

Development of a Corpus-Grounded AI Self-Learning System for Arabic Language Learning

Zainur Rijal Abdul Razak¹, Habibah Ismail², Manal Jusoh@Mohamad Daud³, Yuslina Mohamed⁴, Azni Haslizan Ab Halim⁵

Faculty of Major Language Studies, Universiti Sains Islam Malaysia, Bandar Baru Nilai, Negeri Sembilan, Malaysia

DOI: <https://doi.org/10.47772/IJRISS.2026.100900402>

Received: 23 September 2026; Accepted: 28 September 2026; Published: 13 October 2026

ABSTRACT

This study is kicked off from the fact that the level of Arabic language proficiency nowadays among students at various levels of education, especially in universities, is facing serious problems. They face more critical problems particularly in the four language skills; reading, speaking, listening and writing. Therefore, this study tries to develop a corpus-grounded artificial intelligence (AI) self-learning system for two compulsory Arabic courses at Universiti Sains Islam Malaysia (USIM) to assist solving this problem. Firstly, a corpus consisting of these Arabic courses text was developed. They were collected, cleaned, segmented into pedagogical items, organized with metadata, and represented semantically for retrieval. The resulting pedagogical corpus comprised 524 searchable items, including 434 from *Arabic Language II (Communication)* course and 90 from *Arabic Language III (Social Science)* course. To select the AI development design, 43 authentic grammar, morphology, and reading-comprehension questions were used to compare two locally deployable Arabic-specialised models with a general-purpose model using retrieval-augmented generation (RAG). The configurations achieved 37.2%, 32.6%, and 67.4% accuracy, respectively, supporting the selection of RAG design to be implemented. An AI Self-learning System was successfully developed and ready to use by specific students who take these two compulsory courses. The implemented system uses semantic retrieval to ground generative AI responses in course-specific corpus evidence and provides study assistance with source traceability, dynamic practice generation, and diagnostic support. The study demonstrates an implemented pathway from Arabic educational materials to a functioning corpus-grounded RAG system. The findings establish technical and functional implementation rather than learner learning gains.

Keywords: Arabic corpus; retrieval-augmented generation; generative AI; self-learning; data-driven learning; pedagogical corpus

INTRODUCTION

Corpus linguistics provides methodological procedures for constructing and analyzing systematically organized collections of authentic language data (Sinclair, 1991; McEnery & Hardie, 2012). Its pedagogical applications have developed particularly through data-driven learning (DDL), which enables learners and teachers to investigate authentic language examples and identify linguistic patterns (Boulton, 2017). Meta-analytic and systematic-review evidence indicates that corpus-based approaches are now established in language education, although their implementation depends on task design, learner preparation, guidance, and accessibility (Boulton & Cobb, 2017; Lusta et al., 2023). In line with this recent development, Arabic learner corpora have been developed for research and pedagogical purposes (Alfaifi & Atwell, 2014), while corpus-based activities have been integrated with textbook-based Arabic instruction for vocabulary and grammar learning (Zaki, 2017). Consequently, the pedagogical use of Arabic corpus data is not itself a sufficient research gap.

Generative artificial intelligence (GenAI), however, creates a different relationship between learners and corpus evidence. Large language models (LLM) provide conversational interaction and flexible generation, but their responses are not inherently restricted to a particular curriculum. Corpus resources, by contrast, can preserve

identifiable textual domains and systematic linguistic evidence. Recent scholars therefore increasingly consider corpus tools and GenAI as potentially complementary resources (Crosthwaite & Baisa, 2023; Mizumoto, 2023; Flowerdew, 2025).

For course-specific applications, this complementarity creates a practical development problem. Educational materials must first be transformed into structured, machine-retrievable data before they can function as external evidence for generative AI. This issue is particularly relevant to Arabic because computational processing can involve script, orthographic, morphological, preprocessing, and resource-related complexities (Habash, 2010; Guellil et al., 2021). A second problem concerns how the resulting corpus should be connected to AI. Retrieval-augmented generation (RAG) provides one solution by combining generative models with information retrieved from an external knowledge resource (Lewis et al., 2020; Gao et al., 2023). Rather than depending exclusively on knowledge represented within model parameters, relevant external evidence can be retrieved at the point of interaction.

The present study investigates these interconnected problems through the development of a course-specific Arabic pedagogical corpus and its implementation in an AI self-learning system for two compulsory Arabic courses at USIM namely *Arabic Language II (Communication)* and *Arabic Language III (Social Science)*. It addresses three research questions:

1. What text-processing and corpus-construction requirements arise when Arabic course and assessment materials are prepared for AI-supported retrieval?
2. How can these materials be organized into a course-specific Arabic pedagogical corpus suitable for retrieval by an AI self-learning system?
3. How can the resulting Arabic corpus be integrated through RAG into an AI self-learning system providing course-grounded assistance, practice, and diagnostic functions?

The contribution is therefore not a claim to the first use of corpora in Arabic teaching or the first combination of corpora and GenAI. It concerns an implemented development pathway connecting authentic Arabic educational materials, pedagogical corpus construction, semantic retrieval, design evaluation, RAG, and learner-facing functionality.

LITERATURE REVIEW

Corpus-Based Language Learning

The purpose of corpus construction is to make decisions concerning text selection, organization, processing, and representation that should correspond to the intended analytical or pedagogical function (Sinclair, 1991; McEnery & Hardie, 2012). Within language education, Data-Driven Learning (DDL) applies this principle by making corpus evidence a resource for linguistic investigation (Boulton, 2017). Evidence for DDL has accumulated across different instructional environments. Boulton and Cobb (2017) synthesised 64 studies representing 88 unique samples and found substantial evidence for corpus use in language learning. Lusta et al. (2023), reviewing 89 studies published between 1997 and 2022, likewise demonstrated the breadth of DDL applications while identifying continuing challenges involving learner proficiency, corpus interfaces, task design, and pedagogical guidance.

Learner-corpus research further demonstrates that corpora designed around particular pedagogical purposes can contribute to language analysis, instructional materials, and assessment (Granger, 2015; Granger et al., 2015; Callies & Paquot, 2015). The present corpus differs because it contains curricular rather than learner-produced language, but the underlying methodological principle remains relevant which is the corpus should be designed according to the function it is intended to perform. Arabic corpus pedagogy provides a more specific context. Alfaifi and Atwell (2014) describe an Arabic Learner Corpus and its potential uses for research and Arabic teaching, while Zaki (2017) proposes the integration of corpus activities into textbook-based teaching of Arabic as a foreign language. These studies establish that corpus-supported Arabic pedagogy precedes the present system.

The current study therefore assigns the corpus to a different technological role. The course teaching and assessment materials become the external evidence resource of a generative-AI system, rather than functioning primarily as a concordance resource or learner-language collection. This orientation also clarifies why the corpus should be evaluated in relation to its intended educational function. The present resource was assembled to represent the teaching and assessment domain of two compulsory courses, not Arabic usage as a whole. Its organization therefore privileges pedagogically identifiable units, course provenance, and retrievability. This is consistent with the broader corpus-linguistic principle that design decisions should follow the questions and functions for which a corpus is constructed (Sinclair, 1991; McEnery & Hardie, 2012).

The distinction is important in the context of DDL as the conventional corpus activities often require learners to inspect concordance lines, compare examples, and infer patterns, whereas the present system uses the corpus as the evidence source behind an AI-mediated interface. The corpus remains central, but learners encounter its contents through retrieved passages, generated explanations, and generated practice. Existing DDL research shows that corpus use depends on task design, preparation, and guidance (Boulton, 2017; Boulton & Cobb, 2017; Lusta et al., 2023). These considerations remain relevant even when the interface changes from direct corpus consultation to generative mediation.

Corpus Evidence and Generative AI

Large language Models have expanded opportunities for conversational educational applications but have also intensified concerns about reliability and unsupported generation (Kasneci et al., 2023; Tlili et al., 2023). Hallucination is a recognised problem in natural-language generation, making evidential grounding especially relevant when AI is used for educational purposes (Ji et al., 2023). Corpus resources provide one potential counterbalance because their textual domain can be explicitly identified. Crosthwaite and Baisa (2023) consequently argue that GenAI does not make corpus-assisted DDL obsolete, instead, corpus evidence and generative capabilities can perform complementary functions. Similarly, Mizumoto (2023) proposes integrating DDL and GenAI within a wider ecosystem of learning resources through the Metacognitive Resource Use framework. The framework emphasises metacognitive knowledge and regulation in learners' strategic resource selection. The present study does not test metacognitive behaviour, but this work demonstrates that combining DDL and GenAI is already conceptually established.

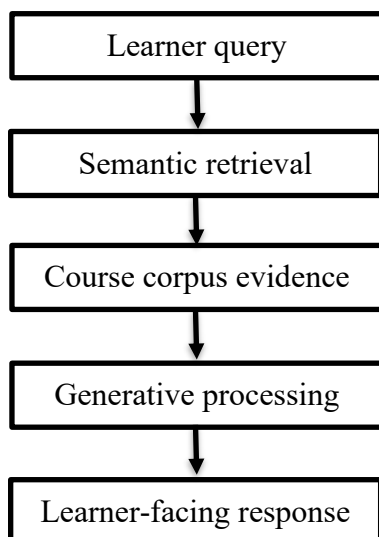
Flowerdew (2025) likewise positions large language models within the continuing development of DDL. Collectively, he means that the defensible research issue is not whether corpora and GenAI can be combined, but how a bounded pedagogical corpus can remain an evidential resource when access to it is mediated through generative AI.

Retrieval-Augmented Generation (RAG)

RAG is crucial in providing a technological mechanism for this integration. Lewis et al. (2020) introduced RAG for knowledge-intensive Natural Language Processing (NLP) by combining parametric generative models with external non-parametric memory. Dense passage retrieval further demonstrates how numerical representations can retrieve semantically relevant passages rather than relying solely on exact lexical matching (Karpukhin et al., 2020). Subsequent RAG research has expanded the design into retrieval, augmentation, and generation components and has examined challenges involving retrieval quality, generation reliability, and system evaluation (Gao et al., 2023). Educational applications now therefore include interactive learning, content generation, and assessment (Li et al., 2025) to meet the demand of this technology.

For course-specific language learning, the important property of RAG is that the external knowledge source can be bound by the curriculum as in chart below:

Figure 1: Workflow of the corpus-grounded AI self-learning system



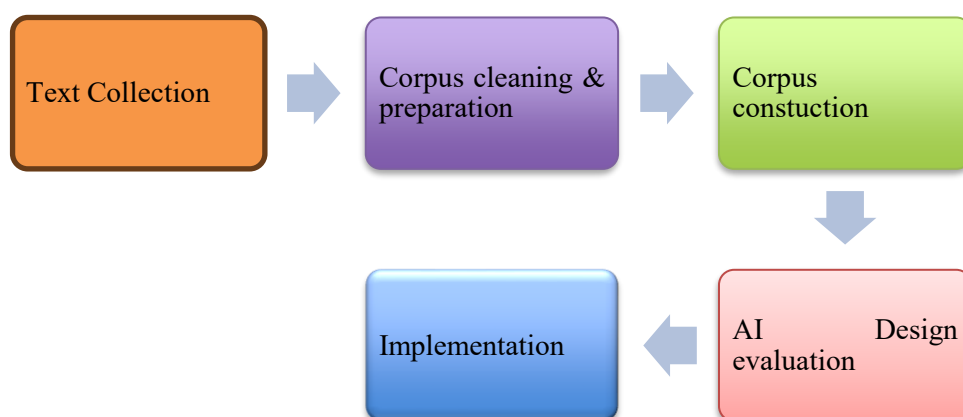
RAG does not guarantee correctness, but it permits the pedagogical corpus to remain external, dynamically searchable, and identifiable during generation. Within this study, retrieval and generation consequently perform distinct functions. Retrieval identifies course material that is semantically relevant to the learner's request, while generation reformulates or uses that material for an answer, practice item, or diagnostic activity. Keeping these functions conceptually separate is important because the presence of relevant evidence does not by itself ensure that the generated response will interpret it correctly. RAG nevertheless makes the evidential pathway inspectable in a way that an unsupported response from model parameters alone would not be (Lewis et al., 2020; Li et al., 2025).

METHODOLOGY

Research Design and Data

This study adopted a development-oriented design comprising five (5) connected stages from text collection, corpus cleaning and preparation, corpus construction, AI development design evaluation to implementation of learner-facing functions.

Diagram 1: Design workflow of the AI Self-Learning System



The data consisted in current AI system exclusively of teaching and assessment materials associated with *Arabic Language II (Communication)* and *Arabic Language III (Social Science)* courses. The initial collection comprised two textbooks and fifteen past examination papers, five (5) from *Arabic Language II (Communication)* and ten (10) from *Arabic Language III (Social Science)*. Collection was therefore bound by the materials actually used for the two courses.

Arabic Text Processing

All Documents were inspected for extractability before corpus construction. Several examination papers produced a character-order problem involving adjoining Arabic characters, particularly *lām* (ل) followed by an alif-family character. Two extraction methods reproduced the problem. Because item boundaries remained identifiable, the affected materials could be structurally processed; uncertain wording was not automatically reconstructed where correction could introduce unsupported forms. Their Arabic content was stored as vector line drawings rather than machine-readable text or conventional scanned images and could not be recovered using the implemented extraction procedures.

The preparation process therefore combined automated extraction with human-supervised inspection and normalisation. Although Arabic NLP literature identifies broader preprocessing challenges (Habash, 2010; Guellil et al., 2021), the specific problems reported here derive from the documents used in this study. Cleaning was conducted at the level needed for later pedagogical segmentation. Each recoverable document was separated into reading passages, questions, answer choices, and grammar exercises so that the system could retrieve meaningful units instead of returning an entire examination paper or textbook section. The character-order problem did not prevent the boundaries between these units from being located, but uncertain wording was left unreconstructed where intervention might have introduced unsupported Arabic forms.

The document represented by a positional visual section in the system required a different response because the internal representation recorded the displayed form of a character rather than its standard identity. By contrast, the eight affected *Arabic Language III (Social Science)* papers contained vector line drawings rather than machine-readable characters or conventional scanned images. Their content was therefore excluded under the implemented procedure. Recovering those materials would require an additional recognition process and remains outside the completed implementation.

Corpus Construction and Retrieval

Recoverable materials were segmented into pedagogically meaningful items, including reading passages, questions and answer options, and grammar exercises while metadata identified course, source document, and content type. Each item was converted into a numerical semantic representation to enable meaning-based retrieval. Semantic retrieval follows the broader principle that dense representations can support relevance matching beyond exact lexical overlap (Karpukhin et al., 2020). Finally, the completed corpus contained 524 searchable items, comprising 434 from *Arabic Language II (Communication)* and 90 from *Arabic Language III (Social Science)*.

As mentioned, the metadata layer records the course, source document, and content type of each item. These labels serve two related purposes, firstly, they retain provenance after the original files have been divided into smaller units and, secondly, they allow later retrieval to be constrained according to the learning task, such as locating grammar material rather than an unrelated reading passage. Semantic representation supplements this categorization by supporting relevance matching when the wording of a learner's question differs from the wording in the corpus.

Development Design Evaluation

During implementation, two small Arabic-specialized models suitable for local deployment were instead evaluated against a larger general-purpose model combined with RAG. A test set of 43 authentic questions covering grammar, morphology, and reading comprehension was assembled from the examination corpus. Because official answer keys were unavailable, reference answers were manually checked. The two Arabic-specialized models achieved 37.2% and 32.6% accuracy, respectively. The RAG configuration achieved 67.4% and performed better across the tested question categories. RAG was therefore selected for implementation, and no dedicated model was trained on the corpus.

The comparison was conducted for development design selection within this project. It was not designed as a general benchmark of Arabic language models but specifically for these two courses. The test set was assembled from authentic examination questions rather than from newly written evaluation items to cover grammar,

morphology, and reading comprehension, and reference answers were manually checked because the source papers did not include official answer keys. The same 43-question set was used across the three configurations. The resulting percentages were used only to decide which approach should support the implemented system, and they were not treated as estimates of general model performance beyond this corpus and task setting.

System Implementation

Under the selected development design, learner input initiates semantic retrieval from the course corpus. Retrieved material is supplied to the generative model as contextual evidence, consistent with the general retrieval–augmentation–generation design described in RAG research (Lewis et al., 2020; Gao et al., 2023). Three principal learning functions were implemented. First, the study assistant retrieves relevant material, produces an answer grounded in it, and displays the source passage. Second, the practice generator retrieves a passage or grammar point and generates a new multiple-choice question, answer, and explanation and lastly, the diagnostic assessment presents questions across grammar, vocabulary, and reading comprehension and reports the learner's weakest area to direct subsequent revision.

The study also evaluates corpus construction, development design selection, and functional implementation before the design was selected. Operationally, the selected design searches for the corpus at the time of each interaction which differs from treating the corpus as material consumed once during model training. In the implemented workflow, retrieved evidence remains external to the generative model and can be displayed to the learner. The same corpus infrastructure supports all three functions, but each uses the evidence differently, for example, the assistant answers from a retrieved passage, the practice generator creates an item from retrieved curricular material, and the diagnostic activity draws on several skill areas before reporting the weakest area. Image below indicates the main interface of the AI Self-Learning System that had been successfully developed:

Image 1: Main interface of the AI Self-Learning System

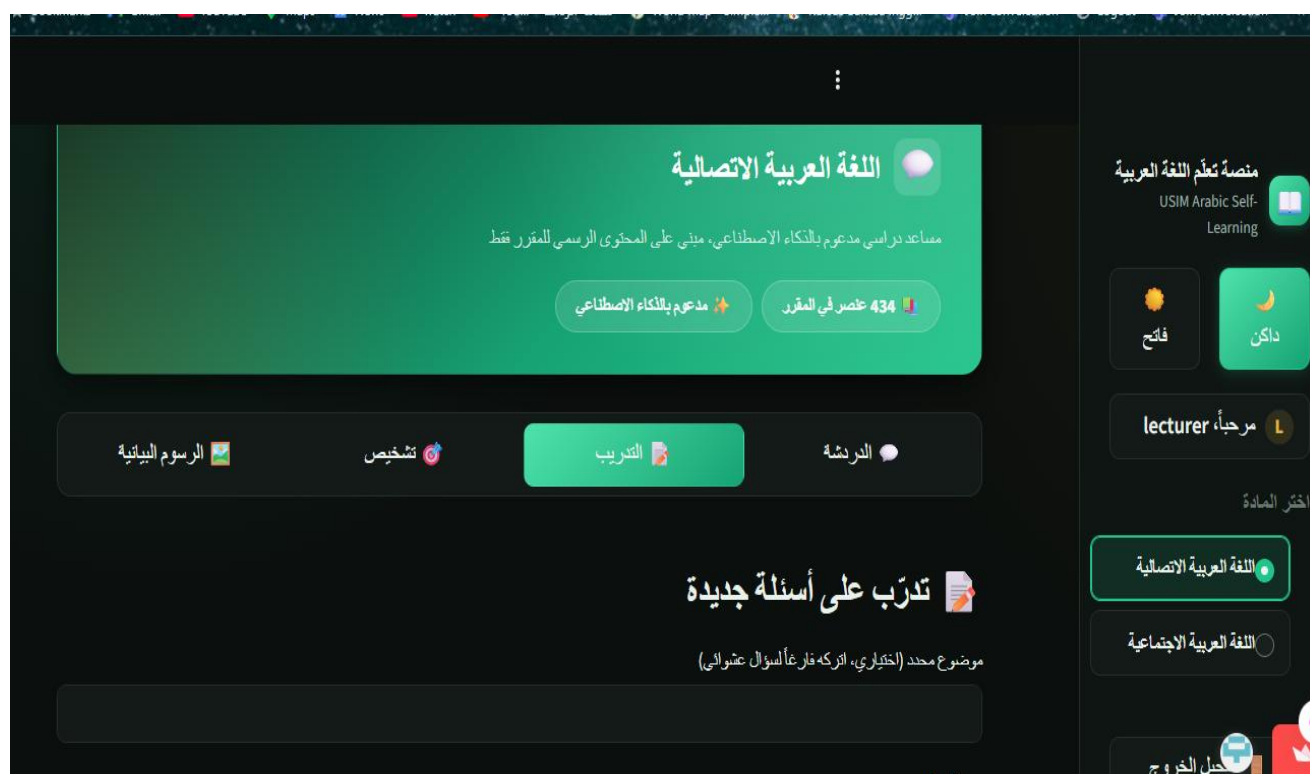


Image 1 above shows the end product of the system development, based on theories and steps discussed in this section. The system is divided into two compulsory Arabic courses, as explained and supported by language activities for the respective courses. Users may need to register for the first time before exploring and interacting with the system, for privacy and security reasons. Responses from the system are limited to information regarding the courses, based on the corpus that has been developed.

Image 2: Example of response from AI Self-Learning System

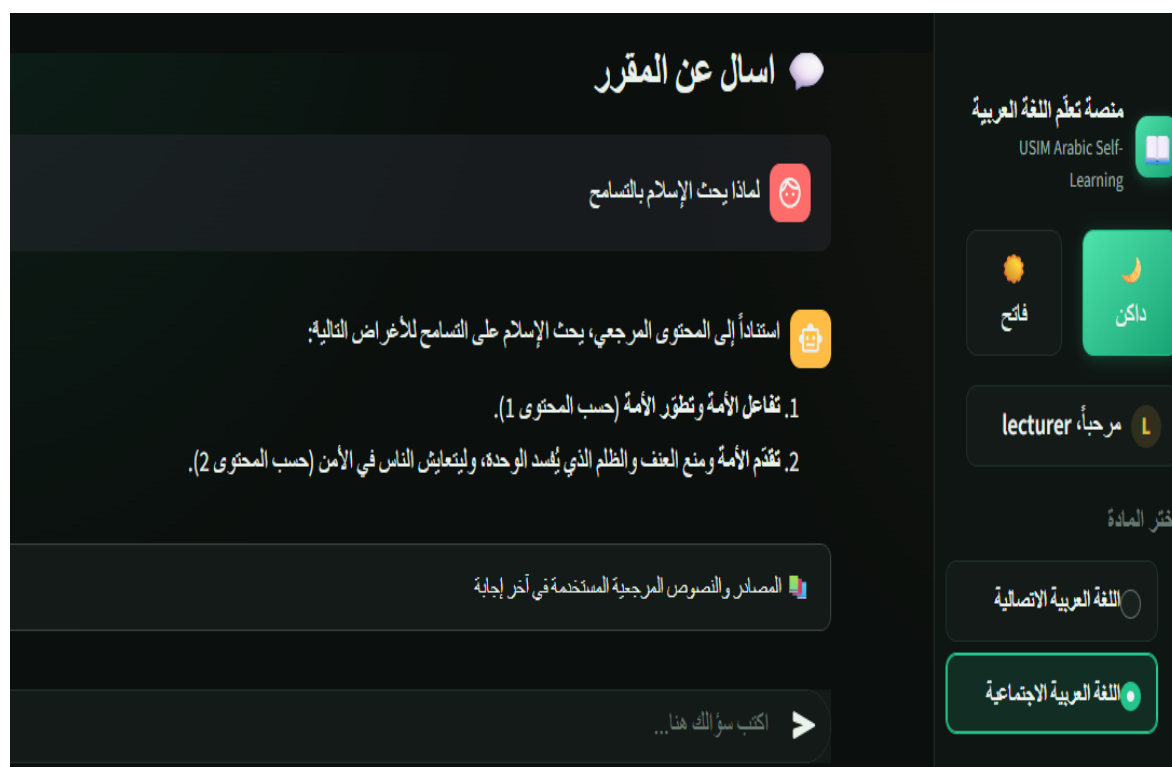


Image 2 above demonstrates an example of a prompt response from the developed AI system. It shows that the prompt is for the sub-corpus of the *Arabic Language III (Social Science)* course, and the answer was generated from the data we used to develop the corpus.

RESULTS

Corpus Preparation and Construction

The study came to findings that show there are three (3) substantive processing constraints emerged from the source documents; first, character-order problems during extraction, second, positional visual section requiring normalization, and third, non-extractable Arabic stored as vector drawings. The latter prevented eight *Arabic Language III (Social Science)* examination papers from being incorporated through the implemented procedure. These results demonstrate the need for human-supervised inspection within this corpus-development process. The three constraints required different decisions rather than a single generic cleaning rule. Consistent character-order errors permitted structural segmentation but did not justify speculative correction of every affected word. Positional section encoding could be normalized because the representation could be mapped to standard characters. Vector-drawn text could not be recovered through the implemented extraction methods and was excluded. This differentiation was necessary to preserve the evidential integrity of the corpus while making as much of the source collection searchable as the available procedure allowed.

Following cleaning and segmentation, the corpus contained 524 searchable pedagogical items where 434 from *Arabic Language II (Communication)* and 90 from *Arabic Language III (Social Science)*. Metadata and semantic representations made these items searchable according to course, content type, and relevance. The completed corpus is consequently both structured and bounded. Structure comes from separating documents into pedagogically meaningful units and attaching metadata that preserves course, source, and content type. Boundedness comes from retaining only material traceable to the collected course documents and from excluding content that could not be recovered reliably. Together, these properties allow the retrieval process to return a focused passage, question, or exercise while maintaining a connection to its curricular origin. They also define what the system cannot answer from the corpus when relevant evidence falls outside the two courses or within the unrecovered examination papers.

AI implementation of self-learning functions

The study shows that the 43-question evaluation differentiated the three tested configurations. The two locally deployable Arabic-specialized models achieved 37.2% and 32.6%, whereas the general-purpose model combined with RAG achieved 67.4% as discussed. This comparison provided the empirical basis for selecting RAG and expend the role of the corpus from merely as a training model to a generative model and corpus-grounded output as targeted form this study. The 67.4% result is evidence for development design selection in this test and should not be interpreted as universal superiority of RAG over Arabic-specialized models.

The corpus–RAG design was successfully implemented as an Arabic self-learning system. The study assistant retrieves relevant course material, generates an answer from that material, and presents its source passage to provide at the end course-specific grounding and source traceability. Meanwhile, the practice generator retrieves existing curricular material and uses it to produce new multiple-choice questions with answers and explanations. The finite corpus consequently functions as reusable grounding evidence for additional practice. The diagnostic function differentiates performance across grammar, vocabulary, and reading comprehension and reports the weakest area to guide revision.

Evaluation of the developed system

To evaluate the system, we presented it to two language and system experts. Each expert assessed its student-friendliness, technical aspects, and informativeness.

Expert 1, an AI technology expert, commented:

"Overall, the system effectively helps students master course content. The chatbot and practice questions are drawn strictly from the official textbook and past examination papers, ensuring every interaction is directly relevant to the syllabus rather than generic knowledge sources. The diagnostic module is particularly useful, clearly identifying student weaknesses in grammar, vocabulary, comprehension, or language expressions. It directs students to focused practice instead of unstructured revision, supporting a more organized, self-paced learning approach. In terms of usability, the interface is simple and easy to use, featuring a single chat window, a one-click question generator, instant feedback on submitted answers, and a progress dashboard that tracks improvement over time. Students require very little technical knowledge to navigate between modules, making the system accessible even to those with limited experience with digital learning tools. Regarding content alignment, the material closely matches the stated course objectives because every question and passage comes directly from the official syllabus and previous examinations. This ensures an authentic learning experience that is relevant to exams and consistent with the intended course outcomes."

This feedback confirms the system's high technical reliability and suitability for university-level student use.

Expert 2 is an Arabic language expert. She commented:

"The platform is suitable for learners at the lower end of CEFR level B2, offering activities that support their language development. It provides varied exercises focusing on text comprehension and grammar, enabling students to strengthen their understanding of meaning and language structures. Clear, straightforward instructions help learners complete each task confidently and independently. The answer options are accompanied by explanations identifying the correct response and clarifying why it is appropriate, helping students recognize mistakes and deepen their understanding. Exercises can also be repeated as often as needed, allowing learners to consolidate knowledge, monitor progress, and gradually build greater language confidence. The expert however, suggested to add information about the background of the content in terms of introduction of the system, objective and language skill level"

DISCUSSION

The findings demonstrate that developing course-specific generative AI for Arabic begins with the computational accessibility and organization of pedagogical evidence. Corpus construction is a purposeful

process rather than simply the accumulation of texts (Sinclair, 1991; McEnery & Hardie, 2012), and the present findings show that educational documents designed for human use cannot necessarily be treated as immediately machine-ready corpus data. Arabic NLP literature provides a broader context by identifying script, orthographic, morphological, preprocessing, and resource-related challenges (Habash, 2010; Guellil et al., 2021). The present study adds a narrower document-level observation including the character-order errors, positional section, and vector-based Arabic directly affected the recoverability of the target educational materials. These are observations from this corpus, not universal properties of Arabic documents.

The 524-item corpus likewise demonstrates the importance of purpose in designing the system, and it is not intended to represent Arabic generally. Its value is its correspondence to two specific courses and the organization and retrievability of their pedagogical material. Its uneven distribution across the two courses must also be understood in relation to source recoverability. The final item counts therefore describe the recoverable and processed corpus, not the total amount of Arabic visible across every collected document. This limitation is part of the corpus record and should accompany interpretation of system coverage. The most consequential architectural result was the transition from model training to retrieval grounding. RAG was developed to combine generative models with external retrieved knowledge (Lewis et al., 2020), while dense retrieval provides a mechanism for identifying semantically related passages (Karpukhin et al., 2020). In the present system, these principles allow the corpus to remain an external resource that is searched during learner interaction rather than becoming solely model-training data.

According to Crosthwaite and Baisa (2023), corpus provenance was an important property in the GenAI era. In the developed study assistant, this property is partly preserved through the sequence of questions, retrieved course evidence, generated response and identifiable source. Traceability should not be conflated with correctness. Neural generation remains vulnerable to hallucination (Ji et al., 2023), and RAG depends on retrieval relevance and the behaviour of the generative model (Gao et al., 2023; Li et al., 2025). The demonstrated advantage is therefore an identifiable evidential pathway rather than elimination of unsupported generation. The development also changes the learner–corpus relationship where DDL conventionally foregrounds learner engagement with corpus evidence (Boulton, 2017; Boulton & Cobb, 2017; Lusta et al., 2023).

The implemented interface therefore reduces the need for a learner to formulate corpus queries or interpret concordance output directly, but it does not remove the need to engage critically with evidence. Displaying the retrieved source text gives the learner an opportunity to compare the generated response with the course material on which it is based. Whether learners actually make this comparison, and whether doing so supports corpus literacy or more effective independent learning, remains an empirical question. This boundary is important because access to a traceable source is a functional affordance, whereas critical source use is a learner behaviour that was not measured in the present development study. This study describes that configuration as corpus-grounded AI-mediated learning and the expression is not presented as an established theoretical category. It captures the fact that corpus evidence remains central even though learner access is mediated by generative AI.

Such mediation presents a potential pedagogical trade-off. DDL research identifies corpus literacy, task design, learner preparation, and instructional support as important considerations (Boulton & Cobb, 2017; Lusta et al., 2023). Natural-language AI interaction may reduce some interface demands, but it can also reduce learners' direct involvement in searching and interpreting corpus evidence. Whether AI mediation increases accessibility while preserving the pedagogical benefits of direct corpus engagement requires empirical investigation. The practice generator extends corpus grounding from explanation to activity creation. Retrieved curricular evidence becomes the basis for new questions, answers, and explanations. This is consistent with educational RAG applications involving content generation and assessment (Li et al., 2025). However, successful generation does not establish the pedagogical quality of the resulting questions.

The feature can retrieve a passage or grammar point associated with a learner-selected topic or with a topic selected for practice and use that evidence to compose a multiple-choice question with an answer and explanation. This expands the available practice beyond the fixed set of source questions without presenting the generated item as an original examination question. Its pedagogical value still depends on the relevance of

retrieval and the quality of generation, which is why functional availability should be distinguished from validated question quality (Li et al., 2025).

The diagnostic component similarly provides a mechanism for targeted revision but does not establish self-regulated learning. Mizumoto's (2023) Metacognitive Resource Use framework emphasises metacognitive knowledge and regulation in strategic resource selection. Demonstrating these processes would require learner behaviour or process data that were not collected. The diagnostic output is therefore best understood as guidance produced by the implemented activity rather than as a validated measurement of proficiency. It identifies relative weakness across the three included areas within the generated exercise and can direct the learner toward subsequent practice. Establishing diagnostic validity would require additional evidence concerning item quality, skill coverage, scoring consistency, and correspondence with independent measures of Arabic performance.

This extends existing Arabic corpus-pedagogy work (Alfaifi & Atwell, 2014; Zaki, 2017) toward a generative-AI-mediated implementation without claiming that corpus-AI integration itself is new. The corpus represents two USIM compulsory courses; eight examination papers could not be incorporated; and the 43-question comparison was intended for design selection rather than comprehensive model benchmarking. Most importantly, functional implementation is not evidence of learning effectiveness and therefore, future work should evaluate retrieval relevance, answer accuracy, generated-question quality, and diagnostic validity before investigating usability, learner interaction, and measurable language-learning outcomes. On the other hand, the retrieval should be examined for relevance to learner queries; displayed sources should be checked for fidelity to the originating corpus items; and generated answers and questions should be assessed for accuracy and pedagogical appropriateness. The diagnostic function requires separate validation to determine whether the reported weakest area corresponds to learner performance. Only after these system-level properties are established should later studies make claims about usability, engagement, autonomy, self-regulation, or improvements in Arabic proficiency.

CONCLUSION

This study developed and implemented a corpus-grounded AI self-learning system from Arabic teaching and assessment materials. Human-supervised processing transformed heterogeneous educational documents into a 524-item semantically searchable pedagogical corpus, while also revealing extraction and representation constraints affecting Arabic document recoverability. Design evaluation subsequently repositioned the corpus from anticipated training data to live grounding evidence. On the 43-question test, the RAG configuration achieved 67.4% accuracy, compared with 37.2% and 32.6% for the two locally deployable Arabic-specialised models tested. This provided the empirical basis for selecting RAG within the project.

The resulting system operationalizes the corpus through course-grounded study assistance with source traceability, dynamic practice generation, and diagnostic support. Its central contribution is therefore an implemented pathway connecting Arabic educational materials and pedagogical corpus to learner-facing functions. The evidence establishes corpus construction, design selection and functional implementation. It does not establish improved proficiency, learner autonomy, engagement, or self-regulated learning because those outcomes require subsequent learner-based evaluation.

ACKNOWLEDGEMENT:

The financial support from the Universiti Sains Islam Malaysia (USIM) [PPPI/PENTADBIR/FPBU/USIM/19824] is gratefully acknowledged. Therefore, researchers extend sincere appreciation to the institutions and affiliated parties whose support facilitated the successful execution of this study.

Conflicts of Interest: We would like to declare that there are no conflicts of interest associated with this publication. All authors have been fully informed and actively participated in the study.

Data Availability: The data that support the development of this system was not made publicly available. Request to access the data should be directed to the main author.

REFERENCES

1. Alfaifi, A. Y. G., & Atwell, E. S. (2014). Arabic learner corpus and its potential role in teaching Arabic to non-native speakers. In Proceedings of the American Association for Corpus Linguistics Conference. Learner Corpus Association.
2. Boulton, A. (2017). Data-driven learning and language pedagogy. In S. L. Thorne & S. May (Eds.), *Language, education and technology: Encyclopedia of language and education* (pp. 181–192). Springer. https://doi.org/10.1007/978-3-319-02328-1_15-1
3. Boulton, A., & Cobb, T. (2017). Corpus use in language learning: A meta-analysis. *Language Learning*, 67(2), 348–393. <https://doi.org/10.1111/lang.12224>
4. Callies, M., & Paquot, M. (2015). Learner corpus research: An interdisciplinary field on the move. *International Journal of Learner Corpus Research*, 1(1), 1–6. <https://doi.org/10.1075/ijlcr.1.1.00edi>
5. Crosthwaite, P., & Baisa, V. (2023). Generative AI and the end of corpus-assisted data-driven learning? Not so fast! *Applied Corpus Linguistics*, 3(3). <https://doi.org/10.1016/j.acorp.2023.100066>
6. Flowerdew, J. (2025). Data-driven learning: From Collins COBUILD Dictionary to ChatGPT. *Language Teaching*, 58(3), 359–376. <https://doi.org/10.1017/S0261444824000144>
7. Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, M., & Wang, H. (2023). Retrieval-augmented generation for large language models: A survey. *arXiv*. <https://doi.org/10.48550/arXiv.2312.10997>
8. Granger, S. (2015). The contribution of learner corpora to reference and instructional materials design. In S. Granger, G. Gilquin, & F. Meunier (Eds.), *The Cambridge handbook of learner corpus research* (pp. 485–510). Cambridge University Press.
9. Granger, S., Gilquin, G., & Meunier, F. (Eds.). (2015). *The Cambridge handbook of learner corpus research*. Cambridge University Press.
10. Guellil, I., Saâdane, H., Azouaou, F., Gueni, B., & Nouvel, D. (2021). Arabic natural language processing: An overview. *Journal of King Saud University—Computer and Information Sciences*, 33(5), 497–507. <https://doi.org/10.1016/j.jksuci.2019.02.006>
11. Habash, N. Y. (2010). *Introduction to Arabic natural language processing*. Morgan & Claypool.
12. Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12), Article 248.
13. Karpukhin, V., Oğuz, 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* (pp. 6769–6781).
14. Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., Stadler, M., Weller, J., Kuhn, J., & Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103. <https://doi.org/10.1016/j.lindif.2023.102274>
15. 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. <https://doi.org/10.48550/arXiv.2005.11401>
16. 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. <https://doi.org/10.1016/j.caeai.2025.100417>
17. Lusta, A. F., Demirel, Ö., & Mohammadzadeh, B. (2023). Language corpus and data driven learning (DDL) in language classrooms: A systematic review. *Heliyon*, 9(12), 1–19. <https://doi.org/10.1016/j.heliyon.2023.e22731>
18. McEnery, T., & Hardie, A. (2012). *Corpus linguistics: Method, theory and practice*. Cambridge University Press.

19. Mizumoto, A. (2023). Data-driven learning meets generative AI: Introducing the framework of Metacognitive Resource Use. *Applied Corpus Linguistics*, 3(3). <https://doi.org/10.1016/j.acorp.2023.100074>
20. Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford University Press.
21. Tlili, A., Shehata, B., Adarkwah, M. A., Bozkurt, A., Hickey, D. T., Huang, R., & Agyemang, B. (2023). What if the devil is my guardian angel: ChatGPT as a case study of using chatbots in education. *Smart Learning Environments*, 10, 15.
22. Zaki, M. (2017). Corpus-based teaching in the Arabic classroom: Theoretical and practical perspectives. *International Journal of Applied Linguistics*, 27(2), 514–541. <https://doi.org/10.1111/ijal.12159>