http://dx.doi.org/10.18623/rvd.v23.5888

DESIGN AND EVALUATION OF A RETRIEVAL-AUGMENTED AI TUTOR FOR ACADEMIC ENGLISH WRITING IN SAUDI HIGHER EDUCATION

CONCEPÇÃO E AVALIAÇÃO DE UM TUTOR DE IA COM RECUPERAÇÃO DE INFORMAÇÕES PARA A REDAÇÃO ACADÊMICA EM INGLÊS NO ENSINO SUPERIOR SAUDITA

Article received on: 12/24/2025
Article accepted on: 3/25/2026

Asad Shafi*
*University Of Exeter, London, United Kingdom
Orcid: https://orcid.org/0009-0009-9277-8330
shafi.asad@gmail.com

The authors declare that there is no conflict of interest

Abstract
There has been growing interest in the role that academic English writing plays in assessment, publishing, and employment opportunities for initiatives that aim at improving national capability in Saudi Arabia [29,30]. However, in parallel with growing popularity of use of generative AI for writing amongst students, validating this AI raises challenge as it can produce unlimited writing fluently with either false references or overwrite the authoring process while masking their logical thinking behind elegant writing. Retrieval-Augmented Generation (RAG) is a technical remedy as it entails generation of text by retrieving knowledge from the source corpus which can be more readily auditable [3,4]. This research paper provides a review of existing literature published in the years 2020-2025 covering various aspects of RAG architecture, retrieval evaluation, LLM adaptation, and academic integrity governance towards designing and evaluating an AI tutor based on the RAG technology for academic English writing. In compliance with PRISMA guidelines [1] for the design-synthesis methodology of systems [2], a methodology strategy adopted for this purpose encompasses failure mode analysis vis-à-vis system controls and performance indicator analysis resulting in (i) the reference architecture consisting of authorized corpus, hybrid retrieval and re-ranking, schema awareness in tutoring, quality gates, and monitoring and (ii) two synthesis tables of design criteria and evaluation standards along with deployment requirements.

Keywords: Retrieval-Augmented Generation. Large Language Models. Academic Writing. EFL. Saudi Higher Education. Evaluation

Resumo
Tem havido um interesse crescente no papel que a redação acadêmica em inglês desempenha na avaliação, na publicação e nas oportunidades de emprego para iniciativas que visam melhorar a capacidade nacional na Arábia Saudita [29,30]. No entanto, paralelamente à crescente popularidade do uso da IA generativa para redação entre os estudantes, a validação dessa IA representa um desafio, pois ela pode produzir textos ilimitados com fluência, contendo referências falsas ou substituindo o processo de autoria, ao mesmo tempo em que mascara o raciocínio lógico por trás de uma redação elegante. A Geração Aumentada por Recuperação (RAG) é uma solução técnica, pois envolve a geração de texto por meio da recuperação de conhecimento do corpus de origem, o que pode ser mais facilmente auditável [3,4]. Este artigo de pesquisa apresenta uma revisão da literatura existente publicada nos anos de 2020 a 2025, abrangendo vários aspectos da arquitetura RAG, avaliação de recuperação, adaptação de LLM e governança da integridade acadêmica, com o objetivo de projetar e avaliar um tutor de IA baseado na tecnologia RAG para a redação acadêmica em inglês. Em conformidade com as diretrizes PRISMA [1] para a metodologia de projeto-síntese de sistemas [2], uma estratégia metodológica adotada para esse fim abrange a análise de modos de falha em relação aos controles do sistema e a análise de indicadores de desempenho, resultando em (i) a arquitetura de referência composta por corpus autorizado, recuperação híbrida e reclassificação, reconhecimento de esquemas na tutoria, controles de qualidade e monitoramento e (ii) duas tabelas de síntese de critérios de projeto e

Metrics. Faithfulness. Governance. Learning Transfer.

padrões de avaliação, juntamente com requisitos de implantação.

Palavras-chave: *Geração Aumentada por Recuperação. Grandes Modelos de Linguagem. Redação Acadêmica. EFL. Ensino Superior Saudita. Métricas de Avaliação. Fidelidade. Governança. Transferência de Aprendizagem.*

1 INTRODUCTION

Being one of the gatekeepers for further progress in Saudi universities, Academic English Writing has influenced the way students progress in their career and conduct academic research and prepare for the professional future [29,30]. Emergence of advanced LLMs had a significant influence on the field in addition to numerous others. Artificial intelligence features are used by the students to assist them in planning, editing, and paraphrasing, understanding opportunities and potential problems of using such means, as they might become guidance tools that can be abused and even cause ethical issues [28]. At the same time, there are challenges from the institutional point of view since prohibition will not work while allowing too powerful capabilities will compromise construct validity because students' papers will be indistinguishable from those done with the help of AI [28]. Technical challenge in the case involves provenance problem since it will not be possible to determine whether something was written by the student themselves or generated by artificial intelligence, as opposed to LLMs that allow retrieval augmented generation, which means using an external knowledge source, retrieving relevant parts and conditioning the model on those [3,4]. As is known from the field of educational question answering, retrieval with reasoning increases effectiveness when the answer requires the involvement of knowledge scattered across several knowledge sources, just like in case with rubrics, samples of writing, and rules of correct grammar [27]. Training should aim at delivering feedback, not solving tasks, in order to allow for learning through evidential analysis without compromising integrity. This paper aims to explore a design problem to present in a Q1/Scopus article: developing a retrieval-augmented AI tutor that provides feedback on academic English writing in accordance with rubrics of Saudi universities, and auditing and testing for auditing and safety

properties and learning transfer. System reliability cannot depend on only one aspect of design but rather include interaction among corpus development, retrieval strategies, quality gates, constraints and monitoring processes, as it is not the case with retrieval-augmented models and learning.

1.1 Aim and objectives

Aim. To review the literature on the subject and provide an overview of the system engineering principles and approach in designing the retrieval-augmented AI tutor for academic English writing and its evaluation according to performance indicators of the period 2020-2025.

Objectives.

O1: To study and address failure modes of retrieval and generation – irrelevant retrieval, fabrication, over-generation, prompt injection, and subpopulation bias – accordingly [13-15,25,28].

O2: To propose architecture of the system that includes evidence corpus with permissions, hybrid retrieval and reranking methods, constraint-based tutoring, and quality gating, etc. [3-5,16-18].

O3: To devise an evaluation protocol for retrieval, consistency, pedagogical impact, learning transfer, fairness and governance measures, including ablation experiments [7,9-11,17].

O4: To determine deployment boundaries according to Vision 2030 standards and policy guidelines of Saudi universities [29,30].

2 LITERATURE REVIEW

The section focuses on the literature from the perspective of mechanisms first – technologies-related decisions that influence reliability and validity of evidence as well as the required methodologies. The issues of retrieval-augmented generation, benchmarking, controllable generation, and education integrity governance are examined.

2.1 Retrieval-augmented generation and retrieval evaluation

It is the architecture of RAG that dictates collaboration between parametric and non-parametric memory components. To get higher results regarding the degree of similarity, embeddings can be optimized via techniques such as DPR to reflect similarities between passage and question [5]. Further investigations revealed that metric heterogeneity may help overcome the issues associated with coverage and relevance (nDCG) [7]. Considering uncertainties of noise in texts as well as ambiguities of questions, it could be reasonable to opt for hybrid methods of retrieving based on dense and lexical signals. While lexical signals refer to rubrics, dense signals include paraphrasing and intentionality. It could be assumed that there is some connection between retrieval-augmented generation and context limitation. Given the fact that guidelines are presented in detail for each of the courses, it would be more appropriate to reduce noise and amount of irrelevant contexts. It means that the task of retrieving is to limit context to illustrative examples. Thanks to the use of RAG and LLM, this problem was overcome through successful implementation of RAG to the guidelines of another course, resulting in high accuracy [27]. Despite different tasks in terms of writing tutoring and QA, it is possible to utilize certain benefits provided by retrieving (particularly, to tackle context scattering as well as unhelpful generation because of guidance).

2.2 Reranking, filtering, and adversarial robustness

Although recall is extremely important, one should not neglect precision as false information would mean credibility of misleading suggestions. Reranking should be done in order to improve ranking of query and passage, taking into account complicated models of relations between them [8]. It is also needed to incorporate both reranking and filtering into the tutoring process. Filters would help to delete unnecessary data based on such criteria as course name, type of assignment, proficiency level of students. Filters may also restrict generation of content by certain policies (for instance, to prohibit creating response by full model). Filters will provide additional layer of protection by excluding prompt injection attacks. From the standpoint of software security recommendations, it is needed to treat draft text as malicious and, thus, sanitize it (remove instructions and

extract delimiters) together with employing 'system-policy-always' strategy. One can say that threats do not depend on the type of interaction. For example, the question to the assignment can be asked to get the correct answer to trigger rubrics. In this case, threat modeling and continuous monitoring should be employed in tutoring system. One should implement techniques such as continuous lifespan monitoring and threat assessments in tutoring systems [17]. Threat assessments can be obtained through the implementation of certain technical measures like prompt injection detection, random auditing, and conservative fallbacks in dangerous situations.

2.3 Language model adaptation and controllable tutoring behavior

With the help of fine-tuning and reinforcement learning, it is possible to increase alignment of the model and minimize harmful outputs. Both approaches can be used to adjust tutor under certain policy [11]. Language model adaptation to the domain can be achieved by means of parameter-efficient fine-tuning (LoRA) as well as quantization techniques to guarantee adaptation despite limited resources [9, 10]. However, although there are positive scaling and fine-tuning effects shown by large language models, the output generated by those does not remain grounded [13, 14]. That is, fluent lies would inevitably appear in the output. Tutor's behavior control cannot be confined solely to proper behavior, certain constraints should be set.

Behavior control of tutor can be guaranteed by introducing schema-constrained tutoring. Tutor's responses need to be generated in compliance with certain schema (feedback, diagnosis, explanations, action items, and sequencing). Moreover, tutor needs to follow the rules dictated by the policy. Structured output makes the assessment of tutor performance easier due to its composition of several elements. Thus, thanks to adaptation tutor is able to respond accordingly while certain constraints and assessments facilitate governance.

2.4 Education integrity: evidence, threats, and transfer measures

According to the results of the study, the development of EFL learning skills received positive response from students in terms of grammar improvement, increase of

vocabulary size, generation speed and creativity [28]. Possible risks were associated with errors, over-reliance on prompts, inability to perform critical reasoning [28]. Systematic review reveals that currently available evidence relates to students' perceptions making it invalid [24]. At the same time, education integrity governance should rely on measurable outcomes rather than cheat detection and punishment [25]. In this regard, tutor's performance should be assessed to discriminate cosmetic gains from real ones.

3 REVIEW METHODOLOGY

PRISMA methodology [1] and literature review criteria are used to make sure that the synthesis can be replicated and relevant to the specified task [2]. The goal of the review is to synthesize knowledge to create design guidelines, i.e., to translate mechanisms and metrics into design guidelines.

3.1 Search strategy and inclusion criteria

The following keywords are applied: retrieval-augmented generation, hybrid retrieval, reranking, automated feedback, writing tutoring, academic integrity, and AI governance. Inclusion criteria are defined in terms of the sources and time frame of their relevancy (2020-2025), upon request. The inclusion criteria are as follows: (i) sources providing design guidelines for retrieval-augmented generator [3-8,27]; (ii) sources providing design guidelines concerning alignment and adaptation of control methods for generator behavior [9-12]; and (iii) sources providing design guidelines for solving the problems of education, integrity, and governance [23-26,28]. Any sources not providing design guidelines and/or evaluation metrics must be excluded. After that, the inclusion is coded by system layers (corpus, retrieval, reranking, generation, gates, and monitoring) and failure modes addressed by the source itself.

3.2 Data extraction and synthesis checks

For each source, the following information is extracted: artifacts used (datasets, benchmarks, prompts); design guidelines (retrieval, chunking, reranking, and

constraints); and evaluation metrics (for retrieval, faithfulness, learning outcomes). For the synthesis process to be successful, it is vital that there is consistency between failure types and "means-to-control". Every failure type has to be addressed by some guidance and metrics. This is where the criterion of "failures-to-control" becomes helpful to mitigate the narrative bias, i.e., to make any claim quantifiable. See Figure 1 and Table 1-2. Interfaces and criteria obtained via the synthesis will be term-locked and iterated further.

4 RESULTS

The following are the results of the literature review containing the design guidelines. The main lesson drawn from the literature review is that the design of writing tutoring must be performed within the framework of a pipeline architecture featuring metric-based control.

4.1 Authorized evidence corpus and term-locked versioning

An authorized corpus must be collected to create a retrieval-augmented generator that can be utilized for writing tutoring. It is important to have an access to evidence, such as rubrics, sample essays evaluated by experts, conventions and citation rules, as well as recommendations on when to use the output generated by the AI. The development process may require term mapping for two languages in the Saudi context. To improve the reproducibility, testability, and auditing of the project (by implementing SDLC security principles [16] and risk management [17]), a term-locked version of the corpus should be created.

4.2 Chunking and metadata engineering

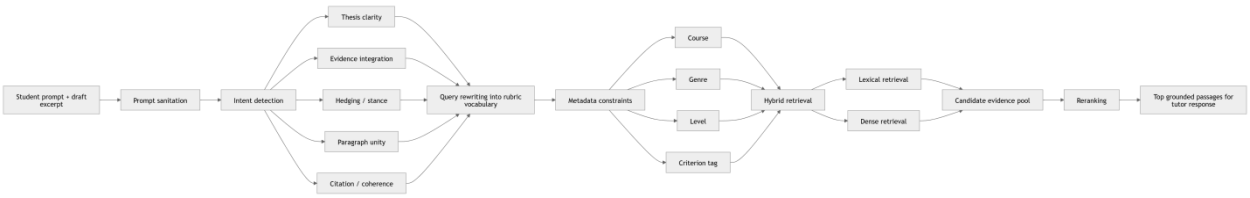
It is crucial to note that the size of chunks cannot be taken into account because it is defined semantically through rubric. For example, chunks associated with cohesion criteria should consist of definition, common mistakes, and examples. Metadata should include information about course, genre, level, and criteria-to-criterion mapping in order

to constrain the process of searching and eliminate the search results which correspond to irrelevant "general advice". This type of metadata allows conducting auditing via the sampling method, thanks to the stratification of the outputs with respect to genres and/or proficiency levels. As far as retrieval evaluation is concerned, there are certain tasks, which require the evaluation of different kinds. In other words, benchmarking should take several writing purposes (structure, citations, hedging, coherence) into consideration [7].

4.3 Hybrid retrieval, intent detection, and query rewriting

Prompts provided by users often include unnecessary information such as drafts or requests for multiple languages. With the help of intent detection, it becomes possible to classify a prompt to the purpose which is connected with one of several rubric criteria (thesis clarity, evidence integration, stance/hedging, paragraph unity, etc.). Intent detection allows applying constraints based on metadata and limiting the output to the necessary results only. Query rewriting aimed at converting user requests into rubric-based vocabulary could be used for refining the retrieval results. Query rewriting should be monitored because it can have negative effects on the evidence selection. Hybrid retrieval technique taking into account both lexical and dense representations makes it possible to find matching rubric-related concepts and paraphrasing the purposes [5,7].

Figura 1



4.4 Reranking, filtering, and abstention rules

Through reranking of results it becomes possible to sort the outcomes based on scoring query-passage pairs and interaction models [8]. Then, the filtering step eliminates materials, which contradict policies and do not relate to the identified purposes. Last but not least, it is vital to mention that abstention is an approach that should not be

overlooked: in cases when no proper outcomes are found, the system needs to ask clarifying questions or offer general rubric assistance. The latter aspect is very important considering hallucination and alignment with risk management recommendations [17].

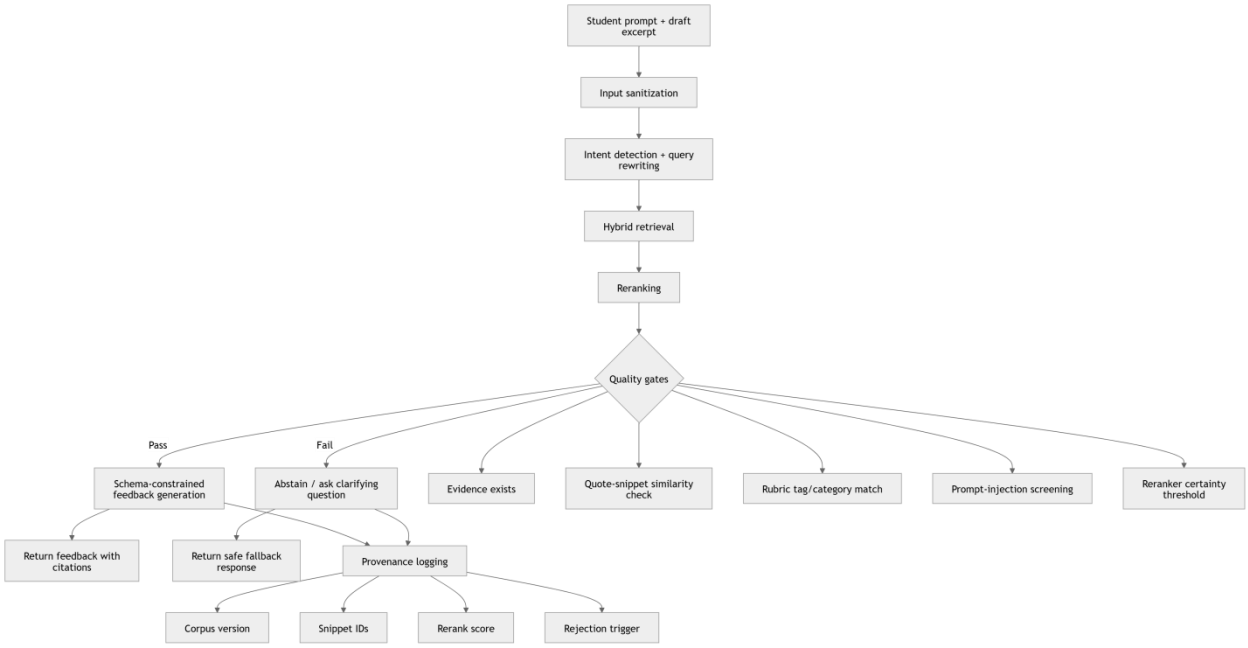
4.5 Schema-constrained tutoring and non-appropriation design

Security considerations make the outputs look like tutorials. Some tips concerning schema are listed below: (1) diagnosis label which corresponds to a rubric criterion; (2) diagnosis explanation supported by evidence; (3) revision plan; and (4) micro-examples, which consist of several sentences. It is very important to ensure that micro-examples constraint is applied because tutorial should produce only micro-examples, but not full-text outputs. This aspect deserves special attention in connection with integrity risks related to inappropriate usage of generative technologies reported by peer reviewers [25] and surveys [28]. Although alignment helps format the output, schema validation makes it happen no matter what the model is [11].

4.6 Quality gates, citation verification, and auditability

First of all, there is an application of the automated quality gates before providing the user with feedback. Automated quality gates may ensure the verification of the existence of evidence, consistency in the similarity of the quotes and text snippets, and compatibility between the inferred category and the relevant tag. There are several cases when feedback cannot be provided because the reasons for rejection include incomplete text snippet or prompt injection. In order to enable the auditability of the entire process, it is necessary to log the provenance data concerning the version of the corpus, id of the text snippets, reranking certainty, and rejection triggers. That means that it would be possible to repeat the series of steps if needed. Moreover, there would be an opportunity to publish the method in a scientific paper since it could be documented. From the point of view of governance frameworks, the principles of accountability and traceability can be based on logging and auditing of tutoring system interactions [17,19–22].

Figure 2



It should be noted what kind of attacks might affect the useful tool such as injection ("ignore all previous rules"), trick ("display rubric in hidden section"), and laundering ("generate it and then refuse"). Consequently, threat modeling is a process of determining assets (draft, logs, rubrics), vectors (prompt, indexing within the corpus, administrative interface), and countermeasures (input sanitization, access control, monitoring). Threat modeling can be optimized by applying the zero trust approach since the inputs are considered potentially hostile and the assumption is used

4.7 Threat modeling for educational RAG tutoring system

accordingly [16,18]. To enhance the reliability of the software, it is necessary to offer the conservative mode for high occurrence rate events as well as for the generation of quotations and diagnoses of the rubric.

4.8 Model selection, adaptation, and cost/latency envelopes

RAG enables an education organization to distinguish the special requirements from the common requirements related to large language models. The instruction-tuned

model suits well if it is possible to extract the context. When it comes to the adaptation, it should be kept in mind the issue of parameter efficiency that would help to get higher probabilities of schema compliance and effective implementation of the rejection policy [9,10]. In this connection, it is necessary to establish the service level objectives (SLOs) such as p95 latency, availability, and token limit. However, it should be noted that, while quantization reduces computation costs, it should be remembered about faithfulness and schema compliance too [10].

Figure 3

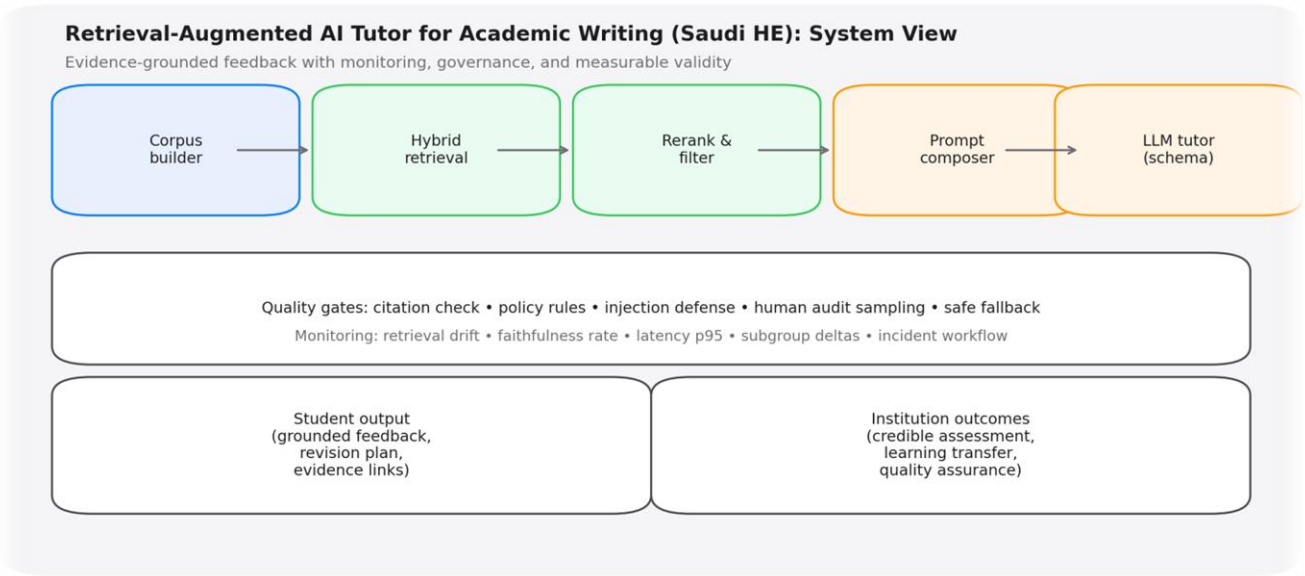


Table 1

Technical layers, measurable indicators, and operational controls for the proposed tutor.

Layer	Design choices (technical)	Measurable indicators	Operational controls
Evidence corpus	Rubrics, exemplars, genre prompts, citation guides, institutional policy; bilingual glossary; versioned snapshots	Coverage per rubric dimension; freshness; leakage risk	Semester lock; redaction; access control; audit trail
Embedding & index	Embedding model selection; ANN index; metadata filters (course, genre, level); chunking aligned to criteria	Index recall@k; update drift; storage cost	Rebuild between terms; regression suite; corpus hash identifiers
Hybrid retrieval	Lexical+dense retrieval; query sanitation; intent detection; per-criterion candidate pools	Recall@5; nDCG@5; latency p95	Caching; throttling; abstention triggers
Reranking	Cross-encoder rerank; diversity penalty; policy filter; injection screening	nDCG gain; false-support rate	Red-team prompts; blocked-request rate

Tutor generation	Schema feedback (diagnosis→reason→plan→micro-examples); evidence links; refusal rules; style controls	Schema compliance; faithfulness; usefulness	Template fallback; human escalation
Monitoring	Citation verification; drift checks; subgroup reporting; incident response	Unsupported claim %; disparity Δ; incident frequency	Release gates; term-locked versions; periodic audit

Table 2

Evaluation matrix with quantitative targets and ablation baselines.

Evaluation dimension	**Recommended method**	**Metrics**	**Illustrative targets**	**Ablations / baselines**
Retrieval relevance	Instructor-labelled query set from assignments	Recall@5; nDCG@5; p95 latency	Recall@5 ≥ 0.85; nDCG@5 ≥ 0.75; p95 ≤ 900 ms	Dense-only vs hybrid; rerank on/off
Faithfulness	Claim-level audit on stratified samples	Citation-supported sentence %; unsupported claim %	≥ 90% supported; ≤ 5% unsupported	Citations enforced vs optional
Pedagogical utility	Blind ratings by writing instructors	Correctness; actionability; non-appropriation	Mean ≥ 4.0/5 on each	Schema vs free-form; example cap vs none
Learning transfer	Delayed post-task (new prompt) design	Effect size (g); retention	g ≥ 0.30; stable after 2–3 weeks	Tutor vs baseline feedback
Equity & fairness	Disaggregated analysis (proficiency, gender, campus)	Δgain; calibration error	Δ ≤ 0.10 SD; investigate outliers	Filter strategies; prompt variants
Governance	Policy logging + disclosure checks	Disclosure rate; violations/1k sessions	≥ 95% disclosure; ≤ 2 violations/1k	Process-visible vs traditional workflow

5 EVALUATION PROGRAMME

The evaluation must differentiate between capacities and safety and educational outcomes. Following the publication standards, the next sub-sections describe the evaluation programme from components and metrics to learning transfer and governance metrics. The last sub-section discusses fairness, as part of evaluation, which involves subgroup analysis and subgroup reporting.

5.1 Retrieval benchmarks and reranker ablations

The component evaluation is conducted using benchmark sets based on actual tasks annotated by the instructor according to the rubric dimensions of relevancy. For retrieval, the relevant metrics would be recall@k and nDCG@k. The efficiency metrics comprise latency and cache hits. As for reranking ablation, dense vs. hybrid retrieval, with and without reranking and intent filtering, are compared to determine contributing factors and identify potential regression from updating the system [5,7,8]. Draft snippets, code-switching prompts and ambiguous queries would help address possible noise in the query.

5.2 Faithfulness measures and claim-level audits

Both automated and human evaluations are required for a faithful evaluation. The automated evaluation includes citation and unmatched quote counts. The human evaluation is done by labelling of stratified samples according to the directives as fully supported, partially supported, or not supported by retrieved documents. See the list of targets in Table 2 tailored for each institution individually. The key takeaway here is that the unsupported claim rate can't be too low when all retrieved documents are irrelevant. Thus, it is critical that faithfulness audits are conducted together with relevance audit of retrieval.

5.3 Instructors' ratings of utility and non-appropriation

Instructors can give feedback about utility and appropriateness of feedback regarding correctness, actionability, and alignment with local rubric. Another metric could be appropriateness, defined as how far the instructors have exceeded the tutoring boundaries by giving students advice beyond just guidance and examples; no ghostwritten paragraphs are allowed. With schema-based generation of feedback, the evaluation of tutors' work is relatively straightforward. In Q1 publications, we would discuss the methodology of human evaluation in details, including the rater training protocol and inter-rater agreement statistics.

5.4 Designs for learning transfer and retention

Learning transfer and retention represent quite a sophisticated programme of evaluation associated with education. One can recommend conducting a design in which the participants will produce a pre-task writing, then revise it with the tutor's assistance, followed by post-task writing of the same genre but with another topic. The outcome measures could be effect sizes and uncertainty intervals for those. The checkup for retention would be conducted at least two weeks later. Such a design will address some of the critiques brought in systematic reviews on the focus on perception and immediate editing as opposed to learning [24].

5.5 Fairness diagnostics and subgroup reporting

For the evaluation of fairness, it is necessary to conduct subgroup reporting and analysis. The differences between subgroups, called deltas, could be calculated for proficiency levels, gender, campus and program types. If the deltas exceed 0.10 SD, one can research the reasons for that, which could be non-representativeness of the corpus or lack of alignment of the policy. To achieve greater fairness in relation to RAG tutors, one needs to add more appropriate genres, fix tagging and tune the policy on-the-fly if necessary. This practice would conform to the recommendation on the engineering approach to governance and fairness [17,19].

5.6 Governance metrics as validity evidence

Some of the indicators could measure the governance of the tutoring service. They are the disclosure rate (proportion of users who admitted the use of tutors), and blocked queries rate. Among others, the metrics could be the rate of human escalation and violation per session (for implementing the recommendation about transferring integrity review to workflow control [25]). There could be transparency prompts and citation source encouragement in the tutoring process.

6 DISCUSSION AND TECHNICAL RESEARCH AGENDA

Of course, technical challenges with high prospects for research in the field are abundant enough. Benchmarking is obviously one of them as there is no reliable retrieval and faithfulness benchmark for cumulative progress in writing tutor. On the contrary, de-identification is a technique, which can be used as a solution to tackle the problem, preserving confidentiality of the draft text at the same time. Paraphrasing with the faithfulness requirement sounds like another exciting challenge, but it will be more productive to combine quotation and entailment scoring methods. Bilingual evidence retrieval is one more challenge for investigation, though not likely as valuable as metadata engineering and standardization as well as corpus preparation. Last but not least, consistent reporting of hashes, versions, and release gates is necessary to ensure reproducibility [17].

7 CONCLUSION

The use of writing tutor based on evidence retrieval may be efficient for writing English in Saudi universities provided that the AI technology pipeline is reliable. Based on the literature of 2020-2025, the generation of texts by AI systems is assisted by external evidence [3,4]; thus, the evidence-based retrieval requires not only quantitative benchmarks but also reranking evaluation [5,7,8]. Pedagogical literature provides an idea that capability infrastructure should concentrate on transfer and governance measurements rather than perceived usefulness only [24,25,28]. In this regard, the paper offers four elements, including alignment with the model, architecture design (Figure 1), 80/20 problem statement (Figure 2), and synthesis tables (Tables 1-2). Future work should focus on creating benchmarks, paraphrasing techniques, and measuring transferability with subgroup analysis and governance measurements.

Addendum: Improvement of reproducibility of writing tutor practical implementation would mean creating a configuration sheet containing IDs for the embedding model, ANN search policy, reranking algorithm, token budget for each field, and refusal policy.

Addendum: The cost estimate may be obtained based on cost breakdown and analysis of latency and sensitivity of the cost structure. One can decrease the costs of implementing evidence retrieval by skipping the reranking step and changing the evidence chunk size.

Addendum: Stratification of data is useful for the design of audit procedures, implying stratification of the draft text based on genre, proficiency, and location. Moreover, rules for auditing have to be specified in advance to provide pre-registration and minimize statistical bias.

Addendum: Logging will allow restricting content exposure by keeping only the signal rather than the whole draft text, e.g., revision numbers, rubric dimensions, citation checks, etc.

REFERENCES

Alawwad HA, Alhothali A, Naseem U, Alkhathlan A, Jamal A. Enhancing textual textbook question answering with large language models and retrieval-augmented generation. *Pattern Recognition*. 2025;162:111332.

Bender EM, Gebru T, McMillan-Major A, Shmitchell S. On the dangers of stochastic parrots: can language models be too big? *FAccT*. 2021.

Bittle K, El-Gayar O. Generative AI and academic integrity in higher education: a systematic review. *Information*. 2025;16(4):296.

Bommasani R, Hudson DA, Adeli E, et al. On the opportunities and risks of foundation models. 2021.

Brown TB, Mann B, Ryder N, et al. Language models are few-shot learners. *NeurIPS*. 2020.

Dettmers T, Pagnoni A, Holtzman A, Zettlemoyer L. QLoRA: efficient finetuning of quantized LLMs. 2023.

European Commission. *Ethical guidelines on the use of AI and data in teaching and learning for educators*. 2022.

European Parliament and Council. *Artificial Intelligence Act*. 2024.

Guu K, Lee K, Tung Z, Pasupat P, Chang M. REALM: retrieval-augmented language model pre-training. *ICML*. 2020.

Hu EJ, Shen Y, Wallis P, et al. LoRA: low-rank adaptation of large language models. 2021.

ISO/IEC. *ISO/IEC 27001: information security management systems—requirements*. 2022.

Karpukhin V, Oguz B, Min S, et al. Dense passage retrieval for open-domain question answering. *EMNLP*. 2020.

Kasneci E, Sessler K, Küchemann S, et al. ChatGPT for good? Opportunities and challenges for education. *Learning and Individual Differences*. 2023;103:102274.

Lewis P, Perez E, Piktus A, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. *NeurIPS*. 2020.

Lo CK. Impact of AI writing tools on student learning: a systematic review. *Computers and Education: Artificial Intelligence*. 2024;5:100163.

Nguyen Thi XH, Hoang Thien HV, Vuong KN, Nguyen TT. Enhancing writing skills through AI-powered tools: perceived benefits and challenges among EFL students. *Discover Education*. 2025;4:472.

NIST. *Privacy framework 1.0*. 2020.

NIST. *Secure software development framework (SSDF)*. 2022.

NIST. *Artificial intelligence risk management framework (AI RMF 1.0)*. 2023.

Nogueira R, Jiang Z, Lin J. Document ranking with a pretrained sequence-to-sequence model. *Findings of EMNLP*. 2020.

OpenAI. *GPT-4 technical report*. 2023.

Ouyang L, Wu J, Jiang X, et al. Training language models to follow instructions with human feedback. *NeurIPS*. 2022.

Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*. 2021;372:n71.

Saudi Ministry of Education. *Higher education and research strategy aligned to Vision 2030*. 2020.

Saudi Vision 2030. *Vision 2030 annual report 2023*. 2024.

Snyder H. Literature review as a research methodology: an overview and guidelines. *Journal of Business Research*. 2020;104:333–339.

Thakur N, Reimers N, Daxenberger J, et al. BEIR: a heterogeneous benchmark for zero-shot evaluation of information retrieval models. 2021.

UNESCO. *AI and education: guidance for policy-makers*. 2021.

UNESCO. *Guidance for generative AI in education and research*. 2023.

Zhai X. ChatGPT user experience: implications for education. *Computers and Education: Artificial Intelligence*. 2022;3:100085.

Authors' Contribution

All authors contributed equally to the development of this article.

Data availability

All datasets relevant to this study's findings are fully available within the article.