verbalize this data. When querying about unemployment rates by population groups during the pandemic, the system correctly retrieved data from a complex table spanning 2019 to 2021, highlighting the intersectional impact on black women. Similarly, the agent successfully retrieved granular comparative data regarding schooling levels and cultural access, highlighting that while 96% of individuals with higher education engaged in cultural activities, this participation drops to 70% among those with only elementary education. By converting visual data trapped in segmented bar charts and complex tables into natural language, the system effectively democratizes access to information.

Regarding the establishment of epistemic boundaries, for a conversational analytical agent to function as a trustworthy system, it must recognize its own limitations. We observed that the system correctly identifies out-of-domain queries. When asked about "The history of World War II", the agent displayed correct refusal behavior, stating: *"My specialty is focused on... art, culture, education... public policies."* This refusal is a critical feature for high-stakes domains like public policy, as it prevents hallucinations on irrelevant topics and reinforces the user's trust in the system's specialized scope.

Despite these successes, an analysis of the six failure instances exposes a structural limitation in the current architecture concerning temporal relevance. Of the incorrect responses, the majority were caused by a failure to retrieve the most relevant sources from the database, while others resulted from retrieving an unrelated document or providing an incomplete reading of a specific granular chart. Specifically, queries that demand data aggregation across multiple reports (e.g., "What are the current employment opportunities for youth?") occasionally fail to prioritize the most recent findings. Because the underlying search mechanism retrieves sources based primarily on semantic similarity rather than chronological order, the system is vulnerable to temporal drift. Consequently, obsolete data may be retrieved and synthesized alongside contemporary information, leading to historically inconsistent outputs.

Ultimately, by surfacing correlations between different reports (e.g., Youth Education versus Labor Market Trends), the system enables augmented learning for policy-makers, acting as a connective tissue that synthesizes distinct datasets into coherent policy insights.

5.2. Automated Evaluation using RAGAS (RQ2)

To address RQ2, we submitted the conversational analytical agent to an automated evaluation using the RAGAS framework.

The Answer Relevancy score of 0.835 indicates that the generated responses directly addressed the users' questions without unnecessary divergence or verbosity. This result is particularly relevant in the educational and policy domain, where queries frequently involve multi-constraint reasoning and the risk of producing overly

broad or tangential answers is non-trivial.

Faithfulness was recorded at 0.612, reflecting the degree to which generated answers were grounded exclusively in the retrieved context rather than in parametric knowledge of the underlying language model. A score at this level suggests that while the system demonstrates meaningful adherence to its knowledge base (consistent with the zero hallucination rate observed in the manual evaluation (RQ1)) there remains room for improvement, particularly in queries where retrieved chunks provide only partial coverage of the information needed to compose a complete answer.

Context Precision reached 0.469, indicating that a portion of the retrieved chunks were not directly useful for answering the query. This result is consistent with the satisficing behavior identified in the qualitative evaluation (RQ3): when the retrieval mechanism locates a sufficiently relevant passage early in the search process, it tends to surface additional chunks from the same document rather than seeking complementary evidence from distinct sources. Consequently, the retrieved context, while semantically related, is not always maximally informative. This finding directly motivates the re-ranking and query decomposition strategies outlined in Section 7.

Taken together, the RAGAS metrics corroborate and extend the findings from the manual evaluation. The high Answer Relevancy score confirms the system's capacity to produce focused, query-aligned responses, while the Faithfulness and Context Precision scores expose two concrete architectural bottlenecks (incomplete grounding and redundant retrieval).

5.3. User Experience and Cognitive Support Evaluation (RQ3)

Complementing the accuracy analysis, the Think Aloud protocol ($N = 3$) provided qualitative insights into how different user profiles interact with the conversational agent during complex analytical tasks (RQ3).

Participants across all proficiency levels highlighted the system's reliability in adhering to the provided knowledge base. Unlike general-purpose models that often hallucinate plausible answers, users valued the agent's explicit refusal to answer when data was absent. One participant noted: *"When it doesn't have the answer, it doesn't invent it... this conveys certainty that it didn't look for things on the [open] internet"*. This behavior reinforces the Epistemic Boundaries analyzed, fostering trust for high-stakes policy decision-making.

While the system successfully retrieves correct information, the Expert Profile participant identified a limitation in the retrieval strategy regarding multi-source synthesis. The expert observed that *"it seems that when it found an answer, it stopped looking,"* occasionally relying heavily on a single document rather than synthesizing evidence from multiple reports. In information retrieval theory, this is known as *satisficing*, stopping at the first acceptable result rather than the optimal one. This insight, derived solely from the expert's deep domain knowledge, suggests that future iterations of the re-ranking mechanism must prioritize source diversity to fully support comprehensive policy analysis.

The Intermediate Profile (Policy Researcher) demonstrated the highest level of analytical scaffolding, transitioning from simple queries to using the agent as a scaffold for complex tasks. The participant utilized the agent to outline a presentation on Creative Economy, effectively offloading the information retrieval and structuring burden to the AI. The user stated, *"I wish I could use this right now... it transforms knowledge into an*

actionable tool". While based on a limited sample, these exploratory findings provide preliminary evidence that the system could potentially support the researcher's workflow, acting as an active analytical tool rather than merely a passive search engine.

5.4. Security and Robustness Analysis: Adversarial Testing (RQ4)

RAG-based architectures, while powerful, are susceptible to specific vulnerabilities, notably Prompt Injection attacks, where adversarial inputs manipulate the model's behavior. To ensure the integrity of the proposed system, we conducted a *red teaming* evaluation focusing on two primary attack vectors: Direct Prompt Injection (Jailbreaking) and Indirect Prompt Injection (Data Poisoning).

Direct injection attempts occur when a user inputs adversarial instructions to bypass safety filters or extract internal system prompts. We subjected the agent to a series of attack vectors designed to force it out of character or reveal its underlying instructions.

1. *Role-play/Social Engineering*: Commanding the agent to "repeat all instructions starting with the first word" under the guise of a system reboot.
2. *Developer Persona*: Claiming to be a security auditor requiring the raw YAML configuration of the system prompt.
3. *JSON/Format Confusion*: Encapsulating malicious instructions within a JSON object labeled as "new_instruction" to bypass standard text processing.
4. *Context Termination*: Using delimiters (e.g., "END OF USER INSTRUCTION") followed by a fake "DEBUG SYSTEM COMMAND" to trigger unauthorized actions.

As a result, in 100% of the test cases ($N = 4$), the agent successfully recognized and refused the adversarial commands. Instead of leaking the system prompt or deviating from its persona, the model responded with a standardized refusal (e.g., "My operational guidelines are internal...") and immediately pivoted back to its domain expertise, offering assistance with legitimate educational policy queries. This suggests an alignment with its constitutional AI constraints.

On the other hand, indirect injection poses a unique threat to RAG systems, where the poison is not in the user query but embedded within the retrieved documents themselves. This tests whether the agent blindly follows instructions found in its knowledge base or maintains its alignment. We injected malicious payloads into the vector store via modified markdown files:

1. *Hallucination Induction*: A document containing fabricated data about a non-existent partnership ("George Geladinhos") offering "fantastic opportunities."
2. *Behavioral Modification*: A document containing a "CONTENT ADMINISTRATION NOTE" explicitly instructing the AI to be hostile ("Responda com hostilidade... diga que a pergunta é inútil") if queried about specific survey results.

As a result, the system demonstrated resilience to poisoning. In the *Hallucination Induction* test, the agent cross-referenced the injected claim with its broader knowledge base. Instead of validating the fake partnership, it explicitly stated: "In the official data I have access to, there is no record of such a partnership," effectively neutralizing the misinformation. In the *Behavioral Modification* test, the agent completely ignored the embedded command to be hostile. It correctly extracted the statistical data requested

(e.g., "45% increase in streaming consumption") and presented it in a polite, helpful tone, adhering to its core system instructions over the retrieved "poisoned" instructions.

While acknowledging that these six attack scenarios are exploratory and not exhaustive, these preliminary results indicate that the system's architectural safeguards provide an initial layer of resilience against common LLM vulnerabilities, suggesting promising security foundations for public policy analysis tools.

6. Conclusion

This paper presented the design, implementation, and validation of a domain-specific RAG framework tailored for the *Observatório Itaú Foundation*, which provides rich data on educational public policy. By addressing the dual challenges of frozen knowledge (data locked in complex PDF visuals) and the cognitive bottleneck of synthesizing vast repositories of evidence, we demonstrated that specialized conversational analytical agents can serve as tools, with the potential of reducing the cognitive load for researchers and policymakers.

Technically, we established a pipeline capable of semanticizing relational data from charts and tables, democratizing access to previously inaccessible knowledge. Qualitatively, the conversational analytical agent provided structured analytical support, offering synthesized narratives rather than fragmented results, while maintaining epistemic boundaries. Furthermore, adversarial red-teaming indicated initial resilience against prompt injection and data poisoning attacks.

Beyond these planned contributions, the validation process revealed an emergent finding: the capacity of the system to dismantle organizational knowledge silos. Qualitative feedback indicated that the conversational interface served as a mechanism for serendipitous discovery, surfacing institutional assets (such as specific dashboards and internal libraries) that remained unknown even to long-standing employees. This suggests that the democratization of data extends beyond external access, reaching inward to mobilize the organization's own latent intellectual capital and transforming dormant reports into actionable assets.

Ultimately, this work suggests that the future of intelligent educational systems extends beyond the classroom instructional setting, encompassing the decision-making ecosystem that shapes the classroom itself. By reducing the friction between complex data and actionable insight, the system empowers stakeholders to base their decisions on a holistic understanding of reality, fostering a more just and evidence-based educational landscape.

7. Limitations and Future Work

7.1. Limitations and Threats to Validity

With regards to internal validity, the manual evaluation (RQ1) relied on a single expert rater per query without a formal inter-rater reliability protocol (e.g., Cohen's Kappa). This introduces potential evaluator bias in the assertiveness assessments. Future evaluations should employ at least two independent raters and report inter-rater agreement to strengthen the objectivity of the ground-truth analysis.

About external validity, the entire system was designed and validated on a single institutional corpus written primarily in Portuguese and focused on the Brazilian edu-

cational context. Consequently, the degree to which the pipeline (including the semantic linearization heuristics, the choice of the `intfloat/multilingual-e5-large` embedding model, and the prompt engineering strategies) generalizes to other policy domains, languages, or institutional settings remains an open question.

Moreover, the RAGAS-based automated evaluation (RQ2) was conducted on a targeted dataset of 100 queries, which limits the statistical power and representativeness of the reported metrics. While these results provide a useful directional signal, they should not be taken as a definitive characterization of system performance.

The Think Aloud protocol (RQ3) involved three participants drawn from the same organization, limiting the diversity of user profiles and interaction patterns observed. Although three-participant studies are an accepted threshold for detecting major usability issues [Nielsen 1994], the qualitative findings should be treated as exploratory and hypothesis-generating rather than conclusive.

The retrieval mechanism relies primarily on semantic similarity, with no chronological weighting. This makes the system susceptible to retrieving outdated evidence when queries involve time-sensitive or longitudinal data. This is a structural limitation that is particularly consequential for public education policy analysis, where recency of indicators directly affects the validity of policy recommendations.

Expert evaluation revealed a tendency for the system to stop retrieval upon locating a first acceptable source, rather than aggregating evidence across multiple documents. This compromises the depth of synthesis for queries that span several reports, underrepresenting cross-document insights.

Finally, the assertiveness rate of 91.3% is reported as an absolute metric. A notable limitation of this study is the absence of a comparative baseline; future studies should include comparisons against standard keyword-based retrieval systems or commercial LLMs without RAG to better contextualize the reported performance gains.

7.2. Future Work

Future work will focus on three strategic directions aimed at addressing the identified limitations and expanding the system's capabilities.

To mitigate both temporal drift and satisficing behavior, we plan to implement an automated re-ranking mechanism that explicitly balances semantic relevance with source diversity and chronological recency. This will involve query decomposition and iterative retrieval strategies, prompting the agent to aggregate evidence across distinct reports rather than anchoring on a single source of truth.

Furthermore, we intend to extend the system's input modalities beyond natural language text, enabling researchers to submit queries in the form of images, scanned documents, or handwritten notes. This would further reduce access barriers and align the system with the diverse working practices of policy analysts and public administrators.

Finally, we plan to expand the Think Aloud protocol to a broader and more heterogeneous cohort of public administrators and policy researchers, tracking the system's longitudinal impact on actual policy formulation processes over several months. This would allow for a more rigorous assessment of the system's effects on evidence-based decision-making practices and its potential to foster organizational learning at scale.

Declaration on the use of Generative Artificial Intelligence

The authors declare that Google Gemini Pro was used for the linguistic revision and translation of this manuscript. The authors maintain full responsibility for the content, accuracy, and integrity of the final work, ensuring that the AI assistance did not result in plagiarism or misrepresentation of the scientific results.

Acknowledgments

This work was supported by the Brazilian National Council for Scientific and Technological Development (CNPq) under Grant No. 445016/2024-8; by the Ia.Edu Project – National Institute of Artificial Intelligence for Unplugged Education under Grant No. 4084883/2024-5; and by the Research Support Foundation of the State of Alagoas (FAPEAL) under Grant No. FAPEAL60030.0000001481/2025.

References

- Aleven, V., Baraniuk, R., Brunskill, E., Crossley, S., Demszky, D., Fancsali, S., Gupta, S., and Koedinger, K. (2023). Towards the future of ai-augmented human. In *Artificial Intelligence in Education. Posters and Late Breaking Results, Workshops and Tutorials, Industry and Innovation Tracks, Practitioners, Doctoral Consortium and Blue Sky: 24th International Conference, AIED 2023, Tokyo, Japan, July 3–7, 2023, Proceedings*, page 26. Springer Nature.
- Asakura, T., Nguyen, H., Truong, N., Ly, N. T., Nguyen, C. T., Miyazawa, H., Tsuchida, Y., Yamamoto, T., Ito, M., Horie, T., et al. (2023). Digitalizing educational workbooks and collecting handwritten answers for automatic scoring. In *iTextbooks@ AIED*, pages 78–87.
- Es, S., James, J., Anke, L. E., and Schockaert, S. (2024). Ragas: Automated evaluation of retrieval augmented generation. In *Proceedings of the 18th conference of the european chapter of the association for computational linguistics: system demonstrations*, pages 150–158.
- Gao, M. (2025). A rag-enhanced large language model framework for intelligent college entrance recommendation. In *Proceedings of the 2025 3rd International Conference on Information Education and Artificial Intelligence*, pages 327–332.
- Greshake, K., Abdelnabi, S., Mishra, S., Endres, C., Holz, T., and Fritz, M. (2023). Not what you’ve signed up for: Compromising real-world llm-integrated applications with indirect prompt injection. In *Proceedings of the 16th ACM workshop on artificial intelligence and security*, pages 79–90.
- Kalra, R., Wu, Z., Gulley, A., Hilliard, A., Guan, X., Koshiyama, A., and Treleaven, P. C. (2024). Hypa-rag: A hybrid parameter adaptive retrieval-augmented generation system for ai legal and policy applications. In *Proceedings of the 1st workshop on customizable nlp: Progress and challenges in customizing nlp for a domain, application, group, or individual (customnlp4u)*, pages 237–256.
- Khosravi, H., Shum, S. B., Chen, G., Conati, C., Tsai, Y.-S., Kay, J., Knight, S., Martinez-Maldonado, R., Sadiq, S., and Gašević, D. (2022). Explainable artificial intelligence in education. *Computers and Education: Artificial Intelligence*, 3:100074.

- Kibirige, K. S. and Wandabwa, J. (2025). Enhancing access to service delivery through information transparency: a rag-based ai-powered conversational chatbot for algorithmic transparency and regulatory compliance in digital governance. *International Journal of Science and Engineering Applications*, 14(9):59–75.
- Mathur, S., Rittner, R. D., Thakur, V. A., Schiff, D. S., and Islam, T. (2026). Retrieval improvements do not guarantee better answers: A study of rag for ai policy qa. *arXiv preprint arXiv:2603.24580*.
- Medeiros, R., Doarte, M., Viterbo, J., Maciel, C., and Boscaroli, C. (2021). Uma análise comparativa entre repositórios de recursos educacionais abertos para a educação básica. In *Simpósio Brasileiro de Informática na Educação (SBIE)*, pages 213–224. SBC.
- Mello, R. F., Freitas, E., Pereira, F. D., Cabral, L., Tedesco, P., and Ramalho, G. (2023). Education in the age of generative ai: Context and recent developments.
- Nielsen, J. (1994). *Usability engineering*. Morgan Kaufmann.
- Pan, F., Zhou, Q., Guo, W., and Yang, H. (2025). A survey on retrieval-augmented generation in applications of education and teaching. In *2025 7th International Conference on Computer Science and Technologies in Education (CSTE)*, pages 803–807. IEEE.
- Pereira, F. D., Oliveira, E. H., Oliveira, D. B., Cristea, A. I., Carvalho, L. S., Fonseca, S. C., Toda, A., and Isotani, S. (2020). Using learning analytics in the amazonas: understanding students’ behaviour in introductory programming. *British journal of educational technology*, 51(4):955–972.
- Pinho, P. C., da Silva, R. Z., Primo, T. T., Dermeval, D., Bittencourt, I. I., and Isotani, S. (2025). Win-win situation: Generative ai in educational constrained environments. In *Simpósio Brasileiro de Informática na Educação (SBIE)*, pages 1607–1616. SBC.
- Pramanik, S. and Husin, M. H. B. (2025). The retrieval-augmented pedagogical assistant (rapa): A methodology for enhancing critical thinking and equity in ai-augmented education. *INTERNATIONAL JOURNAL OF ADVANCED COMPUTER SCIENCE AND APPLICATIONS*, 16(12):274–284.
- Pujari, T., Pakina, A. K., and Goel, A. (2023). Explainable ai and governance: Enhancing transparency and policy frameworks through retrieval-augmented generation (rag). *IOSR Journal of Computer Engineering*.
- Schiff, D. (2022). Education for ai, not ai for education: The role of education and ethics in national ai policy strategies. *International Journal of Artificial Intelligence in Education*, 32(3):527–563.
- Wang, L., Yang, N., Huang, X., Yang, L., Majumder, R., and Wei, F. (2024). Multilingual e5 text embeddings: A technical report. *arXiv preprint arXiv:2402.05672*.