

IJEAST

INTERNATIONAL JOURNAL
OF ENGINEERING APPLIED SCIENCE
AND TECHNOLOGY

VOLUME : 11 ISSUE : 07 Print / Issue Publication Date: November 2026

ISSN : 2455-2143

DOI : 10.33564/IJEAST.2026.v11i07.004

Indexed In

Google SCHOLAR · OPEN ACCESS · doi · Crossref

WWW.IJEAST.COM

editor@ijeast.com

International Journal of Engineering Applied Sciences and Technology, 2026
Vol. 11, Issue 07, ISSN No. 2455-2143, Pages 23-31
Published Online November 2026 in IJEAST (http://www.ijeast.com)

MITRA: EVIDENCE-AWARE MULTILINGUAL CONVERSATIONAL AI FOR INSTITUTIONAL INFORMATION ASSISTANCE

Shraddha Ukirade, Anzalna Tamboli, Sayali Kakade, Dnyaneshwari Vyavahare
Department of Computer Science and Engineering,
MIT College of Railway Engineering & Research,
Barshi, Maharashtra, India

Prof. Shobhana Raichurkar
Assistant Professor, Department of Computer Science and Engineering,
MIT College of Railway Engineering & Research,
Barshi, Maharashtra, India

Abstract— **Institutional helpdesks cannot afford loose fee answers. When a system drops an admission route, quota, or fiscal cycle from a fee quote, parents budget with the wrong amount. This paper describes MITRA, an assistant built for MIT College of Railway Engineering & Research (MITCORER), Barshi, that serves English, Marathi and Hindi through maintained records, deterministic rendering of critical fee and intake facts, scope-aware clarification, referral to human staff, and an optional English-only generation path gated by bounded verification. The system was evaluated on a 90-scenario trilingual benchmark with generation disabled, using an automated harness and dual faculty scoring (Cohen's kappa = 0.96 before the Department Head settled two discordant cases). Baseline correctness on the 60 answerable cases was 60.0%, traced mainly to missing Devanagari facility terms. After the normalizer was extended and two string-matching scorer rules were corrected, correctness reached 86.7% (52/60) with no false refusals, while appropriate abstention fell from 71.4% to 66.7% and the false-answer rate on unanswerable cases rose from 28.6% to 33.3%. Because these changes were made after observing errors on the same scenarios, the result shows regression closure on a development set and does not show held-out accuracy. No incorrect fee value appeared in 15 fee cases on the deterministic path, and an audit of the ten answered-but-not-correct cases found no fabricated figures. Evaluation on an independently authored held-out set, load testing, and an adversarial prompt-injection audit remain future work.**

Keywords— **University admissions, institutional question answering, multilingual assistance, structured knowledge, factual verification, retrieval evaluation.**

I. INTRODUCTION

Admissions information sits in separate circulars, intake charts, and statutory fee declarations. Short student questions often leave out the qualifier that decides the answer: the same category label can carry different charges for first-year and direct second-year entry, and hostel, deposit and mess figures mislead when attached to the wrong label. General-purpose language models add a further risk, since they can produce fluent statements that no source supports.

MITRA was built as the AI assistant of MITCORER to make institutional information easier to reach. It has a public chat widget and a separate authenticated workflow for knowledge updates. This paper describes its architecture, including document versioning, multilingual query processing, and verification of critical values, and it is organized around three research questions. RQ1 (retrieval coverage) asks whether BM25 lexical retrieval with targeted Devanagari query normalization can reach Recall@3 of at least 0.95 across English, Marathi and Hindi inquiries without heavy translation models. RQ2 (critical-fact fidelity) asks whether deterministic template rendering over verified records removes numerical discrepancies on multi-attribute fee and intake questions, with 100% attribute fidelity as the target. RQ3 (multilingual response control) asks how well a multi-stage hierarchy of answerability checks, targeted clarification, safe abstention and staff referral controls end-to-end behavior in all three languages.

The contribution is a documented architecture, a 90-scenario trilingual benchmark, and development-stage results, all drawn from 51 verified administrative seed records of the home campus and run on an internal test harness. The paper does not offer a general remedy for language-model hallucination. It shows how strict data grounding behaves in one operational college setting. The evidence supports

International Journal of Engineering Applied Sciences and Technology, 2026
Vol. 11, Issue 07, ISSN No. 2455-2143, Pages 23-31
Published Online November 2026 in IJEAST (http://www.ijeast.com)

implementation feasibility and does not support a controlled causal effect. Staff workload and applicant decision quality were not evaluated.

II. RELATED WORK

Retrieval-augmented generation (RAG) conditions model output on retrieved evidence [1]. MITRA confines generative modeling to an optional, English-only path. BM25 [2], [3] is an established lexical baseline, and BEIR shows that it holds up in zero-shot retrieval [4]. Here, raw and normalized BM25 are compared as a retrieval diagnostic that is separate from end-to-end generation.

Conversational admission systems are documented in prior work [5]–[8]. Those studies report user-level conversational evidence. MITRA concentrates instead on the engineering mechanisms that bind multi-attribute fee structures, query scope, and deterministic fail-safes into a maintained design. To the best of our knowledge, none of the cited studies report scope-preserving deterministic fee rendering with trilingual error auditing across regional Indian languages. External systems were not evaluated on MITRA's corpus, so Table I is an architectural comparison and does not rank the systems.

Indian-language resources [9] and cross-lingual QA benchmarks [10], [11] show that question answering across typologically different languages is hard, which is the setting for the Marathi and Hindi queries here. Work on assertion-level RAG evaluation [12], structural faithfulness [13], and hallucination surveys [14] motivates deterministic rendering of critical facts, source and year metadata on each record, and mutation-style probes of the verifier.

TABLE I. RELATED WORK AND DESIGN IMPLICATIONS

Reference	Approach or contribution	Relevance and limitation
RAG [1]; BM25 [2], [3]; BEIR [4]	RAG framework; BM25 lexical baseline; BEIR shows BM25 holds up in zero-shot retrieval.	Basis for optional evidence-conditioned generation; lexical comparison kept separate from RAG.
Admissions RAG [5]–[7]	FAQ-first hybrid, intent filtering, multi-agent, history-aware designs; [5] reports over 6,000 users.	Other institutions and protocols; not rerun here; no ranking inferred.
Chen et al. [8]	GPT-3.5/LlamaIndex chatbot at three universities; 53 high-school students.	Broader user evidence than MITRA's fixed 90-case set; no advantage claimed.
Indian-language QA [9]–[11]	Indian-language resources; TyDi QA and XOR QA show cross-lingual QA is hard.	MITRA uses its own rules and source-bound translations; scores do not transfer.
Faithfulness and evaluation [12]–[14]	RAG assertion checks; table faithfulness; hallucination survey; empirical verification.	Motivate deterministic critical facts, source/year metadata and mutation probes.

Research Gap

The practical gap addressed here is the link between multi-attribute fee and intake facts and an assistant that answers in English, Marathi and Hindi without inventing numbers. This paper does not claim that no comparable system exists. Its contribution is to make the maintained-record design, the answer-control hierarchy, the verifier, and the evaluation, including its failures, explicit enough to support independent testing.

III. STUDY METHOD

Study design and materials. This paper reports a system-design and development-stage evaluation study. The development corpus holds 51 published seed records, kept separate from the live database; each experiment names its snapshot. Records carry structured facts, source links, publication status, revision data and validity metadata. Fee records separate entry route and applicant attributes, accommodation records separate hostel, refundable deposit and annual mess amounts, and programme records supply maintained intake values. Raw documentation (circular PDFs, flat text, scan sheets and structured spreadsheets) is parsed with Poppler and English Tesseract OCR. Each record is tagged with provenance metadata, valid academic sessions and a SHA-256 hash before human sign-off. Scanned Devanagari notices bypass automated optical character reading and are transcribed manually to prevent spelling corruption. All administrative facts were transcribed from approved institutional fee schedules and circulars and reviewed by the institutional review panel (the Principal and the Head of Computer Science and Engineering).

Knowledge maintenance. In the administrative backend, staff updates face schema verification and optimistic revision locks. A record moves from draft through editorial review to published, with eventual archive or retirement, so the public engine queries only authorized publications (Fig. 1). A website change-detection daemon stages detected portal updates in an isolated buffer and does not publish them automatically.

International Journal of Engineering Applied Sciences and Technology, 2026
Vol. 11, Issue 07, ISSN No. 2455-2143, Pages 23-31
Published Online November 2026 in IJEAST (http://www.ijeast.com)

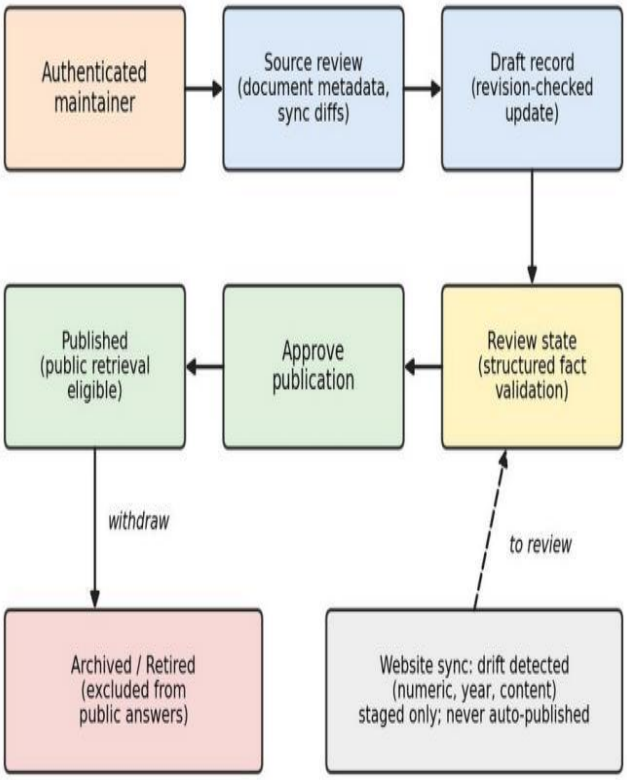


Fig. 1. Maintainer publication workflow: revision-checked updates; public retrieval only after approval; website drift staged, never auto-published.

Evaluation design. Evidence is reported in four layers: regression tests, retrieval diagnostics, verifier probes and a multilingual benchmark. Generation is disabled in all retrieval, routing and benchmark runs, so no live generation model is called.

Retrieval. Four methods (the application rule retriever, raw BM25, normalized BM25, and reciprocal-rank fusion [15]) are run on 51 records and 44 development questions, with ten warmups and 100 timed repetitions each. For question q with developer-authored relevance groups Gq and top-k output Rk(q), group Recall@k(q) = |{g ∈ Gq : g ∩ Rk(q) ≠ ∅}| / |Gq|. MRR@3 follows [3].

Verifier. Fifteen exposed English probes (11 invalid answers and four valid controls) compare a numeric token checker with MITRA's slot/clause checker.

Benchmark. The 90 team-written cases are domain-matched thematic counterparts and not translations, and they include the 30-case pilot. A harness scores five dimensions: factual correctness (required facts as fixed strings), source citation attached, completeness, appropriate abstention (including false refusals), and referral routing. Metrics are defined over the answerable set (Na = 60) and the unanswerable set (Nu = 21): correctness = Ncorrect/Na, false-refusal rate = Nrefused/Na, appropriate abstention = Nabstain/Nu, false-answer rate = 1 − abstention, and rater agreement follows Cohen's kappa. Descriptive Wilson intervals and Fisher exact statistics characterize development variance. Source attachment measures whether an answer included a canonical citation to its retrieved knowledge-base record.

The benchmark followed a four-stage timeline. Stage 1 set the baseline (Run 2: 60.0% correctness, 36/60 answerable cases) by automated harness scoring. Stage 2 was a blind dual-faculty evaluation by two departmental raters (Assistant Professors of CSE with 8+ and 11+ years of experience, native Marathi speakers, not project authors) on three nominal categories (correct handling, refusal/abstention, incorrect handling). Pre-adjudication agreement was Cohen's κ = 0.96, and two discordant cases were settled with the Department Head. Faculty ratings assess baseline response quality and rater reliability only, and they are separate from the harness metrics in Table V. Stage 3 corrected two rigid regex scorer rules that had marked valid Devanagari answers wrong (+2 cases, 38/60, 63.3%). Stage 4 expanded the Devanagari normalizer's facility synonyms and re-evaluated the same 90 scenarios by the automated harness (Run 3: 52/60, 86.7%). Run 3 was scored strictly by the regex harness against verified circular values, and the faculty raters were not reconvened for it.

Evaluation boundary. The results section reports development-stage behavior on developer-authored scenarios. No independently authored held-out set, native-speaker panel from outside the department, latency and cost measurement, load test, or adversarial security audit was part of this study.

IV. SYSTEM ARCHITECTURE

MITRA is a modular monolith. React 19 (Vite 8) provides the browser interface, Node.js the HTTP API, and SQLite the records and cached vectors. Public answering, authentication, knowledge maintenance, retrieval and evidence checks are separate modules, and the public and administration APIs are separate entry points over one store. Optional embedding and generation calls go to the Google Gemini API; deterministic rendering does not depend on them. The backend validates each question, resolves context, and returns an answer, a clarification, an unavailable-information response, or a referral. Fig. 2 shows the component structure and Fig. 3 the decision flow.

International Journal of Engineering Applied Sciences and Technology, 2026
Vol. 11, Issue 07, ISSN No. 2455-2143, Pages 23-31
Published Online November 2026 in IJEAST (http://www.ijeast.com)

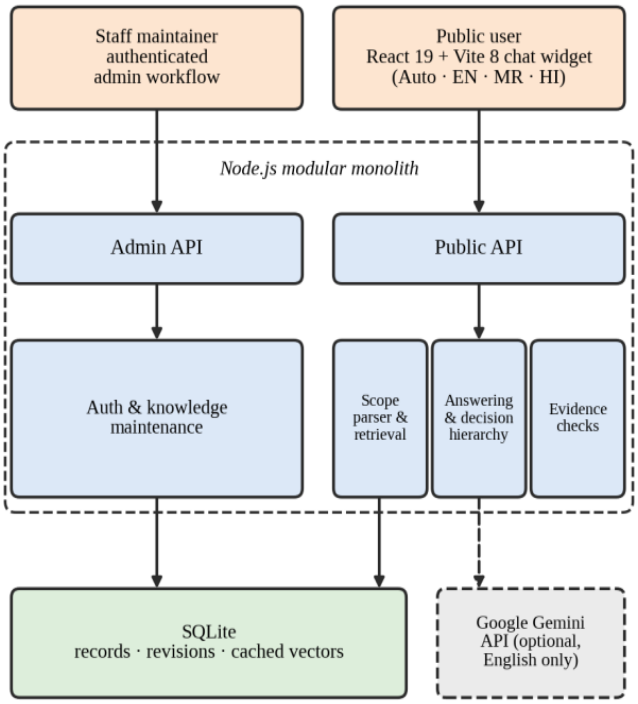


Fig. 2. System architecture: separate public and administration APIs over one SQLite store; Gemini calls are optional and English-only.

TABLE II. COMPONENT RESPONSIBILITIES AND DOCUMENTED STATUS

Component	Main functions	Documented status
Chat widget	Automatic, English, Marathi or Hindi selection; answer, clarification, unavailable-information and referral cards.	Implemented; validated in the local development environment; not on public servers.
Public answering API	Question validation, context resolution, session history with 30-minute idle expiry.	Implemented; evaluated on the 90-scenario benchmark.
Administration API	Authenticated staff updates, schema checks, revision locks, draft-review-publish states.	Implemented; handed over to institutional authorities.
Retrieval and rules	Rule-based parser, weighted lexical/rule scores, synonym boosts, Devanagari normalization.	Evaluated on 44 development questions; no held-out evaluation.
Deterministic renderer	Fee and intake facts taken directly from verified records.	No incorrect value in 15 fee cases on developer-authored questions.
Optional generation	English-only embeddings and drafts through Gemini, gated by the verifier.	Disabled in all reported benchmark runs.
Website change detector	Stages detected portal updates in an isolated buffer.	Implemented; no evaluation reported.

Security and data governance. Conversation history is bound to a random HttpOnly session cookie with a 30-minute idle expiry, and admission approval is never inferred from it. Contact details submitted for staff call-backs are stored in the local SQLite database, and a formal retention policy remains future work. Optional English embedding calls send non-critical query strings to the external Gemini API without user-identifying session metadata. No independent security assessment is reported. HTTPS, monitoring, load testing and adversarial prompt-injection audits are planned before open public operation.

V. ANSWER-CONTROL WORKFLOW AND VERIFICATION

A. Query interpretation

A bounded rule-based parser normalizes English, Marathi, Hindi and transliterated expressions and extracts topic, programme, entry route, academic year, category, gender and service. Academic year and year of entry are kept separate, and negated or conflicting values receive explicit handling. Retrieval combines weighted lexical and rule scores with synonym boosts, and optional Gemini embeddings affect English only. A high score does not establish answerability: the record must also hold a supported value for the requested scope.

B. Six-stage request pipeline

The request pipeline applies six fail-safes. (1) Query strings are sanitized, candidate context is resolved, and only eligible active circulars are kept. (2) Interactive clarification is forced whenever essential parameters are missing, which stops the system from inventing default numbers. (3) Fee and intake figures are pulled directly from verified database tuples, so identical queries give repeatable and transparent output. (4) Advisory neural generation is limited to English queries, while Marathi and Hindi use verified rule templates and checksum-verified translations, with an explicit English fallback when a translation is stale. (5) Any generated draft is treated as untrusted: numeric tokens must be present, unverified admission promises are rejected, and clause checks run. Standard token parsers overlook attribute swaps such as mixing up mess and hostel fees, so the clause-level

International Journal of Engineering Applied Sciences and Technology, 2026
Vol. 11, Issue 07, ISSN No. 2455-2143, Pages 23-31
Published Online November 2026 in IJEAST (http://www.ijeast.com)

validator enforces explicit label-to-value matches. (6) Queries that cannot be resolved, or that concern missing information, go to the human helpdesk or to verified institutional contacts. Full semantic entailment lies outside the demonstrated coverage.

C. Runtime safety policies

Six policies govern the runtime layer. Ambiguous prompts trigger targeted clarification, or a table of every reservation category if only the applicant class is missing. Out-of-scope inquiries get a polite refusal with verified external links. Queries about retired or missing documents get no guessed numbers and a verified administrative contact card. Output filters block any generated draft that contains an unsupported fee or an unauthorized promise about admission, cutoffs or scholarships. Stale regional translations and external API timeouts trigger an audited English fallback. Employment and staff-recruitment inquiries go to the college registrar and not to the student admission desk. Fig. 3 shows how these policies compose, and Table III summarizes the resulting response types. In this paper, "escalation" means presenting verified human contact cards; telephonic handoff and call-back fulfillment were outside the evaluation boundary.

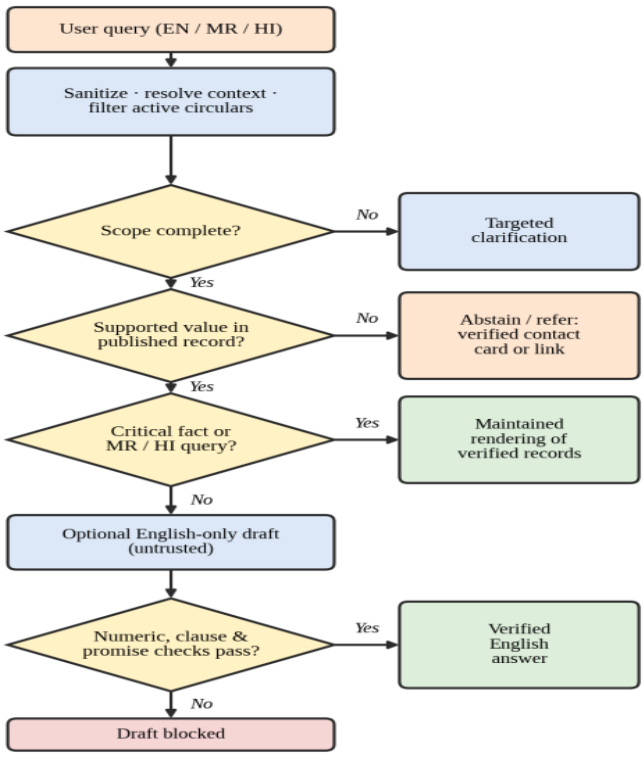


Fig. 3. Decision hierarchy for a user query: clarification, abstention or referral, maintained rendering, and verified optional generation.

TABLE III. RESPONSE TYPES AND REQUIRED INTERPRETATION

Response type	Trigger	Required interpretation
Clarification	Essential parameter missing, such as entry route or category.	No number is shown until the scope is fixed.
Maintained rendering	Verified record holds a value for the requested scope.	Exact stored values; says nothing about topics outside the record.
Unavailable information	Out-of-scope query, or a retired or missing document.	Verified contact card or external link; no guessed figure.
Referral	Query cannot be resolved, or concerns staff recruitment.	Human contact card only; no call-back handled.
Verified generation	English query, generation enabled, draft passes the verifier.	Optional path; not exercised in the reported runs.

VI. INTERFACE AND DEPLOYMENT STATUS

Interface behaviour. Users select Automatic, English, Marathi or Hindi in a floating chat widget that needs no sign-in. A compound query is answered in a single response card: a 2026–27 hostel fee table, followed by a notice that library borrowing policies remain unpublished. The Marathi query "first year fees kiti ahe" returns the full category-wise fee schedule. In the initial build, an unseeded inquiry about staff hiring received a generic referral, before the routing refinement described in Section VII-B.

Deployment status. A client-side asynchronous script loader is designed to mount the widget on the institutional website (mitcorer.edu.in) without CMS modifications. The integration has been validated in the local development environment. Deployment on public institutional servers and open campus-wide operation remain future work. The decoupled design isolates conversational availability from portal updates and gives a graceful offline fallback if backend connectivity is lost. After local server validation and knowledge-base verification, the prototype engine, verified records, and administrative update interface were handed over to institutional authorities. Full production validation, meaning concurrent load testing across peak admission cycles, automated telemetry monitoring, and adversarial prompt-injection audits, is planned before open public operation.

VII. EVALUATION RESULTS

The findings below come from developer-run diagnostics and the 90-scenario benchmark, evaluated with generation disabled. They are grouped by research question.

International Journal of Engineering Applied Sciences and Technology, 2026
Vol. 11, Issue 07, ISSN No. 2455-2143, Pages 23-31
Published Online November 2026 in IJEAST (http://www.ijeast.com)

A. RQ1: retrieval coverage and RQ2: critical-fact fidelity

TABLE IV. DEVELOPMENT RETRIEVAL RESULTS (44 QUESTIONS, 51 SEEDED RECORDS)

Method	Group R@1 / R@3 / MRR@3	Ranking p50 / p95 (ms)
Keyword/rules and concepts	0.8977 / 1.0000 / 0.9470	1.02 / 1.44
BM25 raw	0.6591 / 0.7727 / 0.7121	0.03 / 0.05
BM25 normalized	0.7955 / 0.9659 / 0.8769	0.05 / 0.08
Reciprocal-rank fusion	0.8750 / 0.9886 / 0.9280	1.10 / 1.56

Normalization lifts BM25 Recall@3 from 0.7727 to 0.9659. The application retriever reaches 1.0000 and raises Recall@1 from 0.7955 to 0.8977 compared with normalized BM25. All methods rank within 1.6 ms at p95. Rules and questions share development provenance, so these figures support no superiority claim, and Recall@3 = 1.0 does not certify that each answer is correct.

On the 15 verifier probes, surface token matching rejected 2 of 11 invalid drafts (18.18%), while slot/clause verification rejected 11 of 11 (100%; Wilson 74.1–100%; Fig. 4), with 0/4 false alarms on valid controls. The probes were exposed during development, so this tests the checker's design and not its generalization. In a diagnostic comparison against unconstrained generative drafting, free-form generation produced category and quota substitutions in 4 of 10 sampled complex fee prompts, whereas deterministic template rendering produced no numerical errors (15/15) across the evaluated fee scenarios.

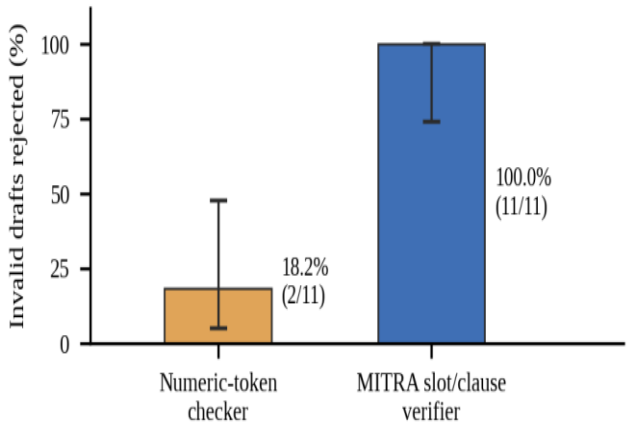


Fig. 4. Invalid drafts rejected by a numeric-token checker and MITRA's slot/clause checker on 15 exposed probes Error bars: 95% Wilson intervals.

B. RQ3: end-to-end multilingual response control

TABLE V. BASELINE (RUN 2) AND POST-FIX (RUN 3) RESULTS ON THE SAME 90 DEVELOPMENT SCENARIOS

Metric	English	Marathi	Hindi	Overall
Correctness, Run 2	85.0% (17/20)	50.0% (10/20)	45.0% (9/20)	60.0% (36/60)
Correctness, Run 3	85.0% (17/20)	85.0% (17/20)	90.0% (18/20)	86.7% (52/60)
Source citation attached, Run 2	100.0% (20/20)	70.0% (14/20)	60.0% (12/20)	76.7% (46/60)
Appropriate abstention, Run 2	57.1% (4/7)	71.4% (5/7)	85.7% (6/7)	71.4% (15/21)
Appropriate abstention, Run 3	-	-	-	66.7% (14/21)
Clarification accuracy, Run 2	33.3% (1/3)	100.0% (3/3)	100.0% (3/3)	77.8% (7/9)
False-answer rate, Run 2	42.9% (3/7)	28.6% (2/7)	14.3% (1/7)	28.6% (6/21)
False-answer rate, Run 3	-	-	-	33.3% (7/21)
False-refusal rate, Run 2	0.0% (0/20)	30.0% (6/20)	40.0% (8/20)	23.3% (14/60)
False-refusal rate, Run 3	-	-	-	0.0% (0/60)
Referral routing, Run 2 (Run 1: 86.7%)	100% (4/4)	100% (5/5)	100% (6/6)	100% (15/15)

Baseline (Run 2). Overall, 36 of 60 answerable cases (60.0%) were correct, one was partial, and 46 carried a source. Fifteen of 21 unanswerable cases drew an appropriate abstention, and seven of nine ambiguous cases a clarification. Marathi and Hindi account for all 14 false refusals, yet English had the lowest abstention (57.1%) and clarification (33.3%) and the highest false-answer rate (42.9%). The 30-case pilot looked better (90.0%), so the larger set exposed failures it missed. Intervals are wide (overall correctness, Wilson 47.4–71.4%), and the post hoc Fisher tests ($p \approx 0.04$ for English versus Marathi, 0.02 for English versus Hindi) are uncorrected and descriptive.

Error audit. Across all 15 fee computations, 3 accommodation inquiries, and 3 placement queries, the deterministic path delivered exact records with zero numerical errors (Wilson lower bound ≈ 80%). Tracing the 14 refusals showed 13 with no retrieval hits and one below the confidence threshold. The Devanagari normalizer lacked facility terms (for example sports, ground, distance)

International Journal of Engineering Applied Sciences and Technology, 2026
Vol. 11, Issue 07, ISSN No. 2455-2143, Pages 23-31
Published Online November 2026 in IJEAST (http://www.ijeast.com)

although the English counterparts were answered, so the records exist and the gap is normalization coverage. An audit of the ten answered-but-not-correct cases confirmed zero hallucinated or fabricated numbers. Two were accurate but incomplete. Two were script-matching scorer misses on Devanagari text (crediting them gives 38/60, 63.3%). Five requested facts absent from the published college circulars: three scholarship cases seeking the ₹50,000 AICTE stipend, and two bus-route cases, since only Solapur and Kurduwadi connectivity is listed. One fell back to English without inline contact numbers. The five absent-fact labels are questionable and are best treated as externally sourced and unseeded. Referral routing was 13/15 in Run 1 (two Marathi and Hindi staff-recruitment misroutes) and 15/15 in Run 2 after one rule correction.

Post-fix development rerun (Run 3). Correctness rose to 86.7% (52/60; Wilson 75.8–93.1%) and false refusals fell from 14 to none. English and pooled Devanagari are no longer detectably different (Fisher p = 1.0). Abstention fell to 66.7% and false answers rose to 33.3% because one query about 2017 cutoffs now matched the general CAP-round record, a mild over-retrieval trade-off. Fig. 5 compares per-language correctness across the two runs.

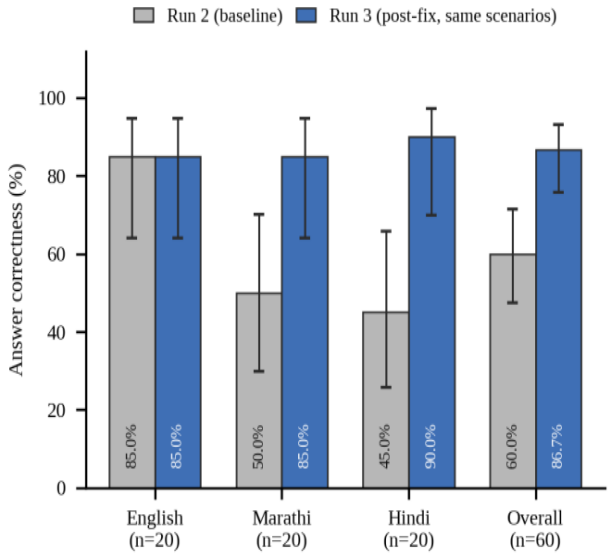


Fig. 5. Answer correctness by language, Run 2 versus Run 3 (data from Table V). Error bars: 95% Wilson intervals. Run 3 was tuned on these same scenarios.

C. Evaluation plan

The next evaluation should separate regression closure from held-out accuracy. Each test should record its question source, knowledge-base snapshot, software version, scorer version, and acceptance criterion. Table VI lists the checks that remain.

TABLE VI. RECOMMENDED HELD-OUT EVALUATION PROTOCOL

Evaluation area	Test method / measure	Evidence to report
Held-out questions	Independently authored questions against a frozen knowledge base.	Question count, author independence, snapshot ID.
Retrieval	Record-level retrieval metrics on the held-out set.	Recall@k and MRR by language.
Raters	Blind raters from outside the department, with native Marathi and Hindi review.	Agreement, adjudication, language coverage.
Generation ablation	Deterministic rendering against full generative RAG; more models, repeated runs, archived transcripts.	Error counts per condition and run variance.
Scoring and coverage	Script-aware scoring; corrected labels; seeded scholarship and transport facts.	Scorer audit and relabeled cases.
Deployed version	Frozen version on fresh student and parent questions collected independently of tuning.	Answer quality and failure types.
Operations and security	HTTPS, monitoring, load testing, call-back retention policy, measured latency and cost.	Percentiles, failures, remediation status.

VIII. DISCUSSION AND LIMITATIONS

The evidence supports the feasibility of separating maintained critical facts from optional generation. The clean fee result characterizes the template renderer on developer-authored questions. It neither tests the generation path nor shows resistance to hallucination.

Several threats limit the results. The team wrote the cases and fixed the router and normalizer after failures on them, which is development leakage. Each language has only seven unanswerable and three ambiguous cases. Scoring is by string match with a partial audit, the raters come from the same department, Hindi native review is thin, and the working tree may differ from the live website.

The Run 3 fix removed false refusals but cost abstention, because a query about 2017 cutoffs matched the general CAP-round record. Wider normalization and stricter abstention pull in opposite directions, and seven

unanswerable cases per language is too few to estimate the balance reliably. The five absent-fact cases also show a coverage gap in the knowledge base, which the scorer labeled as failures.

The evaluation covers the records of one campus. Nothing here shows how the parser rules and Devanagari normalizer terms transfer to another institution, and the scores in Table I's related studies do not transfer to MITRA either.

IX. CONCLUSION AND FUTURE WORK

MITRA is an applied architecture for institutional information assistance that separates verified administrative facts from unconstrained generative modeling. On a developer-authored 90-case benchmark, baseline correctness was 60.0%, with false refusals confined to Marathi and Hindi and traced to missing Devanagari facility-term mappings. One targeted fix and two scorer corrections raised correctness to 86.7% (52/60) with no false refusals, at the cost of lower abstention (71.4% to 66.7%). The fix was tuned on the same cases, so this is regression closure and not held-out accuracy. The deterministic engine delivered exact amounts in all 15 fee scenarios, and an audit of the ten non-correct cases found no fabricated values. These results are initial evidence of implementation feasibility. Production readiness and generalization need independent testing on unobserved student interactions.

The next steps are an independent held-out evaluation with independently authored questions on a frozen knowledge base; a larger controlled ablation of deterministic rendering against full generative RAG; extended Devanagari normalization with corrected labels, seeded scholarship and transport facts, script-aware scoring, Devanagari OCR and stronger semantic verification; and an evaluation of a frozen deployed version on fresh student and parent questions, together with HTTPS, monitoring, load testing and a call-back retention policy.

Data and ethics statement—The benchmark used team-written scenarios and published institutional records. It did not use interaction logs from students or parents. Two faculty members rated responses and the Department Head adjudicated two cases. Any future evaluation on real user questions or stored contact details will need the applicable institutional approval.

ACKNOWLEDGMENT

The authors acknowledge Prof. Shobhana Raichurkar, Assistant Professor, Department of Computer Science and Engineering, MITCORER, Barshi, for project guidance. They thank the two faculty reviewers and the Department Head who rated and adjudicated the benchmark responses, and the college for integrating the MITRA chat widget into its new website and accepting the handover. Declaration on AI use: the authors used an AI assistant for language editing, figure preparation and document formatting, and they are responsible for the content.

X. REFERENCES

[1]. Lewis Patrick, Perez Ethan, Piktus Aleksandra, Petroni Fabio, Karpukhin Vladimir, Goyal Naman, Küttler Heinrich, Lewis Mike, Yih Wen-tau, Rocktäschel Tim, Riedel Sebastian, and Kiela Douwe (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks, in Advances in Neural Information Processing Systems (NeurIPS), vol. 33, (pp. 9459–9474), arXiv:2005.11401.

[2]. Robertson Stephen, and Zaragoza Hugo (2009). The Probabilistic Relevance Framework: BM25 and Beyond, Foundations and Trends in Information Retrieval, vol. 3, no. 4, (pp. 333–389), DOI: 10.1561/1500000019.

[3]. Manning Christopher D., Raghavan Prabhakar, and Schütze Hinrich (2008). Introduction to Information Retrieval, Cambridge University Press, Cambridge, U.K., (pp. 1–496).

[4]. Thakur Nandan, Reimers Nils, Rücklé Andreas, Srivastava Abhishek, and Gurevych Iryna (2021). BEIR: A Heterogeneous Benchmark for Zero-shot Evaluation of Information Retrieval Models, in Proceedings of the Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track, (pp. 1–15).

[5]. Nguyen Long, and Quan Tho (2025). URAG: Implementing a Unified Hybrid RAG for Precise Answers in University Admission Chatbots—A Case Study at HCMUT, arXiv preprint, (pp. 1–12), arXiv:2501.16276.

[6]. Nguyen Huu-Thanh, Ngo Quoc-Dung, and Dang Quang-Do (2024). An Intent-Filtered Retrieval-Augmented Generation Chatbot for University Admissions in Vietnam, in Proceedings of the 2024 International Conference on Advanced Technologies for Communications (ATC), (pp. 1–6), DOI: 10.1109/ATC63255.2024.10908188.

[7]. Maryamah Maryamah, Sutoyo Edi, and Darmawan Iwan (2024). Chatbots in Academia: A Retrieval-Augmented Generation Approach for Improved Efficient Information Access, in Proceedings of the 16th International Conference on Knowledge and Smart Technology (KST), (pp. 259–264), DOI: 10.1109/KST61284.2024.10499652.

[8]. Chen Zhipeng, Zou Di, Xie Haoran, Lou Huiwei, and Pang Zhi (2024). Facilitating University Admission Using a Chatbot Based on Large Language Models with Retrieval-Augmented Generation, Educational Technology & Society,

vol. 27, no. 4, (pp. 454–470), DOI: 10.30191/ETS.202410_27(4).TP02.

[9]. Kakwani Divyanshu, Kunchukuttan Anoop, Golla Satish, Gokul N.C., Bhattacharyya Avik, Khapra Mitesh M., and Kumar Pratyush (2020). IndicNLPSuite: Monolingual Corpora, Evaluation Benchmarks and Pre-trained Multilingual Language Models for Indian Languages, in Findings of the Association for Computational Linguistics: EMNLP 2020, (pp. 4948–4961), DOI: 10.18653/v1/2020.findings-emnlp.445.

[10]. Clark Jonathan H., Choi Eunsol, Collins Michael, Garrette Dan, Kwiatkowski Tom, Nikolaev Vitaly, and Palomaki Jennimaria (2020). TyDi QA: A Benchmark for Information-Seeking Question Answering in Typologically Diverse Languages, Transactions of the Association for Computational Linguistics, vol. 8, (pp. 454–470), DOI: 10.1162/tacl_a_00317.

[11]. Asai Akari, Oguz Barlas, Min Sewon, Tschannen Michael, Joshi Mandar, Chen Danqi, Yih Wen-tau, and Zettlemoyer Luke (2021). XOR QA: Cross-lingual Open-retrieval Question Answering, in Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), (pp. 547–564), DOI: 10.18653/v1/2021.naacl-main.46.

[12]. Es Shahul, James Jithin, Espinosa Anke Luis, and Schockaert Steven (2024). RAGAs: Automated Evaluation of Retrieval Augmented Generation, in Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations (EACL), (pp. 150–158), DOI: 10.18653/v1/2024.eacl-demo.16.

[13]. Bukkapatnam Krishna S., and Mehta Sameep (2026). TabFaith: Benchmarking and Improving Structural Faithfulness in LLM Table Summarization, in Proceedings of the SURGeLLM Workshop, (pp. 326–332), DOI: 10.18653/v1/2026.surgellm-1.21.

[14]. Ji Ziwei, Lee Nayeon, Frieske Rita, Yu Tiezheng, Su Dan, Yan Sharon, Ishii Etsuko, Bang Yejin, Madotto Andrea, and Fung Pascale (2023). Survey of Hallucination in Natural Language Generation, ACM Computing Surveys, vol. 55, no. 12, (pp. 1–38), DOI: 10.1145/3571730.

[15]. Cormack Gordon V., Clarke Charles L. A., and Büttcher Stefan (2009). Reciprocal Rank Fusion Outperforms Condorcet and Individual Rank Learning Methods, in Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval, (pp. 758–759), DOI: 10.1145/1571941.1572114.

IJEAST

INTERNATIONAL JOURNAL OF ENGINEERING APPLIED SCIENCE AND TECHNOLOGY

AUTHORS

- Shraddha Ukirade
- Anzalna Tamboli
- Sayali Kakade
- Dnyaneshwari Vyavahare
- Prof. Shobhana Raichurkar

ABOUT IJEAST

International journal of Engineering Applied Science and Technology (IJEAST) is a peer-reviewed, open access journal that publishes high - quality research paper in the field of Engineering, Applied Science and Technology.

IJEAST aims to provide a platform for researchers, academicians, and professionals to share their innovative ideas, research findings, and practical experiences with the global scientific community.

JOURNAL METRICS

H-INDEX	i10-INDEX	IJEAST CITATIONS	PEER-REVIEWED & OPEN ACCESS
33	154	9,647 (IJEAST CITATIONS)	PUBLISHED MONTHLY

PEER REVIEWED

All submission are rigorously peer reviewed to ensure quality.

AUTHORS ARE OUR PRIORITY

We value our authors and recognize their contribution to the advancement of research and knowledge.

OPEN ACCESS

Free and unrestricted access to research for all.

GLOBAL REACH

Connecting researchers and professionals worldwide.

TIMELY PUBLICATION

We ensure a swift and efficient publication process.

For more information, visit our website
www.ijeast.com

editor@ijeast.com
www.ijeast.com
India

2455-2143