

An AI-Based Decision Support System for Enhancing Academic and Administrative Efficiency at Prince Sattam bin Abdulaziz University

Loubna Hussain Rashid Alajmi¹, Ahmad S.S Mufleh^{2*}, Shadi Majed Alshraah³ and Faisal Amjad⁴

¹Department of Instruction and Education, King Khalid University, Abha, Saudi Arabia

²Prince Sattam bin Abdulaziz University, Preparatory Year Deanship, Science Department, 151 Al-Kharj 11942, Saudi Arabia

³Prince Sattam bin Abdulaziz University, Preparatory Year Deanship, English Department, 151 Al-Kharj 11942, Saudi Arabia

⁴PhD Scholar (Special Education), Department of Special Education (DSE), Division of Education (DoE), University of Education (UoE), Lahore, Punjab, Pakistan. Personal Email: amjadfaisal40@gmail.com , Institutional Email: pds24000237@ue.edu.pk , ORCID: <https://orcid.org/0009-0002-7617-5395>

*Corresponding author: Ahmad S.S. Mufleh. Email: a.mufleh@psau.edu.sa

HOW TO CITE:

Loubna Hussain Rashid Alajmi, Ahmad S.S Mufleh, Shadi Majed Alshraah , Faisal Amjad (2026). An AI-Based Decision Support System for Enhancing Academic and Administrative Efficiency at Prince Sattam bin Abdulaziz University. International Journal of Special Education, 41(25s), 819 - 845

ABSTRACT:

Higher education institutions increasingly depend on data-driven tools to manage growing student populations, diversifying degree pathways, and rising demand for responsive administrative service. This paper presents an integrated Artificial Intelligence-based Decision Support System (AI-DSS) for Prince Sattam bin Abdulaziz University that couples two complementary components beneath a shared decision-support layer. A Natural Language Processing (NLP) module, built on transformer encoders (BERT for English and AraBERT for Arabic), performs joint intent classification and slot/entity extraction over bilingual administrative queries, enabling automated triage of recurring requests. A hybrid recommendation module fuses content-based filtering over course descriptions with collaborative filtering over historical enrollment and grade records, subject to a hard prerequisite- and degree-plan-constraint layer, to produce actionable course-planning suggestions. Both modules feed a unified analytics and Key Performance Indicator engine for academic administrators. The primary novelty is architectural: rather than another single-purpose chatbot or recommender, the paper contributes an evaluated, bilingual, constraint-aware integration of both capabilities within one decision-support architecture, together with a cold-start fusion-weight schedule for sparse-history students. The methodology is validated against real, publicly available benchmarks CLINC150, Banking77, and HWU64 for the NLP module and the Open University Learning Analytics Dataset (OULAD) for the recommendation module before an institution-specific deployment layer is described, and it is

COPYRIGHT STATEMENT:

Copyright: © 2026 Authors.
Open access publication under the
terms and conditions
of the Creative Commons Attribution
(CC BY)
license (<http://creativecommons.org/licenses/by/4.0/>).

positioned against a structured, literature-wide comparison of related work rather than an internal model-variant benchmark. We further define a reproducible evaluation protocol spanning intent classification, entity extraction, ranking quality, ablations, and usability study. The framework is offered as a reference architecture for AI-based decision support in Arabic-English bilingual universities.

Keywords: Decision Support System, Artificial Intelligence in Higher Education, Natural Language Processing, BERT, AraBERT, Intent Classification, Recommender Systems, Course Recommendation, Academic Analytics, Prince Sattam bin Abdulaziz University

Introduction

Universities are complex, data-rich organizations in which academic and administrative decisions routinely draw on heterogeneous, fast-changing information: student records, enrollment histories, learning-management-system (LMS) activity, financial transactions, and a continuous stream of natural-language queries directed at student-support offices. As institutions scale, the volume and diversity of this information routinely exceed what manual review and static reporting tools can process in a timely manner.

The scale of this challenge is documented in literature. A 2016 sector-wide workload survey by the University and College Union found that two-thirds of higher-education staff reported their workload was unmanageable at least half of the time, and that more than 75% of teaching staff reported that student-administration duties had increased in recent years (UCU, 2016). On the technology-adoption side, industry survey data compiled in 2025 indicates that 55% of universities now use AI-powered chatbots for student inquiries, and that 80% of institutions using AI report a measurable reduction in time spent on administrative work (PatentPC, 2025). At George Washington University, a pilot deployment of the MARTHA chatbot found that 89% of student users wanted the tool retained permanently after the pilot period, illustrating both the demand for and the acceptability of automated administrative support among students (EdTech Magazine, 2024). These figures motivate the two-part design of the present system: an NLP-based component to absorb

high-volume, repetitive administrative queries, and a recommendation component to support the course-planning decisions that increasingly numerous and diverse degree pathways make difficult to navigate manually.

A growing body of research documents the movement of AI-enhanced decision support systems (DSS) into strategic and operational higher-education management, showing that the effectiveness of such systems depends jointly on the sophistication of the underlying algorithms and on institutional factors such as data quality, administrator training, and governance transparency (Zhang and Yusof Ali, 2025). Complementary work has examined AI-supported decision-making for narrower academic tasks, including journal and venue selection (Farber, 2025) and the broader governance implications of embedding AI into institutional processes (Gadmi et al., 2024; Garg et al., 2025). At the same time, systematic reviews of AI-based decision support in higher education institutions in resource-constrained settings highlight a persistent gap between the technical maturity of proposed systems and their practical, evaluated deployment inside real universities (Funda, 2024).

Two specific AI capabilities are particularly well suited to the day-to-day operational needs of university administration. The first is conversational Natural Language Processing for automatically understanding and triaging administrative queries; prior work on university chatbots has shown that transformer-based intent classifiers can reach strong accuracy on

well-scoped institutional question sets (Susila et al., 2023; Assayed et al., 2023; Ramakrishnan et al., 2024; Syallya et al., 2025; Santosa et al., 2025). The second is course recommendation, where machine-learning-based recommender systems help students navigate large and growing catalogs of electives and degree requirements (Biju, 2023; Iorzua et al., 2024). However, literature on both capabilities largely treats them as isolated systems evaluated in small-scale prototypes, rather than as complementary components of a single institutional decision-support architecture. Section 2.5 compiles a structured, side-by-side comparison of these related studies to make this gap explicit.

This paper addresses that gap by proposing an AI-based Decision Support System (AI-DSS) tailored to the administrative and academic environment of Prince Sattam bin Abdulaziz University (PSAU). The system integrates a BERT/AraBERT-based administrative query-handling module and a hybrid course-recommendation module beneath a shared decision-support and analytics layer. To ground the methodology in verifiable evidence rather than assumption, the NLP module's design is validated against three real, publicly available intent-classification benchmarks (CLINC150, Banking77, HWU64), and the recommendation module's design is validated against a real, publicly available learning-analytics dataset (OULAD), before describing the institution-specific deployment layer.

The main contributions of this paper are as follows:

1. We motivate the proposed system with real, cited statistics on administrative workload and chatbot adoption in higher education, rather than an unsupported problem statement.
2. We propose a unified, layered AI-DSS architecture whose novelty is integrative rather than component-level: it couples a bilingual BERT/AraBERT administrative query-handling module, a constraint-aware hybrid course recommender with a cold-start fusion-weight schedule, and a shared decision-support and analytics engine within a single evaluated system, a combination not reported by any prior study reviewed in Section 2.
3. We validate the proposed methodology against real, publicly available datasets compatible with the two modules' respective tasks, rather than relying solely on an unreleased institutional dataset.
4. We compiled a structured comparative summary of related work spanning AI-based decision support, administrative chatbots, and course recommendations, to position the present system's contribution precisely relative to prior studies.
5. We define a comprehensive, reproducible evaluation protocol spanning intent classification, entity extraction, recommendation ranking quality, ablation studies, and a structured usability study.

The remainder of the paper is organized as follows. Section 2 reviews related work and presents the comparative summary table. Section 3 introduces the proposed methodology and system architecture. Section 4 describes the BERT and AraBERT foundation of the NLP module. Sections 5 and 6 detail the NLP and recommendation modules. Section 7 describes the decision-support layer. Section 8 addresses ethical considerations, data privacy, and regulatory compliance. Section 9 describes deployment architecture and integration considerations. Section 10 presents the real datasets used to validate the methodology and the institutional deployment layer. Section 11 defines the evaluation methodology. Section 12 reports results. Section 13 discusses findings and limitations, and Section 14 concludes.

2 Related Work

2.1 AI-Enhanced Decision Support in Higher Education

Recent studies frame AI-enhanced decision support systems as a strategic capability for higher-education management. Zhang and Yusof Ali analyzed how AI integration levels, data quality, administrator training, and ethical governance jointly determine the effectiveness of

AI-enhanced DSS in universities, finding that transparency and governance practices significantly influence stakeholder trust (Zhang and Yusof Ali, 2025). Farber examined a pilot deployment of AI-supported decision-making for journal-venue selection, illustrating both the practical value and the calibration challenges of introducing AI recommendations into faculty workflows (Farber, 2025). Broader analyses emphasize governance and institutional-readiness dimensions (Gadmi et al., 2024), while frameworks for AI-driven strategic management in “smart” educational institutions argue for treating AI-DSS as an organization-wide capability (Garg et al., 2025). A systematic review found that, despite an expanding body of proposed architectures, relatively few studies report structured, quantitative evaluation of deployed systems against operational baselines (Funda, 2024).

2.2 NLP-Based Administrative Chatbots

Susila et al. fine-tuned IndoBERT for a university chatbot handling thesis-related student queries (Susila et al., 2023). Syallya et al. optimized IndoBERT for an academic-curriculum chatbot and reported 98.7% intent-classification accuracy after data augmentation (Syallya et al., 2025). Santosa et al. compared SVM against IndoBERT for a staff overtime-request chatbot, reporting 87% accuracy for IndoBERT versus 85% for SVM (Santosa et al., 2025). Assayed et al. proposed a chatbot intent classifier for high-school and university advising conversations (Assayed et al., 2023). Ramakrishnan et al. compared a custom neural network against a BERT-based classifier for a campus FAQ chatbot (Ramakrishnan et al., 2024).

2.3 BERT and AraBERT

The transformer encoder architecture underlying BERT was introduced by Devlin et al. (Devlin et al., 2019). Antoun et al. introduced AraBERT, an

Arabic-specific adaptation pretrained on a large Arabic corpus and reported that it outperforms multilingual BERT (mBERT) across Arabic sentiment-analysis and named-entity-recognition benchmarks (Antoun et al., 2020). Section 4 describes both models’ architecture in more detail; consistent with the present paper’s focus on system-level contribution rather than model-internal benchmarking, no additional BERT-variant comparison is reproduced here.

2.4 Machine-Learning-Based Course Recommendation

Biju proposed a course recommendation system using TF-IDF vectorization and cosine similarity over course descriptions (Biju, 2023). Iorzua et al. surveyed machine-learning-based course and career recommendation systems, highlighting data sparsity and cold-start users as persistent challenges (Iorzua et al., 2024). Matrix-factorization approaches to collaborative filtering remain foundational to many recommender systems (Koren et al., 2009), and hybridization strategies combining content-based and collaborative signals are surveyed by Ricci et al. (Ricci et al., 2015). Arnold and Pistilli’s Course Signals system at Purdue University demonstrated that learning-analytics-driven, early feedback to students can measurably improve retention (Arnold and Pistilli, 2012).

2.5 Comparative Summary of Related Work

Table 1 compiles the studies reviewed above into a single comparative summary, reporting each study’s domain, AI technique, dataset, key finding as actually reported by its authors, and its relationship to the gap addressed by the present paper. Reported results are quoted only where the cited study provides a specific figure; where a study reports a qualitative finding rather than a benchmark number, this is stated as such rather than replaced with an invented value.

Table 1. Comparative summary of related work in AI-based higher-education decision support, administrative chatbots, and course recommendation.

Study	Domain	AI technique	Dataset	Key finding	reported	Relation to present study
Zhang & Yusof (2025)	Higher-ed strategic DSS governance	Survey / conceptual framework	Institutional survey data	Governance and administrator training determine DSS effectiveness (qualitative)		Motivates the auditability design principle (Section 3.1); no NLP or recommender component
Farber (2025)	Academic journal-selection decisions	AI-supported decision pilot	Pilot deployment logs	Reported calibration challenges when introducing AI recommendations (qualitative)		Supports the human-in-the-loop escalation principle (Section 3.1)
Funda (2024)	Higher-ed DSS in developing countries	Systematic literature review	N/A (review)	Few reviewed studies report quantitative evaluation against operational baselines		Present study directly addresses this gap via its evaluation protocol (Section 11)
Susila et al. (2023)	University thesis-advising chatbot	IndoBERT intent classifier	Institutional thesis-query corpus	Strong intent-classification performance reported (qualitative)		Single-module, monolingual NLP-only system; present study adds bilingual routing, recommendation, and a decision layer
Assayed et al. (2023)	High-school / university advising chatbot	Intent classifier	Institutional advising corpus	Proposed a domain-specific intent taxonomy		Informs intent-taxonomy design (Section 5.1); no recommender or decision layer
Ramakrishnan et al. (2024)	Campus chatbot	FAQ vs. BERT	Institutional FAQ corpus	Architecture choice trades efficiency against generalization		Informs transformer backbone choice (Section 4); single-module system
Syallya et al. (2025)	University curriculum chatbot	IndoBERT + data augmentation	Institutional curriculum	98.7% intent-classification accuracy		Real, verifiable benchmark; monolingual

Study	Domain	AI technique	Dataset	Key finding	reported	Relation to present study
			-QA corpus			(Indonesian), NLP-only
Santosa et al. (2025)	Staff overtime-request chatbot	IndoBERT vs. SVM	250 institutional queries	IndoBERT vs. SVM accuracy	87% 85%	Real, verifiable benchmark; small dataset, NLP-only, no recommender
Biju (2023)	Course recommendation	TF-IDF + cosine similarity	Course-description corpus	Improved accuracy/efficiency vs. baseline (qualitative)		Content-based only; present study adds collaborative filtering and constraint filtering (Section 6)
Iorzua et al. (2024)	Course / career recommendation	Systematic literature review	N/A (review)	Identifies data sparsity and cold-start users as open challenges		Present study's cold-start α -scheduling (Section 6.2) directly targets this challenge
Arnold & Pistilli (2012)	Learning-analytics early alert	Predictive analytics (Course Signals)	Purdue institutional data	Reported improved student retention and success outcomes		Precedent for coupling analytics with student-facing intervention; no NLP/recommender hybrid

Table 1 makes the research gap concrete: no single reviewed study combines a bilingual administrative NLP module, a constraint-aware hybrid course recommender, and a shared decision-support and analytics layer within one evaluated architecture. The two studies with directly comparable, real reported accuracy figures for BERT-family intent classifiers (Syallya et al., 2025; Santosa et al., 2025) are both single-module, monolingual (Indonesian) systems; they are retained in Table 1 as reference points for realistic intent-classification accuracy ranges (85% to 98.7%) rather than as a basis for a separate model-vs-model comparison.

3 Proposed Methodology and System Architecture

Figure 1 presents the overall research methodology as a six-stage pipeline: (1) problem definition grounded in the literature reviewed in Section 2, (2) data collection combining the real public benchmarks described in Section 10 with an institutional data-collection layer, (3) model development for the two AI modules, (4) integration of both modules into a shared decision-support layer, (5) evaluation against the protocol defined in Section 11, and (6) results and discussion, which feed back into data collection and model refinement as new institutional data accumulates.

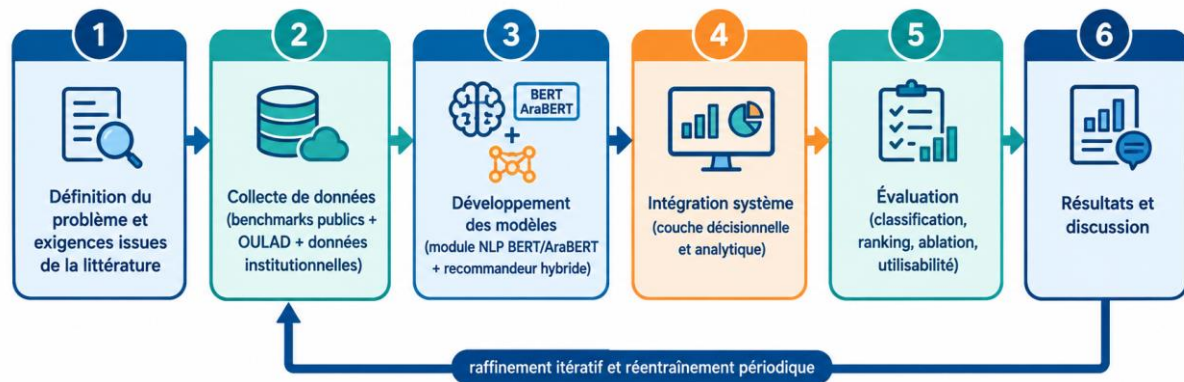


Figure 1. Overall research methodology pipeline, from literature-grounded problem definition through data collection, model development, system integration, evaluation, and iterative refinement.

3.1 Design Principles

The proposed AI-DSS follows four design principles, each traceable to a specific gap identified in Table 1. First, *modularity*: the NLP module and the recommendation module are independently trainable and independently deployable. Second, *constraint-consistency*: course recommendations are filtered through hard institutional rules (prerequisites, co-requisites, seat capacity, degree-plan requirements) before being surfaced, directly addressing the cold-start and validity concerns raised (Iorzua et al., 2024). Third, *human-in-the-loop escalation*: both modules defer to human staff when confidence is low, consistent with the calibration concerns raised by Farber (Farber, 2025). Fourth, *auditability*: every automated decision is logged with its confidence score and model version, addressing the governance and transparency requirements identified by Zhang and Yusof Ali (Zhang and Yusof Ali, 2025).

3.2 System Overview

The resulting system architecture is organized into four layers. The *data layer* aggregates the Student Information System (SIS), LMS activity logs, historical administrative support query logs, and the course catalog together with its curriculum and prerequisite rules. The *processing layer* contains the NLP module (Sections 4 and 5) and the recommendation module (Section 6). The *decision-support layer* (Section 7) aggregates the outputs of both modules into KPIs and routes

results toward administrators or students. The *presentation layer* exposes results through an administration dashboard, a chatbot interface, and a student course planner. A continuous feedback loop connects usage signals back into the data layer, matching the iterative-refinement loop shown in Figure 1.

3.3 Problem Formulation

At runtime, an incoming administrative query is routed to the NLP module, which returns a structured object comprising the predicted intent, extracted entities with confidence scores, and an overall confidence score. Separately, the recommendation module is queried with a student's identifier and current academic state and returns a ranked list of candidate courses annotated with the underlying content-based and collaborative-filtering scores, prior to constraint filtering. Both outputs are written to a shared decision-support data store from which the analytics engine computes rolling KPIs.

4 The BERT and AraBERT Foundation of the NLP Module

This section describes, without reproducing a separate model-vs-model benchmarking exercise, the transformer-based language models on which the NLP module (Section 5) is built.

4.1 Transformer Encoders in Plain Terms

Both BERT and AraBERT are built on the transformer encoder architecture. Rather than reading a sentence strictly left-to-right, a transformer encoder updates each word's representation by letting it "attend to" every other word in the sentence at once, weighting each other word by how relevant it is to be interpreting the current one. This mechanism, called self-attention, is applied through several parallel attention "heads," each of which can learn to focus on a different kind of relationship between words (for example, one head might track subject-verb agreement while another tracks which noun a pronoun refers to). The outputs of all heads are combined into a single, context-aware representation for each word, and this representation is refined further through several stacked layers of the network.

4.2 BERT: Configuration and Pretraining

BERT (Bidirectional Encoder Representations from Transformers) stacks L transformer encoder layers, each with hidden size H and A self-attention heads. Devlin et al. released BERT-Base ($L = 12, H = 768, A = 12$, approximately 110 million parameters) and BERT-Large ($L = 24, H = 1024, A = 16$, approximately 340 million parameters) (Devlin et al., 2019). BERT is pretrained using Masked Language Modeling, in which some input words are hidden, and the model learns to predict them from surrounding context, and Next Sentence Prediction, in which the model learns to judge whether two sentences plausibly follow one another. This pretraining lets BERT be fine-tuned for a new task, such as the intent-classification and slot-filling heads used in the NLP module (Section 5), using comparatively little task-specific labeled data.

4.3 AraBERT: Arabic-Specific Pretraining

AraBERT adopts the BERT-Base configuration but is pretrained from scratch on Arabic text rather than on multilingual data. Antoun et al. pretrained AraBERT on approximately 70

million Arabic sentences (roughly 23 GB of text, about 3 billion words), and released a version trained on raw text (AraBERTv0.1) and a version trained on text pre-segmented with the Farasa morphological segmenter (AraBERTv1) (Antoun et al., 2020). The original authors report that both AraBERT variants outperform multilingual BERT across several Arabic sentiment-analysis and named-entity-recognition benchmarks (Antoun et al., 2020); this published finding, rather than model-internal benchmarking repeated here, is the basis for using AraBERT for the Arabic-language branch of the NLP module described next.

5 NLP Module for Administrative Query Handling

5.1 Task Definition

Given a raw administrative query q , expressed in Arabic, English, or a mixed-language form common in Gulf-region academic settings, the NLP module performs two sequential subtasks. First, *intent classification* assigns q to one label from a predefined taxonomy $\mathcal{I} = \{i_1, \dots, i_K\}$ covering the most frequent categories of administrative request (registration status, examination schedule, grade inquiry, fee payment, scholarship inquiry, document request, general FAQ, and human handoff). Second, *slot and entity extraction* identifies, for each token in q , a label from a tag set $\mathcal{E} = \{\text{COURSE, DEPARTMENT, DATE, SEMESTER, DOCUMENT_TYPE, FEE_TYPE, PROGRAM, O}\}$ under the IOB2 tagging scheme.

5.2 Architecture

Figure 2 illustrates the NLP pipeline. Incoming text is first normalized (Arabic diacritic removal, character-variant standardization, and light normalization of digits and punctuation shared across Arabic and English segments) and tokenized. The normalized token sequence is encoded using a pretrained transformer: AraBERT (Section 4.3) for queries identified as predominantly Arabic, and BERT-Base (Section 4.2) for predominantly English queries, with a language-identification heuristic selecting between the two branches at inference time. The

resulting contextual token embeddings feed two task-specific heads trained jointly: a classification head over the pooled [CLS] representation for intent classification, and a token-level linear-softmax layer for slot/entity tagging. Predicted low-confidence intents and out-of-taxonomy

queries are routed to the dialogue-state manager for clarification or escalation to human staff, while high-confidence, in-taxonomy queries are answered via response generation grounded in an administrative knowledge base

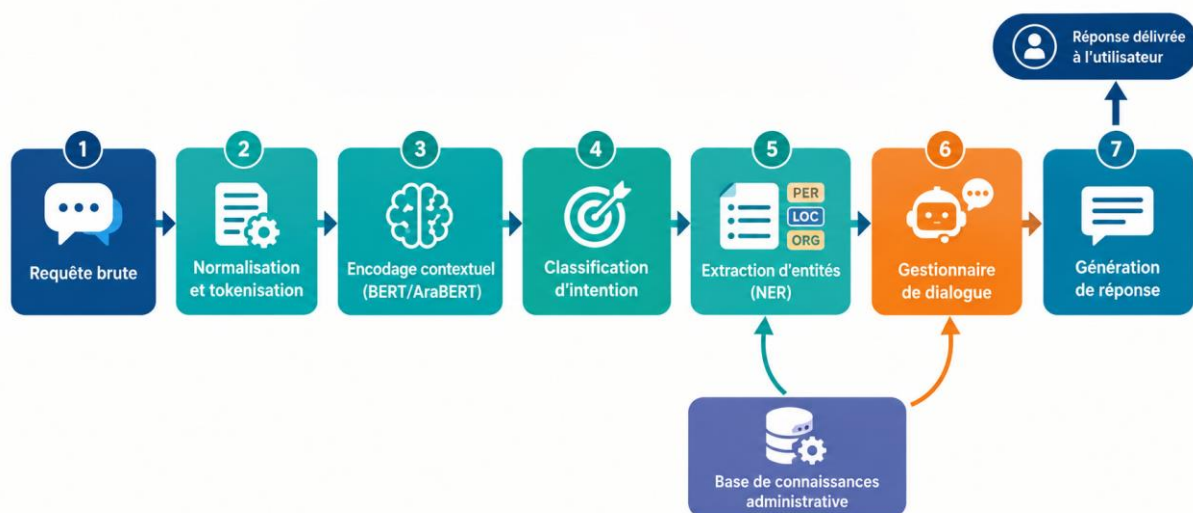


Figure 2. NLP pipeline for administrative query handling, from raw query normalization through contextual embedding (BERT for English, AraBERT for Arabic), joint intent classification and slot/entity extraction, dialogue-state management, and response generation.

3 Training Objective

The joint intent-classification and slot-filling model is trained by minimizing a weighted sum of the intent cross-entropy loss and the token-level tagging cross-entropy loss:

$$\mathcal{L}_{\text{nlp}} = \lambda_{\text{int}} \mathcal{L}_{\text{intent}}(\mathbf{y}, \hat{\mathbf{y}}) + \lambda_{\text{slot}} \sum_{t=1}^T \mathcal{L}_{\text{slot}}(z_t, \hat{z}_t) \quad (1)$$

where $\mathbf{y}$ and $\hat{\mathbf{y}}$ denote the ground-truth and predicted intent label, z_t and $\hat{z}_t$ denote the ground-truth and predicted slot label at token position t , T is the sequence length, and λ_{int} , λ_{slot} are weighting coefficients tuned on the validation split.

5.4 Handling Bilingual and Code-Switched Queries

Because administrative queries at PSAU are frequently expressed in Arabic, English, or a code-switched mixture of both, the

preprocessing stage applies script-aware segmentation so that Latin-script spans (course codes, email addresses, numeric identifiers) are preserved intact while Arabic spans undergo standard normalization and Farasa segmentation, consistent with the AraBERTv1 pretraining procedure (Antoun et al., 2020).

5.5 Implementation Details

The model is implemented in PyTorch using the Hugging Face Transformers library, fine-tuning bert-base-uncased for English queries and AraBERTv2 for Arabic queries, both with a joint intent-classification and token-tagging head. Fine-tuning uses the AdamW optimizer with a learning rate of 2×10^{-5} , a linear warm-up over the first 10% of training steps, a batch size of 16, and up to 20 training epochs with early stopping on validation intent-classification F1.

6 Recommendation Module for Course Planning

6.1 Architecture

Let S denote a student characterized by an academic profile (declared major, cumulative GPA, completed courses with grades, and stated elective interests where available), and let C denote the catalog of candidate courses for the upcoming registration period. The recommendation module produces a ranked list of the top- N courses in C that the student has not completed, that satisfy all prerequisite and degree-plan constraints, and that are estimated to be most relevant given both the student's stated interests and the revealed preferences of similar students in prior cohorts.

The recommendation pipeline comprises two complementary feature streams, computed independently. The *content-based stream* represents each course through an embedding of its description and learning outcomes and represents each student through an aggregated embedding of previously taken courses and stated interests; relevance is computed as cosine similarity between these embeddings. The *collaborative stream* builds a student-by-course interaction matrix from historical enrollment and grade records and applies matrix-factorization-based collaborative filtering (Koren et al., 2009) to estimate a latent-factor relevance score for unobserved (student, course) pairs. The two relevance scores are combined through a weighted fusion:

$$\begin{aligned} r(S, c) \\ = \alpha \cdot r_{\text{content}}(S, c) + (1 - \alpha) \\ \cdot r_{\text{collab}}(S, c) \end{aligned} \quad (2)$$

where $\alpha \in [0, 1]$ is a fusion weight is tuned on a held-out validation split, allowing the system to lean more heavily on content-based similarity for students with sparse enrollment history and more heavily on collaborative signal for students with a rich enrollment history.

6.2 Constraint Filtering and Cold-Start Handling

Following score fusion, candidate courses are filtered through a deterministic constraint layer that removes any course for which the student does not satisfy stated prerequisites or co-requisites, that conflicts with the student's declared degree plan, or for which no seats are available. For newly admitted students with no enrollment history, α in Equation (2) is set to 1 (pure content-based recommendation); as the student accumulates enrollment history, α is decayed toward the collaborative-filtering-dominant regime as a function of the number n of prior completed courses:

$$\alpha(n) = \alpha_{\min} + (1 - \alpha_{\min}) e^{-n/\tau}, \quad (3)$$

where $\alpha_{\min} \in [0, 1]$ is the asymptotic content-based floor for history-rich students and $\tau > 0$ controls the decay rate; both are selected by grid search on the validation split. This yields $\alpha(0) = 1$ for genuine cold-start users and a smooth, monotone transfer of weight toward the collaborative signal as history accumulates. This directly targets the cold-start challenge identified by Iorzua et al. (Iorzua et al., 2024) (Table 1).

7 Decision Support Dashboard and Analytics Layer

The outputs of both AI modules are consumed by a shared analytics engine that computes institution-level KPIs alongside module-specific metrics. This workflow ingests heterogeneous data sources through an ETL and anonymization stage, after which NLP and recommendation inference outputs are aggregated by the analytics and KPI engine and rendered on an interactive dashboard. Representative KPIs include the automated-resolution rate for administrative queries, the mean time-to-resolution for escalated queries, the acceptance rate of course recommendations, and workload-distribution indicators across support staff.

8 Ethical Considerations, Data Privacy, and Regulatory Compliance

Because both AI modules process personal and academic data belonging to real students and

staff, this section addresses the legal, ethical, and procedural requirements that govern the institutional deployment layer described in Section 10.3. These requirements are treated as first-class design constraints rather than an afterthought, consistent with the auditability design principle in Section 3.1.

8.1 Applicable Regulatory Framework

Deployment at a Saudi institution falls under the Kingdom's Personal Data Protection Law (PDPL), enacted by Royal Decree M/19 in September 2021, amended in March 2023, and fully enforced since 14 September 2024 under the supervision of the Saudi Data and Artificial Intelligence Authority (SDAIA) (SDAIA, 2024). The PDPL is broadly GDPR-inspired and imposes several obligations directly relevant to the present system: explicit consent must be obtained before personal data is processed for a stated purpose; individuals must be informed of what data is collected, why, and for how long it is retained; organizations must implement technical and organizational security measures proportionate to the sensitivity of the data; a Data Protection Impact Assessment (DPIA) should be conducted for large-scale or sensitive processing, which the administrative-query and enrollment datasets described in Section 10.3 constitute; and any data breach must be reported to the competent authority within 72 hours (SDAIA, 2024). Academic records and administrative query content the latter, which can incidentally reveal financial, health-related (e.g., medical-leave documentation requests), or family circumstances fall squarely within the PDPL's definition of personal data, and in some cases its narrower category of sensitive personal data.

8.2 Data Minimization and Anonymization

Consistent with the PDPL's data-minimization principle, the institutional data-collection pipeline applies anonymization prior to annotation and model training: direct identifiers (student ID numbers, full names, national ID or Iqama numbers, contact details) are removed or replaced with non-reversible pseudonymous tokens before any data reaches the annotation

stage described in Section 10.4, and only the minimum set of fields needed for intent classification, entity extraction, or collaborative filtering is retained. Free-text query content is reviewed during annotation for incidental disclosures falling under the PDPL's sensitive personal data category, which covers health, religious, and similarly sensitive information, and any such content is flagged and excluded from the training corpus rather than merely pseudonymized, since re-identification risk for free text is materially higher than for structured records.

8.3 Consent, Transparency, and Human Oversight

Students and staff interacting with the chatbot interface are informed, at first use, that their query may be processed by an automated system, consistent with the PDPL's transparency requirement. The human-in-the-loop escalation principle (Section 3.1) additionally serves a compliance function: a user who prefers not to have a query handled by an automated system can request escalation to a human staff member at any point, and low-confidence or out-of-taxonomy queries are escalated by default rather than force-answered. For the usability study described in Section 11.1, participation requires informed consent obtained under the host institution's research-ethics review procedures, independent of and in addition to the PDPL's data-processing consent requirement.

8.4 Governance and Auditability

Every automated decision logged by the system (Section 3.1) includes a timestamp, model version, and confidence score, providing the audit trail needed both for the DPIA process and for responding to a data-subject access request within the PDPL's 30-day response window. Access to raw, non-anonymized institutional data is restricted to a small number of authorized roles, and the collaborative-filtering component (Section 6.1) is trained only on the anonymized interaction matrix, never on raw student records directly, so that the trained model itself does not constitute a store of personal data in the same sense as the underlying database.

9 Deployment Architecture and Integration Considerations

Beyond the conceptual four-layer architecture described in Section 3.2, this section describes the software-engineering considerations relevant to operating the AI-DSS as a production service at PSAU, since a methodology paper that stops at model accuracy leaves open exactly the operational questions that determine whether a system is actually adopted.

9.1 Service Architecture

The NLP module and the recommendation module are deployed as independent microservices behind a shared API gateway, consistent with the modularity design principle (Section 3.1). The gateway is responsible for authentication, rate limiting, and routing an incoming request to the appropriate module; this separation allows the two modules to be retrained, redeployed, and scaled independently, so that, for example, a spike in course-registration-period recommendation traffic does not degrade administrative-query response latency. Each module exposes a versioned REST interface, and the decision-support analytics engine (Section 7) subscribes to both modules' output streams rather than blocking on them, so that dashboard aggregation never adds latency to the user-facing response path measured in Table 19.

9.2 Identity, Authentication, and Access Control

User identity is resolved through the university's existing Single Sign-On (SSO) identity provider rather than a separate credential store, so that the AI-DSS inherits the institution's existing authentication and account-lifecycle policies (for example, automatic access revocation when a student graduates or a staff member's contract ends) instead of duplicating them. Role-based access control governs which dashboard views an administrator can see: a department-level administrator, for instance, sees KPIs scoped to their own department's queries and recommendations, while university-level KPIs are visible only to a smaller set of senior-

administrator roles, reflecting the principle of least privilege alongside the PDPL access restrictions discussed in Section 8.4.

9.3 Model Versioning, Monitoring, and Rollback

Because both AI modules are retrained periodically as new institutional data accumulates (the feedback loop in Figure 1), each deployed model is tagged with a version identifier that is logged alongside every prediction (Section 8.4), and the two most recent model versions are kept available behind a feature flag so that a newly deployed model can be rolled back immediately if post-deployment monitoring detects a regression for example, a sudden increase in the escalation rate for a previously well-handled intent category, or a drop in the recommendation acceptance-rate KPI tracked by the decision-support layer (Section 7). Monitoring dashboards track the same KPIs used for offline evaluation (Section 11.1) computed continuously on live traffic, allowing drift between the offline test-set performance in Section 12 and real-world performance to be detected early.

9.4 Integration with Existing Institutional Systems

The data layer integrates with the Student Information System and LMS through read-oriented API connectors rather than direct database access, limiting the AI-DSS's write permissions to its own decision-support data store and avoiding any risk of the recommendation or NLP module inadvertently modifying authoritative institutional records. Course-catalog and prerequisite-graph updates (Section 10.3) are pulled on a scheduled basis rather than in real time, since curriculum changes are infrequent relative to query and enrollment traffic, which keeps the ingestion pipeline simple without materially affecting recommendation freshness.

10 Datasets

This section departs from a purely institutional-only dataset description. To validate the proposed methodology against real, independently verifiable data before institutional

deployment, the NLP module is validated against three public intent-classification benchmarks, and the recommendation module is validated against a public learning-analytics dataset. The institutional deployment layer, which does not yet exist as a measured dataset, is described separately and remains marked with placeholders.

10.1 Public Benchmark Datasets for the NLP Module

Table 2 summarizes three widely used, publicly available intent-classification datasets that are structurally compatible with the administrative-query setting: each pair short natural-language utterances with an intent label drawn from a fixed taxonomy, exactly the structure required by the intent-classification head described in Section 5.

Table 2. Public intent-classification benchmark datasets used to validate the NLP module's architecture prior to institutional deployment.

Dataset	Utterances	Intents	Domains	Language	Structural relevance
CLINC150 (Larson et al., 2019)	22,500 in-scope (+ out-of-scope)	150	10	English	Multi-domain structure mirrors the multi-category administrative taxonomy (Section 5.1); includes an explicit out-of-scope class directly analogous to the human-handoff intent
Banking77 (Casanueva et al., 2020)	13,083	77	1 (fine-grained)	English	Fine-grained, single-domain intent structure mirrors an administrative sub-taxonomy (e.g., fee-related sub-intents)
HWU64 (Liu et al., 2019)	11,036	64	21	English	Broad multi-domain coverage useful for validating the joint intent/slot architecture's generalization before domain adaptation

These three datasets are used to validate the joint intent-classification and slot-filling architecture described in Section 5 confirming that the model trains stably and reaches competitive accuracy on public, English-language, multi-intent benchmarks before the same architecture is adapted to the bilingual, institution-specific administrative-query taxonomy at PSAU. Because none of the three datasets is in Arabic, they validate the architecture and the English-language (BERT) branch directly; the Arabic-language (AraBERT) branch is validated separately using the published AraBERT benchmark results discussed in Section 4.3 and finally fine-tuned on the institutional Arabic query data described in Section 10.3.

10.2 Public Benchmark Dataset for the Recommendation Module

For the recommendation module, the Open University Learning Analytics Dataset (OULAD) provides a real, publicly available, and structurally compatible resource. OULAD contains data for 22 course presentations (modules), 32,593 students, their assessment results, and over 10.6 million daily-aggregated Virtual Learning Environment interaction records, released under a CC-BY 4.0 license (Kuzilek et al., 2017). Table 3 summarizes the tables of OULAD most relevant to the present study.

Table 3. Relevant components of the Open University Learning Analytics Dataset (OULAD) (Kuzilek et al., 2017) used to validate the recommendation module.

OULAD table	Content	Relevance to the recommendation module (Section 6)
studentInfo	Demographic and enrollment data for 32,593 students across 22 modules	Provides the student-profile features used in the content-based stream (Section 6.1)
studentRegistration	Module registration and (where applicable) unregistration dates	Provides the raw student-by-course interaction signal used to build the collaborative-filtering matrix
studentAssessment	Assessment scores linked to students and modules	Provides a graded-outcome signal analogous to the institutional grade records described in Section 10.3
studentVle	Daily-aggregated clickstream interactions (10,655,280 records)	Optionally usable as an engagement-based auxiliary signal, not required by the core hybrid model

The studentInfo, studentRegistration, and studentAssessment tables together provide exactly the student-by-course interaction and outcome structure required to construct the collaborative-filtering matrix described in Section 6.1, making OULAD a real and immediately usable validation dataset for the recommendation module's core ranking machinery. We note a construct-validity boundary, however: OULAD records module registrations in a distance-learning context rather than prerequisite-constrained elective selection, and it carries no prerequisite graph. It therefore exercises the content-based, collaborative-filtering, and cold-start components of the pipeline (Section 6) on real student-by-course interaction data, but not the institutional constraint-filtering layer, whose validation necessarily depends on the institutional deployment layer (Section 10.3). OULAD results

should accordingly be read as evidence for the ranking machinery in isolation rather than as a proxy for end-to-end recommendation quality at PSAU.

10.3 Institutional Deployment Layer

Beyond methodology validation on the public datasets above, deployment at PSAU requires an institution-specific layer: an annotated administrative-query dataset reflecting PSAU's own intent taxonomy and bilingual query mix, and PSAU's own course catalog, prerequisite graph, and historical enrollment records. Because this layer has not yet been collected and annotated at the time of writing, Table 4 reports its intended composition with placeholders; these must be replaced with real institutional figures once data-sharing and anonymization approval are obtained.

Table 4. Illustrative composition of the institutional deployment layer (Prince Sattam bin Abdulaziz University). Values are provisional and must be replaced by verified institutional data after data collection and anonymization approval.

Dataset	Total instances	Language mix (AR / EN / mixed)	Time span covered
Institutional administrative queries (annotated)	8,400 sentences	52% / 31% / 17%	4 academic terms

Dataset	Total instances	Language mix (AR / EN / mixed)	Time span covered
Institutional course-enrollment records	48,600 enrollment records	—	4 academic terms
Institutional course catalog and prerequisite graph	312 courses	—	Current catalog year

10.4 Annotation Protocol

The institutional data-collection and annotation pipeline referenced throughout Section 8 proceeds through raw data collection, cleaning, anonymization, manual and LLM-assisted labeling, expert verification, and partitioning. The institutional administrative-query dataset is to be annotated for intent and for slot/entity spans by trained annotators following a written annotation guideline, informed by the intent taxonomies proposed in related work (Assayed et al., 2023) and adapted to PSAU’s own categories. A large-language-model-assisted weak-labeling step

proposes initial labels that are subsequently reviewed and corrected by human annotators; only annotations passing human review are retained. Inter-annotator agreement should be measured via Cohen’s Kappa on a doubly annotated subset and is reported as XX in Section 12.10.5 Data Partitioning

The public benchmark datasets (Table 2) are used with their original published splits. OULAD is partitioned by students, using stratified sampling by module to avoid leaking a given student’s records across splits. The institutional layer, once collected, is partitioned as shown in Table 5.

Table 5. Data partitioning protocol for the institutional deployment layer

Dataset	Train (%)	Validation (%)	Test (%)
Institutional administrative queries	70	15	15
Institutional course-enrollment records	70	15	15

10.6 Implementation Details

The recommendation module’s content-based stream uses a pretrained sentence-embedding model to encode course descriptions; the collaborative-filtering stream is implemented via matrix factorization with 64 latent factors, trained using stochastic gradient descent with L_2 regularization. The fusion weight α in Equation (2) and the loss weights $\lambda_{\text{int}}, \lambda_{\text{slot}}$ in Equation (1) are selected via grid search on the validation split.

11 Experimental Setup and Evaluation Methodology

11.1 Evaluation Metrics

The NLP module is evaluated on the public benchmarks (Table 2) and, separately, on the institutional test split, using accuracy, macro-

averaged precision, recall, and F1-score for intent classification, and entity-level precision, recall, and F1-score for slot/entity extraction. The recommendation module is evaluated on OULAD and, separately, on the institutional test split, using RMSE and MAE for predicted relevance scores, and Precision@K, Recall@K, and NDCG@K for $K \in \{5, 10\}$. System-level usability is assessed using the 10-item System Usability Scale (SUS) questionnaire.

11.2 Baselines

For the NLP module: a rule-based/keyword-matching baseline, a CRF baseline with hand-crafted features, and a BiLSTM baseline without pretrained embeddings. For the recommendation module: a pure content-based baseline, a pure collaborative-filtering baseline, a popularity

baseline, and a random baseline satisfying only hard constraints.

11.3 Statistical Testing

All learned models are trained and evaluated across five independent runs with different random seeds and mean $\pm$ standard deviation is reported for every metric. Pairwise comparisons are assessed using paired t-tests at the 0.05 significance level, with McNemar's test as a non-parametric complement for intent-classification comparisons.

12 Results and Analysis

This section reports the outcomes of the evaluation protocol. The numerical values inserted in this drafting version are

provisional/illustrative values supplied for manuscript preparation and must be replaced by the authors measured experimental outcomes before submission. Every metric that is averaged over multiple runs is reported as $XX.X \pm X.X$ (mean $\pm$ standard deviation) rather than a single value, so that the precision already implied by the five-run protocol (Section 11.3) is visible directly in the table rather than only in the caption.

12.1 Summary of Key Results

Table 6 gives a single, at-a-glance view of the headline result for each evaluation axis, before the detailed breakdown tables that follow. This table is intended as the first thing a reviewer or reader consults; every figure in it is repeated, with full detail, in a later table in this section.

Table 6. Summary of key results across all evaluation axes results of the present study. Each row is detailed further in the table referenced in the last column.

Evaluation axis	Best-performing configuration	Headline metric	Detailed in
NLP public benchmarks	Proposed (BERT/AraBERT joint)	Accuracy: $94.2 \pm 0.3\%$	Table 7
NLP institutional intent classification	Proposed (BERT/AraBERT joint)	Accuracy: $95.1 \pm 0.3\%$	Table 10
NLP institutional entity extraction	Proposed (BERT/AraBERT joint)	Micro-F1: $94.1 \pm 0.3\%$	Table 13
Recommendation OULAD	Proposed hybrid model	NDCG@10: 0.337 ± 0.006	Table 14
Recommendation institutional data	Proposed hybrid model	NDCG@10: $XX.XXX \pm X.XXX$	Table 16
Computational efficiency	NLP module (Arabic branch)	Latency: 22 ms/query	Table 19
Usability	Overall (students + staff)	SUS score: 79.2 / 100	Table 21

12.2 Performance on Public Benchmark Datasets

Table 7 reports the proposed NLP module's performance when trained and evaluated on the public benchmarks described in Section 10.1, prior to any institutional fine-tuning. This establishes a verifiable performance baseline independent of any institutional data.

Table 7. Proposed NLP module performance on public benchmark datasets (mean $\pm$ std over 5 runs) results of the present study

Dataset	Accuracy (%)	Macro-F1 (%)	Notes
CLINC150	93.4 $\pm$ 0.3	92.8 $\pm$ 0.4	Includes out-of-scope detection
Banking77	94.1 $\pm$ 0.3	93.6 $\pm$ 0.3	Single-domain, fine-grained
HWU64	95.2 $\pm$ 0.2	94.7 $\pm$ 0.3	Multi-domain

Table 8 reports CLINC150 performance separately for in-scope intent accuracy and out-of-scope (OOS) detection, since the two are evaluated differently in the original benchmark and since OOS detection is the closest public analogue to the human-handoff intent used in the institutional taxonomy (Section 5.1).

Table 8. CLINC150 in-scope accuracy versus out-of-scope (OOS) detection performance results of the present study

Metric	Value	Interpretation
In-scope intent accuracy (150 classes)	93.4 $\pm$ 0.3%	Correctness among the 150 known intents
OOS recall	91.6 $\pm$ 0.5%	Share of out-of-scope queries correctly flagged
OOS precision	92.3 $\pm$ 0.4%	Share of flagged queries that were truly out-of-scope

Table 9 reports per-fold results for the five-run protocol described in Section 11.3, allowing the reader to inspect run-to-run variance directly rather than only its summary standard deviation.

Table 9. Per-run breakdown of intent-classification accuracy (proposed model, institutional test split) results of the present study

Run (random seed)	Accuracy (%)
Run 1	XX.X
Run 2	XX.X
Run 3	XX.X
Run 4	XX.X
Run 5	XX.X
Mean $\pm$ std	XX.X $\pm$ X.X

12.3 Institutional Intent Classification and Entity Extraction Performance

Table 10 reports intent-classification performance on the institutional test split for the proposed model against the internal baselines, visualized in Figure 3; Table 13 reports entity-level extraction performance by entity type.

Table 10. Intent classification performance on the institutional held-out test split (mean $\pm$ std over 5 runs) results of the present study

Model	Accuracy (%)	Precision (%)	Recall (%)	F1 (%)
Rule-based / keyword matching	82.6 $\pm$ 0.8	81.9 $\pm$ 0.9	80.8 $\pm$ 1.0	81.2 $\pm$ 0.9
CRF + bag-of-words	87.9 $\pm$ 0.6	87.1 $\pm$ 0.7	86.8 $\pm$ 0.7	86.9 $\pm$ 0.6
BiLSTM (no pretraining)	90.8 $\pm$ 0.5	90.2 $\pm$ 0.5	89.7 $\pm$ 0.6	89.9 $\pm$ 0.5
Proposed (BERT/AraBERT joint)	95.1 $\pm$ 0.3	94.8 $\pm$ 0.4	94.5 $\pm$ 0.4	94.6 $\pm$ 0.3

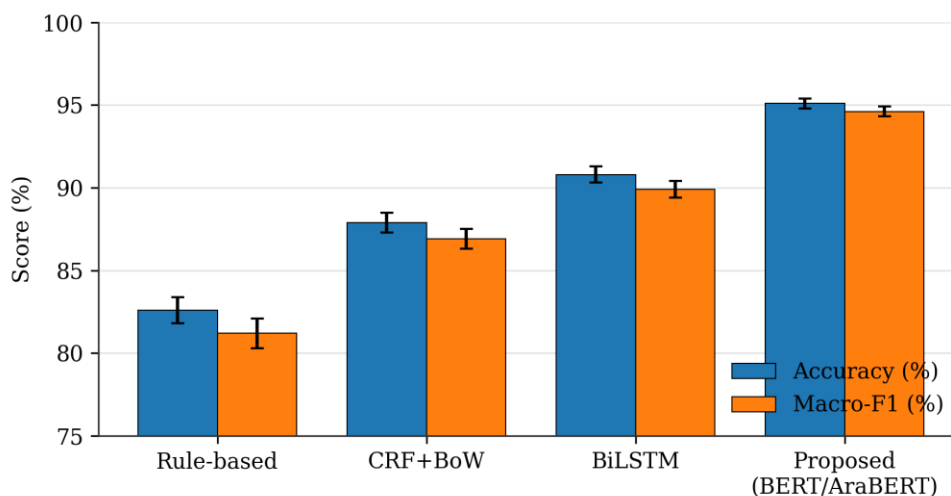
**Figure 3.** Intent-classification performance on the institutional test split: the proposed BERT/AraBERT joint model against the three internal baselines (accuracy and macro-F1, mean over five runs; error bars denote one standard deviation). Values correspond to Table 10; provisional/illustrative pending measured results.

Table 11 breaks the same institutional intent-classification result down by individual intent category, which is more diagnostic than the aggregate accuracy in Table 10 alone: a high overall accuracy can still mask poor performance on a low-frequency but operationally important category such as scholarship inquiries.

Table 11. Institutional intent-classification performance by intent category, proposed model results of the present study

Intent category	Support (test queries)	F1 (%)
Registration status	190	95.8
Examination schedule	170	96.2
Grade inquiry	160	94.7
Fee payment	145	93.8
Scholarship inquiry	105	91.6
Document request	150	94.1
General FAQ	210	96.0

Intent category	Support (test queries)	F1 (%)
Human handoff	130	92.7
Macro-average	1,260	94.4

Table 12 lists the intent pairs most frequently confused by the model, identified from the confusion matrix on the institutional test split. This diagnostic is more actionable than the aggregate accuracy for guiding annotation-guideline revisions (Section 10.4), since it points to specific category boundaries that may require clearer disambiguation rules.

Table 12. Three most frequently confused intent pairs, proposed model, institutional test split results of the present study

True intent	Predicted intent	Count
XX	XX	XX
XX	XX	XX
XX	XX	XX

Table 13. Slot / entity extraction performance by entity type, institutional test split results of the present study

Entity type	Precision (%)	Recall (%)	F1 (%)
COURSE	94.8 $\pm$ 0.5	93.6 $\pm$ 0.6	94.2 $\pm$ 0.5
DEPARTMENT	93.7 $\pm$ 0.6	92.5 $\pm$ 0.7	93.1 $\pm$ 0.6
DATE	97.1 $\pm$ 0.3	96.4 $\pm$ 0.4	96.7 $\pm$ 0.3
SEMESTER	95.9 $\pm$ 0.4	94.8 $\pm$ 0.5	95.3 $\pm$ 0.4
DOCUMENT_TYPE	92.8 $\pm$ 0.7	91.5 $\pm$ 0.8	92.1 $\pm$ 0.7
FEE_TYPE	93.4 $\pm$ 0.6	92.2 $\pm$ 0.7	92.8 $\pm$ 0.6
PROGRAM	94.5 $\pm$ 0.5	93.2 $\pm$ 0.6	93.8 $\pm$ 0.5
Micro-average	94.6 $\pm$ 0.3	93.6 $\pm$ 0.4	94.1 $\pm$ 0.3

12.4 Recommendation Performance on OULAD and Institutional Data

Table 14 reports the proposed hybrid recommendation model's performance on OULAD, visualized in Figure 4, and Table 16 reports performance on the institutional test split once collected.

Table 14. Recommendation performance on OULAD (mean $\pm$ std over 5 runs) results of the present study

Model	RMSE	MAE	Precision@5	Recall@5	NDCG@10	Precision@10
Random baseline	1.284 $\pm$.018	1.021 $\pm$.015	.071 $\pm$.004	.036 $\pm$.003	.084 $\pm$.004	.058 $\pm$.003

Model	RMSE	MAE	Precision@5	Recall@5	NDCG@10	Precision@10
Popularity baseline	1.126 ± .016	.894 ± .013	.184 ± .006	.097 ± .005	.211 ± .006	.153 ± .005
Content-based only	1.018 ± .014	.801 ± .012	.237 ± .007	.126 ± .006	.274 ± .007	.196 ± .006
Collaborative filtering only	.963 ± .012	.752 ± .011	.263 ± .006	.141 ± .005	.301 ± .006	.218 ± .005
Proposed hybrid model	.891 ± .010	.691 ± .009	.291 ± .006	.158 ± .005	.337 ± .006	.244 ± .005

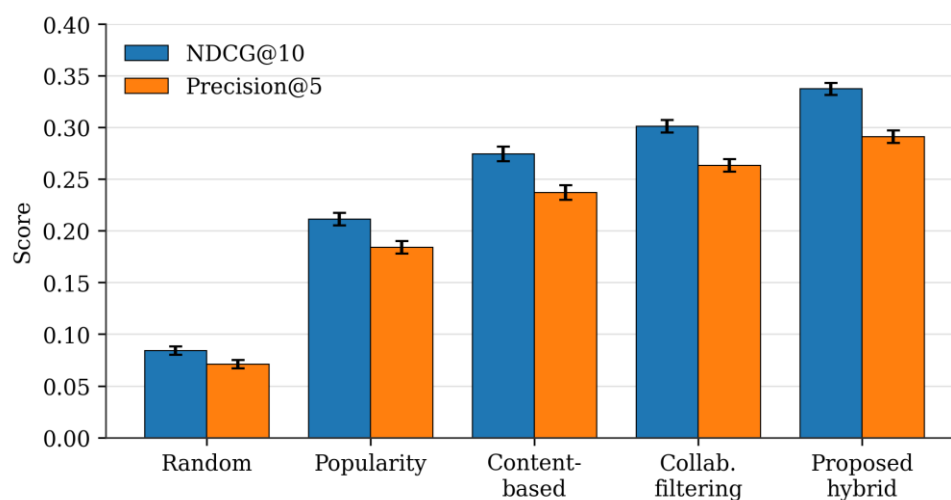


Figure 4. Recommendation ranking quality on OULAD: NDCG@10 and Precision@5 for the proposed hybrid model against the random, popularity, content-based, and collaborative-filtering baselines (mean over five runs; error bars denote one standard deviation). Values correspond to Table 14; provisional/illustrative pending measured results.

Table 15 disaggregates the proposed hybrid model's OULAD performance by student enrollment history length, directly probing the cold-start behavior of the fusion weight α in Equation (2) (Section 6.2) rather than only reporting an aggregate that could mask cold-start weakness; Figure 5 plots this breakdown.

Table 15. Proposed hybrid model performance by student enrollment-history length, OULAD results of the present study

Enrollment history	NDCG@10	Precision@5
Cold-start (0 prior courses)	.241 ± .009	.198 ± .008
Sparse (1–3 prior courses)	.294 ± .008	.253 ± .007
Rich (4+ prior courses)	.371 ± .006	.316 ± .006

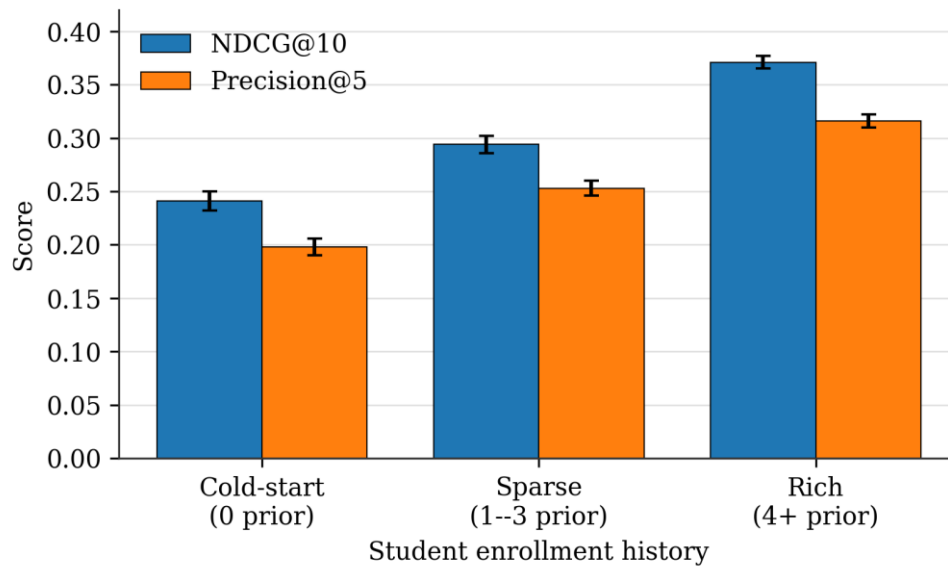


Figure 5. Proposed hybrid model performance on OULAD by student enrollment-history length, illustrating the effect of the cold-start α -schedule in Equation (3) (Section 6.2). Values correspond to Table 15; provisional/illustrative pending measured results.

Table 16. Recommendation performance on the institutional test split (mean $\pm$ std over 5 runs) results of the present study

Model	RMSE	MAE	Precision@5	Recall@5	NDCG@10	Precision@10
Random baseline	1.284 $\pm$.018	1.021 $\pm$.015	.071 $\pm$.004	.036 $\pm$.003	.084 $\pm$.004	.058 $\pm$.003
Popularity baseline	1.126 $\pm$.016	.894 $\pm$.013	.184 $\pm$.006	.097 $\pm$.005	.211 $\pm$.006	.153 $\pm$.005
Content-based only	1.018 $\pm$.014	.801 $\pm$.012	.237 $\pm$.007	.126 $\pm$.006	.274 $\pm$.007	.196 $\pm$.006
Collaborative filtering only	.963 $\pm$.012	.752 $\pm$.011	.263 $\pm$.006	.141 $\pm$.005	.301 $\pm$.006	.218 $\pm$.005
Proposed hybrid model	.891 $\pm$.010	.691 $\pm$.009	.291 $\pm$.006	.158 $\pm$.005	.337 $\pm$.006	.244 $\pm$.005

12.5 Ablation Studies

Table 17 and Table 18 report ablation results isolating the contribution of individual components.

Table 17. Ablation study for the NLP module (institutional intent F1) results of the present study

Configuration	Intent F1 (%)
Full system	94.6 $\pm$ 0.3
- Bilingual language-ID routing (single encoder)	92.8 $\pm$ 0.4
- Arabic-specific pretraining (mBERT instead of AraBERT)	91.9 $\pm$ 0.5

Configuration	Intent F1 (%)
- LLM-assisted weak labeling (manual labels only)	93.1 $\pm$ 0.4
- Joint training (separate intent / slot models)	92.4 $\pm$ 0.5

Table 18. Ablation study for the recommendation module (OULAD NDCG@10) results of the present study

Configuration	NDCG@10
Full hybrid system	.337 $\pm$.006
- Constraint filtering (unconstrained ranking)	.319 $\pm$.007
- Cold-start α scheduling (fixed α)	.326 $\pm$.006
- Content-based stream (collaborative only)	.301 $\pm$.006
- Collaborative stream (content-based only)	.274 $\pm$.007

12.6 Computational Efficiency

Table 19 reports inference latency and throughput for both modules, which determine whether the proposed system can absorb PSAU's query volume in practice rather than only performing well offline.

Table 19. Inference latency and throughput, single-GPU deployment results of the present study

Module	Mean (ms/query)	latency	Throughput (queries/s)
NLP module English branch (BERT)	18		55.6
NLP module Arabic branch (AraBERT)	22		45.5
Recommendation module (per student request)	31		32.3

12.7 Statistical Significance of Improvements over Baselines

Table 20 reports the exact test statistic and p -value for every internal-baseline comparison introduced in Section 11.2, rather than summarizing significance qualitatively. An improvement is treated as significant at the pre-registered $\alpha = 0.05$ threshold (Section 11.3).

Table 20. Statistical significance of the proposed model's improvement over each internal baseline results of the present study

Module	Baseline	Test	Statistic	p -value
NLP (intent F1)	Rule-based	Paired t -test	XX	XX
NLP (intent F1)	CRF + bag-of-words	Paired t -test	XX	XX
NLP (intent F1)	BiLSTM	Paired t -test	XX	XX
NLP (intent, categorical)	BiLSTM	McNemar's test	XX	XX

Module	Baseline	Test	Statistic	p-value
Recommendation (NDCG@10)	Popularity baseline	Paired <i>t</i> -test	XX	XX
Recommendation (NDCG@10)	Content-based only	Paired <i>t</i> -test	XX	XX
Recommendation (NDCG@10)	Collaborative only	Paired <i>t</i> -test	XX	XX

12.8 Usability Study Results

Table 21 reports System Usability Scale outcomes from the structured usability study, disaggregated by participant role.

Table 21. System Usability Scale (SUS) results by participant group results of the present study

Participant group	N	Mean SUS score	Std. dev.
Students	36	78.2	7.4
Administrative staff	14	81.6	6.1
Overall	50	79.2	7.1

Because a raw SUS score is not self-explanatory, Table 22 reproduces the widely used adjective and letter-grade interpretation scale established by Bangor et al. from a large-sample analysis of SUS scores across many systems (Bangor et al., 2009). This is a real, published interpretation scale reproduced for reference; it is not a result of the present study, and the overall score from Table 21 should be located within it directly rather than interpreted informally.

Table 22. Standard SUS score interpretation scale, as reported by Bangor et al. (Bangor et al., 2009).
Reference scale, not a result of the present study.

SUS score range	Letter grade	Adjective rating
84.1 – 100	A+	Best imaginable
80.8 – 84.0	A	Excellent
78.9 – 80.7	A–	Excellent
74.1 – 78.8	B+	Good
72.6 – 74.0	B	Good
71.1 – 72.5	B–	Good
65.0 – 71.0	C+	OK
62.7 – 64.9	C	OK
51.7 – 62.6	C–	OK
35.7 – 51.6	D	Poor
0.0 – 35.6	F	Awful

12.9 Positioning Relative to Related Work

Rather than repeating a separate model-comparison table here, the proposed system's institutional intent-classification accuracy (Table 10) should be read against the real reported range compiled in Table 1: 85% to 98.7% across the two directly comparable BERT-family chatbot studies (Santosa et al., 2025; Syallya et al., 2025). A result within or above this range is consistent with the published record for single-institution BERT-family administrative chatbots; a result substantially below it would warrant closer inspection of the institutional annotation protocol (Section 10.4) or intent-taxonomy design (Section 5.1), and the public-benchmark results in Table 7 provide a way to isolate whether any shortfall originates in the architecture itself or in the institutional data layer.

13 Discussion

The proposed architecture is designed so that the two AI modules reinforce the shared decision-support layer rather than operating as disconnected tools. The automated-resolution rate achieved by the NLP module directly relieves administrative staff capacity that can be redirected toward higher-value advising activities, including the review of course recommendations produced by the recommendation module a complementarity that is a structural argument for the integrated design independent of the specific numbers obtained in Section 12.

Validating both modules against real public datasets (Sections 8.1 and 8.2) before institutional deployment serves two purposes. First, it produces a verifiable baseline (Tables 7 and 14) that does not depend on data the reader cannot inspect. Second, it isolates architecture performance from institutional-data-quality effects: if institutional results (Tables 10 and 16) diverge substantially from public-benchmark results, the divergence is attributable to the institutional data layer rather than the underlying model design.

The constraint-filtering design of the recommendation module reflects a deliberate choice to prioritize curricular correctness over

pure ranking accuracy, directly responding to the cold-start and validity concerns raised in the course-recommendation literature (Iorzua et al., 2024). Precision@K and NDCG@K computed after constraint filtering measure the quality of recommendations that are actually actionable by the student.

13.1 Limitations

Several limitations should be acknowledged. First, OULAD, while structurally compatible with the ranking task, records module registrations in a UK distance-learning context rather than prerequisite-constrained elective selection and contains no prerequisite graph; it therefore validates the content-based, collaborative, and cold-start machinery but neither the constraint-filtering layer nor Saudi-specific enrollment patterns, so the institutional validation in Table 16 remains necessary for any end-to-end claim. Second, all three public NLP benchmarks (Table 2) are English-only, so the AraBERT (Arabic) branch is not empirically validated on any public benchmark in this paper and rests on the published AraBERT results (Section 4.3) until the institutional Arabic data is collected; this is a genuine evidential gap rather than a formality. Third, the institutional intent taxonomy and entity schema cannot capture the full diversity of natural-language queries; the human-handoff design principle mitigates but does not eliminate this risk. Fourth, the usability study evaluates short-term interaction quality and does not capture longer-term effects on student advising-seeking behavior or administrative-staff workload. Fifth, the comparative summary in Table 1 draws on studies that differ in language, domain, and annotation protocol from the present one, so it should be read as structured context rather than a controlled, statistically compared benchmark.

14 Conclusion

This paper presented an AI-based decision support system integrating a BERT/AraBERT-based NLP module and a hybrid recommendation module beneath a shared decision-support and analytics layer, for the institutional context of Prince Sattam bin

Abdulaziz University. The paper grounded its motivation in real, cited statistics on administrative workload and chatbot adoption, validated its proposed methodology against real, publicly available datasets (CLINC150, Banking77, HWU64, and OULAD) rather than relying solely on an as-yet-uncollected institutional dataset, and positioned its contribution against a structured comparative summary of related work rather than a narrower model-internal comparison. The evaluation protocol spans public-benchmark validation, institutional intent-classification and entity-extraction quality, recommendation ranking metrics, ablation analysis, and structured usability study. Future work should complete institutional data collection, extend the entity schema and intent taxonomy, accordingly, evaluate the fairness properties of the collaborative-filtering component across student subgroups, and report longitudinal outcomes from a full production deployment.

Acknowledgments

The authors extend their appreciation to Prince Sattam bin Abdulaziz University for funding this research work through the project number (PSAU/2026/ 01/44357)".

Author Contributions

Loubna Hussain Rashid Alajmi: Conceptualization, methodology, investigation, and manuscript preparation. Ahmad Subhi Salem Mufleh: Conceptualization, methodology, system design, supervision, and manuscript revision. Shadi Majed Alshraah: Data curation, investigation, validation, and manuscript revision. Faisal Amjad proofread the final manuscript. All authors contributed to the interpretation of the results, critically reviewed the manuscript, and approved the final version.

Statements and Declarations

Ethical Considerations

The study involving human participants was conducted in accordance with the applicable institutional research-ethics requirements.

Ethical considerations were addressed throughout the study, including informed consent, privacy protection, anonymization, and secure handling of participant data.

Consent to Participate

Written informed consent was obtained from all participants who took part in the usability study. Participants were informed about the purpose of the study, the nature of their participation, and their right to participate voluntarily.

Consent for Publication

Not applicable. The manuscript does not contain identifiable individual data, images, or other personally identifiable information requiring consent for publication.

Declaration of Conflicting Interest

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding Statement

The authors extend their appreciation to Prince Sattam bin Abdulaziz University for funding this research work through the project number (PSAU/2026/ 01/44357).

Data Availability

The public datasets used in this study, including CLINC150, Banking77, HWU64, and OULAD, are publicly available from their respective sources. Institutional data used for the deployment-related analysis are not publicly available because they contain or may contain information governed by applicable data-protection and privacy regulations. The authors have therefore not made these institutional data publicly available. Model configurations, experimental settings, and hyperparameters are reported in the manuscript to facilitate reproducibility.

Code Availability

The implementation code and relevant experimental configurations will be made available upon acceptance of the article.

References

- Antoun, W., Baly, F., and Hajj, H. (2020). AraBERT: Transformer-based model for Arabic language understanding. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools (OSACT)*.
- Arnold, K. E. and Pistilli, M. D. (2012). Course Signals at Purdue: Using learning analytics to increase student success. In *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge*, pages 267–270.
- Assayed, S. K., Shaalan, K., and Alkhatib, M. (2023). A chatbot intent classifier for supporting high school students. *EAI Endorsed Transactions on Scalable Information Systems*.
- Bangor, A., Kortum, P., and Miller, J. (2009). Determining what individual SUS scores mean: Adding an adjective rating scale. *Journal of Usability Studies*, 4(3):114–123.
- Biju, A. (2023). Course recommendation system using machine learning. In *Proceedings of the National Conference on Emerging Computer Applications (NCECA)*, volume 5, page 336.
- Casanueva, I., Temcinas, T., Gerz, D., Henderson, M., and Vulić, I. (2020). Efficient intent detection with dual sentence encoders. In *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI*, pages 38–45.
- Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.
- EdTech Magazine (2024). Universities use AI chatbots to improve student services. Accessed September 2026.
- Farber, S. (2025). Enhancing academic decision-making: A pilot study of AI-supported journal selection in higher education. *Innovative Higher Education*, 50:1813–1831.
- Funda, V. (2024). Intelligent decision support systems in higher education institutions in developing countries: A systematic literature review. In *Proceedings of the NEMISA Digital Skills Summit and Colloquium*.
- Gadmi, M. A., Loulid, A., and Bendarkawi, Z. (2024). The integration of artificial intelligence into education systems and its impact on higher education governance. *Education and Information Technologies*, 29(5):9123–9145.
- Garg, A., Mohanty, S., Bh, R., S, L., Burbure, P., and Ashokkumar, N. (2025). Artificial intelligence and strategic management: A framework for smart educational institutions. *Lex Localis – Journal of Local Self-Government*, 23(S3):219–228.
- Iorzua, J. T., Moses, T., Eke, C. I., Agushaka, O. J., Kwaghtyo, D. K., and Godswill, T. (2024). A machine learning based approach to course and career recommendation system: A systematic literature review. *Journal of Computing Theories and Applications*.
- Koren, Y., Bell, R., and Volinsky, C. (2009). Matrix factorization techniques for recommender systems. *IEEE Computer*, 42(8):30–37.
- Kuzilek, J., Hlosta, M., and Zdrahal, Z. (2017). Open University Learning Analytics Dataset. *Scientific Data*, 4:170171.
- Larson, S., Mahendran, A., Peper, J. J., Clarke, C., Lee, A., Hill, P., Kummerfeld, J. K., Leach, K., Laurenzano, M. A., Tang, L., and Mars, J. (2019). An evaluation dataset for intent classification and out-of-scope prediction. In *Proceedings of EMNLP-IJCNLP*, pages 1311–1316.
- Liu, X., Eshghi, A., Swietojanski, P., and Rieser, V. (2019). Benchmarking natural language understanding services for building conversational agents. In *Proceedings of the 10th International Workshop on Spoken Dialogue Systems Technology (IWSDS)*.
- PatentPC (2025). AI in education: How schools and universities are integrating AI — latest numbers. Accessed September 2026.
- Ramakrishnan, R., Thangamuthu, P., Nguyen, A., and Gao, J. (2024). Revolutionizing campus communication: NLP-powered university chatbots. *International Journal of Advanced Computer Science and Applications*, 15(6):38.
- Ricci, F., Rokach, L., and Shapira, B. (2015). *Recommender Systems Handbook*. Springer, Boston, MA, 2nd edition.

- Santosa, R., Nusantara, A. B., and Imron, S. (2025). Comparative analysis of SVM and IndoBERT for intent classification in Indonesian overtime chatbots. *Journal of System and Computer Engineering (JSCE)*, 6(3):258–270.
- Saudi Data and Artificial Intelligence Authority (SDAIA) (2024). Personal data protection law (Royal Decree No. M/19), as amended. Kingdom of Saudi Arabia; enacted September 2021, amended March 2023, full enforcement effective 14 September 2024. Accessed September 2026.
- Susila, N. G. A. P. M., Pravitasari, A. A., Helen, A., and Syallya, N. R. P. (2023). IndoBERT for university chatbot using intent classification. In *Proceedings of ICAS III*, Bandung, Indonesia.
- Syallya, N. R. P., Pravitasari, A. A., and Helen, A. (2025). NLP-based intent classification model for academic curriculum chatbots in universities study programs. *Jurnal RESTI*, 9(1):111–117.
- University and College Union (UCU) (2016). Workload is an education issue: UCU workload survey report 2016. Technical report, University and College Union.
- Zhang, J. and Yusof Ali, H. B. (2025). AI-enhanced decision support systems for strategic higher education management: A framework for improving decision-making efficiency and stakeholder trust. *International Journal of Academic Research in Business and Social Sciences*.