

Evolutionary Paradigms in University Knowledge Management: A Comparative Analysis of Intent-Based Architectures and Vectorized Semantic Retrieval

Bojan Papaz, Singidunum University, Miloš Mravik, Singidunum University

Abstract - This research examines the transition of conversational artificial intelligence within higher education administrative systems, specifically focusing on the shift from traditional intent-based architectures to vectorized semantic retrieval models. By analyzing the current implementation at Singidunum University, which utilizes the Rasa framework and a Dual Intent and Entity Transformer (DIET) pipeline, the study identifies a "knowledge bottleneck" when dealing with unstructured regulatory corpora. The investigation compares this deterministic model with emerging vector database technologies such as Pinecone and Weaviate, which employ high-dimensional embeddings and Cosine Similarity to retrieve information based on semantic proximity. The results suggest that while intent-based systems provide high precision for frequent queries, a hybrid model incorporating Retrieval-Augmented Generation (RAG) is essential for navigating complex institutional statutes.

Keywords - Artificial Intelligence; Vector Databases; Knowledge Management; Natural Language Understanding; Higher Education Administration

I. INTRODUCTION

The global landscape of higher education is undergoing a profound digital transformation, necessitated by the increasing complexity of administrative regulations and the rising expectations of a digitally native student population. Conversational artificial intelligence has moved beyond the role of simple automated responders, evolving into sophisticated administrative ecosystems that must navigate intricate legal frameworks.

The core challenge within these systems is effective knowledge management: the capacity to not only decipher the linguistic intent of a user query but to retrieve accurate, verified, and institutionally aligned information from a diverse array of sources. At institutions like Singidunum University, current implementations rely heavily on intent-based architectures where Natural Language Understanding (NLU) [1] models map student inquiries to predefined answers stored in structured databases and accessed via RESTful APIs.

While these systems demonstrate high precision for standard queries such as enrollment dates or tuition costs as they frequently encounter a "knowledge bottleneck" [2] when faced with the extensive, unstructured regulatory corpus of a modern university. This corpus includes statutes, rulebooks, and specialized regulations often trapped in static PDF formats,

which are difficult for traditional intent-labeling systems to process. The limitations of such models have prompted an investigation into more flexible, semantic-driven retrieval systems utilizing vector databases like Pinecone [3] or Weaviate [4]. By representing regulatory text as high-dimensional vectors, these systems allow for retrieval based on the underlying meaning of a query rather than exact keyword matches, representing a significant shift in institutional digital strategies.

The following analysis evaluates the architectural foundations of existing systems, the mathematical theory behind vectorized retrieval, and the operational advantages of a hybrid model that combines deterministic reliability with exhaustive semantic depth.

II. METHODS OF ANALYSIS

The comparative analysis performed in this study was structured around the evaluation of two distinct knowledge management paradigms. The first is the existing intent-based architecture deployed at Singidunum University, which utilizes the Rasa framework [5] for message processing and dialogue management. [6] This system was analyzed based on its NLU pipeline configuration, specifically the use of the DIET [6,7] classifier and mBERT [7] embeddings. Performance was assessed through historical macro-F1 scores and the operational efficacy of the Rasa Action Server [8] in handling synchronous API requests [9] for structured Q&A content.

The second paradigm involves the proposed transition to a vectorized semantic retrieval model. This approach was evaluated by examining the technical requirements for document ingestion, chunking, and indexing of the university's regulatory corpus. The study reviewed several vector database candidates: Pinecone, Weaviate, and Qdrant [10], comparing their latency, throughput, and support for hybrid search techniques. Furthermore, the research investigated the implementation of a Retrieval-Augmented Generation (RAG) pipeline, which utilizes a vector database to provide context to a Large Language Model (LLM) for generating natural language responses grounded in official statutes.

Data for the comparison was synthesized from technical documentation, university acts, and performance benchmarks associated with each architecture. The analysis also considered security and privacy [11] requirements, including compliance

with GDPR standards for data pseudonymization and the ethical implications of using generative AI in an administrative context. The goal was to establish a framework for a hybrid knowledge management system that maximizes accuracy while minimizing the frequency of system fallbacks during high-stress administrative periods, such as the June enrollment cycle.

III. RESULTS OF COMPARATIVE ANALYSIS

The investigation yielded detailed technical and operational insights into the two primary knowledge management models. The results are categorized into the mechanics of the current intent-based system, the mathematical and structural foundations of vectorized retrieval, and a technical comparison of their performance metrics. [12]

A. Architectural Foundations of Intent-Based Systems

The current system at Singidunum University is built on the Rasa framework, which provides a modular architecture for separating natural language understanding from dialogue logic. [6, 12] This system is specifically tuned for the Serbian language, which presents unique linguistic challenges such as complex morphology and the frequent omission of diacritics in informal digital communication.

The NLU pipeline utilizes a multi-layered approach to feature extraction. [13] Sparse features are generated using the CountVectorsFeaturizer, which processes words and character n-grams. The use of character-level n-grams (char_wb) is essential for maintaining robustness against typographical errors and the lack of Serbian diacritics (e.g., using "c" instead of "č" or "ć"). These sparse features are complemented by dense features from mBERT, which provide contextual embeddings that allow the system to differentiate between identical words used in different syntactic contexts. [14] The classification and entity extraction tasks are handled by the DIET model, a transformer-based architecture that learns both tasks simultaneously to improve overall accuracy. [15] This pipeline achieves macro-F1 scores above 0.85 for the 65 most common administrative intents. [16]

TABLE I. SUMMARY OF THE CURRENT ARCHITECTURAL STATE OF THE SINGIDUNUM ADMINISTRATIVE CHATBOT

ID	Summary of the current architectural state	
	Feature	Current Intent-Based System Subhead
1	NLU Model	DIETClassifier + mBERT
2	Language Support	Serbian (Robust to diacritics/lapses)
3	Knowledge Source	Structured Q&A API (Editorially controlled)
4	Dialogue Logic	RulePolicy + TEDPolicy
5	Latency (P50)	~180 ms (Warm cache)
6	Primary Limitation	Returns fallback for unmapped regulatory content

B. The Knowledge Bottleneck and Fallback Mechanism

Despite the precision of the DIET-based pipeline, the system faces a critical "knowledge bottleneck" because its operational success is binary: it can either answer a query for which an intent has been explicitly defined and trained, or it must trigger a fallback. A FallbackClassifier [17, 18] is employed to monitor the confidence scores of NLU predictions; if the score for the top intent falls below the 0.45 threshold, or if there is excessive ambiguity between two possible intents, the system defaults to a standard failure response. [17] While this prevents the delivery of incorrect information, it reveals a fundamental "wall" in the architecture: as the number of intents scales to accommodate the long-tail of university statutes, the system becomes computationally expensive to manage, and label correlation errors increase. [18]

This limitation is most evident when students inquire about specific regulatory articles sequestered in unstructured PDF documents, such as the Regulation on Conditions and Procedures for Student Enrollment or the Rulebook on Disciplinary Responsibility. Because the current structured Q&A API [9] typically contains only high-level, "pre-canned" summaries for frequently asked questions, it fails to retrieve the nuanced legal data required for specialized inquiries. Consequently, complex procedures such as the recognition of foreign qualifications or specific departmental operational protocols which often trigger fallbacks even though the information is legally defined within the university's internal documents. [19] Furthermore, the system is susceptible to "semantic drift," where the meaning of student queries shifts seasonally, often outpacing the manual and front-loaded effort required for administrative staff to update the intent-based model. This creates persistent information silos where the institution's official "source of truth" remains inaccessible to the automated agent.

C. Vectorized Semantic Retrieval and Mathematical Theory

To overcome the knowledge bottleneck, the research evaluated a paradigm shift toward vectorized semantic retrieval. In this model, the entire regulatory corpus is transformed into high-dimensional vectors, or embeddings, using deep learning models such as Sentence-Transformers. These embeddings place the text into a multi-dimensional space (typically 768 to 1536 dimensions) where the semantic relationship between different segments of text can be calculated mathematically. [20]

The primary metric for determining similarity in this space is Cosine Similarity [21], which measures the angle between a query vector and a document vector. The formula is expressed as follows:

$$\text{similarity} = \cos(\theta) = \frac{A \cdot B}{\|A\| \|B\|}$$

This approach is highly effective for institutional regulations because it prioritizes the thematic direction of the text over keyword frequency. A student's question about "transferring credits" can be semantically matched to a regulatory article about the "recognition of foreign degrees" even if the specific word "transfer" does not appear in the document.

D. Comparative Analysis of Vector Databases

The implementation of semantic retrieval requires a specialized database to store and query these high-dimensional vectors. The research analyzed three leading candidates for institutional deployment: Pinecone, Weaviate, and Qdrant. Pinecone is notable for its serverless architecture and low latency, making it ideal for high-traffic environments. Weaviate, however, offers superior flexibility for enterprise knowledge management through its support for hybrid search, which combines vector similarity with traditional keyword-based (BM25) search. This hybrid approach is particularly valuable for university regulations that contain specific article numbers or legal terms.

TABLE II. COMPARATIVE ANALYSIS OF VECTOR DATABASE CANDIDATES FOR INSTITUTIONAL DEPLOYMENT

Vector Database	Comparative analysis of vector database candidates	
	Key Advantages	Operational Context
Pinecone	Serverless, zero-ops, <50ms P95 latency	Fast-moving startups, rapid prototyping
Weaviate	Hybrid search, native multi-tenancy, GraphQL	Complex enterprise knowledge management
Qdrant	Rust-based, high throughput, in-memory speed	Real-time applications with custom infra

The technical performance comparison between the current intent-based API and the proposed vector database search reveals significant differences in throughput and accuracy for unmapped content. Vector search is optimized for scale, utilizing Approximate Nearest Neighbor (ANN) algorithms to find relevant matches among millions of documents in milliseconds.

TABLE III. TECHNICAL PERFORMANCE COMPARISON OF INTENT-BASED VS. VECTOR-BASED RETRIEVAL

Metric	Technical performance comparison	
	Intent-Based API (Current)	Vector Database Search
Query Time (P95)	~480 ms (Pipeline + API)	< 50-70 ms (Vector Lookup)
Throughput	Moderate (Rasa Action Server limited)	High (5,000 - 10,000 QPS)
Memory Usage	High (mBERT loading)	Efficient (Optimized vector stores)
Accuracy (Unmapped)	0% (Fallback)	~70-90% (Semantic Match)

E. Transitioning to a Hybrid Knowledge Management Model

The results indicate that the most effective strategy for university administration is a hybrid model that integrates the deterministic precision of intent-based APIs with the expansive coverage of vectorized retrieval. In this architecture, the intent-based system serves as the primary layer for high-confidence, frequently asked questions (FAQs). When the NLU confidence falls below the established threshold, the query is automatically routed to a Retrieval-Augmented Generation (RAG) pipeline.

In the RAG workflow, the vector database retrieves the most relevant passages from the university statutes. These passages are then provided as context to a generative LLM, which constructs a natural language response. This grounding ensures that the model's output is based strictly on institutional regulations, effectively eliminating the risk of hallucinations. Furthermore, the system can provide direct citations to the source material, such as "Article 15 of the Rulebook on Enrollment," thereby increasing transparency and student trust.

F. Operational Targets for Implementation

To ensure the success of the hybrid system, several operational targets have been established, focusing on uptime, accuracy, and user satisfaction. These targets reflect the rigorous requirements of educational administration, particularly during peak periods like the enrollment windows.

TABLE IV. OPERATIONAL TARGETS FOR THE HYBRID KNOWLEDGE MANAGEMENT SYSTEM

Metric	Operational targets for the hybrid knowledge system	
	Goal	Implementation Strategy
Uptime	99.9%	Cloud-native hosting with high availability
Hallucination Rate	< 5%	Vector-grounded responses with strict prompt engineering
Response Accuracy	> 92%	Hybrid search combining vector and keyword retrieval
User Satisfaction	> 8.0/10	Intuitive natural language interface with citations

IV. CONCLUSION OF THE STUDY

The comparative analysis of knowledge management models for higher education administration demonstrates that a hybrid approach is the most effective solution for bridging the gap between precision and comprehensive coverage. While intent-based architectures using DIET and mBERT provide a high level of accuracy for common administrative tasks, they are inherently limited by the need for explicit manual modeling. The integration of vectorized semantic retrieval through databases like Pinecone or Weaviate allows for the efficient processing of unstructured regulatory documents, effectively eliminating the current "knowledge bottleneck."

The implementation of a Retrieval-Augmented Generation pipeline ensures that AI-generated responses are grounded in the university's official statutes, meeting the high standards for accuracy, security, and ethics required in institutional governance. By leveraging these advanced retrieval techniques, universities can provide students with transparent, cited, and reliable information, even for complex and rare inquiries. This evolutionary shift not only improves the success rate of user queries but also significantly reduces the maintenance burden on administrative staff. Ultimately, the move toward a hybrid, semantic-driven model positions the university as a leader in digital transformation, ensuring that its "source of truth" remains accessible and effective in an increasingly complex educational environment.

ACKNOWLEDGEMENT

The authors express their gratitude to Singidunum University for providing the data and technical environment required for this comparative study. Support for this research was facilitated by the institutional commitment to digital transformation and the ongoing development of conversational AI systems for administrative support. Special thanks are extended to the administrative staff and technical teams whose work on the existing Rasa-based chatbot served as the foundational baseline for this analysis.

REFERENCES

- [1] Rasa, "Tuning Your NLU Model," 2025. [Online]. Available: <https://legacy-docs-oss.rasa.com/docs/rasa/tuning-your-model/>.
- [2] X. L. D. Y. a. J. T. H. Chen, "A Survey on Dialogue Systems: Recent Advances and New Frontiers," *ACM SIGKDD Explorations Newsletter*, vol. 19, pp. 25-35, 2017.
- [3] P. Systems, "Pinecone Documentation," 2025. [Online]. Available: <https://docs.pinecone.io/>.
- [4] W. B.V., "Weaviate Documentation," 2025. [Online]. Available: <https://weaviate.io/developers/weaviate>.
- [5] Rasa, "Knowledge Base Actions," 2025. [Online]. Available: <https://rasa.com/docs/reference/integrations/action-server/knowledge-bases/>.
- [6] J. E. M. M. a. A. N. V. Vlasov, "Dialogue Transformers," arXiv, 2019.
- [7] J. Devlin, M.-W. Chang, K. Lee and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," arXiv / NAACL-HLT 2019, 2018-2019.
- [8] Rasa, "Introduction to Rasa Action Server," 2025. [Online]. Available: <https://rasa.com/docs/action-server/>.
- [9] Rasa, "HTTP API," 2025. [Online]. Available: <https://legacy-docs-oss.rasa.com/docs/rasa/reference/http-api/>.
- [10] Qdrant, "Qdrant Documentation," 2025. [Online]. Available: <https://qdrant.tech/documentation/>.
- [11] R. Kamp and N. Weinbaum, "The myth of anonymization: Why AI needs a new privacy paradigm," 26 March 2025. [Online]. Available: <https://iapp.org/news/a/the-myth-of-anonymization-why-ai-needs-a-new-privacy-paradigm>.
- [12] s.-l. developers, "sklearn.feature_extraction.text.CountVectorizer — parametar analyzer," 2023. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.CountVectorizer.html.
- [13] J. F. N. P. a. A. N. T. Bocklisch, "Rasa: Open Source Language Understanding and Dialogue Management," 15 December 2017. [Online]. Available: <https://arxiv.org/abs/1712.05181>.
- [14] H. Face, "bert-base-multilingual-cased - model card," 2024. [Online]. Available: <https://huggingface.co/google-bert/bert-base-multilingual-cased>.
- [15] Rasa, "Custom Actions," [Online]. Available: <https://rasa.com/docs/reference/primitives/custom-actions/>.
- [16] M. Mantha, "Introducing Dual Intent and Entity Transformer (DIET) - state-of-the-art for intent and entity recognition," 9 March 2020. [Online]. Available: <https://rasa.com/blog/introducing-dual-intent-and-entity-transformer-diet/>.
- [17] Rasa, "Default Actions," 2025. [Online]. Available: <https://rasa.com/docs/reference/primitives/actions/>.
- [18] Rasa, "FallbackClassifier," 2025. [Online]. Available: https://legacy-docs-oss.rasa.com/docs/rasa/reference/rasa/nlu/classifiers/fallback_classifier/.
- [19] Rasa, "Fallback and Human Handoff," [Online]. Available: <https://legacy-docs-oss.rasa.com/docs/rasa/fallback-handoff/>.
- [20] Rasa, "TEDPolicy (Transformer Embedding Dialogue)," 2025. [Online]. Available: https://legacy-docs-oss.rasa.com/docs/rasa/reference/rasa/core/policies/ted_policy/.
- [21] C. D. Manning, P. Raghavan and H. Schütze, Introduction to Information Retrieval, Cambridge: Cambridge University Press, 2008.