

# Design, Technical Audit, and Beta Demonstration of an AI-Assisted Decision-Support System for Supplier-Side Public Procurement Screening

Breiner Navarro-Vergara, José Isaías Salas-Hernández\*

Fundación Universitaria Cafam, Bogotá, Colombia

## Abstract

Public procurement portals expose large volumes of opportunities, but potential suppliers still spend substantial effort deciding which notices merit deeper bid preparation. AI applications for buyer-side procurement, monitoring, supplier assessment, document analysis, and, recently, supplier-side opportunity detection have been researched. This paper reports on the development, code audit, and beta testing of a web-based decision support system designed to screen suppliers within Colombia's SECOP II environment. The functions examined during the audit included live retrieval from SECOP II, deterministic selection based on UNSPSC codes and lexical business rules, the construction of institution-specific profiles from procurement registration data, support for several selectable large language models, JSON output validated against a predefined schema, rule-based priority ranking, and, finally, a human-led Bid/No-Bid decision. Code audit determines that this prototype does not use a persistent database, authentication function, retrieval-augmented generation, embedding-based search, formal explainability algorithm, or autonomous agent loop. Functional demonstration (Gemini 2.5 Flash) of end-to-end batch process time for ten SECOP II processes take approximately 40-50 s; each of the ten model responses conforms to the required JSON format; six processes are labeled VIABLE, three processes labeled REQUIRES ADJUSTMENT, and one process labeled NOT VIABLE with a mean applicability score of 74%. These figures are treated as functional evidence rather than decision accuracy. A structured 36-study evidence matrix is supplied as supplementary material. The contribution is a reproducible characterization of a rule-constrained, human-validated supplier-side screening workflow and its current technical boundaries.

## Keywords

Public procurement, Supplier-side decision support, Large language models, Pre-bid screening, Human-in-the-loop, SECOP II, UNSPSC, Technical audit

## 1. Introduction

Public procurement is an attractive market for firms, but participation is costly before a bid is ever written. A complex array of different notices, eligibility criteria, technical requirements, deadlines, contract values, and types of procurements must be interpreted by the suppliers to decide if the opportunity is worth investigating further or not. Supplier-side research reveals that ambiguity, excessive information, feelings of unfairness and unresponsiveness result in a higher level of dissatisfaction, no response and lower bidding effort [1]. There is also large-scale European evidence indicating that the quality of the procurement and regulatory framework influences SME participation. These findings are, however, contextual and should not be turned into universal Bid/No-Bid rules [2].

Artificial intelligence is increasingly used across procurement workflows. Prior work covers institutional process support, drafting and preparation [3, 4], supplier evaluation [5], standardized procurement data [6], semantic analysis of procurement objects [7], corruption-risk indicators in unstructured tender documents [8], and human-AI requirements in public-sector procurement [9]. Taken together, these studies show that procurement text and metadata can support technically rigorous AI-assisted analysis. Most of this literature, however, places the public buyer or auditor at the center of the workflow.

---

ICAIW 2026: Workshops at the 9th International Conference on Applied Informatics 2026, October 13–17, 2026, Orlando, United States

\*Corresponding author.

✉ jose.salas@unicafam.edu.co (J. I. Salas-Hernández)

ORCID 0009-0005-7743-0628 (B. Navarro-Vergara); 0000-0002-6601-2720 (J. I. Salas-Hernández)



© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

The closest technical antecedent to the present study is the Artificial Intelligence Procurement Assistant (AIPA), which combines LLM-based agents with tender and funding portals, including TED, to derive search parameters from a user profile, filter opportunities, and support preliminary review before a human decision [10]. Taken together, these precedents make it difficult to sustain a broad claim of novelty grounded solely in the use of an LLM to search public tenders from the supplier’s side. A more circumscribed set of questions remains: how can supplier-specific institutional information be made operational within an actual screening workflow; how are deterministic rules and generative analysis combined; and what technical constraints become visible when a system is audited rather than described only in conceptual terms? To address these questions, this paper examines a beta web application connected to Colombia’s SECOP II public procurement data.

The study does not present the prototype as an autonomous agent. Source-code inspection shows a bounded decision-support pipeline initiated by a human user: live opportunity retrieval, rule-based filtering, institution-specific context injection, generative analysis, structured validation and prioritization, followed by human review and the final Bid/No-Bid decision. The paper addresses three research questions:

1. RQ1. What architecture and division of responsibilities among deterministic rules, generative AI, and human judgment are implemented in the beta system?
2. RQ2. How does the system operationalize supplier-specific institutional information for pre-bid screening of public procurement opportunities?
3. RQ3. What functional behavior, technical constraints, and reproducibility limitations are revealed by the source-code audit and beta demonstration?

The paper contributes a code-grounded system characterization, a comparison with the closest supplier-side antecedent, and a functional demonstration whose claims are deliberately limited to what the current evidence supports. A broader expert-agreement study, manual-time comparison, and UN-SPSC ablation are reserved for subsequent evaluation rather than inferred from a small demonstration.

## 2. Related Work

### 2.1. AI in public procurement and the supplier-side problem

AI applications in public procurement are heterogeneous. Institutional studies report opportunities to reduce administrative work and accelerate selected procurement tasks [3, 4]. Risk-oriented work combines structured outcomes with unstructured text and evaluates contextual transformer models on procurement documents [8]. Other studies employ deep learning techniques to find similarities in descriptions of contracted objects [7] or focus on data architectures for interoperable structures in procurement analytics [6]. Research on human-AI requirements and explainability in public sector procurement has been discussed [9]. While these papers relate to the technical aspects of the current system, they do not provide a general framework for supplier-side opportunity screening. From the suppliers’ perspective, the task is not simply about classifying documents. The decision precedes proposal preparation and involves both retrieval and organizational fit. Kelly et al. [1] show how ambiguity, redundant information, and procedural friction affect supplier experience. Hoekman and Taş [2] show that SME participation is related to procurement policy and contract design. They give insights into the reasons for the importance of screening costs, but do not implement an AI assistant for the supplier.

### 2.2. Supplier-side opportunity screening and the AIPA benchmark

AIPA [10] is therefore the appropriate benchmark. It already demonstrates that an LLM-based assistant can use a supplier/user profile and public tender sources to support opportunity discovery and preliminary filtering. The present beta should not be positioned as the first supplier-side AI procurement

assistant. The relevant distinction lies in the specific representation of supplier capability, the rule-constrained screening logic tied to SECOP II and UNSPSC, and the explicit audit of the implemented boundary between deterministic logic, generative output, and human action.

**Table 1**

Feature-level comparison with the closest supplier-side antecedent.

| Dimension               | AIPA [10]                                     | Current beta (audited)                                                 |
|-------------------------|-----------------------------------------------|------------------------------------------------------------------------|
| Public source           | TED and European tender/funding portals       | SECOP II via Socrata (p6dx-8zbt)                                       |
| Supplier representation | User profile used to derive search parameters | RUP-derived financial, experience, UNSPSC and exclusion profile        |
| Candidate screening     | LLM-assisted opportunity filtering            | UNSPSC + lexical deterministic gate, then LLM analysis                 |
| LLM output              | Preliminary opportunity assessment            | Strict 18-field JSON with viability, applicability, risks, and actions |
| Human role              | Preliminary review before human decision      | Final Bid/No-Bid decision remains human                                |
| Technical boundary      | LLM-based assistant                           | No RAG, embeddings, vector search, XAI or autonomous loop              |

Retrieval and structured-output research provide useful comparators without implying that those techniques are implemented here. RAG pipelines combine lexical and dense retrieval, reranking, and generation with measurable trade-offs between latency and quality [11]. Template-guided LLM pipelines can impose structured outputs without guaranteeing semantic correctness [12], and taxonomy-driven retrieval can reduce false positives in specialized domains [13]. The current beta uses none of these retrieval-augmentation techniques; its institutional context is inserted directly into a fixed prompt, and candidate retrieval is rule-based.

### 2.3. Human–AI complementarity and oversight

Human control is not a cosmetic addition. McKinley et al. [5] find that humans are more aligned with GAI on compliance signals than on competitive differentiation in supplier evaluation. They also show scoring volatility over repeated LLM evaluations. Cognitive forcing function research shows that explanations alone do not deter overreliance [14]. A recent HITL review highlights the importance of loop placement, granularity of interactions, and timing [15]. Work on user-centered XAI has long insisted that explanations be tailored to the task at hand and to what users actually need [16]. Research on complementarity points in a similar direction, suggesting that deferring to human judgment is advisable when model confidence is low or when the structure of the task undermines reliable automation [17].

These findings support a conservative division of labor in the present system. The machine retrieves, filters, structures, and proposes a viability assessment; the analyst interprets ambiguous evidence and decides whether to bid. Colombian legal scholarship on the reservation of humanity in public administrative discretion [18] is relevant background for accountable human judgment, but it does not directly regulate a private supplier’s internal Bid/No-Bid decision. Here, human control is justified as a socio-technical design choice for uncertainty management, traceability, and accountability, rather than as a legal requirement imposed on the supplier.

## 3. Research Method

### 3.1. Evidence base and supplementary matrix

This work is a combination of a structured evidence review, a source code audit, and a functional beta demo. We use the literature review as a way to situate our artifact, to identify other works that made similar decisions, and to set boundaries on what constitutes legitimate novelty. The paper includes a literature review, but the main contribution is the source code audit and functional beta demo. In the main paper, we cite only those sources which directly support a claim, design decision, or interpretation.

Supplementary Material S1: Supplementary Material S1 – Literature Analysis Matrix (36 studies). The review package keeps the spreadsheet in the same folder as this manuscript; a persistent repository URL/DOI should replace the local link in the camera-ready version.

### 3.2. Source-code audit

A technical audit was performed according to the following steps:

1. Project structure and dependencies inspection;
2. SECOP II data path reconstruction;
3. Deterministic filtering and UNSPSC logic inspection;
4. Institutional profile representation inspection;
5. LLM provider/model router, prompt and output schema review;
6. Validation, prioritization, interface actions and cache behavior tracing;
7. Identification of missing components.

Artifacts' characterization and empirical evaluation were separated based on FEDS [19]. The current paper reports the audit and a bounded functional demonstration rather than an expert-ground-truth performance study.

### 3.3. Functional beta demonstration

The beta demonstration used the deployed Streamlit interface and a batch of ten SECOP II processes evaluated with Gemini 2.5 Flash. The demonstration was used to verify end-to-end execution, JSON validity, dashboard generation, and process-level presentation. It was not designed to establish classification accuracy, causal time savings, or superiority of one UNSPSC granularity. Those claims require an independent reference set, repeated runs, and controlled comparison conditions that are outside the scope of the current submission.

The evaluation was structured in two complementary levels. The first corresponded to a functional demonstration on ten real SECOP II processes, used to check the end-to-end execution of the system, the application of deterministic rules, the generation of structured outputs, and the presentation of results for human review. The second level incorporated an external validation of the same processes. Two professionals responsible for the regular search and review of opportunities in SECOP independently evaluated each case, without previously consulting the classification generated by the system, and assigned a feasibility category –VIAABLE, REQUIRES ADJUSTMENTS or NOT VIAABLE– accompanied by a justification. In addition, an academic external to the development reviewed the relevance of the criteria used for screening and the separation between AI-assisted recommendation and final human decision. This strategy allows the output of the system to be contrasted with independent expert judgment, but its results should be interpreted as evidence of agreement in a beta demonstration of limited scope and not as a generalizable estimate of accuracy for all public procurement.

## 4. Audited System Architecture

### 4.1. Implementation stack and persistence

Application-Python uses Streamlit for the user interface. Principal modules of the application: Application (contains main logic), AI-backend (uses libraries: pandas to operate on tables, sodapy to interact with Socrata, Groq and Google GenAI SDK to call LLMs, python-dotenv to manage credentials, and openpyxl to export to Excel). No persistent application database. Results from analysis are cached in memory for the active session and lost after restart.

This is the layer of the system responsible for connecting to the source of the data, in this case the Colombian Open Data Portal using Socrata. It is implemented using a combination of queries to the SECOP II dataset. Instead of scraping the webpage or doing manual downloads, it gets live data. It

uses a 90-day historical window and an over-retrieval strategy to allow for later filtering of irrelevant records while not limiting the number of records too early. In the event of API errors, bounded retries and backoff strategies are used.

## 4.2. SECOP II acquisition and deterministic candidate retrieval

Before calling any LLM, candidate objects are filtered out. Each call to the relevance assessment function includes three components:

1. A two-digit UNSPSC segment prefix
2. An exact eight-digit UNSPSC commodity match with accredited codes for the institution,
3. Lexical inclusion/exclusion criteria for the object being procured.

We set up 12 broad segment prefixes, but we have 35 exact commodity codes as accredited experience. There is a +15-point bonus for exact code matches. All objects associated with construction, cleaning, physical security, manufacturing, and other expressly prohibited activities are removed or heavily restricted prior to generative analysis. This is deterministic logic and therefore not semantic matching via embeddings or vector search.

## 4.3. Institution-specific supplier representation

In this case, the profile for the supplier is based on an anonymized representation of a Colombian Registro Único de Proponentes (RUP) embedded in the code. The profile consists of a set of financial information, experience, accredited UNSPSC codes, and domains that differentiate high competence, medium competence, and out-of-profile areas. The financial context consists of a liquidity ratio of 1.36 and an indebtedness ratio of 0.59. The experience block shows over 50 complete contracts and 17 years of organizational path. The audited master profile uses these fields as structured text embedded in the prompt, instead of obtaining them from a database.

## 4.4. LLM routing, prompting and structured output

The generative layer is multi-provider and selectable at runtime. The audited catalogue includes Groq-hosted LLaMA 3.3 70B Versatile as the default option, LLaMA 4 Scout 17B, DeepSeek R1 Distill LLaMA 70B, and Google Gemini 2.5 Flash. Temperature, top-p, and maximum token parameters are not fixed in the code and so use provider defaults. While this provides more freedom for experimentation in the beta, it reduces strict reproducibility between multiple runs. Provider-specific delays and retry logic are implemented to remain within the operational limits of the deployed services.

The main prompt defines the model’s role as a senior Colombian public procurement analyst. The prompt is made up of the institutional profile, selected fields from SECOP II, competitive signals (e.g., invited suppliers, received responses), and explicit business rules that are not left to be interpreted by the model (e.g., objects out of profile force NOT VIABLE with low applicability; exact UNSPSC match adds predefined bonus). The business logic present in the deterministic pre-filter is also present inside the prompt, making it a double gate around the LLM, which prevents an unconstrained generative decision.

The required output is a strict JSON object with 18 fields, including process identifier, contracting entity, object, UNSPSC code and category, viability class, applicability percentage, financial score, competition level, strengths, risks, improvement actions, recommendation, estimated value, duration, procurement modality, closing date, SECOP link, and exact RUP-code match. A validation function checks required fields and rejects incomplete responses. A second prompt can generate a short executive summary for the analyzed batch.

The audited version of the prototype does not explicitly state the temperature, top-p, and max\_tokens parameters, so the calls to the model use the vendor’s default settings. Consequently, the reproducibility of the study must be distinguished at two levels. Procedural reproducibility is documented by an

identified data source, explicit deterministic rules, a fixed institutional profile, a documented prompt, a defined JSON schema, and a verifiable execution pipeline. However, the exact reproducibility of generative output cannot be assumed, since external LLM services can introduce variability between executions even when the analysed process, context, and prompt are kept constant. For this reason, the results of the functional run are reported as beta version observations under the available configuration and not as deterministic outputs. Explicit fixing of generation parameters or systematic stability evaluation between executions is recognized as a requirement for future performance benchmarking.

#### 4.5. Human decision and verified system boundary

The dashboard displays viability filters, minimum applicability filters, UNSPSC categories, model selection, processing delay, aggregate KPIs, urgent processes, executive summaries and process-level cards. It exposes processes with viability filters sorted by their combined applicability and processing time to close scores (after validation). Users can examine strengths, risks, and recommended actions, access the relevant SECOP II records, and export results. The LLM generates a provisional classification—VIABLE, REQUIRES ADJUSTMENT, or NOT VIABLE—but the analyst retains the final Bid/No-Bid decision. The audit further defines the system’s boundaries. The beta does not incorporate a persistent database, role-based authentication, retrieval-augmented generation, an embedding store, vector similarity search, a formal XAI method, an autonomous planning loop, or persistent agent memory. Credentials are kept in environment variables rather than in source code. All these absences are not treated as defects hidden from the reader; they delimit the current contribution and help distinguish the system from more elaborate RAG or Agentic architectures.

### 5. Results of the Technical Audit and Beta Demonstration

#### 5.1. Verified implementation

The audit confirms six operational stages: real-time recovery of SECOP II; deterministic filtering of candidates; construction of the specific context of the institution; selectable LLM inference; validation and classification of structured output; and the final human review. Figure 1 summarizes the verified data flow and explicitly separates the implemented components from the missing capabilities. This distinction is important because labels like "agent," "semantic matching," or "RAG" would exaggerate what the code implements.

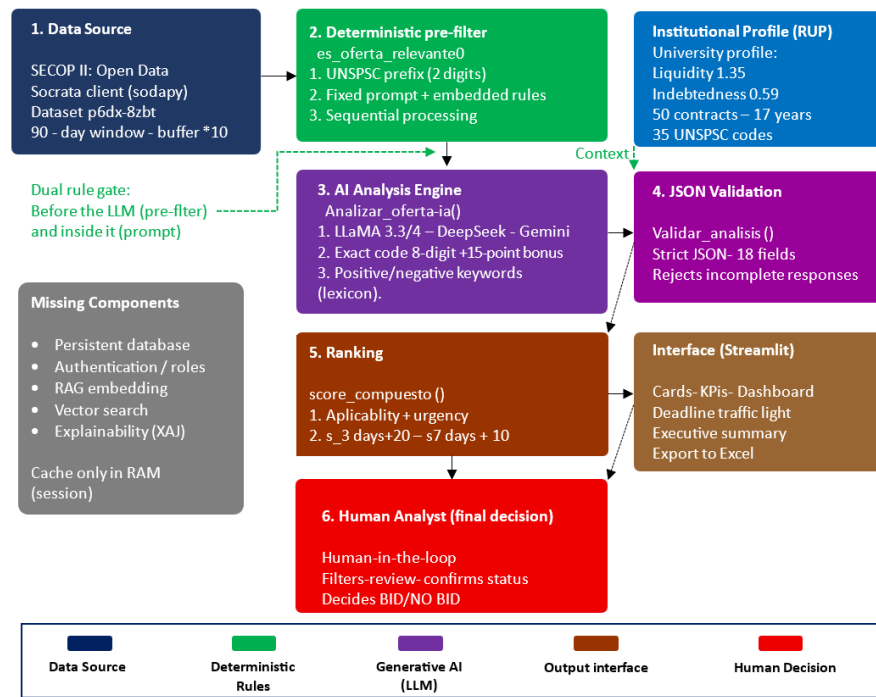
#### 5.2. End-to-end functional demonstration

The functional beta run with Gemini 2.5 Flash analyzes ten SECOP II processes. The interface estimates about 40–50 seconds for the batch, and all ten model responses pass the required structured output validation. The dashboard displays six processes as VIABLE, three as REQUIRES ADJUSTMENT, and one as NOT VIABLE, with an average value of 74% and one process marked as urgent. Figure 2 shows the aggregate dashboard and process-level review interface used by analysts.

These numbers represent end-to-end execution and interface behavior; It does not determine the quality of the decision. There was no independent evaluation expert reference set for this presentation, so the distribution of the feasibility classes cannot be interpreted as accuracy, accuracy, remembering, F1, or human-AI agreement. Similarly, the batch time of 40–50 s is a functional observation of display time, not a controlled estimate of the productivity increases relative to manual screening.

#### 5.3. Verified technical constraints

The code audit also exposes reproducibility and deployment constraints. The organizational profile is hard-coded, so adding another supplier would require modifications at the code level. Since caching is session-only, historical audit trails cannot be retained. The lack of authentication also restricts deployment beyond demonstration purposes. Because sampling parameters are not fixed, model defaults



**Figure 1:** Audited end-to-end architecture. The diagram separates deterministic rules, generative analysis, structured validation, interface functions, and the final human decision, while also listing verified absent components.

introduce stochasticity. Finally, the application depends on external LLM services and sequential calls, which limits batch scale and makes latency sensitive to provider availability.

**Table 2**

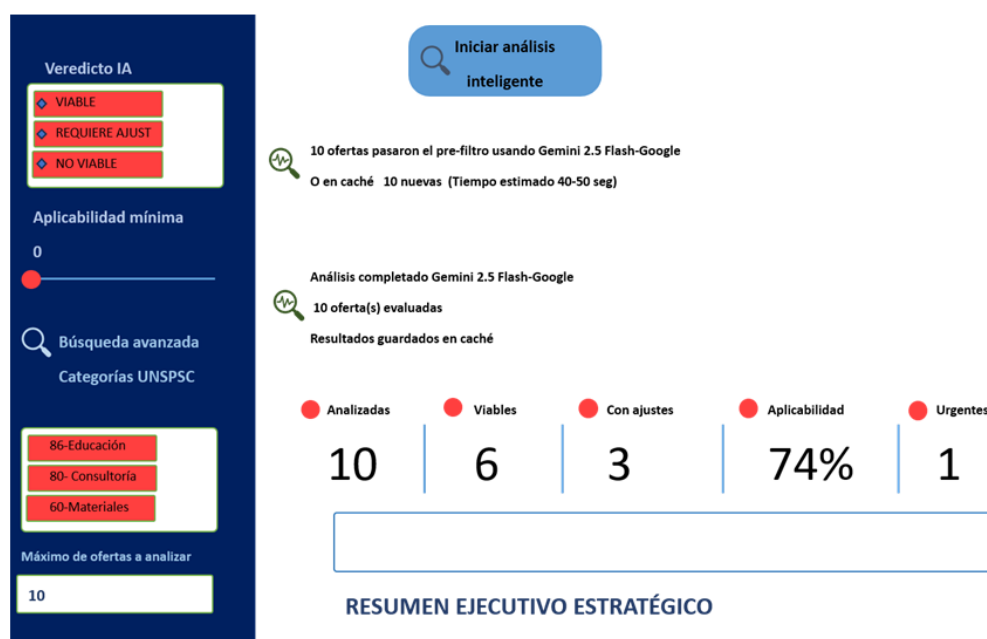
Technical boundary established by the source-code audit.

| Confirmed in the beta             | Not implemented                 |
|-----------------------------------|---------------------------------|
| SECOP II / Socrata live retrieval | persistent database             |
| UNSPSC deterministic filtering    | Authentication / roles          |
| RUP-derived context               | RAG                             |
| Selectable LLMs + fixed prompt    | Embeddings / vector search      |
| Schema-validated JSON             | Formal XAI                      |
| Human final decision              | Autonomous planning / memory    |
| Session cache / Excel export      | Production monitoring / backups |

## 6. Discussion

### 6.1. Contribution relative to AIPA

The contribution of beta should be interpreted in relation to AIPA rather than as an artificial basis for "no prior system by the provider". AIPA already combines LLM support, a vendor-oriented profile, and sources of public opportunity [10]. The current system adds a different combination of constraints: a RUP-derived capability profile, SECOP II/UNSPSC-specific recovery, deterministic exclusion and exact code rules, and a schema-validated output whose feasibility recommendation remains subject to human review. The practical value of the research lies in documenting how these components are connected and where the boundaries lie.



**Figure 2:** Beta interface during the ten-process demonstration. Aggregate dashboard showing batch status and viability distribution.

## 6.2. Rule-constrained generative analysis and human control

The architecture also illustrates a useful middle ground between unconstrained generative analysis and fully hard-coded screening. Deterministic rules remove clearly irrelevant cases and encode institutional exclusions before model invocation, which reduces unnecessary LLM calls and limits the decision space. The LLM then takes on the synthesis of heterogeneous fields and generates a structured assessment. As similar rules are repeated in the prompt, the model is subject to a further constraint at inference. This two-layer design enhances traceability, though it also gives rise to a maintenance risk: any institutional policy change needs to be synchronized across code and prompt.

The human contribution here is not merely formal; it is substantive. As earlier studies have suggested, LLM evaluations tend to align reasonably well with human judgments when it comes to compliance signals, yet they remain unstable when assessing competitive differentiation [5], and that overreliance is not solved merely by displaying an explanation [14]. The beta therefore presents strengths, risks, actions, and source links for inspection rather than converting the model output directly into an irreversible decision. This design is consistent with broader HITL and human-centered AI literature [15, 16, 17] and with trustworthy public-sector AI principles [20, 9].

## 6.3. UNSPSC granularity as an empirical design question

UNSPSC granularity remains an open design issue. The two-digit segment prefix increases breadth, while the eight-digit exact match is more specific. None of the 36 reviewed studies demonstrates that the two-digit prefix is superior to full-code matching. The current implementation should therefore be described as an engineering choice intended to preserve candidate recall, not as an evidence-based best practice. A future ablation can compare prefix-only, exact-code, lexical, and combined retrieval on a frozen SECOP II window.

## 6.4. Practical implications and next evaluation step

The same caution applies to practical benefits. The beta version is intended to keep attention away from clearly irrelevant opportunities, while the pre-filter is expected to reduce the number of paid or limited LLM calls. Even so, this article stops short of claiming a confirmed reduction in analyst time, improved tender success, or higher proposal quality. Evidence for such outcomes would require controlled measurement. Any subsequent assessment should therefore rely on a fixed corpus, blind expert evaluations, model replications, comparisons of manual and assisted timing, and a systematic analysis of disagreements. If these results do not provide significant benefits, the contribution should remain a repeatable technical characteristic, not a claim of efficacy.

## 7. Threats to Validity and Limitations

The study has some limitations: The audited profile is one institution, thus limiting generalizability and making the hard-coded capability rules organization-specific. Retrieval errors due to SECOP II data completeness and classification practices are out of the LLM’s control. The demonstration only covers ten processes and one observed model run. It does not allow for

statistical analysis of decision quality, latency, or stability. Model updates on the provider’s side and unfixed sampling parameters impede repeatability. The deployed Streamlet beta version lacks persistence, authentication, and production monitoring. Evidence from the literature was used to position and constrain claims. However, this paper is not a systematic review contribution, so the full 36-study extraction matrix is provided as an additional file rather than included in the paper body.

## 8. Conclusion

The current implementation of the supplier-side public procurement prototype does not incorporate such future elements; instead, this paper documents that fact. It retrieves live SECOP II records, it applies a hybrid UNSPSC and lexical pre-filter, it inserts an RUP-derived institutional profile into a fixed prompt, it uses a selectable LLM, it validates an 18-field JSON response, it ranks opportunities, and leaves the final Bid/No-Bid decision to the analyst. That same audit also confirms that the system lacks RAG, embeddings, vector search, formal XAI, persistent storage, and autonomous agent behavior.

A ten-process functional demonstration verified end-to-end execution and structured output, although it was not treated as a proxy for accuracy. The paper thus offers three contributions: a reproducible technical characterization, a transparent account of the system’s boundaries, and a supplier-side implementation positioned against the closest prior work. A subsequent evaluation can test expert agreement, repeated-run stability, manual time savings, and UNSPSC retrieval trade-offs on a fixed corpus.

## Data, Code, and Supplementary Materials Availability

The complete literature extraction is provided as [Supplementary Material S1 – Literature Analysis Matrix](#) (36 studies). The source code is not publicly released in this review version because of institutional intellectual-property and credential-security constraints.

The paper reports on the audited architecture, prompt logic, model catalogue, and implementation boundaries needed to interpret the results. For the camera-ready version, the local supplementary link should be replaced by a persistent repository URL/DOI if institutional conditions permit.

## Acknowledgments

This work was developed within the framework of the project “AI and Data Analytics-Driven Intelligent Management Model to Enhance the Competitiveness and Sustainability of MSMEs in Innovation, Strate-

gic Marketing, OHS, and Environmental Management: An Applied Study in the Bogotá-Cundinamarca Metropolitan Region” (Code FacingCBT2025-2-68), supported by the Grupo Interdisciplinario de Investigación para la Sostenibilidad y Competitividad Empresarial UNICAFAM. Principal Investigator: José Isaías Salas Hernández.

The authors acknowledge CODEX, the Data Science, Artificial Intelligence, and Digital Innovation Student Research Group, for its academic support and for providing a collaborative environment for the development and discussion of the prototype.

## Declaration on Generative AI

During the preparation of this manuscript, the authors used generative AI and AI-assisted tools, including ChatGPT (OpenAI), Claude (Anthropic), Gemini Pro (Google), Microsoft Copilot, and Llama-based models (Meta), solely to improve the readability, clarity, grammar, and linguistic precision of selected sections of the manuscript. These tools were not used for data analysis, interpretation of results, generation of scientific findings, or methodological decision-making. All AI-assisted content was critically reviewed, verified, and edited by the authors to ensure accuracy, consistency, and scientific integrity. The authors take full responsibility for the final content of the publication.

## References

- [1] S. Kelly, D. Marshall, H. Walker, J. Israilidis, Supplier satisfaction with public sector competitive tendering processes, *Journal of Public Procurement* 21 (2021) 183–205. doi:10.1108/JOPP-12-2020-0088.
- [2] B. Hoekman, B. K. O. Taş, Procurement policy and sme participation in public purchasing, *Small Business Economics* 58 (2022) 383–402. doi:10.1007/s11187-020-00414-z.
- [3] M. Iders-Bankovs, V. Politika, J. Pundure, et al., Public procurement in age of ai: Challenges and opportunities, in: *Engineering for Rural Development*, 2025. doi:10.22616/ERDev.2025.24.TF198.
- [4] T. Hyppänen, J. Vilko, Ai potential in public procurement tasks, in: *ICECER 2025*, IEEE, 2025. doi:10.1109/ICECER65523.2025.11401049.
- [5] F. A. McKinley, A. E. Dohmen, V. E. Castillo, Do humans and gai see eye to eye? implications of llm scoring volatility in supplier evaluations, *Journal of Business Logistics* 47 (2026) e70072. doi:10.1111/jbl.70072.
- [6] A. Al-Sulaiti, A standardised public procurement data architecture, *International Journal of Procurement Management* 26 (2026) 1–23. doi:10.1504/IJPM.2026.155566.
- [7] W. E. M. Lima, et al., A deep learning approach for identifying similarities in contracted object description in public procurement, *Procedia Computer Science* (2025). doi:10.1016/j.procs.2025.07.124.
- [8] N. Peppes, T. Alexakis, E. Daskalakis, E. Adamopoulou, Ai-driven corruption risk indicator detection: A comparative evaluation of transformer-based nlp models in unstructured procurement data, *Information* 17 (2026) 329. doi:10.3390/info17040329.
- [9] M. A. Abbasi, T. Mikkonen, S. Pirinen, A. Koski, Human-ai interaction requirements in public sector procurements, in: *Lecture Notes in Business Information Processing*, volume 584, 2026, pp. 343–352. doi:10.1007/978-3-032-26363-6\_26.
- [10] M. Waseem, T. Das, T. Paloniemi, M. Koivisto, E. Räsänen, M. Setälä, T. Mikkonen, Artificial intelligence procurement assistant: Enhancing bid evaluation, in: *Lecture Notes in Business Information Processing*, volume 500, 2024. doi:10.1007/978-3-031-53227-6\_8.
- [11] X. Wang, et al., Searching for best practices in retrieval-augmented generation, in: *Proceedings of EMNLP 2024*, 2024, pp. 17716–17736. doi:10.18653/v1/2024.emnlp-main.981.
- [12] M. A. Mridul, I. Sloyan, A. Gupta, O. Seneviratne, Ai4contracts: Llm & rag-powered encoding of

- financial derivative contracts, in: *IJCAI 2025*, 2025, pp. 9305–9312. doi:10.24963/ijcai.2025/1034.
- [13] H. Pan, Q. Zhang, M. Adamu, E. C. Dragut, L. J. Latecki, Taxonomy-driven knowledge graph construction, in: *Findings of ACL 2025*, 2025, pp. 4295–4320. doi:10.18653/v1/2025.findings-acl.223.
- [14] Z. Bućinca, M. B. Malaya, K. Z. Gajos, To trust or to think: Cognitive forcing functions can reduce overreliance on ai in ai-assisted decision-making, *Proceedings of the ACM on Human-Computer Interaction* 5 (2021) 188. doi:10.1145/3449287.
- [15] K. Lazaros, A. G. Vrahatis, S. Kotsiantis, Human-in-the-loop artificial intelligence: A systematic review of concepts, methods, and applications, *Entropy* 28 (2026) 377. doi:10.3390/e28040377.
- [16] S. Hong, W. Park, Developing user-centered system design guidelines for explainable ai: A systematic literature review, *Artificial Intelligence Review* 58 (2025) 386. doi:10.1007/s10462-025-11363-y.
- [17] A. Arnaiz-Rodriguez, A. Corvelo Benz, P. Thejaswi, N. Oliver, M. Gomez-Rodriguez, Towards human-ai complementarity in matching tasks, 2026, pp. 36–59. doi:10.1007/978-3-032-19099-4\_3.
- [18] C. A. Díez Díez, Inteligencia artificial en la contratación pública: instrumento y objeto de adquisición, *Ius et Veritas* (2025). doi:10.18800/iusetveritas.202502.009.
- [19] J. Venable, J. Pries-Heje, R. Baskerville, Feds: A framework for evaluation in design science research, *European Journal of Information Systems* 25 (2016) 77–89. doi:10.1057/ejis.2014.36.
- [20] I. G. B. Sampaio, et al., Trustworthy ai in public administration: The pro-trust pipeline, *Government Information Quarterly* 43 (2026) 102157. doi:10.1016/j.giq.2026.102157.