

Integrating Semantic Clustering-Based Recommendation Models and RAG Generative AI Chatbots for Regional Innovation Analytics

Alfi Fadliana¹, Dea Kayla Putri Darusman², Dinda Ayu Permatasari³✉

^{1, 2, 3} Applied Data Science Study Program, Department of Informatics and Computer Engineering, Politeknik Elektronika Negeri Surabaya, Indonesia

✉ **Corresponding Author** : Dinda Ayu Permatasari (e-mail: dindapermatasari020@gmail.com)

Article Information	ABSTRACT
Article History Received : July 29, 2026 Revised : August 14, 2026 Published : October 02, 2026	Regional research and innovation agencies are responsible for monitoring large volumes of innovation data submitted by local government units. Non-standardized submissions, manually matched collaboration opportunities, and a growing data request burden have left this data underutilized for policymaking. This study aimed to design, implement, and evaluate an integrated platform for regional innovation monitoring at the Regional Research and Innovation Agency (BRIDA) of East Java Province, using a design and development research approach. The platform combines a clustering based recommendation model that groups semantically similar innovations using multilingual sentence embeddings, an AI chatbot built on Retrieval-Augmented Generation (RAG) for querying the innovation database, and descriptive analytical reports on innovation maturity. Tested on 643 innovation records, the platform achieved a 100 percent functional pass rate across 35 test scenarios, while the chatbot achieved mean faithfulness and relevance scores of 4.0 out of 5.0 under an LLM as a judge evaluation. These results suggest that embedding based clustering can reveal previously hidden collaboration opportunities, while a generative chatbot can meaningfully reduce manual search workload. To the authors' knowledge, this is the first platform integrating clustering based recommendation with RAG based querying for regional innovation monitoring, offering a replicable design for similar settings.
Keywords: <i>Agglomerative Clustering;</i> <i>Innovation Clustering;</i> <i>Regional Innovation;</i> <i>Retrieval-Augmented Generation</i>	

INTRODUCTION

Regional research and innovation agencies play a pivotal role in shaping research-based policy within provincial governments. They are tasked with monitoring and evaluating regional innovation initiatives that span governance, public services, and technology adoption (Directorate General of Fiscal Balance, 2020, Aminah & Wardani, 2018). Their day-to-day operations involve managing large volumes of regional innovation data, encompassing hundreds of initiatives proposed by Regional Government Units (OPD) and community stakeholders alike. This dataset constitutes a strategic asset for mapping regional competitiveness and steering innovation policy.

The innovation data received is often non-standardized and inconsistent, which complicates the monitoring of innovation trends and the periodic evaluation of maturity levels. Identifying potential collaboration between innovations is likewise hindered, since matching hundreds of records is still done manually, even though structured collaboration mechanisms are known to encourage local governments toward greater innovation (Saksono, 2020). Beyond monitoring and matching, staff at the Regional Research and Innovation Agency must also

manually sift through this data to respond to routine inquiries from government units and the public, a workload that keeps growing as the number of submitted innovations increases. Left unaddressed, this inefficiency will continue to hamper data driven strategic decision making at the leadership level.

The challenge of collaborative matching can be addressed by clustering unlabeled text such as innovation descriptions using agglomerative clustering combined with multilingual sentence embeddings. This approach places semantically similar texts close together in vector space, even when the wording differs lexically (Reimers & Gurevych, 2020, Artetxe & Schwenk, 2019). Unlike centroid based methods such as k-means, agglomerative clustering does not require the number of clusters to be fixed in advance. It instead produces a merging hierarchy that can be evaluated at varying levels of granularity, making it well suited to administrative text corpora whose natural grouping patterns are not yet known. This property makes it possible to identify innovations with a sufficiently high degree of thematic similarity, which can then be recommended as collaboration candidates.

Beyond the collaboration challenge, the manual reporting burden can also be eased by combining Large Language Model (LLM) with Retrieval-Augmented Generation, an approach that grounds generated text in external knowledge retrieved through dense passage retrieval (Hambarde & Proença, 2023), thereby reducing hallucination and improving answer accuracy within specific domains (Lewis et al., 2020, Zhao et al., 2026, Kasneci et al., 2023). While a purely parametric language model generates responses based solely on knowledge acquired during training, a Retrieval-Augmented Generation based system retrieves relevant data from an external source at inference time and incorporates it as context before generating a response, keeping the output traceable to verifiable source documents. This approach has been shown to outperform conventional FAQ based chatbots in academic service settings (Suherwin et al., 2026), pointing to its potential in other institutional contexts that likewise depend on domain specific internal knowledge (Li et al., 2025, Swacha & Gracel, 2025).

Descriptive analytics, embedding based clustering, and Retrieval-Augmented Generation have each, individually, been applied fairly extensively across various domains. However, existing regional innovation monitoring efforts in Indonesia have so far relied primarily on static descriptive dashboards or manually curated reports, without embedding based mechanisms for surfacing latent collaboration opportunities or retrieval based tools for reducing the manual burden of responding to data inquiries. To the best of the authors knowledge, no prior study has integrated these three components into a single operational platform for a regional research and innovation agency, where the underlying data is institutional in origin, multilingual, and only partially standardized. While Retrieval-Augmented Generation chatbots have been evaluated in academic (Suherwin et al., 2026) and general public service (Pujono et al., 2024) contexts, none of these systems operate on institutional innovation data characterized by non-standardized, multilingual, and heterogeneous submission formats, nor are they paired with an embedding-based collaboration recommendation mechanism.

The main contribution of this study is therefore an integrated regional innovation monitoring platform for the Regional Research and Innovation Agency of East Java Province, combining embedding based clustering that recommends collaboration among semantically similar innovations, a retrieval based chatbot that eases the manual workload of searching and reporting on innovation data, and descriptive analytical reports that explain trends and maturity levels across the full dataset.

RESEARCH METHODS

This study employed a design and development research approach, encompassing the stages of needs analysis, system design, implementation, and evaluation. The needs analysis stage involved identifying the types and structure of innovation data managed by Regional Research and Innovation Agency (BRIDA) of East Java Province, as well as defining the functional requirements for the descriptive analytics, collaboration recommendation, and RAG chatbot

modules. The processing flow architecture connecting all platform components is presented in Figure 1.

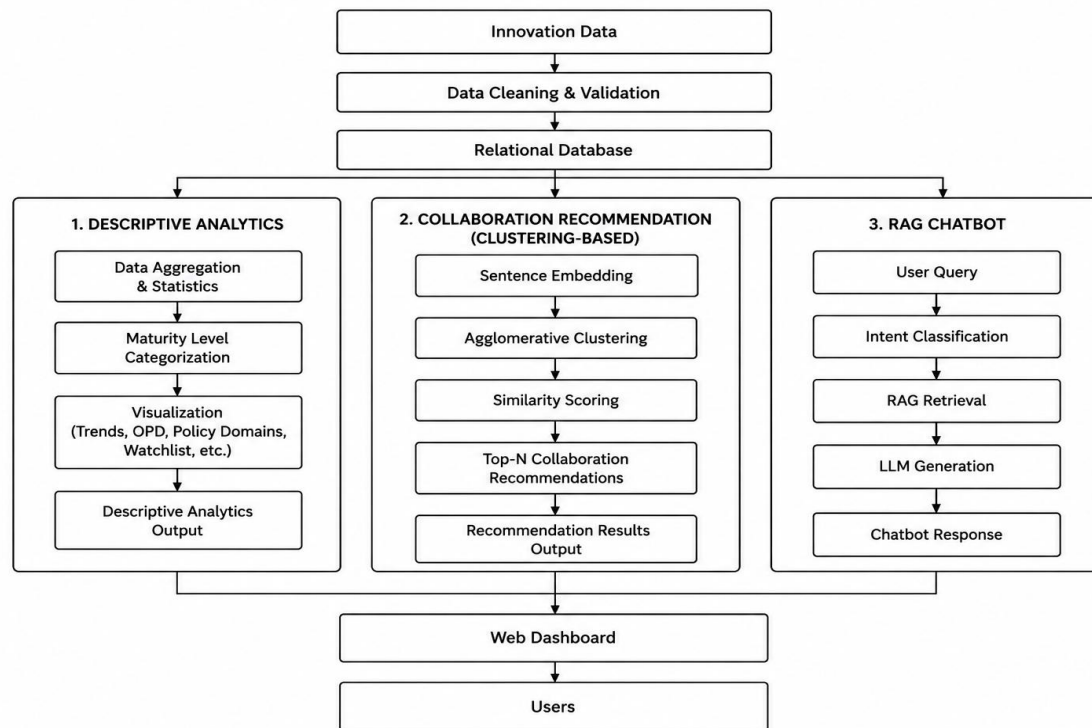


Figure 1. System Processing Flow Architecture

Figure 1 shows that the cleaned innovation data is stored in a relational database serving as a single source for three parallel processing paths. The first path, descriptive analytics, aggregates the data to produce maturity level categorizations along with visualizations of trends, OPDs, and policy domains. The second pipeline, the collaborative recommendation model, converts innovation descriptions into sentence embeddings before grouping them through agglomerative clustering and assigning similarity scores to generate ranked collaboration candidates. The third pipeline, a RAG-based chatbot, processes user queries through intent classification, semantic retrieval, and response generation by a Large Language Model. These three outputs are integrated into a single web dashboard accessible to both BRIDA staff and the general public. The implementation uses a FastAPI back end with a PostgreSQL database, along with a React and TypeScript based user interface.

The dataset attributes used in this study include the innovation title, free-text description, composite maturity score (range 0-111), development stage, innovation type (Digital, Non-Digital, or Technology), responsible OPD, primary policy domain and submission and implementation dates.

The study population consisted of 679 data points. These data then underwent several stages of cleaning, including the removal of duplicates based on innovation titles and the removal of incomplete or invalid entries such as empty descriptions or incorrect geolocation formats. After this stage, the sample size was reduced to 643 data points.

Descriptive Analytics

Each innovation record has a pre calculated maturity score, following a regional innovation index framework that combines two assessment aspects across seven variables and thirty five indicators (Directorate General of Fiscal Balance, 2020) into a composite score ranging from 0 to 111. This module groups these scores into three levels as formulated in Equation (1).

$$M(x) = \begin{cases} \text{Less Innovative,} & x < 45 \\ \text{Innovative,} & 45 \leq x < 65 \\ \text{Highly Innovative,} & x \geq 65 \end{cases} \quad (1)$$

This clustering component is specifically designed to support innovation collaboration recommendations. Each cluster that is formed serves as a set of candidates, from which the most similar innovation pairs are presented to users as concrete collaboration proposals, helping to address the difficulty of identifying recommendation opportunities. This also helps identify data that has reached the implementation stage but still has a maturity score below 45, allowing it to be followed up with more targeted policies.

Clustering-Based Recommendation Model

Each innovation description is first converted into a dense vector representation using sentence embeddings, building on the Siamese-network approach for computing sentence-level similarity through vector comparison (Reimers & Gurevych, 2019), specifically the trained multilingual sentence embedding model paraphrase-multilingual-mpnet-base-v2 (Reimers & Gurevych, 2020). From this representation, semantically similar innovations are grouped using agglomerative clustering with Ward linkage.

Compared to centroid based methods such as k-means, agglomerative clustering was selected for three main reasons. First, the number of clusters does not need to be specified in advance, which is suitable for this task since the actual number of natural collaboration groups among BRIDA innovations is not known beforehand. Second, the resulting merging hierarchy can be truncated at any level of granularity, allowing the number of clusters to be adjusted post hoc based on the silhouette evaluation described below. Third, k-means clustering performance depends heavily on centroid initialization and requires the number of clusters to be fixed a priori, a limitation that becomes more pronounced for high dimensional text embeddings (Petukhova et al., 2024). Empirical comparisons on large text datasets further show that agglomerative based clustering techniques tend to yield larger silhouette score improvements than k-means based techniques under equivalent embedding conditions (Sutrakar & Mogre, 2025), supporting its suitability for the semantically diverse and non-standardized innovation descriptions used in this study.

The clustering process itself is performed in the Euclidean space of the embeddings. To determine the optimal number of clusters k , the average silhouette coefficient based on the cosine distance (Ikotun et al., 2025) is used, tested over a range of k from 2 to 10. The choice of the cosine metric at this stage is based on its suitability for capturing the semantic similarity structure among embeddings, regardless of the distance metric used during the clustering process, as formulated in Equation (2).

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))} \quad (2)$$

where $a(i)$ denotes the average distance between sample i and all other points in the same cluster, while $b(i)$ denotes the minimum average distance between sample i and points in another cluster. For each formed cluster, the cosine similarity is calculated pairwise between innovation embeddings as formulated in Equation (3).

$$\text{CosSim}(A, B) = \frac{A \cdot B}{\|A\| \|B\|} \quad (3)$$

Pairs with a similarity score of at least 0.50 in each cluster are retained as collaboration candidates and categorized into three recommendation levels presented to users. High potential is assigned to pairs with a similarity above 0.90, moderate collaboration potential to pairs between

0.80 and 0.90, and low potential to pairs at or below 0.80, supporting interpretability for decision makers. This threshold based classification follows an approach commonly used in embedding based recommendation systems, where candidates are presented based on their cosine similarity scores (Riski et al., 2026). It should be noted, however, that these scores are not entirely free of bias, as they can be influenced by term frequency effects inherent in the embeddings on which the calculations are based (Sun et al., 2022).

Generative AI Integration

The generative component of this platform is designed using the Retrieval-Augmented Generation approach, with the goal of ensuring that generated responses are grounded in verifiable data rather than relying solely on the language model's internal knowledge (Lewis et al., 2020, Zhao et al., 2026). Each incoming query is first classified based on its intent, as illustrated in Figure 1, determining whether the request is routed to the semantic retrieval path or directly to the descriptive analytics layer.

The successfully retrieved data is then used as context for Gemini 2.5 Flash Lite to formulate the final answer in natural language. This model was chosen for its lightweight characteristics in terms of both latency and inference cost, making it suitable for production needs. This retrieval based approach aligns with previous findings showing that this method can reduce the risk of hallucination while delivering better results compared to template based or FAQ based approaches in equivalent public service contexts (Suherwin et al., 2026, Pujiono et al., 2024).

Evaluation Instrument and Data Analysis

For the clustering model, internal validity was assessed using the average cosine-based silhouette coefficient at the previously selected optimal value of k , complemented by the Davies-Bouldin Index to cross-check cluster separation from a different distance perspective. The cosine metric was retained for evaluation despite clustering being performed in Euclidean space with Ward linkage, since cosine similarity more directly reflects the semantic proximity intended by the sentence embedding representation. To further support the internal validity results, a subset of formed clusters was manually reviewed by inspecting sample innovation descriptions within each cluster, confirming that grouped items generally shared comparable policy themes despite the moderate silhouette score. No ground-truth cluster labels were available for this dataset, which limits the evaluation to internal and qualitative validity rather than external validation against a labeled benchmark.

Meanwhile, the generative artificial intelligence component was evaluated using the LLM-as-a-judge approach (Chang et al., 2024). A number of user queries representing all intent categories supported by the system were submitted to the chatbot; subsequently, each generated response was evaluated by a Large Language Model, deliberately chosen to be different from the model that generated the response itself to avoid self-preference bias. The evaluation is conducted across two dimensions based on the RAG evaluation framework such as RAGAS (Papageorgiou et al., 2025): faithfulness to the retrieved data and relevance to the user's query, each using a five-point scale. The evaluation results for these two dimensions are then aggregated to obtain a comprehensive picture of the chatbot's answer quality.

RESULTS AND DISCUSSION

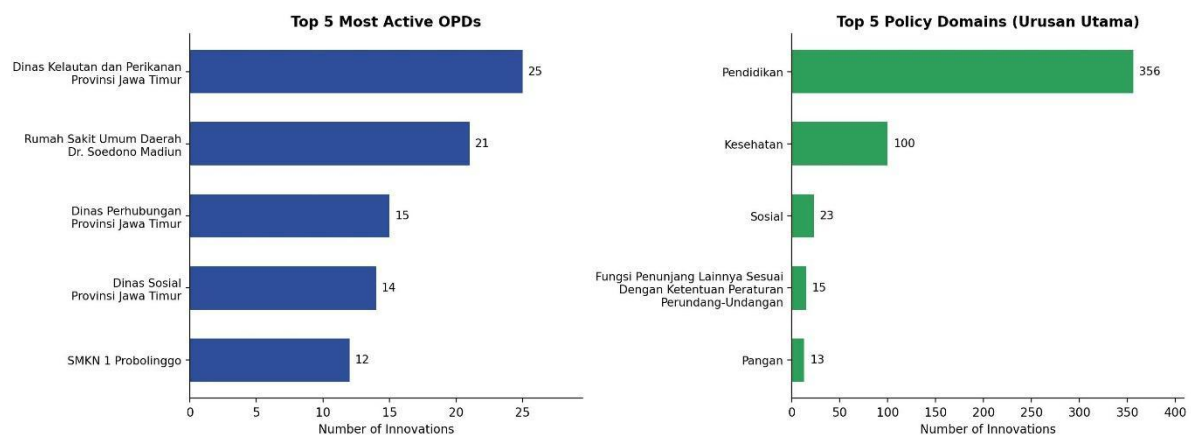
The observed dataset comprises 643 innovation records, with an average maturity score of 83.8. Of these records, 215 (33.4%) were categorized as digital innovations, 99 (15.4%) as technology-based innovations, and the remaining 329 (51.2%) as non-digital innovations. Applying the maturity categorization formula $M(x)$, which aligns with the labels previously calculated on the source data, yielded a level distribution summarized in Table 1.

Table 1. Distribution of Innovation by Maturity Level

Maturity Level	Total Data	Percentage (%)
Less Innovative ($x < 45$)	102	15.9
Innovative ($45 \leq x < 65$)	43	6.7
Highly Innovative ($x \geq 65$)	498	77.3
Total	643 (avr = 83.8)	100.0

The high proportion of records classified as Highly Innovative (77.3%) reflects a structural characteristic of the dataset rather than a genuinely even distribution of innovation quality: innovations submitted to BRIDA have typically already undergone internal institutional review prior to data entry, meaning that only initiatives considered viable tend to reach the monitored dataset. Records with markedly low maturity scores are therefore more likely to represent stalled initiatives or incomplete documentation than a representative cross-section of all regional innovation activity. This concentration at the upper tier limits the discriminative value of the maturity categorization for distinguishing quality within the Highly Innovative tier itself, and reinforces the importance of the watchlist mechanism described below, which specifically targets records that combine low maturity with advanced implementation status.

Based on this distribution, 102 innovations were identified as “Less Innovative” (maturity score < 45 , Table 2) that have reached the “Implementation” stage, including several with a score of 0, indicating stalled or poorly documented implementation, thereby requiring targeted policy follow-up. These descriptive findings reinforce the pattern of concentration of innovation activities in a small number of OPDs and policy domains, as illustrated in Figure 2, which depicts the five most active OPDs and the five most represented policy domains (key areas).

**Figure 2.** Most Active OPDs and Top 5 Policy Domains (N=643)

To develop the collaborative recommendation model, the first step was to determine the optimal number of clusters (k) by evaluating agglomerative clustering solutions for k values ranging from 2 to 10 on the sentence-embedding representation. The average cosine-based silhouette coefficient was used as the primary selection criterion, complemented by the Davies-Bouldin Index (DBI) as a corroborating internal validity metric. The resulting scores for both metrics are listed in Figure 3.

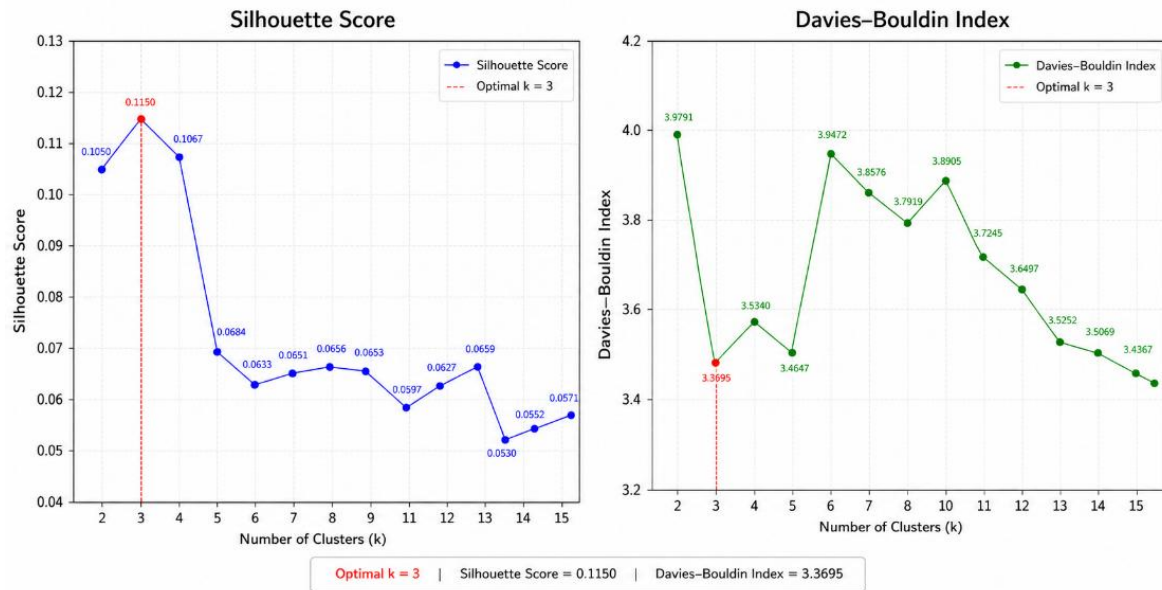


Figure 3. Silhouette Coefficient and Davies-Bouldin Index by Number of Clusters (k)

The highest silhouette coefficient was achieved when $k = 3$ (0.115), which was determined to be the optimal number of clusters. This selection is further corroborated by the Davies-Bouldin Index (DBI), which also reached its lowest value at $k = 3$ (3.3695) among all tested k , indicating the most favorable balance of cohesion and separation across both independent metrics. The absolute value of the silhouette score is nonetheless moderate, reflecting blurred semantic boundaries in the descriptions of administrative innovations, which tend to be brief and share similar vocabulary despite belonging to different policy domains. DBI values across the tested range also remained relatively high overall (3.37 to 3.98), suggesting that while $k = 3$ yields the best available separation, no clustering configuration achieved sharply delineated groups. Taken together, the agreement between silhouette and DBI at $k = 3$ strengthens confidence in this choice, while the moderate absolute values of both metrics indicate that the resulting clusters should still be interpreted as relative rather than sharply separated groupings.

This finding is consistent with values reported in the broader clustering literature for short, domain-specific text: Ikotun et al. (2025) note that silhouette scores below 0.2 are common in text clustering tasks where documents share substantial domain vocabulary despite belonging to distinct topical categories, a pattern directly applicable here given that most OPD innovations describe public-service delivery using similar administrative terminology regardless of their underlying policy domain. The modest score should therefore not be read as a weakness specific to this implementation, but as an expected characteristic of clustering short administrative text.

Within the formed clusters, potential collaboration pairs were ranked based on cosine similarity. The three highest-ranked pairs were SMANSKA AWARD paired with SMANSKA Love to Save (both from SMAN 1 Kwadungan, Ngawi) at 99% similarity classified as direct replication potential; SIJALINMAJA paired with SIGAP: Protecting and Safeguarding the People of East Java During Christmas and New Year's (both from the East Java Provincial Public Order Agency) at 91% similarity also classified as direct replication potential; and SI CINKELING from Ihya Ulumuddin Vocational High School Banyuwangi paired with GEMILANG from State Senior High School 1 Wongsorejo Banyuwangi at 89% similarity classified as development collaboration potential.

The top two pairs come from the same government unit, which is a limitation in the current implementation without a cross-unit filter during the ranking stage, pairs with high similarity within a single institution will appear alongside inter-organizational candidates that are actually different. These intra-unit pairs (Rank 1: 99%) may still indicate replication opportunities, but they do not represent collaboration in the intended sense. Rank 3 is the only genuinely cross-unit case,

as it connects two different schools in Banyuwangi with an 89% similarity, confirming that the model can generate thematically aligned innovations across institutions. Future implementations should therefore incorporate a cross-unit filter during the candidate screening stage.

The following section reports on the performance of the RAG-based chatbot, which fulfils the third objective, reducing the workload associated with manual queries. The Retrieval-Augmented Generation (RAG)-based chatbot was evaluated to determine whether a two-stage retrieval pipeline could generate answers sufficiently grounded in BRIDA's internal innovation data. This pipeline combines keyword-based SQL searches with cosine similarity using multilingual sentence embeddings. This two-stage design is crucial in this institutional context, as the same innovation concept can be described using different terminology across government agencies. This situation could potentially make keyword-based or embedding-based retrieval alone suboptimal when used in isolation. To assess the quality of the answers, a test set consisting of 21 queries was compiled, with three queries for each category to obtain a more stable estimate of the system's performance for each type of question. Each response was evaluated by an independent Large Language Model using a scale of one to five, covering faithfulness to the extracted records and relevance to the user's query (Chang et al., 2024). A summary of the results is presented in Table 2.

Table 2. LLM-as-a-judge Evaluation Scores by Query Intent Category

Intent Category	Number of Queries	Faithfulness (avr)	Relevance (avr)
Collaboration	3	4.0	3.7
OPD Rating	3	4.0	4.7
Policy Domain Rating	3	4.0	3.7
Stages of Innovation	3	4.0	4.0
Maturity Level	3	4.0	3.7
Trend	3	4.0	3.7
Statistics	3	4.0	4.7
Overall Average	21	4.0	4.0

The overall average faithfulness score of 4.0 out of 5.0 indicates that the retrieved data is consistently incorporated into the generated responses, rather than being replaced by the model's parametric knowledge. The overall average relevance score of 4.0 out of 5.0 indicates that intent classification successfully routes the majority of queries to the appropriate pipeline. However, categories such as Policy Domain Ranking and Maturity Level received lower scores, suggesting that aggregate queries sometimes produce factually accurate responses but do not directly address the quantitative dimensions of the question an area for improvement in prompt engineering.

A faithfulness score of 4.0 is consistent with findings by Pujiono et al. (2024), who reported that RAG-based chatbots in public service contexts consistently outperform parametric-only generation on factual grounding metrics. The relatively lower relevance scores for aggregate-oriented categories such as Policy Domain Rating and Maturity Level (3.7) compared to lookup-oriented categories such as OPD Rating and Statistics (4.7) align with a known limitation of intent-routing systems. In such systems, queries requiring numerical aggregation are harder to satisfy through retrieval alone without a dedicated query-to-SQL pipeline, making this area a key direction for future improvement.

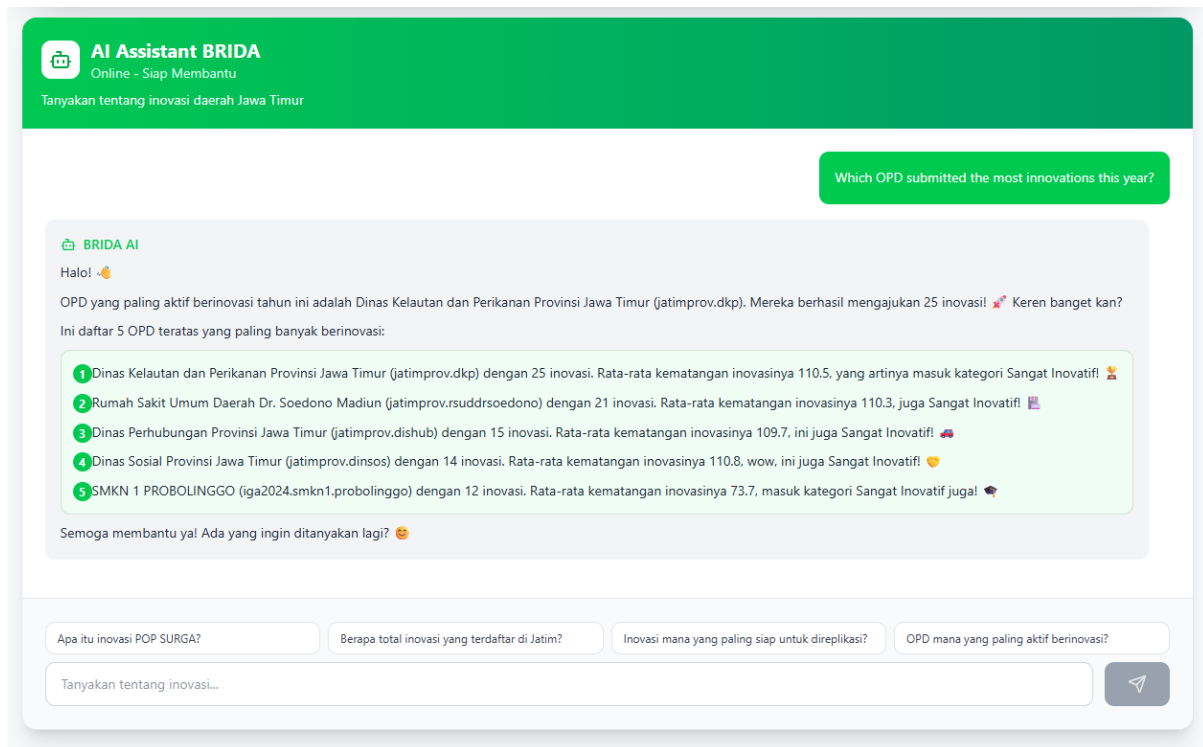


Figure 4. AI Chatbot Interface

Functional Testing

System testing was conducted using black-box testing, a method that assesses whether the system's output aligns with the expected results for a given input, without examining the internal code (Alagarsamy et al., 2024; Zimazanya et al., 2025). This approach was prioritized because functional correctness from the user's perspective is the primary focus of this validation phase.

A total of 35 test scenarios were defined across seven modules, included the landing page and real-time visitor counter, summary cards and trend charts on the dashboard, the analytics module, which includes year filters, reset, PDF export, and four types of charts, an innovation map covering marker rendering, choropleth tooltips, transportation layer toggles, a top-5 ranking panel with year filters, marker detail views, profile video links, a collaboration analysis tab, and collaboration line visualizations; an AI recommendation module covering the main recommendations view, detailed analysis pages, PDF export, and custom exploration, an RAG-based chatbot, and a data management module that includes admin authentication, invalid credential rejection, search and filtering, and add, edit, delete, and Excel/CSV export operations.

All 35 scenarios produced outputs consistent with the expected results, with a functional pass rate of 100%. Table 3 summarizes the results by module.

Table 3. Functional Testing Results

No	Module	Test Scenario	Pass	Failed
1	Landing Page & Visitor Counter	2	2	0
2	Dashboard	5	5	0
3	Analytics	8	8	0
4	Innovation Map	8	8	0
5	AI Recommendations	5	5	0
6	RAG-based Chatbot	1	1	0
7	Data Management	6	6	0
Total		35	35	0

The scope of testing included authentication, data manipulation, real-time visualization, AI-generated output, and file export. This confirmed that the system met all defined functional requirements for all implemented features. However, black-box testing does not capture the system's behavior when users experience concurrent loads or when the system receives invalid input. These two aspects will guide future evaluations.

Evaluation System

Several advantages are evident from this integrated implementation. The combination of embedding-based clustering with generative artificial intelligence significantly reduces the time required to identify collaboration candidates, a process that was previously done manually by reviewing hundreds of spreadsheet rows. This process automates both similarity calculations and narrative generation, while consolidating recommendations and chatbot responses into a single interface reduces the effort required to match clustering outputs with raw data.

Limitations also emerged. Collaboration narratives are highly dependent on the length and specificity of innovation descriptions, as reflected by the relatively low silhouette coefficient (0.115 at $k = 3$), this is not a model failure, but rather semantic boundary overlap in the source data (Reimers & Gurevych, 2020), thus requiring a standardized input protocol. Innovation activity also remains concentrated in a small number of OPDs, consistent with the literature indicating that resource disparities among government units result in uneven innovation output (Directorate General of Fiscal Balance, 2020).

This system also relies on the Google Gemini API, which limits the number of requests on the free plan; consequently, high usage may temporarily delay the Auto Insight and chatbot features. The lower relevance scores observed for aggregate-oriented intent categories in Table 3 further suggest that the current prompt structure does not optimally handle quantitative queries, pointing to prompt engineering as a targeted area for improvement independent of the retrieval mechanism itself. Manually entering data one by one in the data management feature also makes it difficult for staff when they need to upload large batches of innovation data.

There are three proposed improvements. First, BRIDA should establish structured input guidelines for OPDs when submitting innovation descriptions, including specifying a minimum length and required fields. This step will directly improve clustering accuracy and narrative coherence without altering the underlying model. Second, upgrading to the paid tier of the Gemini API will eliminate the request limit constraint that affects the availability of the chatbot and Auto Insight. Third, adding a batch import function that can accept Excel or CSV files to the data management feature will allow large-scale dataset updates to be completed in a single operation.

CONCLUSION

This study demonstrates that agglomerative clustering based on multilingual sentence embeddings can generate collaboration candidates not immediately apparent in administrative innovation data, even with less clearly defined semantic boundaries, achieving an optimal silhouette coefficient of 0.115 at $k = 3$. The RAG-based chatbot, evaluated using the LLM-as-a-judge framework across seven intent categories, achieved average faithfulness and relevance scores of 4.0 and 4.0 out of 5.0, respectively, thereby confirming that retrieval-based generation reduces reliance on parametric knowledge for domain-specific institutional queries. Functional testing across 35 scenarios yielded a 100% pass rate, validating the soundness of the underlying platform on which these AI components operate.

System evaluation identified input data quality as the primary constraint, both for clustering coherence and for the depth of chatbot responses, as well as API rate limits as a practical barrier to production deployment. These results carry practical implications for regional government agencies facing similar data management challenges: embedding-based clustering can surface collaboration candidates that manual review would miss, while a RAG-based chatbot can redistribute the query-handling burden away from staff without requiring users to learn database

query languages. The replicability of this design is supported by the use of open-source and low-cost components throughout, such as agglomerative clustering, multilingual sentence embeddings, and FastAPI. This approach reduces the infrastructure barrier for other agencies considering adoption. Future research should prioritize cross-unit filtering at the recommendation stage, the implementation of structured input guidelines for innovation descriptions, and an evaluation of the system's resilience to out-of-domain queries.

ACKNOWLEDGEMENTS

The authors would like to thank the Regional Research and Innovation Agency (BRIDA) of East Java Province for providing access to the regional innovation dataset and for its institutional support throughout the development of this research, as well as to the Electronic Engineering Polytechnic Institute of Surabaya facilitating this research through its Internship Program.

REFERENCES

- Alagarsamy, S., Tantithamthavorn, C., & Aleti, A. (2024). A3Test: Assertion-augmented automated test case generation. *Information and Software Technology*, 176, 107565. <https://doi.org/10.1016/j.infsof.2024.107565>
- Aminah, S., & Wardani, D. K. (2018). Readiness analysis of regional innovation implementation. *Jurnal Bina Praja*, 10(1), 13–26. <https://doi.org/10.21787/jbp.10.2018.13-26>
- Artetxe, M., & Schwenk, H. (2019). Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7, 597–610. https://doi.org/10.1162/tacl_a_00288
- Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., Chen, H., Yi, X., Wang, C., Wang, Y., Ye, W., Zhang, Y., Chang, Y., Yu, P. S., Yang, Q., & Xie, X. (2024). A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, 15(3), Article 39. <https://doi.org/10.1145/3641289>
- Directorate General of Fiscal Balance. (2020). *Kebijakan inovasi daerah* [PDF]. Kementerian Keuangan Republik Indonesia & Kementerian Dalam Negeri Republik Indonesia. <https://djpk.kemenkeu.go.id/wp-content/uploads/2020/11/Salinan-kebijakan-inovasi-daerah-DJPK.pdf>
- Hambarde, K. A., & Proença, H. (2023). Information retrieval: Recent advances and beyond. *IEEE Access*, 11, 76581–76604. <https://doi.org/10.1109/ACCESS.2023.3295776>
- Ikotun, A. M., Habyarimana, F., & Ezugwu, A. E. (2025). Cluster validity indices for automatic clustering: A comprehensive review. *Heliyon*, 11(2), e41953. <https://doi.org/10.1016/j.heliyon.2025.e41953>
- Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., Stadler, M., Weller, J., Kuhn, J., & Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103, 102274. <https://doi.org/10.1016/j.lindif.2023.102274>
- Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems*, 33, 9459–9474
- Li, Z., Wang, Z., Wang, W., Hung, K., Xie, H., & Wang, F. L. (2025). Retrieval-augmented generation for educational application: A systematic survey. *Computers and Education: Artificial Intelligence*, 8, 100417. <https://doi.org/10.1016/j.caeai.2025.100417>
- Papageorgiou, G., Sarlis, V., Maragoudakis, M., Magnisalis, I., & Tjortjis, C. (2025). Evaluating faithfulness in agentic RAG systems for e-governance applications using LLM-based

- judging frameworks. *Big Data and Cognitive Computing*, 9(12), 309. <https://doi.org/10.3390/bdcc9120309>
- Pujiono, I., Agtyaputra, I. M., & Ruldeviyani, Y. (2024). Implementing retrieval-augmented generation and vector databases for chatbots in public services agencies context. *JITK (Jurnal Ilmu Pengetahuan dan Teknologi Komputer)*, 10(1), 216–223. <https://doi.org/10.33480/jitk.v10i1.5572>
- Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, 3982–3992. <https://doi.org/10.18653/v1/D19-1410>
- Reimers, N., & Gurevych, I. (2020). Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 4512–4525). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2020.emnlp-main.365>
- Riski, F., Auliasari, K., & Orisa, M. (2026). Job vacancy recommendation system based on text description analysis using word embedding and cosine similarity. *bit-Tech*, 8(3), 3718–3729. <https://doi.org/10.32877/bt.v8i3.3690>
- Saksono, H. (2020). Center for innovation: Collaborative media towards innovative local government. *Nahkoda: Journal of Governance Science*, 19(1), 1–16
- Suherwin, Zainuddin, M., & Rachmat. (2026). Comparing FAQ-based and retrieval-augmented generation chatbots for academic service question answering. *G-Tech: Jurnal Teknologi Terapan*, 10(3), 1093–1105. <https://doi.org/10.70609/g-tech.v10i3.10019>
- Sun, X., Meng, Y., Ao, X., Wu, F., Zhang, T., Li, J., & Fan, C. (2022). Sentence similarity based on contexts. *Transactions of the Association for Computational Linguistics*, 10, 573–588. https://doi.org/10.1162/tacl_a_00477
- Swacha, J., & Gracel, M. (2025). Retrieval-Augmented Generation (RAG) chatbots for education: A survey of applications. *Applied Sciences*, 15(8), 4234. <https://doi.org/10.3390/app15084234>
- Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., Du, Y., Yang, C., Chen, Y., Chen, Z., Jiang, J., Ren, R., Li, Y., Tang, X., Liu, Z., Liu, P., Nie, J.-Y., & Wen, J.-R. (2026). A survey of large language models. *Frontiers of Computer Science*, 20(12), 2012627. <https://doi.org/10.1007/s11704-026-60308-3>
- Zimazanya, B., Rizky Supriyatna, M. A., & Anwar, C. (2025). Functionality testing of web-based library information systems using the blackbox testing method using the ISO/IEC 29119 standard. *Journal of Information Systems and Business Technology*, 1(4), 85–93
- Petukhova, A., Matos-Carvalho, J. P., & Fachada, N. (2025). Text clustering with large language model embeddings. *International Journal of Cognitive Computing in Engineering*, 6, 100–108. <https://doi.org/10.1016/j.ijcce.2024.11.004>
- Sutrakar, V. K., & Mogre, N. (2025). An improved deep learning model for word embeddings based clustering for large text datasets. arXiv:2502.16139