iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** IX **Month of publication:** September 2026

DOI: https://doi.org/10.22214/ijraset.2026.84814

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue IX Sep 2026- Available at www.ijraset.com

PERA-Tutor: A Personalized Explainable Retrieval-Augmented Tutor for Educational Question Answering

Sabyasachi Saha
Techno Exponent, India

Abstract: ***Retrieval-augmented generation (RAG) is attractive for educational question answering because external evidence can improve provenance, while learner models can adapt explanations to prior knowledge and misconceptions. However, apparent gains can be overstated when synthetic benchmarks leak topic labels or make unanswerable queries trivially separable. This paper introduces PERA-Tutor, a personalized and explainable retrieval-augmented tutoring architecture, and reports a deterministic retrieval-stage evaluation that can be reproduced without a language model. The benchmark contains 240 curriculum chunks, 1,200 questions (900 answerable and 300 unanswerable), and 1,500 retrieval stress cases. With a BM25 first stage and a 25-document candidate pool, the prespecified pedagogical re-ranking point (semantic/topic/difficulty/misconception weights 0.55/0.20/0.15/0.10) increases Recall@5 from 0.550 to 0.624 versus semantic-only ranking (paired delta = 0.074, exact McNemar p = 1.09e-7) and increases Difficulty Fit@5 from 0.625 to 0.744 (delta = 0.119, Wilcoxon p = 2.18e-49). Yet a post-hoc no-topic reallocation yields only a 0.011 Recall@5 gain (p = 0.477), because every answerable template contains its gold topic name. Masking that surface form reduces PERA Recall@5 to 0.090. Separately, raw BM25 top scores distinguish answerable from unanswerable queries with AUC = 0.944 overall and 0.917 for in-corpus-subject negatives. A five-fold calibrated retrieval-score baseline (B0) correctly abstains on all 300 unanswerable queries while accepting 817 of 900 answerable queries. These findings support difficulty-aware ranking as a mechanism for pedagogical fit, but not a general claim of retrieval or hallucination superiority. They also establish two requirements for future end-to-end evaluation: leakage-resistant questions and verification baselines that outperform calibrated retrieval-score abstention.***

Keywords: retrieval-augmented generation; intelligent tutoring systems; personalized learning; explainable AI; BM25; hallucination; abstention; benchmark leakage

I. INTRODUCTION

Large language models (LLMs) can produce fluent explanations and interactive dialogue, making them appealing as tutoring interfaces. In education, however, fluency is insufficient: a useful tutor must ground substantive claims, adapt explanations to the learner, expose evidence in a way that can be inspected, and decline questions for which the available instructional corpus is inadequate. Prior work on LLMs in education emphasizes both the potential for personalized support and the need for fact checking, human oversight, and pedagogically deliberate deployment [13,14]. Explainable-AI research in education similarly argues that transparency requirements depend on the stakeholder and learning context rather than on generic interpretability alone [15].

Retrieval-augmented generation addresses part of this problem by conditioning generation on externally retrieved evidence [1]. Retrieval itself is modular: sparse rankers such as BM25 remain strong and transparent baselines, while dense retrievers can improve semantic matching in open-domain settings [2,3]. Recent RAG evaluation frameworks have therefore moved beyond end-answer accuracy toward component-level diagnosis of context relevance, answer faithfulness, citation support, and robustness [5-8]. This decomposition is particularly important for educational systems, where a retrieval failure can be mistaken for a generation failure and where a seemingly successful metric can reflect benchmark construction rather than generalizable behavior.

PERA-Tutor (Personalized Explainable Retrieval-Augmented Tutor) was designed to connect four functions: evidence retrieval, learner-aware re-ranking, evidence-constrained generation, and claim-level verification/explanation. The original experimental plan compared a plain LLM, standard RAG, personalized RAG, and the full PERA-Tutor system. In the present paper, the empirical scope is intentionally narrower: only the deterministic retrieval subsystem is evaluated. No LLM generation results, hallucination rates, faithfulness scores, human trust ratings, or learning-gain claims are reported. This restriction is methodologically useful because the retrieval stage can be reproduced exactly and audited for benchmark artifacts before expensive end-to-end experiments are interpreted.

The study makes four contributions:

- A modular PERA-Tutor architecture that separates learner modeling, retrieval, pedagogical re-ranking, generation, explanation, and verification, allowing each stage to be evaluated independently.
- A paired evaluation of semantic-only versus pedagogically weighted ranking on 900 answerable questions, including weight sweeps, adversarial strata, and a 1,500-case retrieval stress set.
- A benchmark leakage audit showing that the apparent recall gain is dominated by a topic-name feature that is effectively oracle-like on the templated benchmark, while the difficulty component produces the intended improvement in pedagogical fit.
- A calibrated B0 retrieval-score abstention baseline showing that unanswerability is already highly separable without a verifier; future verification claims must therefore beat this baseline rather than a non-abstaining RAG system.

A. Research Questions

The empirical questions in this retrieval-stage paper are:

- RQ1: Does pedagogical re-ranking change gold-evidence retrieval and learner-level difficulty fit relative to semantic-only ranking?
- RQ2: Which re-ranking terms drive any observed gains, and how sensitive are the gains to explicit topic-name leakage?
- RQ3: How do semantic-only and pedagogically weighted retrieval behave across adversarial and stress-test strata?
- RQ4: How well can a calibrated raw retrieval-score threshold separate answerable from unanswerable questions, including unanswerable items whose subject is present in the corpus?

II. RELATED WORK

A. Retrieval-augmented question answering

RAG combines parametric generation with a retrievable non-parametric memory [1]. In the original formulation, retrieved passages provide task-relevant evidence and provenance. BM25 remains an important sparse retrieval baseline because its term-based behavior is transparent and efficient [2].

Dense Passage Retrieval demonstrated that learned dual-encoder representations can substantially improve top-k retrieval on several open-domain QA benchmarks [3], motivating hybrid lexical-dense pipelines. More recent RAG work includes self-reflective retrieval and generation strategies, but the present study deliberately avoids attributing downstream generation benefits to retrieval metrics alone [4].

B. RAG evaluation, hallucination, and abstention

RAGAS, ARES, and RAGChecker illustrate a shift toward modular RAG evaluation, with metrics that diagnose retrieval relevance, groundedness/faithfulness, and generation quality rather than reducing performance to a single answer score [5-7]. Hallucination research likewise distinguishes fluent generation from factual support [9-12]. For knowledge-bounded systems, unanswerability is a separate capability: a system should reject or qualify a request when the indexed corpus does not provide enough evidence. UAEval4RAG formalizes this concern and shows that answerable and unanswerable behavior must be evaluated jointly [8]. The present B0 baseline follows this logic at the retrieval stage: before assigning value to an expensive verifier, one should establish what a calibrated retrieval-score threshold already achieves.

C. Personalization and explainability in education

Educational LLM research highlights opportunities for responsive explanation and individualized support, while repeatedly warning that generated content can be unreliable and requires appropriate oversight [13,14]. XAI-ED frames explainability in educational systems as stakeholder- and purpose-dependent [15].

Learner modeling has a longer history in intelligent tutoring and knowledge tracing, including neural approaches such as Deep Knowledge Tracing [16] and large-scale interaction resources such as EdNet [17]. PERA-Tutor uses a deliberately simple learner state in the current retrieval study - prior-knowledge level and misconception rate - so that the effect of personalization features can be inspected directly rather than hidden inside a learned learner model.

III. PERA-TUTOR ARCHITECTURE

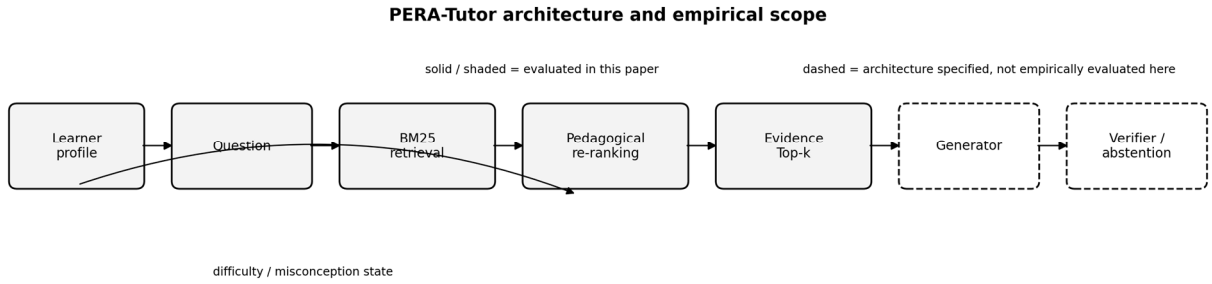


Figure 1. PERA-Tutor architecture. Solid/shaded components are evaluated in this paper; dashed components remain part of the specified end-to-end architecture but are not empirically evaluated here.

Figure 1 separates the architecture into an evaluated retrieval path and a planned generative path. The learner profile contains a prior-knowledge score and misconception rate. The question is first passed to a lexical retriever; a learner-aware scoring function then re-orders a fixed candidate pool. Retrieved evidence is intended to condition the generator, after which claims can be linked to supporting chunks and checked by a verifier. The generator/verifier side is architectural specification only in this paper.

A. Learner State

Each student has a synthetic prior_knowledge_score in [0,100] and a misconception_rate in [0,1]. For retrieval personalization, prior knowledge is mapped to an ordinal learner level L_u in {1,2,3}: beginner for scores below 45, intermediate for 45-74, and advanced for 75 or above. The misconception rate is used only when a candidate chunk is explicitly marked as a misconception-oriented instructional chunk.

B. First-stage Retrieval

The evaluated first stage is BM25 over the concatenation of subject, topic, and chunk text. This is a deliberate correction to the earlier research blueprint, which specified a hybrid dense-plus-lexical stage but was not executed in the available environment. The submitted paper therefore reports the configuration that was actually run: lexical BM25 only. A dense or hybrid retriever is left as future validation, not silently substituted into the results.

For each query q, the top 25 BM25 documents form the candidate pool P(q). Within that pool, raw BM25 scores are min-max normalized to obtain s_sem(q,c). If all scores are equal, s_sem is set to one for all candidates.

C. Pedagogical re-ranking

Topic inference is performed from the query string alone. A topic receives a perfect match when its lower-cased surface form is a substring of the query; otherwise a SequenceMatcher-style fuzzy similarity is evaluated over the query and local word spans. The best topic is accepted when the similarity is at least 0.72. For a candidate chunk c, s_topic is 1.0 for an inferred topic match, 0.5 for a subject-only match, and 0 otherwise.

For chunk difficulty L_c in {1,2,3}, learner-fit is:

$$s\_diff(u,c) = 1 - |L\_c - L\_u| / 2 \quad \textbf{(1)}$$

Chunks with a non-level label (for example, Misconception) receive a neutral difficulty score of 0.5. The misconception term s_mis(u,c) equals the learner misconception rate when c is a misconception-oriented chunk and zero otherwise. The final re-ranking score is:

$$S(q,c,u) = \lambda_s s\_sem + \lambda_t s\_topic + \lambda_d s\_diff + \lambda_m s\_mis \quad \textbf{(2)}$$

The prespecified PERA operating point is (lambda_s, lambda_t, lambda_d, lambda_m) = (0.55, 0.20, 0.15, 0.10). Semantic-only ranking uses (1,0,0,0). Because the weight sweep is monotonic through the strongest tested pedagogical setting, the prespecified point is treated as a design point, not as an optimized or tuned optimum.

D. End-to-end system matrix and B0 baseline

The complete architecture retains the original system matrix for future end-to-end experiments, augmented with B0. Table 1 distinguishes the systems that are specified from those empirically evaluated in this paper.

ID	Retrieval	Personalization	Explanation	Verification	Status in this paper
A: Plain LLM	No	No	No	No	Not run
B0: score-threshold RAG	Yes	No	Citation only	Threshold only	Retrieval diagnostic run
B: RAG	Yes	No	Citation only	No	Retrieval ranking run
C: RAG + personalization	Yes	Yes	Citation	No	Retrieval ranking run
D: PERA-Tutor	Yes	Yes	Source + rationale + confidence	Yes	Architecture specified; not run
D1: no personalization	Yes	No	Yes	Yes	Not run
D2: no verification	Yes	Yes	Yes	No	Not run
D3: no explanation	Yes	Yes	No	Yes	Not run

Table 1. Experimental system matrix. Only deterministic retrieval-stage computations are reported as empirical results.

B0 abstains when the raw first-stage BM25 top score is below a calibrated threshold tau:

$$B0(q) = \text{answer if } \max_c BM25(q,c) \geq \tau; \text{ otherwise abstain} \quad \textbf{(3)}$$

The B0 threshold is selected inside each training fold to maximize balanced accuracy, with F1 and then the higher threshold used as tie breakers. Importantly, B0 uses the raw BM25 score, not the final personalized re-ranking score, because the latter is contaminated by the benchmark's topic feature.

IV. BENCHMARK AND EXPERIMENTAL PROTOCOL

A. Synthetic Research Package

The controlled research package was generated with seed 42 for method development and ablation testing. The present experiment uses 240 knowledge-base chunks, 1,200 QA benchmark questions, and 1,500 retrieval stress items. Of the QA questions, 900 are answerable from the knowledge base and have an expected evidence chunk; 300 are explicitly unanswerable and should trigger abstention or qualification. The answerable subset includes ordinary, ambiguous, distractor, and false-premise strata. The larger package also contains misconception cases, multi-turn sessions, claim-evidence pairs, adaptive assessment items, learning-progress records, synthetic teacher-rating rows, and contradictory-evidence cases. Those additional tables support future pipeline evaluation but are not treated as observed human or LLM outcomes here.

Component	Size	Role in this paper
Knowledge base	240 chunks	Retrieval corpus and gold evidence units
QA benchmark	1,200 questions	900 answerable + 300 unanswerable
Answerable adversarial strata	900 questions	633 none; 103 ambiguous; 94 distractor; 70 false premise
Retrieval stress set	1,500 items	Cross-subject, lexical overlap, near-topic, noisy, underspecified, unanswerable
Claim-evidence pairs	2,000	Reserved for future verifier evaluation
Misconception cases	1,200	Reserved for future tutoring evaluation
Multi-turn sessions	2,500 turns / 500 sessions	Reserved for future dialogue evaluation
Contradictory evidence	600 cases	Reserved for future conflict-aware verification

Table 2. Research package components and empirical scope.

B. Weighting Conditions

Four ranking weight vectors are evaluated on the 900 answerable questions: semantic only (1,0,0,0), mild pedagogical (0.70,0.15,0.10,0.05), the prespecified PERA point (0.55,0.20,0.15,0.10), and strong pedagogical (0.40,0.25,0.20,0.15). Student-query pairing is held constant across conditions. Existing interaction mappings are used where present; remaining questions are paired deterministically with a student profile using seed 42.

C. Metrics

Retrieval quality is reported as Recall@1, Recall@5, Recall@10, and mean reciprocal rank (MRR), using the expected gold chunk as the target. Difficulty Fit@5 measures the mean compatibility of the top-five level-labelled chunks with the paired learner:

$$\text{Fit@5} = \text{mean\_c in top5 } [1 - |L_c - L_u| / 2] \quad \textbf{(4)}$$

For the retrieval stress set, the evaluation records whether the gold chunk appears in the top five, the mean number of annotated distractors in the top five, and whether the top-ranked item is a distractor. For answerability, the raw BM25 top score is evaluated by ROC AUC and by held-out threshold classification.

D. Statistical analysis

Metric means are accompanied by 95% nonparametric bootstrap intervals (4,000 resamples) where available. Paired binary retrieval outcomes are compared with exact McNemar tests; Difficulty Fit@5 is compared with the Wilcoxon signed-rank test. Answerable versus unanswerable score separation is quantified with the Mann-Whitney U statistic expressed as AUC. The three recall-family McNemar p-values remain significant under Holm correction (adjusted p approximately 0.031, 2.17e-7, and 4.96e-24 for Recall@1, Recall@5, and Recall@10, respectively). B0 uses deterministic five-fold stratified cross-validation with seed 42.

E. Benchmark leakage and sensitivity audits

Two post-hoc audits test whether the re-ranking gain depends on the explicit topic surface form. First, the topic term is removed and its weight is reallocated to difficulty and misconception features, yielding (0.55,0,0.30,0.15). Second, every answerable query's literal gold topic surface form is replaced by the phrase 'this concept' before retrieval. The latter is a counterfactual sensitivity test, not a realistic student-query benchmark; its purpose is to measure dependence on a known template field. A direct template audit finds that 900 of 900 answerable questions contain the gold topic name verbatim.

V. RESULTS

A. Weight sweep and paired retrieval effects

Weighting	Recall@1	Recall@5	Recall@10	MRR	Difficulty Fit@5
semantic only (system B)	0.166 [0.141, 0.191]	0.550 [0.517, 0.583]	0.860 [0.837, 0.881]	0.354 [0.333, 0.375]	0.625 [0.610, 0.639]
mild pedagogical	0.176 [0.151, 0.200]	0.596 [0.563, 0.628]	0.917 [0.899, 0.934]	0.386 [0.365, 0.406]	0.736 [0.729, 0.743]
paper setting (system C)	0.192 [0.166, 0.219]	0.624 [0.592, 0.656]	0.949 [0.934, 0.963]	0.394 [0.373, 0.415]	0.744 [0.737, 0.751]
strong pedagogical	0.219 [0.192, 0.247]	0.633 [0.601, 0.663]	0.970 [0.959, 0.981]	0.409 [0.387, 0.432]	0.751 [0.746, 0.757]

Table 3. Re-ranking weight sweep on 900 answerable questions. Brackets are 95% bootstrap intervals where reported.

At the prespecified PERA point, Recall@5 increases from 0.550 to 0.624 (absolute delta 0.074), while Difficulty Fit@5 increases from 0.625 to 0.744 (delta 0.119). The paired McNemar test for Recall@5 yields 113 questions correct only under PERA weighting versus 46 correct only under semantic-only ranking ($p = 1.09e-7$). Recall@1 and Recall@10 also increase significantly. The difficulty-fit difference is highly significant by Wilcoxon signed-rank test ($p = 2.18e-49$).

The sweep also reveals a design warning: the strongest pedagogical weighting tested achieves the highest Recall@1, Recall@5, Recall@10, MRR, and difficulty fit. Therefore (0.55,0.20,0.15,0.10) is not an empirical optimum on this benchmark and should not be described as tuned. Its value in this study is as a prespecified operating point for decomposition and sensitivity analysis.

Metric	B-only correct	C-only correct	Delta (C-B)	p
Recall@1	45	69	+0.027	0.0308
Recall@5	46	113	+0.074	1.09e-07
Recall@10	0	80	+0.089	1.65e-24
Difficulty fit@5	-	-	+0.119	2.18e-49

Table 4. Paired semantic-only (B weighting) versus prespecified PERA (C weighting) comparisons.

B. Performance by Adversarial Stratum

Stratum	n	Semantic R@5	PERA R@5	Delta
ambiguous	103	0.563	0.660	+0.097
distractor	94	0.415	0.553	+0.138
false_premise	70	0.629	0.629	+0.000
none	633	0.559	0.629	+0.070

Table 5. Recall@5 by answerable adversarial stratum.

The largest Recall@5 gain occurs in the distractor stratum (+0.138), followed by ambiguous questions (+0.097). The false-premise stratum shows no change. These differences are descriptive stratum analyses rather than separately powered hypothesis tests.

C. Retrieval Stress Test

Stress type	n	B gold@5	C gold@5	B top1 distractor	C top1 distractor
cross_subject	230	0.552	0.570	0.009	0.009
lexical_overlap	231	0.528	0.584	0.009	0.000
near_topic	243	0.568	0.572	0.012	0.012
noisy_query	248	0.597	0.657	0.004	0.004
unanswerable	300	0.000	0.000	0.020	0.020
underspecified	248	0.077	0.056	0.008	0.008

Table 6. Retrieval stress results. B = semantic-only ranking; C = prespecified PERA weighting.

PERA weighting improves gold@5 most clearly for noisy queries (0.597 to 0.657) and lexical-overlap cases (0.528 to 0.584), with smaller changes for cross-subject and near-topic cases. The underspecified stratum moves in the opposite direction (0.077 to 0.056), indicating that pedagogical features can reinforce a guessed topic even when the query itself is insufficiently specific. The unanswerable stress items have no gold chunk by construction, so gold@5 is zero for both systems.

D. Term decomposition and topic leakage

Condition	Recall@5	Difficulty Fit@5	Delta vs matched semantic
Semantic only - original	0.550	0.625	-
PERA - original	0.624	0.744	+0.074
No-topic reallocation - original	0.561	0.783	+0.011
Semantic only - topic masked	0.086	0.720	-
PERA - topic masked	0.090	0.784	+0.004

Table 7. Topic-term decomposition and counterfactual topic-masking sensitivity analysis.

Removing the topic feature while preserving a unit-sum weight vector changes Recall@5 from 0.550 to 0.561, a non-significant +0.011 difference (exact McNemar p = 0.477). In contrast, the original PERA weighting produces +0.074. Thus the learner-dependent difficulty and misconception terms do not explain the gold-retrieval gain. Their strongest measured contribution is pedagogical alignment: the no-topic reallocation reaches Difficulty Fit@5 = 0.783, compared with 0.625 for semantic-only ranking.

The topic-masking audit is more severe. Topic inference accuracy falls from 1.000 to 0.000, semantic-only Recall@5 falls from 0.550 to 0.086, and PERA Recall@5 falls from 0.624 to 0.090. The original-versus-masked PERA paired difference is -0.534 (482 original-only successes versus one masked-only success; p = 3.88e-143). Because masking also removes a highly discriminative lexical cue from BM25, this test does not estimate performance on natural student language. It does, however, demonstrate that the current benchmark is structurally dependent on the topic surface form and is unsuitable for claims of robust semantic generalization.

E. Answerability separation and B0 abstention

Diagnostic	Value
Mean raw BM25 top score - answerable	10.739
Mean raw BM25 top score - unanswerable	5.292
ROC AUC - all 300 unanswerable	0.944
ROC AUC - 128 in-corpus-subject unanswerable	0.917
ROC AUC - 172 out-of-corpus-subject unanswerable	0.964
Mean 5-fold calibrated threshold	6.645
B0 held-out balanced accuracy	0.954
B0 unanswerable abstention accuracy	1.000 (300/300)
B0 answerable acceptance	0.908 (817/900)
B0 false accepts / false abstentions	0 / 83

Table 8. Raw BM25 answerability diagnostic and cross-validated B0 score-threshold baseline.

Raw BM25 score alone separates answerable from unanswerable queries with AUC = 0.944 (Mann-Whitney p = 4.21e-118). Restricting negatives to the 128 unanswerable questions whose subject is represented in the knowledge base still yields AUC = 0.917; therefore the separation is not explained solely by out-of-domain subject names. In five-fold held-out evaluation, the selected threshold is 6.645 in every fold. B0 abstains correctly on all 300 unanswerable questions and accepts 817 of 900 answerable questions, producing balanced accuracy = 0.954.

A separate diagnostic using the final personalized re-ranking score produces even stronger answerability separation (AUC = 0.998), but that quantity is not used as the B0 baseline because it contains the same topic feature identified as benchmark leakage. The raw BM25 result is the more conservative comparison.

VI. DISCUSSION

A. What the Results Support

The cleanest positive result is pedagogical fit. Difficulty-aware re-ranking changes which evidence variants are surfaced to a learner and increases Difficulty Fit@5 by approximately 0.12 at the prespecified operating point. This is aligned with the purpose of the difficulty term: it is not expected to identify the canonical topic more accurately; it is expected to choose among relevant instructional chunks in a way that better matches the learner. The larger fit under the no-topic reallocation further supports this interpretation.

The measured Recall@5 gain, by contrast, should not be presented as evidence that personalization improves semantic retrieval. The gold topic name appears verbatim in all answerable question templates, topic inference is perfect, and removing the topic term eliminates most of the recall gain.

On this benchmark, the topic term functions as a near-oracle feature even though the code never reads the gold label directly. A reviewer sampling a small number of questions could discover this quickly; treating the effect as a limitation rather than obscuring it makes the study substantially more defensible.

B. Implications for abstention and verification

The B0 result changes the evidentiary standard for the planned verifier. A comparison between a verifier-equipped system and a RAG system that is forced to answer every query would be weak because the benchmark is already highly separable by retrieval score. The appropriate hypothesis is instead: does claim-level verification improve unsupported-claim detection and unanswerable handling beyond a calibrated score threshold, at an acceptable latency and token cost? This aligns with recent work that treats unanswerability as a first-class RAG evaluation dimension [8].

B0 is not a substitute for verification. It can only say that a query looks weakly supported by retrieval; it cannot determine whether a generated answer introduces a wrong number, reverses a relationship, adds an unsupported historical detail, or contradicts retrieved evidence. The research package contains claim-evidence and contradictory-evidence cases specifically for those future tests. The present result simply establishes that verification must demonstrate incremental value over a cheap baseline.

C. Benchmark Design Implications

The topic-masking sensitivity analysis shows why educational RAG benchmarks should decouple natural question wording from metadata used in retrieval scoring. Regenerated questions should paraphrase, imply, or omit the canonical topic label; include multi-topic and cross-concept questions; vary discourse style and learner errors; and contain in-corpus unanswerable queries whose lexical overlap matches answerable items. Public validation on resources such as SciQ, OpenBookQA, and ARC can provide an external check, although each dataset measures a different form of science QA and none is a direct substitute for a tutoring dialogue study [18-20].

The stress set also suggests that ambiguity deserves special treatment. Pedagogical weighting helps on several noisy or distractor-heavy cases but degrades underspecified queries. A production tutor should therefore consider asking a clarification question when query specificity is low rather than forcing the re-ranker to commit to a topic.

D. Weight Selection

Because metrics improve monotonically across the tested sweep, the current lambda values should be described as prespecified, not optimized. A future protocol should tune weights only on a development split and lock them before test evaluation. More importantly, weight selection should be multi-objective: maximizing gold Recall@k can conflict with difficulty alignment, diversity, and abstention calibration. Pareto-style analysis is more appropriate than optimizing a single retrieval metric.

VII. THREATS TO VALIDITY

Synthetic benchmark validity. The benchmark is controlled and privacy-safe but contains template regularities. The 100% verbatim topic inclusion is a concrete instance of construct leakage. Results from the original wording therefore do not estimate performance on unconstrained student questions.

Retrieval configuration. The evaluated first stage is BM25 only. The earlier blueprint specified a hybrid dense-plus-lexical retriever, but no dense embedding model was executed for these results. Dense retrieval could change both absolute recall and sensitivity to topic masking. The paper therefore reports the lexical configuration exactly and makes no hybrid-retrieval claim.

Post-hoc diagnostics. The no-topic reallocation, topic masking, and B0 analysis were motivated by observed benchmark behavior. They are diagnostic analyses rather than preregistered confirmatory tests. The next benchmark version should predefine these checks and reserve an untouched test split.

No generation or human evaluation. This study does not execute systems A-D as language-generation systems and cannot establish reductions in generated hallucination, improvements in citation correctness, explanation usefulness, learner trust, latency-cost trade-offs, or learning gain. Synthetic teacher ratings and learning-progress rows in the research package are pipeline placeholders and are not used as empirical evidence.

Threshold transfer. B0's threshold is corpus-, tokenizer-, and score-scale-dependent. Its strong held-out result may partly reflect templated lexical distributions and should not be assumed to transfer to a new corpus. It is a benchmark-specific baseline, not a universal abstention threshold.

Counterfactual masking. Replacing a topic phrase with 'this concept' creates unnatural questions and removes lexical evidence from both BM25 and the topic feature. The large masked performance drop therefore quantifies sensitivity to the surface form; it should not be interpreted as an estimate of real-student performance.

VIII. END-TO-END VALIDATION PROTOCOL

The architecture is designed for a second-stage study after benchmark regeneration. The following protocol is recommended before making claims about hallucination reduction or tutoring quality:

1) Regenerate the QA benchmark so topic names are not deterministically exposed; reserve development and test splits by topic or concept family to limit template leakage.
2) Re-run retrieval with the intended hybrid lexical+dense stack and report BM25-only, dense-only, and hybrid ablations.
3) Execute A, B0, B, C, D, D1, D2, and D3 with fixed model versions, prompts, temperature, context length, and retrieval top-k. Archive exact outputs.
4) Evaluate faithfulness, answer relevance, citation correctness, hallucination flags, abstention correctness, personalization alignment, latency, and token cost. Use claim-level evidence annotations for verifier evaluation rather than relying only on an LLM judge.
5) Compare D against B0 on unanswerable handling. The key verification hypothesis should require statistically significant incremental benefit over retrieval-score abstention, not merely over a system that always answers.
6) Add blinded human evaluation for explanation usefulness and learner-level appropriateness, report inter-rater reliability, and avoid learning-gain claims until a real learner study is conducted.
7) Validate on at least one public educational QA set and one hallucination/unanswerability stress resource to test transfer beyond the synthetic corpus.

This protocol converts the current negative findings into design controls. In particular, the topic audit prevents a misleading retrieval claim, while B0 prevents a misleading verification claim.

IX. ETHICS, PRIVACY, AND EDUCATIONAL USE

The research package uses synthetic learner profiles and synthetic instructional interactions rather than real student personally identifiable information. This reduces privacy risk during pipeline development but does not eliminate ethical obligations for deployment. A real educational system should minimize learner data collection, protect records, disclose the use of automated generation, provide inspectable evidence, and preserve teacher or qualified human oversight for high-stakes decisions. Personalization should not be treated as a diagnosis of ability, and confidence scores should not be presented as guarantees of correctness. Questions involving medical, legal, or other high-stakes personalized advice should not be answered solely from a general educational tutor.

X. REPRODUCIBILITY

The retrieval experiment is deterministic with seed 42. The first stage uses BM25 with k1 = 1.5 and b = 0.75; candidate pool size is 25. Topic inference uses exact substring matching followed by SequenceMatcher-based fuzzy matching with threshold 0.72. The weight vectors, learner-level mapping, bootstrapping procedure, and paired tests are specified in Sections 3-4. The companion code writes machine-readable CSV/JSON results for the weighting sweep, paired tests, stratum analysis, stress tests, answerability diagnostics, B0 cross-validation, and leakage audit. Reproducibility requires reporting the lexical-only configuration as evaluated rather than substituting results from an unexecuted dense model.

XI. CONCLUSION

PERA-Tutor proposes a modular educational RAG architecture in which retrieval, learner adaptation, explanation, and verification can be tested separately. The deterministic retrieval-stage study yields a useful but deliberately constrained conclusion. Pedagogical re-ranking substantially improves the difficulty match of retrieved instructional chunks, which supports its role as a learner-alignment mechanism. The apparent gold-retrieval improvement is not robust evidence of personalization, because it is driven primarily by verbatim topic leakage in the synthetic question templates. In parallel, a raw BM25 score threshold already provides strong answerability separation and perfect abstention on the 300 unanswerable benchmark items under five-fold held-out evaluation, establishing a demanding low-cost baseline for future verification experiments.

Accordingly, this paper does not claim that PERA-Tutor reduces generated hallucinations or improves learning outcomes. Its contribution is to expose the retrieval-stage conditions that must be satisfied before those claims can be tested credibly: leakage-resistant questions, externally validated retrieval, a calibrated abstention baseline, claim-level verifier evaluation, and real human assessment. In this form, the negative diagnostics are not failures of the framework; they are controls that make the next end-to-end study scientifically stronger.

REFERENCES

[1] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Kuttler, M. Lewis, W.-t. Yih, T. Rocktaschel, S. Riedel, and D. Kiela, 'Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks,' Advances in Neural Information Processing Systems, vol. 33, pp. 9459-9474, 2020.
[2] S. Robertson and H. Zaragoza, 'The Probabilistic Relevance Framework: BM25 and Beyond,' Foundations and Trends in Information Retrieval, vol. 3, no. 4, pp. 333-389, 2009, doi: 10.1561/1500000019.
[3] V. Karpukhin, B. Oguz, S. Min, P. Lewis, L. Wu, S. Edunov, D. Chen, and W.-t. Yih, 'Dense Passage Retrieval for Open-Domain Question Answering,' in Proc. EMNLP, pp. 6769-6781, 2020, doi: 10.18653/v1/2020.emnlp-main.550.
[4] A. Asai, Z. Wu, Y. Wang, A. Sil, and H. Hajishirzi, 'Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection,' in Proc. ICLR, 2024.
[5] S. Es, J. James, L. Espinosa Anke, and S. Schockaert, 'RAGAs: Automated Evaluation of Retrieval Augmented Generation,' in Proc. EACL System Demonstrations, pp. 150-158, 2024, doi: 10.18653/v1/2024.eacl-demo.16.
[6] J. Saad-Falcon, O. Khattab, C. Potts, and M. Zaharia, 'ARES: An Automated Evaluation Framework for Retrieval-Augmented Generation Systems,' in Proc. NAACL-HLT, pp. 338-354, 2024, doi: 10.18653/v1/2024.naacl-long.20.
[7] D. Ru et al., 'RAGChecker: A Fine-grained Framework for Diagnosing Retrieval-Augmented Generation,' Advances in Neural Information Processing Systems, vol. 37, Datasets and Benchmarks Track, 2024, doi: 10.52202/079017-0692.
[8] X. Peng, P. K. Choubey, C. Xiong, and C.-S. Wu, 'Unanswerability Evaluation for Retrieval Augmented Generation,' in Proc. ACL, pp. 8452-8472, 2025, doi: 10.18653/v1/2025.acl-long.415.
[9] Z. Ji et al., 'Survey of Hallucination in Natural Language Generation,' ACM Computing Surveys, vol. 55, no. 12, Art. 248, pp. 1-38, 2023, doi: 10.1145/3571730.
[10] J. Li, X. Cheng, X. Zhao, J.-Y. Nie, and J.-R. Wen, 'HaluEval: A Large-Scale Hallucination Evaluation Benchmark for Large Language Models,' in Proc. EMNLP, pp. 6449-6464, 2023, doi: 10.18653/v1/2023.emnlp-main.397.
[11] P. Manakul, A. Liusie, and M. Gales, 'SelfCheckGPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models,' in Proc. EMNLP, pp. 9004-9017, 2023, doi: 10.18653/v1/2023.emnlp-main.557.
[12] S. Lin, J. Hilton, and O. Evans, 'TruthfulQA: Measuring How Models Mimic Human Falsehoods,' in Proc. ACL, pp. 3214-3252, 2022, doi: 10.18653/v1/2022.acl-long.229.
[13] E. Kasneci et al., 'ChatGPT for good? On opportunities and challenges of large language models for education,' Learning and Individual Differences, vol. 103, 102274, 2023, doi: 10.1016/j.lindif.2023.102274.
[14] A. Tlili et al., 'What if the devil is my guardian angel: ChatGPT as a case study of using chatbots in education,' Smart Learning Environments, vol. 10, Art. 15, 2023, doi: 10.1186/s40561-023-00237-x.
[15] H. Khosravi, S. Buckingham Shum, G. Chen, C. Conati, Y.-S. Tsai, J. Kay, S. Knight, R. Martinez-Maldonado, S. Sadiq, and D. Gasevic, 'Explainable Artificial Intelligence in education,' Computers and Education: Artificial Intelligence, vol. 3, 100074, 2022, doi: 10.1016/j.caeai.2022.100074.
[16] C. Piech et al., 'Deep Knowledge Tracing,' Advances in Neural Information Processing Systems, vol. 28, pp. 505-513, 2015.
[17] Y. Choi et al., 'EdNet: A Large-Scale Hierarchical Dataset in Education,' in Artificial Intelligence in Education, LNCS 12164, pp. 69-73, 2020, doi: 10.1007/978-3-030-52240-7_13.
[18] J. Welbl, N. F. Liu, and M. Gardner, 'Crowdsourcing Multiple Choice Science Questions,' in Proc. W-NUT, pp. 94-106, 2017, doi: 10.18653/v1/W17-4413.
[19] T. Mihaylov, P. Clark, T. Khot, and A. Sabharwal, 'Can a Suit of Armor Conduct Electricity? A New Dataset for Open Book Question Answering,' in Proc. EMNLP, pp. 2381-2391, 2018, doi: 10.18653/v1/D18-1260.
[20] P. Clark et al., 'Think you have Solved Question Answering? Try ARC, the AI2 Reasoning Challenge,' arXiv:1803.05457, 2018.

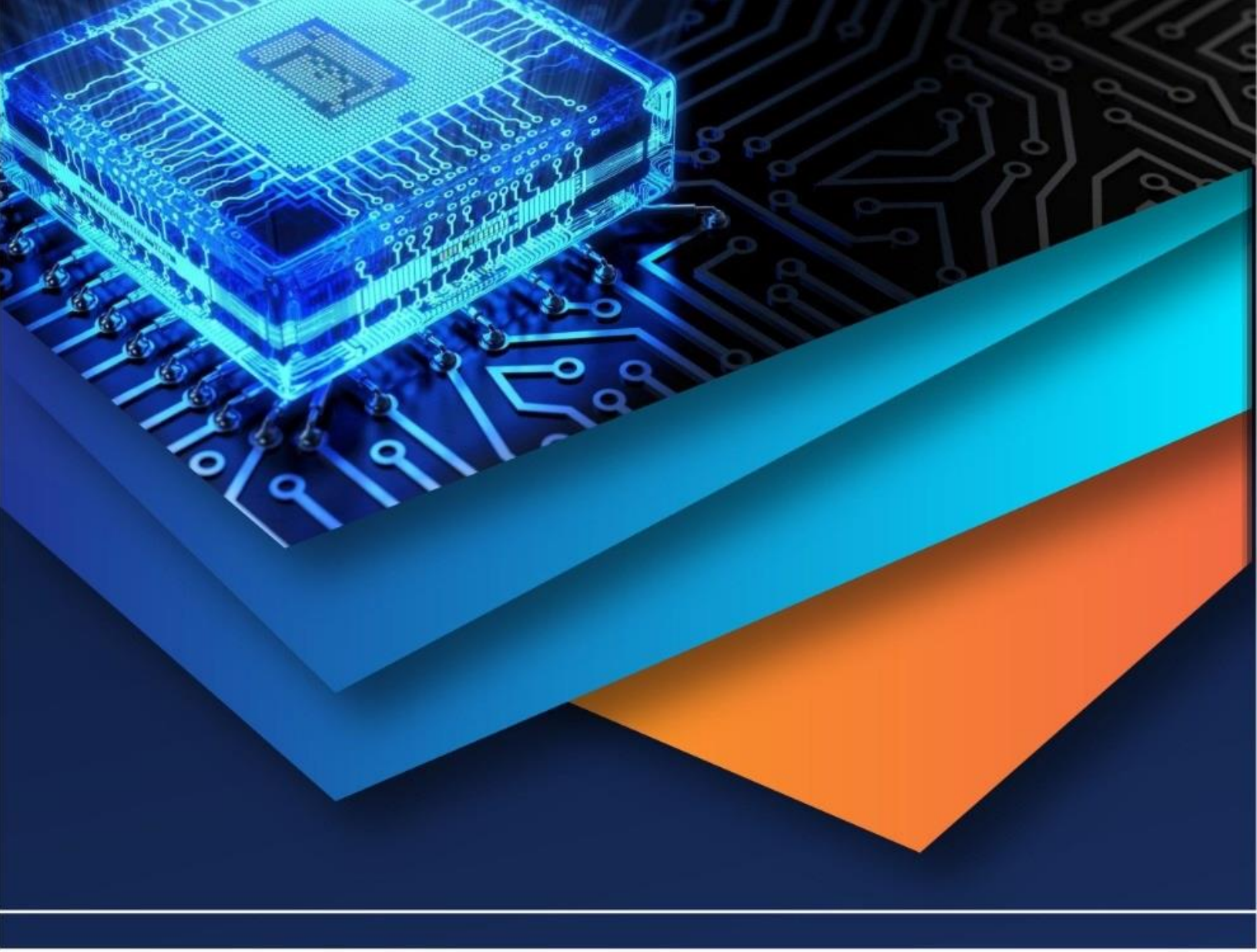
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue IX Sep 2026- Available at www.ijraset.com

Appendix A. Complete Retrieval Diagnostics

This appendix records the exact lexical-only diagnostics used in the paper so that later dense/hybrid re-runs can be compared without overwriting the original evidence.

Quantity	Observed value
Answerable questions	900
Unanswerable questions	300
Stress items	1,500
Candidate pool	25
PERA Recall@5	0.624444
Semantic Recall@5	0.550000
PERA - semantic Recall@5	+0.074444
No-topic Recall@5	0.561111
No-topic - semantic Recall@5	+0.011111 (p = 0.476882)
PERA Difficulty Fit@5	0.744074
Semantic Difficulty Fit@5	0.624814
Topic inference accuracy - original	1.000
Topic inference accuracy - masked	0.000
PERA Recall@5 - masked	0.090000
Raw BM25 AUC - all unanswerable	0.944074
Raw BM25 AUC - in-corpus subject	0.916944
B0 balanced accuracy	0.953889
B0 unanswerable abstention	1.000
B0 answerable acceptance	0.907778

Table A1. Exact observed diagnostics reported to six decimals where useful.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY