

Artificial Intelligence-Based ESP (English For Specific Purposes) Tutor Applications For Science And Engineering Undergraduate Students In Saudi Arabia

Asad Shafi

University of Prince Mugrin, Madinah, Saudi Arabia.

Abstract

While the emergence of artificial intelligence brings changes to the domain of academic assistance, the presence of academic communication continues in the contexts of foreign language learning, information studies, and human computer interaction. This work considers the usage of AI-driven computing applications used by university students for their assignments related to English for Specific Purposes in science and engineering courses at Saudi higher education establishments. Though the emphasis is made on scientific studies focused on advances in terms of better searching algorithms and language models, no attention is given to chatbot systems that provide students with generic tips on how to write essays. The purpose of this study is to discuss important features involved in designing, teaching, evaluating, and the contextual application of AI tutoring tools. Academic journals and papers written within the scope of interest from 2020 until 2025 were obtained from scholarly indexes. Next, the sources were subject to thematic narrative analysis. Thus, even though the capability of open-ended text production seems promising, computing designs based on fine-tuning language models, embeddings, vector searches, segmenting/indexing approaches, hybrid searches, re-ranking, and prompt handling appear the most reliable way to build an AI tutor in terms of contextualized comments on grammar, coherence, argumentative writing, citation, and writing style. Moreover, it can be stated that hallucination prevention, appropriate source referencing, fairness, security, and Arabic-English dual language tutoring are key principles for building AI tutors. However, as observed from the literature review, there are still inconsistencies concerning criteria and measuring the assessment of writing skills, autonomy, and management.

Keywords: ESP writing; retrieval-augmented generation; large language models; academic writing support; Saudi higher education; vector databases; AI tutors; Arabic-English context.

1. Introduction

The ability to write in English for specific purposes is highly significant in undergraduate science and engineering programs, as students have to articulate their disciplinary knowledge through reports, proposals, technical explanations, project writing, and research tasks. Given that Saudi Arabia is increasingly adopting English-mediated materials and focusing on improving the digital capabilities of its citizens as well as innovation and competitiveness, there is a need for students to develop their proficiency in English academic writing. Students experience challenges especially with overall writing accuracy and coherence due to their limited ability to apply appropriate vocabulary and compose well written writing tasks (Mohamed, 2024; Alsofyani & Barzanji, 2025). AI systems are rapidly emerging and have brought about discussions regarding how such support systems could be implemented effectively. Although the potential of using generic chatbots for brainstorming, paraphrasing, and superficial revision has proven so far to be significant, this type of assistance has also been found to be associated with a number of fundamental drawbacks, including unreliable facts in generated text, inadequate citation practices, poor alignment with course goals, and a tendency to produce new texts rather than supporting students' ideas (Kohnke et al., 2023; Levine et al., 2024; Warschauer & Xu, 2024). It should be pointed out that these issues become critical when considering ESP writing in the fields

of science and engineering, where writing is evaluated according to set criteria, among which are fluency, accuracy of technical vocabulary, adherence to academic conventions, and validity of proposed arguments. Hence, this review focuses on ESP writing in terms of architectural principles in relation to the development of AI-based computer-assisted applications. Namely, it is expected that efficient tutoring systems for academic writing will increasingly adopt retrieval-augmented generation, or similar architectures based on grounded pipelines (Lewis et al., 2020; Li et al., 2025). These systems integrate language models with databases and repositories containing curated corpora, rubrics, examples, technical vocabulary glossaries, and other necessary information to provide context-awareness, and most importantly, transparent feedback. The purpose of this paper is to conduct a literature review on the topic of ESP tutoring using AI-based computer applications. Specifically, it seeks to address four goals, namely to provide: an overview of available AI architectures used for tutoring; an explanation of how these architectures allow the generation of feedback across various dimensions of academic writing; highlights of special deployment issues pertaining to Saudi universities and bilingual writing practices; and a discussion of the gaps in evaluating these technologies as well as promising directions for future research.

2. Background / Related Literature

The main characteristic distinguishing ESP writing from other types of language writing is the specific context, genre, and vocabulary of ESP writing. For example, in various fields, writers need to describe certain processes and procedures, argue for and against some decisions, compare alternative variants, and summarize information. Apart from the ability to form correct grammatical structures, it is also necessary to properly organize writing and to employ certain rhetorical devices, attitude, citation and referencing practices. Thus, the application of any kind of automated proofreading tools would not be adequate in this case. All pre-generative AI writing tools are developed in the framework of such areas as writing assessment automation, grammar correction, generation of templates, and provision of corpus-based feedback. Undoubtedly, the above approaches have contributed greatly because of the fact that they have helped eliminate a number of problems related to style, grammar, and other issues associated with writing. At the same time, little effort has been made in terms of teaching people how to write effectively. Nowadays, modern language models have created new opportunities to address this problem and generate feedback for learners in the form of examples, explanations, and other forms of assistance. However, even the increase in volume of generated text does not necessarily bring positive results if learners fail to define the limits of their knowledge (UNESCO, 2023; Levine et al., 2024). Accordingly, two roles of AI generators should be distinguished. Firstly, they may serve as a fast means of learning and generating texts. Secondly, they could be used as an instrument which provides learners with opportunities for productive struggling and improvement of skills (Ali et al., 2023; Wang et al., 2024). It would be ideal if learners received immediate personal assistance; nevertheless, it is also crucial to consider the need to provide them with reliable feedback and help them develop in-depth knowledge. For instance, RAG is essential because it solves one of the major problems associated with LLM operation, namely, its inability to generate guaranteed accurate answers. By employing RAG, it is possible to ensure that generated answers could be based on information from reliable sources, including assignment instructions, examples, and specific vocabulary (Lewis et al., 2020; Li et al., 2025). Therefore, it is necessary for feedback to correspond with local expectations. The existing body of literature about generative AI could be summarized by defining its key subcomponents that would guarantee a tutor success: dense retrieval, vector indexing, chunking, re-ranking, and

prompt design (Karpukhin et al., 2020; Pan et al., 2024; Lee & Palmer, 2025; Qian, 2025). Meanwhile, scientific studies devoted to the use of generative AI in writing instruction mostly concentrate on such issues as revision uptake, quality of examples and feedback, and explanations (Nair et al., 2024; Maurya et al., 2025). However, usually, these components are discussed separately. This literature review aims to fill this gap in relation to Saudi higher education establishments.

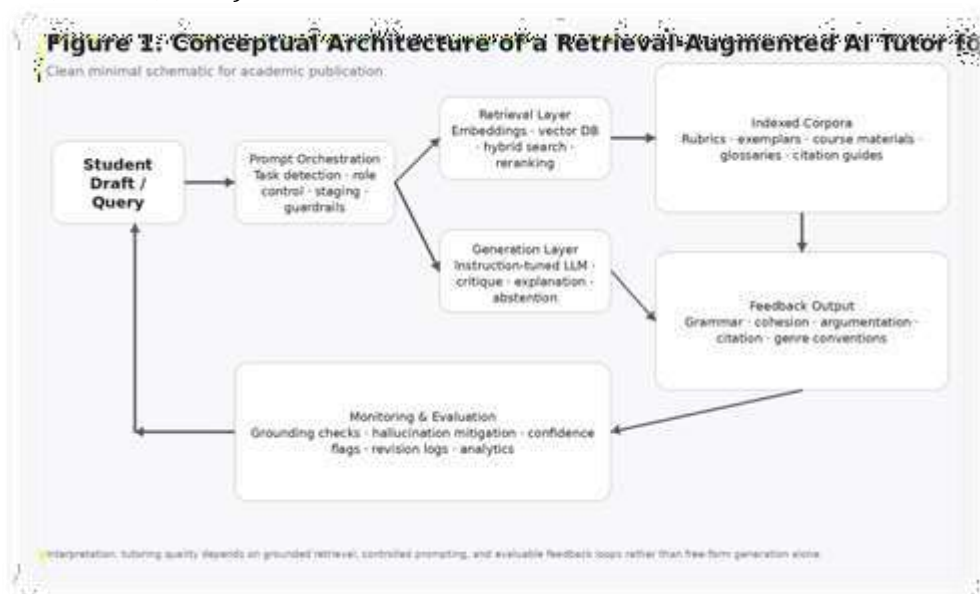
3. Methodology of the Review

The chosen approach to review the cross-disciplinary topics included topics such as natural language processing, educational technology, academic writing, and higher education policies. The aim of the present paper is to investigate how artificial intelligence-powered computers could assist in developing ESP academic writing skills for science and engineering undergraduate students studying in Saudi Arabia using the example of retrieval-enhanced tutoring systems and their features. The main research questions that guide the development of content for the paper include the following. Firstly, technical aspects of the AI-powered tutors that are able to generate writing feedback are examined, namely choosing language models, knowledge retrieval, embeddings, vector database, reranking, prompt generation, and hallucinations reduction. Secondly, it is determined if these systems encompass the whole writing process or just some elements, such as grammar checking, coherence, argumentation, citation styles, and editing for different types of writing. Thirdly, the relevant contextual factors in Saudi Arabian higher education settings (privacy, ethics, fairness, and multilingualism) are considered. Finally, the AI tools are evaluated, and the weaknesses associated with scientific studies and assessments of their performance are revealed. In order to find the sources necessary for carrying out the paper, the following databases are utilized: Scopus, Web of Science, ACM Digital Library, IEEE Xplore, ACL Anthology, ScienceDirect, SpringerLink, ERIC, and international education policies. The key search terms that are used when browsing databases to gather sources are the following: "retrieval augmented generation", "language model tutor", "AI writing feedback", "ESP academic writing", "engineering students", "Saudi higher education", "Arabic essay scoring", "vector database", "reranking", and "prompt generation". Backward and forward citations are applied to check references' appropriateness and eliminate any duplications from the list of literature used to create this paper. The inclusion criteria of the sources for this paper are the following. First of all, they should be reliable and have been published between 2020 and 2025. Second, sources addressing one topic of interest are desirable. Preferably, the sources included in the list should be scholarly articles or reports from prestigious conferences or systematic reviews, as well as educational policy documents. The exclusion criteria include the following aspects: duplication, lack of credibility, relevance to academic writing or tutoring, and the role of AI in education in general as it is not relevant to this paper conceptually and technologically. Thematic analysis rather than data synthesis has been adopted in this paper. Given the significant heterogeneity of the selected sources in terms of methodology, subject matter scope, technical complexity, and meta-analysis do not seem appropriate for this paper. In this respect, articles selected are grouped into five groups: architecture, knowledge retrieval and indexing design, writing functionalities, assessment, and Saudi Arabia context in tutoring applications.

4. Architecture of Retrieval-Augmented AI Tutors

Unlike the standard definition of an AI tutor as a simple function, an ESP writing tutor should be treated as an architectural entity that consists of several layers. Initially, it is possible to mention the frontend layer that receives submissions from the learner. It can

be either a draft, a certain prompt or a particular question. Secondly, it is worth mentioning the orchestration layer that identifies the nature of the task, clarifies what the learner intends and chooses the corresponding retrieval tool based on the query performed on the indexed data. Then, the retrieval layer performs searching through the indexed data. Next, synthesis takes place depending on the data submitted by the two previous layers. Lastly, the monitoring layer involves tracking and logging the output, citing materials used and keeping track of revision and confidence processes (Lewis et al., 2020; Li et al., 2025).



One of the aspects to pay attention to when discussing instruction-tuned language models for tutoring is generation as it allows performing a wide range of tasks relating to tone, identity and procedure of writing. However, generation is not the only skill expected from such an AI tutor. Controlled dialogues, explanations, no action in cases of lacking knowledge, and critical feedback with no paraphrasing are also needed. Currently, it is evident that an efficient tutor is supposed to demonstrate the skills of instructing and providing feedback by asking questions, offering help with understanding the topic better, checking for comprehension, and so forth. When speaking to the ESP academic writing tutor, this means teaching it how to act as a guide, an analyst, and a citation coach. Furthermore, the ability of self-rankers to combine retrieval and critique with generation proves its validity (Asai et al., 2024). In this regard, embedding and vectorization are extremely important since the latter allows generating dense vector representations of documents that allow identifying semantically similar passages regardless of having the same wording. The knowledge base of the ESP tutor can consist of assignment descriptions, departmental rubrics, glossaries, citations guidelines, and samples of work in different genres. By using vectorization for approximate searches, a tutor can access information within certain latency, filtering the results by course names, genres or topics (Pan et al., 2024). The major benefit of dense retrieval in ESP tutoring is that semantically similar passages are identified irrespective of awkward wording and incomplete usage of specialized vocabulary. However, there is also one drawback – dense retrieval fails to detect all lexical similarities since academic texts involve many terms related to the topic, specific sections and specialized vocabulary. For this reason, hybrid retrieval appears more relevant for ESP tutoring in particular. While dense retrieval is the best method for detecting semantically similar passages, keyword-based search (sparse) is much more

accurate regarding detecting certain lexical elements (Karpukhin et al., 2020). Thus, in ESP tutoring, an approach consisting of dense and keyword-based search followed by reranking can be implemented. For example, while dense retrieval allows to find passages with semantically similar guidance, a sparse method is necessary to find passages that include all scientific terms such as "uncertainty", "calibration" or "finite element analysis". Dense and sparse techniques together with generation have proven to be highly efficient in hybrid architectures (Doan et al., 2024; Yu et al., 2024). Lastly, reranking has become one of the key factors influencing the efficiency of both retrieval and generation. Redundancy in the knowledge base of the tutor and vagueness in the queries submitted by the learner lead to retrieving semantically and lexically similar passages in larger quantities. One of the solutions is late-interaction approaches like ColBERT and ColBERTv2 (Khattab & Zaharia, 2020; Santhanam et al., 2022). Yet, reranking becomes crucial for another reason – choosing the most appropriate passages for the subsequent process of generation. Besides, chunking and indexing can also create some challenges related to the effectiveness of retrieval. Firstly, the created rubrics and models for ESP tutors cannot fit retrieval processes in terms of being appropriate. Taking into account the hierarchical structure of the models and indexing, as well as the distribution in their coarse-grained form across multiple pages, retrieved documents would be too large to use in the context of a particular task. Secondly, concerning fine-grained indexing, there is no doubt that it cannot provide enough context since, even though it provides relevant technical documents, it will not show how to perform the task properly. Therefore, contemporary science suggests including techniques of adaptive chunking and mix-of-granularity in the process of retrieval since every task requires a specific context (Zhong et al., 2025). In terms of a writing tutor, for instance, a technique of hierarchical chunking and parent-child indexing should be used. As for sentence-level chunking related to hedging in the discussion of results, an instructor will receive adequate information through the parent document due to the broader context analysis. This approach will be helpful while analyzing cohesiveness of paragraphs and other rhetorical aspects in the document. Orchestration might also serve as an important aspect of the tutor's architecture. Orchestration refers to the process of using documents for performing various actions. While prompt design could be regarded as the art of wording, there are also many other complicated factors that need to be taken into consideration, such as the model's purpose, appropriate task decomposition, restrictions of using evidences, required output format, etc. According to Lee & Palmer (2025) and Qian (2025), structured prompting plays a great role in a model's control by means of task decomposition and prompt-objective alignment. Possible orchestration pipelines could include determining genre and the purpose of the document, analyzing writing difficulty level, retrieving relevant documents, grammar and cohesiveness analysis, argumentation analysis, citation analysis, and giving instructions for changing student's text and its rewriting. Hence, orchestration will help prevent students from overwhelming documents or giving them instructions like "refer only to retrieved documents" and "make the needed changes in the text before rewriting the passage". Also, modularization might prove itself as an effective way of generating feedback. Indeed, while analyzing sentence accuracy, discourse cohesiveness, citation analysis, and genre guidance, one cannot neglect the evidences needed. To begin with, it is impossible to correct grammatical mistakes without taking into consideration local context analysis. Paragraph analysis should take into account a paragraph as a unit. Citation analysis involves knowledge of citation rules and stylistic norms. Genre guidance requires analyzing conventions of each part of the document. Regardless of context, prompts will result in poor feedback generation effects. Nevertheless, orchestrating will help collect any necessary data for generating feedback (Nair et al.,

2024). Furthermore, when considering the choice of the architecture of the tutor, the mitigation of hallucinations should be taken into account as well. Indeed, models of language have the capability of generating references, inferring some facts from context, and suggesting incorrect writing approaches. Sources of hallucinations are due to the probabilistic nature of language generation, inadequate grounding, and rewards for fluency (Huang et al., 2025; Zhang et al., 2025). Mitigation approaches include retrieval grounding, confidence-awareness responses, and verification. According to research related to fluency benchmark testing, fluent models generate inaccurate information; hence, educators should concentrate on creating grounded models (Lin et al., 2022). In addition, choosing an appropriate architectural component for a tutor is necessary to prioritize either the quality of the generated answers or the educational experience. Even though the generated answers provided by the rewriting model are appropriate, the problem lies in the student's inability to assimilate them and put them into practice. Thus, explanations, critique, and recommendations should become the integral parts of the educational system. It is evident that self-directed learning and AI-based writing tools that allow reflection and rewriting rather than substitution are more useful for students' development (Ali et al., 2023; Wang et al., 2024). Thus, the design of the tutor should include the following principles: minimal number of rewrites, changes' justification, and several rewriting options.

Design layer	Main options	Strengths for ESP writing	Risks / trade-offs	Best-fit use in science and engineering ESP
Base LLM role	General assistant; instruction-tuned tutor; critique-oriented tutor	Instruction-tuned and critique-oriented models support explanation, revision guidance, and controlled feedback	General assistants may overwrite student voice or generate unsupported advice	Use a tutor-role model configured to diagnose, explain, and scaffold rather than ghostwrite
Knowledge grounding	Standalone LLM; RAG with local corpus; RAG with institutional corpus plus web-disabled mode	Corpus-grounded RAG improves factuality, citation alignment, and genre relevance	Poor corpus design leads to weak retrieval and misleading feedback	Use local curated corpora: rubrics, model reports, glossaries, citation guides, assignment briefs
Embeddings	General multilingual embeddings; domain-adapted embeddings; hybrid semantic +	Semantic retrieval helps with paraphrased student queries and discipline concepts	Embeddings may blur exact terminology or section-specific intent	Combine multilingual semantic indexing with course, genre, and discipline metadata

Design layer	Main options	Strengths for ESP writing	Risks / trade-offs	Best-fit use in science and engineering ESP
	metadata indexing			
Vector storage	Basic vector index; ANN vector DB; vector DB with metadata filters and versioning	Fast retrieval at scale and traceable source segmentation	Weak metadata control can surface irrelevant or outdated materials	Use vector DB with filters for course, week, genre, source type, and approved version
Chunking strategy	Fixed-size; semantic chunking; hierarchical; parent-child; mix-of-granularity	Better context control for long rubrics, model texts, and section-level writing guidance	Chunks that are too small lose rhetorical context; too large waste prompt space	Prefer hierarchical or mix-of-granularity chunking for reports, lab write-ups, and proposal genres
Initial retrieval	Dense retrieval; sparse lexical retrieval; hybrid retrieval	Hybrid retrieval balances semantic similarity with exact technical terms	Dense-only retrieval may miss formulaic phrasing; sparse-only retrieval may miss paraphrases	Use hybrid retrieval for terminology-heavy engineering and science writing
Reranking	None; cross-encoder; late interaction; LLM-based reranking	Improves relevance of retrieved passages entering the prompt	Extra latency and complexity	Use reranking when feedback must align tightly with rubric criteria and target genre
Prompt orchestration	Single prompt; chained prompts; tool-routed multi-stage prompts	Multi-stage orchestration supports diagnosis, feedback categorization, and revision sequencing	Over-complex workflows may increase latency or instability	Use staged prompts: task detection, evidence retrieval, feedback by category, revision request

Design layer	Main options	Strengths for ESP writing	Risks / trade-offs	Best-fit use in science and engineering ESP
Feedback scope	Grammar only; grammar + cohesion; full writing analytics including argumentation, citation, and genre	Broader scope better matches ESP writing demands	Larger scope increases need for grounded evidence and careful prioritization	Use modular feedback covering grammar, cohesion, argumentation, citation, and genre conventions
Hallucination control	None; citation-required output; abstention rules; post-generation verification	Reduces fabricated sources and overconfident advice	Strict controls may shorten outputs or increase “unable to answer” cases	Require source-grounded feedback and abstention when no adequate evidence is retrieved
Learner interaction model	One-shot correction; iterative coaching; reflective tutoring with revision loop	Iterative coaching promotes uptake, autonomy, and transfer	Students may still seek direct rewriting if guardrails are weak	Use reflective tutoring that asks for revision before offering polished reformulation
Deployment model	Public API chatbot; institution-managed system; privacy-preserving local or hybrid deployment	Institution-managed models better support governance and auditability	Local systems require infrastructure and maintenance	Prefer institution-governed deployment for Saudi higher education contexts involving student writing data

5. Applications to Academic English Writing

When implemented in the academic English writing domains, the retrieval-augmented system will function as an assistant not only while editing texts, but at all stages of text composition. During the planning stage, the tutor will assist students to understand the prompts for writing, distinguish the genre category, and structure the text with respect to its communicative purpose concerning a certain lab report, design report, proposal, or technical memorandum. While creating a draft, the AI tutor will support students to comprehend how to formulate purposes, positions, and arguments for separate paragraphs instead of generating the whole text. Revision assistance will consist of finding repeated errors and linking them to certain rubrics. A staged assistance

approach will work better than a generative one because it maintains learners' agency and makes the process of the tutors' actions understandable (Warschauer & Xu, 2024; Li et al., 2025). Dimensional specificity of feedback seems crucial when applying retrieval to language learning situations. While it will be possible for the tutor to identify problems with grammatical structures and usage of articles, prepositions, tenses, and punctuation, it would not be reasonable to provide excessive comments on grammar use as it may be, for example, less complicated than other aspects of writing. It is important to pay attention to such characteristics of written discourse such as cohesion and coherence in professional texts. Learners should learn how to develop logical connections between various parts of their texts and construct links between claims and ideas. Using the retrieval-augmented tutor, learners will become familiar with the various types of cohesion and compare them to models found in particular discourses. Regarding argumentation, the tutor will be able to analyze every point made by students, its limitations, difference between claims, and so forth. Citation practices with the use of AI systems are also worth paying attention to because it is essential to learn how to attribute sources and integrate citations and quotations in texts. The retrieval-augmented tutoring system has more advantages than general chatbots with regards to genre conventions. In such courses as science and engineering, it will be necessary to assist learners in terms of comprehending the ways in which texts meet the criteria of a particular part of a given scientific paper. Is the abstract informative? Is the problem space set up in the introduction? Is the procedure described adequately in the 'Methods' section? Are the results analyzed in the 'Discussion'? This kind of feedback depends on the conventions related to certain genres and, therefore, requires the use of retrieval of specific examples or explanations. Research has shown the high level of satisfaction of ChatGPT users who receive both explanations and examples of texts in language learning situations. However, some students require additional feedback about the nature of conventions (Levine et al., 2024; Lo et al., 2024). Finally, it should be noted that the proposed approach has the potential for metacognitive support. Using the retrieval-augmented tutoring system, it will be possible to teach revision strategies and encourage critical thinking skills among students. To be more precise, students may be asked to assess the necessity of revision, analyze two strategies of revision, and justify the presence of certain criteria in a rubric. It is important to create such possibilities in order to promote learners' ability to implement the right strategies, because in order to succeed in writing, it is not only vital to be aware of a model, but to be able to repeat it (Nair et al., 2024; Maurya et al., 2025). In ESP, it means providing learners with both suggestions and explanations.

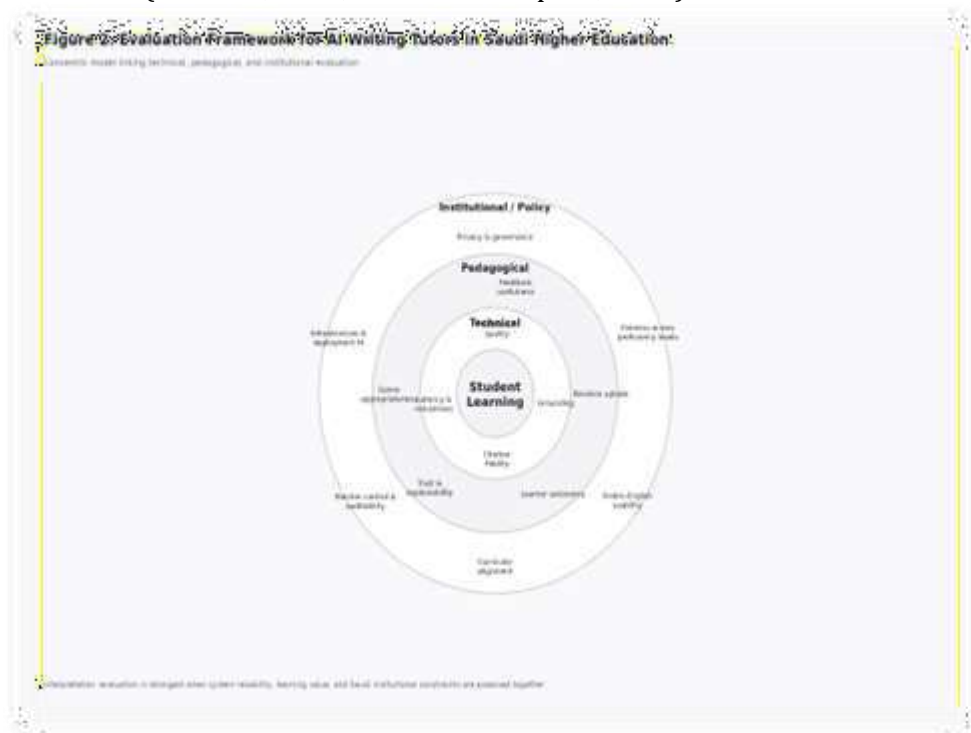
6. Saudi Higher Education Context

The Saudi context is one of the primary reasons for why this particular issue appears meaningful and relevant. The development strategies included digital transformation, preparation for a knowledge economy, and improvement of higher education, which explains the pertinence of the proposed improvement in the Saudi context (Saudi Vision 2030, 2025). Nonetheless, students who study in the English-language, use English-language software, and produce their coursework in English in predominantly an Arabic-speaking environment. Thus, developing a tutor to support bilingual learners is challenging since although learners generate their ideas in Arabic, conduct searches using cross-language queries, and write their essays in English, they cannot comprehend such intentions; thus, a bilingual tutor may appear inferior to a monolingual one. In addition, fairness issues might surface. It should be noted that the script language has an impact on the test-takers behavior as not all Saudi science and engineering students are equal in terms of familiarity with conversational English, discourse of science and

engineering, and AI-driven writing tools (Ghazawi & Simpson, 2025). This means that when working on their essays in a bilingual environment, students may experience difficulties with testing because the writing tutor is intended to help them practice before taking a real exam. Therefore, great care should be exercised when introducing this solution, including retrieving information and the evaluation of such tools. Lastly, there is also data security and data management issues. Any draft papers, coursework, and other educational materials are confidential and personal; therefore, data management will be crucial in this context. For this reason, it will be necessary for Saudi universities to specify how data will be handled in terms of access to the database and involvement of any third parties in data processing. This point needs special attention in light of the recent guidelines issued by UNESCO concerning the responsible use of AI in education (UNESCO, 2023). Other unaddressed issues relate to integrating it into such a curriculum and evaluating it (Sobaih et al., 2024; Alsofyani & Barzanji, 2025).

7. Evaluation Framework and Research Gaps

There are multiple ways of assessing AI-based tutors. Regarding the technological aspect of the assessment, the following parameters should be considered: information retrieval efficiency, ranking the sources correctly, citing sources, speed, robustness to prompts, and tendency to provide false answers. Regarding the pedagogic component, the following criteria need to be measured: accuracy, effectiveness, quality of explanation, user engagement, and promotion of independent thinking. From an institutional standpoint, evaluation would cover issues of data privacy, evaluation fairness, tutor control, and alignment of the tutoring system with the curriculum. Current research into evaluating tutors advocates for considering these criteria in detail (Maurya et al., 2025). Nonetheless, inconsistencies in evaluation are obvious based on recent research (Seo et al., 2025; Ghazawi & Simpson, 2025).



From this literature review, several research gaps become apparent. First, there is no specialized benchmark available for evaluating tutors for ESL writing in science and

engineering. Second, insufficient research focuses on the bilingual aspect of the tool because it is quite relevant. Third, scholars seem to pay less attention to retrieving-based writing tutors than conversational chatbots. Fourth, the focus of current research seems to lie in correcting the syntax of writing, but not long-form writing. Fifth, many studies into adoption mainly explore its acceptance, while other criteria need addressing. Finally, little is known about independent writing practices.

Gap area	Current limitation in the literature	Evaluation criteria needed	Future direction
Genre-aware tutoring	Many studies assess generic chatbot use rather than discipline-specific ESP writing support	Accuracy of genre diagnosis; section-level appropriateness; disciplinary terminology control	Build benchmarks for lab reports, technical reports, design proposals, abstracts, and engineering discussion sections
Retrieval quality in education	Educational RAG studies often discuss architecture broadly but underreport retrieval behavior in writing tasks	Precision@k, reranking gain, source coverage, rubric alignment, evidence traceability	Compare dense, sparse, hybrid, and reranked pipelines specifically for writing-feedback tasks
Chunking for long instructional documents	Rubrics, exemplars, and writing guides are long and structurally complex	Retrieval completeness, context preservation, prompt efficiency, section fidelity	Test hierarchical and mix-of-granularity chunking on ESP corpora
Feedback usefulness	Output fluency is often treated as a proxy for pedagogical value	Actionability, clarity, explanation quality, revision uptake, student comprehension	Evaluate whether feedback leads to improved revision rather than better-looking machine text
Argumentation and evidence use	Grammar feedback is more common than support for claims, support, and citation practice	Argument coherence, evidence relevance, citation fidelity, source integration	Develop rubric-linked feedback modules for argumentation and source use
Hallucination mitigation	Many systems still allow unsupported or fabricated guidance	Grounding rate, unsupported-claim rate, abstention quality, citation verification	Combine retrieval grounding, confidence labeling, and post-generation checking
Arabic-English mediation	Cross-lingual and bilingual writing workflows remain underexplored	Cross-lingual retrieval quality, terminology mapping, bilingual usability, fairness	Develop Arabic-English corpora and evaluation protocols for Saudi university writing contexts

Gap area	Current limitation in the literature	Evaluation criteria needed	Future direction
Fairness and learner diversity	Limited analysis of performance across proficiency levels and educational backgrounds	Equity across proficiency bands, accessibility, bias auditing, consistency across cohorts	Audit systems across gender, major, proficiency, and prior AI experience without overclaiming causal effects
Privacy and governance	Adoption research often emphasizes acceptance more than institutional control	Data minimization, audit logs, retention policy, instructor oversight, vendor transparency	Create university-level governance frameworks for AI writing support
Teacher-AI collaboration	Systems often frame AI as a student-only tool	Instructor control, rubric injection, feedback override, review traceability	Design human-in-the-loop tutor systems with configurable pedagogical policies
Long-form writing evaluation	Short answers and surface correction are easier to assess than extended academic writing	Document-level coherence, rhetorical progression, consistency across sections	Build evaluation tasks for extended reports and project documentation
Educational impact measurement	Many studies rely on perceptions or short-term outcomes	Transfer, autonomy, reflective use, responsible AI literacy, sustained writing improvement	Run longitudinal, review-method-informed research programs linking technical design to writing development

8. Future Research Agenda

Future work should move beyond general-purpose LLMs towards tutor ecosystems that would operate in domain-specific and institution-controlled settings. To start with, an important step will be the development of Arabic-English retrieval benchmarks and corpora designed specifically for writing within the genres of science and engineering in Saudi higher education. Secondly, future research should involve architecture comparisons between chunking, retrieval, and orchestration strategies applied to specific cases of writing feedback. Finally, further research will benefit from the incorporation and analysis of more varied assessments, including writing revisions, generalizability between assignments, reflection, and teachers' interpretation of feedback. Collaborative human-AI approaches deserve consideration as well. Tutors will need to have architectures allowing teachers to configure them, including specifying rubrics, sources of information, analysis of feedback trails, and severity thresholds. On the other hand, there should also be future work involving lighter deployment strategies, such as on-premises installations and privacy-preserving options, which would be relevant to institutions unable to rely on commercial APIs. Finally, as pointed out in Lee and Palmer (2025) and Qian (2025), prompt engineering as an educational problem will require designing tutors that encourage training students how to query AI

and not to comply with prompt designs blindly. Therefore, a better approach for future work will be based on developing tutoring systems for augmentation.

9. Conclusion

Educationally credible computer programs based on AI techniques will prove effective in the context of ESP writing if they use LLM fluency, retrieval strategy, structure of prompts, modular feedback, and evaluation strategies. When it comes to Saudi Arabian undergraduate students writing about science and engineering topics, the key challenge will not be based on the availability of text generation technologies, but in the design of systems that enable writing support in a disciplinary manner, clearly and equitably. According to the results of the review of the literature, the most likely way to create such a system will be through the use of retrieval-augmented architectures. Nonetheless, mere technical sophistication will neither suffice nor be successfully implemented.

References

1. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems*, 33, 9459–9474.
2. Karpukhin, V., Oguz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., & Yih, W.-t. (2020). Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 6769–6781). Association for Computational Linguistics.
3. Khattab, O., & Zaharia, M. (2020). ColBERT: Efficient and effective passage search via contextualized late interaction over BERT. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 39–48). ACM.
4. Santhanam, K., Khattab, O., Saad-Falcon, J., Potts, C., & Zaharia, M. (2022). ColBERTv2: Effective and efficient retrieval via lightweight late interaction. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 3715–3734). Association for Computational Linguistics.
5. Lin, S., Hilton, J., & Evans, O. (2022). TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 3214–3252). Association for Computational Linguistics.
6. Ali, F., Choy, D., Divaharan, S., Tay, H. Y., & Chen, W. (2023). Supporting self-directed learning and self-assessment using TeacherGAIA, a generative AI chatbot application: Learning approaches and prompt engineering. *Learning: Research and Practice*, 9(2), 135–147.
7. Kohnke, L., Moorhouse, B. L., & Zou, D. (2023). ChatGPT for language teaching and learning. *RELC Journal*, 54(2), 537–550.
8. UNESCO. (2023). Guidance for generative AI in education and research. UNESCO.
9. Asai, A., Wu, Z., Wang, Y., Sil, A., & Hajishirzi, H. (2024). Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *Proceedings of the International Conference on Learning Representations (ICLR 2024)*.
10. Pan, J. J., Wang, J., & Li, G. (2024). Survey of vector database management systems. *The VLDB Journal*, 33, 1591–1615.
11. Warschauer, M., & Xu, Y. (2024). Generative AI for language learning: Entering a new era. *Language Learning & Technology*, 28(2), 1–4.

12. Levine, S., Beck, S. W., Mah, C., Phalen, L., & Pittman, J. (2024). How do students use ChatGPT as a writing support? *Journal of Adolescent & Adult Literacy*. Advance online publication.
13. Wang, C., Li, Z., & Bonk, C. J. (2024). Understanding self-directed learning in AI-assisted writing: A mixed methods study of postsecondary learners. *Computers and Education: Artificial Intelligence*, 6, 100247.
14. Mohamed, A. M. (2024). Exploring the potential of an AI-based chatbot (ChatGPT) in enhancing English as a foreign language teaching: Perceptions of EFL faculty members. *Education and Information Technologies*, 29, 3195–3217.
15. Sobaih, A. E. E., Elshaer, I. A., & Hasanein, A. M. (2024). Examining students' acceptance and use of ChatGPT in Saudi Arabian higher education. *European Journal of Investigation in Health, Psychology and Education*, 14(3), 709–721.
16. Lo, C. K., Yu, P. L. H., Xu, S., Ng, D. T. K., & Jong, M. S.-Y. (2024). Exploring the application of ChatGPT in ESL/EFL education and related research issues: A systematic review of empirical studies. *Smart Learning Environments*, 11, 50.
17. Lee, S., Cai, Y., Meng, D., Wang, Z., & Wu, Y. (2024). Unleashing large language models' proficiency in zero-shot essay scoring. In *Findings of the Association for Computational Linguistics: EMNLP 2024* (pp. 181–198). Association for Computational Linguistics.
18. Nair, I. J., Tan, J., Su, X., Gere, A., Wang, X., & Wang, L. (2024). Closing the loop: Learning to generate writing feedback via language model simulated student revisions. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing* (pp. 16636–16657). Association for Computational Linguistics.
19. Yu, Y., Ping, W., Liu, Z., Wang, B., You, J., Zhang, C., Shoeybi, M., & Catanzaro, B. (2024). RankRAG: Unifying context ranking with retrieval-augmented generation in LLMs. In *Advances in Neural Information Processing Systems*.
20. Doan, N. N., Härmä, A., Celebi, R., & Gottardo, V. (2024). A hybrid retrieval approach for advancing retrieval-augmented generation systems. In *Proceedings of the 7th International Conference on Natural Language and Speech Processing (ICNLSP 2024)* (pp. 397–409). Association for Computational Linguistics.
21. Zhong, Z., Liu, H., Cui, X., Zhang, X., & Qin, Z. (2025). Mix-of-granularity: Optimize the chunking granularity for retrieval-augmented generation. In *Proceedings of the 31st International Conference on Computational Linguistics* (pp. 5756–5774). Association for Computational Linguistics.
22. Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Yang, Q., Gao, Y., Yu, J., Zhou, Y., & others. (2025). A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Computing Surveys*, 58(1), Article 9.
23. Zhang, Y., Li, Y., Cui, L., Cai, D., Liu, L., Fu, T., Huang, X., Zhao, E., Zhang, Y., Chen, Y., Shi, F., Xie, C., Fang, H.-S., & others. (2025). A survey on hallucination in large language models. *Computational Linguistics*, 51(4), 1373–1416.
24. Li, Z., Wang, Z., Wang, W., Hung, K., Xie, H., & Wang, F. L. (2025). Retrieval-augmented generation for educational application: A systematic survey. *Computers and Education: Artificial Intelligence*, 8, 100417.
25. Maurya, K. K., Srivatsa, K. A., Petukhova, K., & Kochmar, E. (2025). Unifying AI tutor evaluation: An evaluation taxonomy for pedagogical ability assessment of LLM-powered AI tutors. In *Proceedings of the 2025 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)* (pp. 1234–1251). Association for Computational Linguistics.

26. Seo, H., Kim, J., Lee, H., & Kim, Y. (2025). Large language models as evaluators in education: Verification of feedback consistency and accuracy. *Applied Sciences*, 15(2), 671.
27. Ghazawi, R., & Simpson, E. (2025). How well can LLMs grade essays in Arabic? *Computers and Education: Artificial Intelligence*, 8, 100389.
28. Alsofyani, A. H., & Barzanji, A. M. (2025). The effects of ChatGPT-generated feedback on Saudi EFL learners' writing skills and perception at the tertiary level: A mixed-methods study. *Journal of Educational Computing Research*, 63.
29. Lee, D., & Palmer, E. (2025). Prompt engineering in higher education: A systematic review to help inform curricula. *International Journal of Educational Technology in Higher Education*, 22, 7.
30. Qian, Y. (2025). Prompt engineering in education: A systematic review of approaches and educational applications. *Journal of Educational Computing Research*, 63(7-8), 1782-1818.