Learning, Media and Technology in Arabic Education
Volume 2, Issue 1, 1-12.
e_ISSN: -
https://lmtae.unikhams.ac.id/index.php/lmtae/index

Bridging Arabic NLP and Language Learning: A Critical Review of LLM-Based Approaches, Challenges, and Pedagogical Opportunities

Muh. Sabilar Rosyad*
Universitas Islam Negeri Sunan Kalijaga Yogyakarta, INDONESIA

Laila Farah Fitria
Universitas Islam Negeri Sunan Kalijaga Yogyakarta, INDONESIA

Article Info

Article history:

Received: Mar 10, 2026
Revised: Apr 22, 2026
Accepted: Jun 2, 2026

Keywords:

Arabic Language Learning; Arabic Natural Language Processing; Computer-Assisted Language Learning (CALL); Large Language Models (LLMs); Second Language Acquisition (SLA).

Abstract

Recent advances in Arabic Natural Language Processing (NLP) have been largely driven by transformer-based architectures and Large Language Models (LLMs), which outperform traditional approaches across tasks such as sentiment analysis, machine translation, text classification, and question answering. Nevertheless, Arabic NLP research remains predominantly technocentric and insufficiently connected to Second Language Acquisition (SLA) theories and pedagogical frameworks, thereby limiting its educational applicability. This study aims to systematically synthesize recent developments in LLM-based Arabic NLP and critically examine their potential integration into Arabic language learning, particularly within SLA and Computer-Assisted Language Learning (CALL) frameworks. A Systematic Literature Review guided by the PRISMA framework was conducted using Scopus-indexed studies published between 2025 and 2026. The selected studies were analyzed through thematic synthesis across five dimensions: model architecture, task domain, data strategies, linguistic focus, and pedagogical relevance. The findings demonstrate the growing dominance of LLMs, including AraBERT, AraGPT2, and ALLAM, alongside increased reliance on data augmentation, synthetic corpus generation, and dialect-specific adaptation. Evaluation practices are also beginning to extend beyond conventional performance metrics toward trustworthiness, safety, reasoning, and robustness. However, the review identifies a near absence of SLA-informed applications, Arabic LLM-based CALL systems, and human-centered evaluation involving usability, learner engagement, trust, and cognitive load. Despite this gap, LLMs offer substantial pedagogical potential through adaptive feedback, contextualized input, interactive dialogue, and personalized learning support. The study concludes that Arabic NLP requires stronger interdisciplinary integration with SLA and CALL to transform technological capabilities into learner-centered, pedagogically grounded, and empirically validated language-learning applications.

To cite this article: Rosyad, M. S., & Fitria, L. F. (2025). *Bridging Arabic NLP and language learning: A critical review of LLM-based approaches, challenges, and pedagogical opportunities*. *Learning, Media and Technology in Arabic Education, 2*(1), 1–12.

INTRODUCTION

The rapid advancement of artificial intelligence (AI) has fundamentally transformed language education by enabling more adaptive, interactive, and personalized learning experiences. Within this transformation, Natural Language Processing (NLP) has become one of the most influential technologies for supporting language learning through applications such as automated feedback, intelligent tutoring, machine translation, dialogue systems, and text generation. Recent breakthroughs in transformer-based architectures and Large Language Models (LLMs) have substantially expanded these capabilities by enabling contextual language understanding and human-like language generation. In Arabic language processing, this technological evolution has accelerated with the emergence of Arabic-specific LLMs, including AraBERT, AraGPT2, and ALLAM, which have demonstrated remarkable performance across diverse NLP tasks (Khader et al., 2025; Saiful Bari et al., 2025). Consequently, Arabic NLP has evolved into one of the fastest-growing

* **Corresponding author:**
Muh. Sabilar Rosyad, Universitas Islam Negeri Sunan Kalijaga Yogyakarta, INDONESIA
muhammad.rosyad@uin-suka.ac.id

research domains within artificial intelligence, offering significant opportunities to reshape Arabic language education.

The increasing sophistication of Arabic NLP is particularly important because Arabic remains one of the most linguistically complex languages for both computational processing and second language learning. Rich morphology, orthographic ambiguity, and extensive dialectal diversity continue to present substantial challenges for developing intelligent language learning systems capable of supporting learners across different proficiency levels (Boulesnam & Boucetti, 2025; Dahou et al., 2025). At the same time, learners of Arabic increasingly require digital learning environments that provide immediate feedback, contextual interaction, and personalized instructional support beyond conventional classroom settings. Advances in LLM technology create new possibilities for addressing these needs through conversational agents, automated writing assistance, intelligent assessment, and adaptive learning pathways. These developments indicate that Arabic NLP is no longer merely a computational research topic but has become increasingly relevant to the future of Arabic language education.

In parallel with these educational demands, Arabic NLP research has experienced remarkable technological progress during the past few years. Transformer-based models have consistently achieved superior performance in sentiment analysis, question answering, machine translation, and text classification, outperforming earlier rule-based and conventional machine learning approaches (Alrashidi & Mathkour, 2026; Alrayzah et al., 2026; Ferroud et al., 2026). The development of specialized benchmark datasets, including AraEventCoref and Arabic WikiTableQA, has further strengthened evaluation practices and accelerated methodological innovation within the field (Aldawsari & Dawood, 2025; Alsolami & Alrayzah, 2025). Recent applications have also expanded into hate speech detection, cyberbullying identification, legal document classification, bilingual summarization, and aspect-based sentiment analysis, demonstrating the versatility of LLM-based Arabic NLP across numerous linguistic tasks (Alshahrani & Aksoy, 2025; Hithnawi et al., 2025; Tahtah et al., 2025; Galhom et al., 2025; Bakr et al., 2025). Collectively, these studies confirm that Arabic NLP has reached a high level of technological maturity supported by increasingly sophisticated models and evaluation resources.

Despite these impressive technological achievements, the educational utilization of Arabic NLP remains considerably underdeveloped. Existing studies predominantly emphasize improvements in algorithmic performance using technical evaluation metrics such as accuracy, precision, recall, and F1-score, while paying relatively little attention to pedagogical effectiveness and language acquisition processes. Moreover, ongoing research addressing dialect adaptation, transfer learning, multilingual corpora, and synthetic data generation primarily seeks to improve computational performance rather than facilitate meaningful language learning experiences (Essameldin et al., 2025; Shang et al., 2025; Hamed et al., 2025; Shi & Agrawal, 2025; Almeman, 2025; Nashir et al., 2025). As a result, technological innovation has progressed far more rapidly than its pedagogical implementation. This imbalance raises an important question regarding how the capabilities of modern Arabic LLMs can be translated into meaningful learning experiences that support Arabic as a foreign or second language.

Previous studies generally converge into two dominant streams. The first stream focuses on developing increasingly powerful Arabic NLP models through advances in transformer architectures, benchmark construction, and task-specific optimization across sentiment analysis, machine translation, question answering, summarization, and text classification (Ferroud et al., 2026; Galhom et al., 2025; Alrayzah et al., 2026; Bakr et al., 2025). The second stream concentrates on overcoming linguistic challenges inherent to Arabic, particularly dialect diversity, low-resource settings, code-switching, and corpus development through transfer learning, data augmentation, and dialect-aware adaptation strategies (Boulesnam & Boucetti, 2025; Dahou et al., 2025; Essameldin et al., 2025; Shang et al., 2025; Hamed et al., 2025; Shi & Agrawal, 2025; Almeman, 2025; Nashir et al., 2025). Although these studies collectively establish a robust technological foundation for Arabic NLP, they predominantly evaluate computational efficiency rather than educational value. Consequently, the literature provides limited insight into how these technological advances can be systematically incorporated into language teaching and learning.

A further limitation is that existing literature rarely integrates Arabic NLP with established educational theories, particularly Second Language Acquisition (SLA) and Computer-Assisted

Language Learning (CALL). Research discussing artificial intelligence in Arabic education generally remains conceptual and seldom synthesizes empirical findings from Arabic NLP to explain how LLM capabilities may support learner interaction, comprehensible input, feedback generation, or language production (Guessoum et al., 2026). Likewise, current evaluation frameworks primarily assess model performance while overlooking learner-centered dimensions such as engagement, communicative competence, usability, trust, and cognitive load. These limitations reveal a critical research gap: although Arabic NLP has become technologically mature, its pedagogical integration remains insufficiently conceptualized and empirically examined. Therefore, a comprehensive synthesis that bridges technological innovation with language learning theory is urgently needed to guide future research and educational practice.

Based on these considerations, this study aims to systematically synthesize recent developments in LLM-based Arabic NLP while critically examining their implications for Arabic language learning through the perspectives of SLA and CALL. Specifically, this review analyzes methodological trends, application domains, emerging technological challenges, and pedagogical opportunities reported in recent literature. Beyond mapping technological developments, this study proposes a conceptual framework that connects LLM capabilities with language acquisition principles to support learner-centered Arabic language education. Theoretically, the study contributes to strengthening interdisciplinary discourse between artificial intelligence, Arabic NLP, and language education. Practically, the findings provide directions for researchers, educators, and educational technology developers in designing more pedagogically meaningful AI-supported Arabic learning environments.

METHOD

This study employed a qualitative Systematic Literature Review (SLR) to critically synthesize recent developments in Large Language Model (LLM)-based Arabic Natural Language Processing (NLP) and examine their pedagogical implications for Arabic language learning from the perspectives of Second Language Acquisition (SLA) and Computer-Assisted Language Learning (CALL). The SLR approach was selected because it enables a comprehensive synthesis of fragmented evidence, facilitates the identification of research trends, and reveals conceptual gaps across interdisciplinary studies (David & Gelbard, 2025; Kitchenham et al., 2009; Snyder, 2019). The review followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA 2020) framework to ensure transparency, methodological rigor, and replicability throughout the identification, screening, eligibility, and inclusion processes (Page et al., 2021). The review was conducted between May and June 2026, focusing exclusively on studies published during 2025–2026 to capture the latest developments in transformer-based architectures and LLM applications in Arabic NLP.

The unit of analysis comprised peer-reviewed scientific publications rather than human participants. Literature was retrieved exclusively from the Scopus database because of its comprehensive coverage of high-quality international publications in artificial intelligence, computational linguistics, and educational technology. The search strategy employed combinations of keywords, including *"Arabic NLP," "Arabic LLM," "Large Language Models," "Transformer,"* and *"Arabic Language Processing,"* using Boolean operators to optimize retrieval. Eligible studies were required to be Scopus-indexed, published between 2025 and 2026, and explicitly investigate Arabic NLP, transformer-based models, or LLM applications across tasks such as sentiment analysis, machine translation, question answering, text classification, summarization, dataset development, or dialogue systems. Studies unrelated to Arabic language processing, focused on non-textual modalities, or lacking sufficient methodological information were excluded to ensure the quality and relevance of the synthesized evidence.

The literature selection was performed systematically through four sequential stages following the PRISMA protocol. Initially, potentially relevant publications were identified through structured database searches, after which duplicate records were removed. Titles and abstracts were subsequently screened against the predefined eligibility criteria before full-text articles were assessed for methodological quality and conceptual relevance. Publications satisfying all inclusion criteria were retained for analysis and documented within the PRISMA flow diagram presented in

Figure 1. This systematic procedure minimized selection bias while ensuring that the final corpus represented the current state of research on LLM-based Arabic NLP.

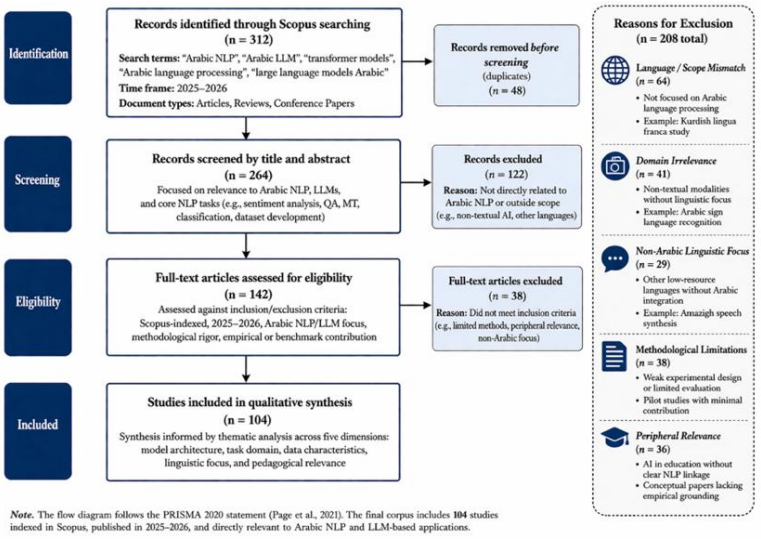

Figure 1. PRISMA Flow Diagram of Study Selection

Data were collected using a structured literature extraction protocol that served as the primary research instrument. Each selected publication was systematically coded according to five analytical dimensions: (1) model architecture, including machine learning, deep learning, transformer-based models, and LLMs; (2) application domain, including sentiment analysis, question answering, machine translation, summarization, and text classification; (3) data characteristics, including benchmark datasets, corpus development, and data augmentation strategies; (4) linguistic focus, distinguishing Modern Standard Arabic from dialectal Arabic; and (5) pedagogical relevance, examining the extent to which each study addressed language learning, SLA, or CALL. The coding framework was developed based on previous Arabic NLP surveys and thematic synthesis methodology, while methodological validity was strengthened through explicit inclusion criteria, standardized coding procedures, and repeated verification of extracted information to ensure analytical consistency across all reviewed studies.

The collected data were analyzed using thematic synthesis, which enables the systematic identification of recurring patterns, dominant research themes, methodological trends, and conceptual gaps across heterogeneous studies (Thomas & Harden, 2008). Coding results were iteratively compared and organized into broader themes describing the technological evolution of Arabic NLP, emerging research directions, linguistic challenges, evaluation paradigms, and pedagogical opportunities. These themes were subsequently interpreted through the theoretical perspectives of SLA and CALL to explain how advances in LLM-based Arabic NLP may contribute to language learning. Throughout the review process, the study adhered to principles of academic integrity by accurately representing original findings, providing appropriate citation for all referenced works, and conducting the synthesis through critical interpretation rather than textual reproduction, thereby ensuring transparency, credibility, and ethical compliance.

RESULTS AND DISCUSSION

Results

The findings reveal that the current landscape of Arabic NLP is characterized by significant advances in LLM-based architectures alongside emerging opportunities and persistent gaps in educational applications. To facilitate a comprehensive presentation of the findings, the results are organized into two main themes: (1) the development of LLM-based Arabic NLP and (2) pedagogical opportunities and research gaps related to Arabic language learning. The following sections present the principal findings representing the overall patterns identified across the reviewed studies.

Large Language Models in Arabic Natural Language Processing

The systematic review demonstrates that transformer-based architectures and Large Language Models (LLMs) have become the dominant paradigm in contemporary Arabic Natural Language Processing (NLP). Across the selected studies, transformer-based models, including AraBERT, AraGPT2, ByT5, and ALLAM, consistently replace conventional lexicon-based and classical machine learning approaches in a wide range of Arabic language processing tasks. The findings further show an increasing adoption of fine-tuning, transfer learning, and domain adaptation techniques to improve model performance across diverse linguistic contexts, particularly in low-resource and domain-specific settings. In addition to architectural advances, the reviewed literature reveals substantial expansion in the range of NLP applications, indicating that LLMs are increasingly employed beyond single-task solutions. A synthesis of the principal findings is presented in Table 1.

Table 1. Synthesis of LLM-Based Arabic NLP: Architectures, Tasks, Data Strategies, and Evaluation Trends

Dimension	Subcategory	Key Findings
Model Architecture	Transformer-based Models	Dominance of BERT, GPT, and variants (AraBERT, AraGPT2, ByT5) as state-of-the-art architectures
	Large Language Models (LLMs)	Shift toward foundation models capable of multi-task learning and contextual understanding
	Hybrid & Fine-tuned Models	Use of fine-tuning, transfer learning, and knowledge distillation for domain adaptation
Task Applications	Sentiment Analysis	Transition from polarity detection to aspect-based sentiment analysis
	Text Classification	Wide use in cyberbullying, legal, and domain-specific classification tasks
	Question Answering (QA)	Emergence of few-shot and table-based QA systems
	Machine Translation (MT)	Transformer-based NMT significantly improves translation accuracy
	Other Applications	Fake review detection, hate speech detection, summarization
Data Strategies	Dataset Development	Growth of specialized datasets (e.g., Quranic treebank, event coreference datasets)
	Data Augmentation	Use of LLM-based augmentation and hybrid techniques (e.g., SMOTE + AraGPT2)
	Synthetic Data Generation	Automated corpus construction using LLMs for low-resource scenarios
Dialect Processing	Dialect Identification	Increasing focus on Saudi, Moroccan, and Jordanian dialects
	Low-resource Dialects	Adaptation of LLMs for dialect-specific tasks (e.g., Atlas-Chat)
Evaluation Trends	Performance Metrics	Continued reliance on accuracy, precision, recall, and F1-score
	Trust & Safety	Emergence of evaluation frameworks addressing trustworthiness and safety
	Reasoning & Robustness	Increasing attention to reasoning capabilities and robustness of LLM outputs

Table 1 summarizes five major dimensions characterizing current developments in Arabic NLP research. In terms of model architecture, the reviewed studies consistently report a transition from conventional task-specific models to transformer-based foundation models capable of supporting multiple NLP tasks within a unified framework. Fine-tuned and hybrid architectures also

appear frequently, reflecting the increasing emphasis on adapting pretrained models to specialized linguistic domains and low-resource environments. This architectural evolution is accompanied by broader model scalability and greater flexibility across diverse Arabic language processing tasks.

The analysis also reveals that task applications remain concentrated in several major domains. Sentiment analysis, text classification, question answering, and machine translation constitute the most frequently investigated applications, while fake review detection, hate speech detection, and text summarization have emerged as additional areas of implementation. Among these applications, sentiment analysis has progressed from simple polarity classification toward aspect-based sentiment analysis, whereas question answering systems increasingly incorporate few-shot learning and table-based reasoning. Machine translation research similarly demonstrates extensive adoption of transformer-based neural architectures for Arabic language translation.

Another prominent finding concerns the growing importance of data-oriented strategies in Arabic NLP. The reviewed studies report continuous expansion of specialized datasets, including benchmark corpora designed for specific linguistic tasks and domains. Data augmentation has become a common strategy for improving model performance, particularly in low-resource settings, while synthetic corpus generation is increasingly employed to overcome limitations in annotated Arabic data. These approaches collectively indicate an increasing emphasis on improving both the quantity and diversity of linguistic resources available for Arabic NLP research.

Dialect processing also represents one of the fastest-growing research areas identified in this review. Current studies increasingly address dialect identification and dialect-specific adaptation using transformer-based architectures and LLMs, with particular attention given to Saudi, Moroccan, and Jordanian Arabic. Alongside advances in dialect processing, evaluation practices have expanded beyond conventional performance indicators. Although accuracy, precision, recall, and F1-score remain the dominant evaluation metrics, recent studies increasingly incorporate additional dimensions such as trustworthiness, safety, reasoning capability, and robustness. These developments indicate a broader evaluation landscape in which model performance is assessed using both technical and reliability-oriented criteria.

Discussion

The findings indicate that the dominance of transformer-based architectures and Large Language Models (LLMs) marks a substantive shift in Arabic Natural Language Processing (NLP), rather than a simple improvement in computational performance. This transition is conceptually important because it moves Arabic NLP from isolated task-specific systems toward more flexible models capable of contextual understanding, transfer across domains, and multi-task learning. Such a pattern is consistent with recent evidence showing that models such as AraBERT, AraGPT2, and ALLAM outperform conventional lexicon-based and machine-learning approaches across sentiment analysis, translation, classification, and question-answering tasks (Khader et al., 2025; Saiful Bari et al., 2025; Alrashidi & Mathkour, 2026; Alrayzah et al., 2026; Ferroud et al., 2026). The significance of this development is particularly strong in Arabic, where rich morphology, orthographic ambiguity, and dialect variation impose challenges that are less pronounced in many high-resource languages (Boulesnam & Boucetti, 2025; Dahou et al., 2025). In theoretical terms, the findings extend the notion of foundation models by showing that their value in Arabic depends not only on model scale, but also on linguistic adaptation and contextual sensitivity. Therefore, the current landscape should be understood as an LLM-centered ecosystem whose progress is inseparable from the structural characteristics of Arabic itself.

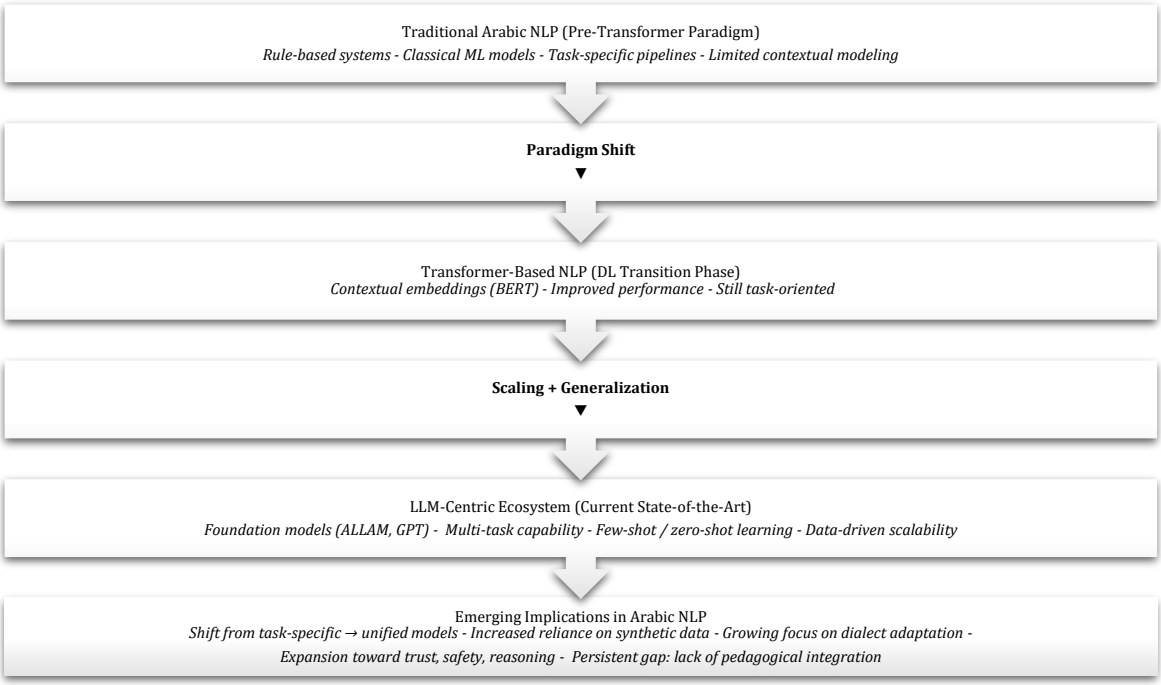

Figure 2. Paradigm Shift in Arabic NLP: From Task-Specific Models to LLM-Centric Ecosystems

As illustrated in Figure 2, the evolution of Arabic NLP has occurred through successive stages, beginning with rule-based systems, progressing through machine learning and deep learning, and culminating in LLM-driven architectures. However, the reviewed evidence also reveals a paradox: although LLMs are designed as general-purpose models, their applications in Arabic NLP remain largely fragmented across separate tasks such as sentiment analysis, cyberbullying detection, legal classification, machine translation, and fake-review detection (Hithnawi et al., 2025; Tahtah et al., 2025; Mohawesh et al., 2025). This fragmentation suggests that the technological generality of LLMs has not yet produced equivalent conceptual integration at the application level. One possible explanation is that current research incentives continue to privilege benchmark-specific gains over cross-task transfer, interoperability, and authentic use contexts. Studies on domain-specific term extraction and Arabic machine-generated text detection further support this pattern because they improve technical specialization without necessarily demonstrating broader pedagogical or communicative relevance (Al-Thubaity, 2025; Al-Shaibani & Ahmed, 2026). Consequently, the findings modify the dominant narrative of LLM maturity by showing that architectural unification does not automatically result in functional or educational integration.

A second major issue concerns the growing dependence on data augmentation, synthetic corpus generation, and specialized benchmark construction. These strategies have become necessary because Arabic remains under-resourced relative to English, particularly in dialectal, domain-specific, and pedagogically annotated datasets. Hybrid augmentation using AraBART, AraGPT2, and Borderline-SMOTE, automated multidialectal corpus generation, and newly developed resources such as AraEventCoref and Quranic treebanks demonstrate clear progress in expanding the available data infrastructure (Abdhood et al., 2025; Almeman, 2025; Aldawsari & Dawood, 2025; Nashir et al., 2025). Nevertheless, the reliance on synthetic data introduces methodological risks because generated corpora may reproduce model bias, overrepresent Modern Standard Arabic, or simplify sociolinguistic variation. This concern becomes more pronounced when dialects are treated merely as classification labels rather than socially situated language varieties. The findings therefore support data-centric approaches, but they also extend them by emphasizing representativeness, cultural validity, and educational usefulness as essential quality criteria rather than viewing data quantity alone as sufficient.

The increasing attention to dialect processing further demonstrates that linguistic diversity has become a central concern in Arabic NLP. Research on Saudi Arabic classification, Moroccan Arabic adaptation, and comparative dialect modeling confirms that transformer-based systems can

be adjusted to underrepresented varieties through fine-tuning and dialect-specific corpora (Aftan et al., 2026; Shang et al., 2025; Essameldin et al., 2025). At the same time, these advances remain weakly connected to Arabic language education, where dialect awareness can influence listening comprehension, sociolinguistic competence, and learners' ability to navigate differences between Modern Standard Arabic and everyday speech. This separation between computational dialect modeling and instructional use limits the practical value of current innovations. It also reveals that linguistic authenticity has not yet been translated into pedagogical design, despite its importance for communicative competence. Thus, the study extends previous dialect-processing research by positioning dialect adaptation not only as a technical challenge, but also as a curricular and pedagogical requirement for future Arabic learning systems.

Another important finding is the gradual expansion of evaluation beyond accuracy, precision, recall, and F1-score toward trustworthiness, safety, reasoning, and robustness. Frameworks such as AraTrust, AraSafe, and AraReasoner represent an important methodological advance because they recognize that high-performing models may still produce unreliable, unsafe, or poorly reasoned outputs (E. A. Alghamdi et al., 2025; Mubarak et al., 2025; Hasanaath et al., 2025). However, these approaches remain predominantly system-centered and do not yet capture how learners and teachers experience AI-generated feedback, explanations, or dialogue. In educational settings, model quality also depends on whether users understand the output, trust the system appropriately, and can distinguish helpful guidance from misleading responses. Studies on Arabic machine-generated text detection reinforce this concern by demonstrating the need to evaluate not only output authenticity, but also how such outputs are interpreted and used (Al-Shaibani & Ahmed, 2026). The findings therefore broaden current evaluation theory by arguing that pedagogical validity, usability, and learner interaction should complement technical reliability in any assessment of LLM-based Arabic learning tools.

The most critical gap identified in this review is the limited integration between Arabic NLP and established theories of Second Language Acquisition (SLA). Core SLA perspectives emphasize comprehensible input, meaningful interaction, feedback, and opportunities for output, all of which align closely with the generative and interactive capacities of LLMs. Question-answering systems, summarization tools, dialogue generation, and text correction can potentially support these processes, yet current applications are rarely framed in explicitly pedagogical terms (Alrayzah et al., 2026; Bakr et al., 2025; Ibrahim et al., 2026). Domain-specific applications in medicine and other specialized fields further demonstrate that LLMs can generate contextualized language resources, but these systems are generally designed for classification or professional assistance rather than instructional scaffolding (Ahmed et al., 2025; Ouali et al., 2025). This finding does not contradict SLA theory, but rather reveals that existing Arabic NLP research has not operationalized it. The theoretical contribution of this study lies in repositioning LLMs as potential mediators of input, interaction, feedback, and output rather than treating them only as computational engines.

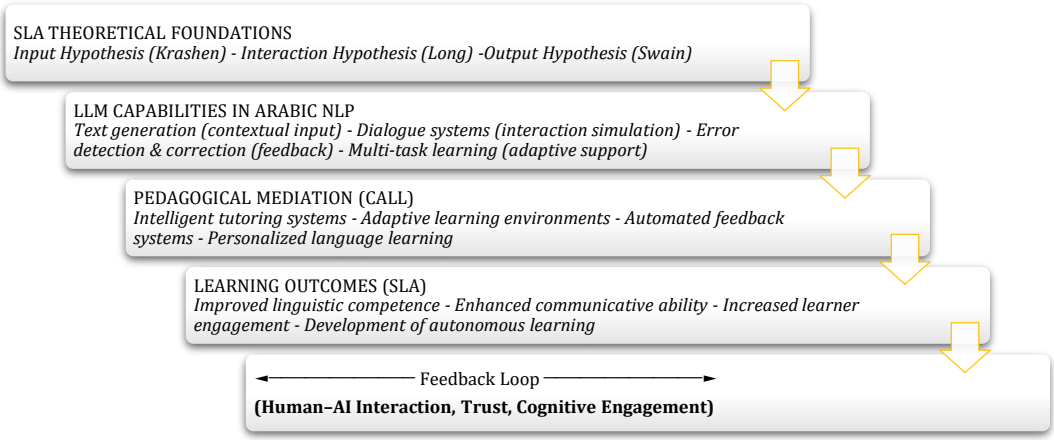

Figure 3. Conceptual Framework for Integrating LLM-Based Arabic NLP into SLA-Oriented CALL Ecosystems

Figure 3 translates this theoretical position into a conceptual CALL ecosystem in which learner input, LLM processing, and pedagogical output are connected through an iterative feedback loop. The proposed framework extends conventional Computer-Assisted Language Learning by placing adaptive language generation, automated correction, classification, and personalized recommendation within a single architecture. It also responds to the limitations of previous studies, which typically isolate question answering, text generation, or classification without linking them to learner goals or instructional progression (Alrayzah et al., 2026; Hithnawi et al., 2025; Mohawesh et al., 2025). In practical terms, such a framework could support intelligent tutoring, dialect-aware interaction, domain-specific Arabic learning, and real-time formative feedback. Nevertheless, its implementation requires careful attention to explainability, cultural appropriateness, privacy, and the possible dominance of Modern Standard Arabic in training data. Accordingly, the framework should be viewed not as a finished model, but as a research agenda requiring empirical validation through learner outcomes, usability studies, and classroom-based evaluation.

Taken together, the findings position Arabic NLP at a point of technological maturity but pedagogical underdevelopment. The field has made substantial progress in model capability, data construction, dialect adaptation, and evaluation, yet these advances have not been matched by systematic integration with SLA, CALL, or human-centered educational design. This imbalance explains why current research demonstrates strong latent pedagogical affordances but provides limited evidence of actual learning impact. The study therefore fills an important gap by connecting technical trends with a theoretically grounded model of Arabic language learning. Future research should test LLM-based Arabic CALL systems through experimental and design-based studies that examine proficiency, communicative competence, learner engagement, trust, and cognitive load. Such work would strengthen the global position of Arabic educational technology by moving the field from technological possibility toward empirically validated pedagogical practice.

CONCLUSION

This study concludes that contemporary Arabic Natural Language Processing is increasingly dominated by transformer-based architectures and Large Language Models, particularly in sentiment analysis, machine translation, question answering, text classification, dataset development, and dialect processing. However, these technological advances remain weakly connected to Arabic language learning, as most studies continue to prioritize technical performance rather than pedagogical outcomes, learner interaction, and educational effectiveness. The review therefore identifies a clear gap between the maturity of Arabic NLP and its integration with Second Language Acquisition and Computer-Assisted Language Learning. To address this gap, the study proposes a conceptual framework that positions LLMs as tools for generating comprehensible input, supporting interaction, providing adaptive feedback, and facilitating learner output. Although the review is limited to recent Scopus-indexed publications and the proposed framework has not yet been empirically tested, the findings provide a foundation for future research on learner-centered, dialect-aware, explainable, and pedagogically validated Arabic CALL systems.

ACKNOWLEDGMENT

The authors would like to express their sincere gratitude to all researchers whose work contributed to this systematic literature review. Their valuable studies provided the foundation for the synthesis and critical analysis presented in this paper. The authors also appreciate the constructive insights from colleagues and reviewers whose feedback helped improve the quality and clarity of this manuscript. Finally, the authors acknowledge the use of scholarly databases and digital research resources that facilitated the literature identification and review process.

AUTHOR CONTRIBUTIONS STATEMENT

Muh. Sabilar Rosyad developed the study concept, designed the analytical framework, conducted data collection and analysis, reviewed relevant literature, and led the writing of the discussion and conclusion. He also ensured overall coherence and finalized the manuscript for

publication. Laila Farah Fitria handled the translation, ensuring linguistic accuracy while preserving the original meaning and academic nuance in the target language.

REFERENCES

Abdhood, S. F., Omar, N., & Tiun, S. (2025). A novel data augmentation framework for Arabic multi-label text classification using AraBART, AraGPT2, and Borderline-SMOTE. *IEEE Access, 13*, 169769–169778. https://doi.org/10.1109/ACCESS.2025.3609462

Aftan, S., Zhuang, Y., Aseeri, A. O., & Shah, H. (2026). A survey of natural language processing for classification of Saudi Arabic dialect: Advancements, opportunities, and challenges. *Lecture Notes of the Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering, 623*, 105–124. https://doi.org/10.1007/978-3-031-92625-9_8

Ahmed, S., Allam, A., Hamdi, A., & Mohammed, A. (2025). Arabic symptom classification and diagnosis using transformer models and LLM-based augmentation. In *2025 3rd International Conference on Intelligent Methods, Systems, and Applications (IMSA)* (pp. 18–23). IEEE. https://doi.org/10.1109/IMSA65733.2025.11167759

Al-Shaibani, M. S., & Ahmed, M. (2026). Arabic machine-generated text detection: Stylometric analysis and cross-model evaluation. *Expert Systems with Applications, 305*, Article 130644. https://doi.org/10.1016/j.eswa.2025.130644

Al-Thubaity, A. (2025). A novel dataset for Arabic domain-specific term extraction and comparative evaluation of BERT-based models for Arabic term extraction. *ACM Transactions on Asian and Low-Resource Language Information Processing, 24*(9). https://doi.org/10.1145/3748323

Aldawsari, M., & Dawood, O. (2025). AraEventCoref: An Arabic event coreference dataset and LLM benchmarks. *ACM Transactions on Asian and Low-Resource Language Information Processing, 24*(7). https://doi.org/10.1145/3743047

Alghamdi, E. A., Masoud, R. I., Alnuhait, D., Alomairi, A. Y., Ashraf, A., & Zaytoon, M. (2025). AraTrust: An evaluation of trustworthiness for LLMs in Arabic. In *Proceedings of the International Conference on Computational Linguistics* (pp. 8664–8679).

Almeman, K. (2025). Automated building of a multidialectal parallel Arabic corpus using large language models. *Data, 10*(12), Article 208. https://doi.org/10.3390/data10120208

Alrashidi, F., & Mathkour, H. I. (2026). An empirical study of transformer-based neural machine translation for English to Arabic. *Information, 17*(2), Article 198. https://doi.org/10.3390/info17020198

Alrayzah, A., Alsolami, F., & Saleh, M. (2026). AraFastQA: A transformer model for question answering for the Arabic language using few-shot learning. *Computer Speech & Language, 95*, Article 101857. https://doi.org/10.1016/j.csl.2025.101857

Alshahrani, E. S., & Aksoy, M. S. (2025). Adversarially robust multitask learning for offensive and hate speech detection in Arabic text using transformer-based models and RNN architectures. *Applied Sciences, 15*(17), Article 9602. https://doi.org/10.3390/app15179602

Alsolami, F., & Alrayzah, A. (2025). Arabic WikiTableQA: Benchmarking question answering over Arabic tables using large language models. *Electronics, 14*(19), Article 3829. https://doi.org/10.3390/electronics14193829

Bakr, M. A., Hany, M., Osama, A., & Gamal, N. (2025). A transformer-driven bilingual approach to Arabic text summarization. In *2025 3rd International Conference on Intelligent Methods, Systems, and Applications (IMSA)* (pp. 112–117). IEEE. https://doi.org/10.1109/IMSA65733.2025.11167880

Boulesnam, I., & Boucetti, R. (2025). Arabic language characteristics that make its automatic processing challenging. *International Arab Journal of Information Technology, 22*(4), 814–831. https://doi.org/10.34028/iajit/22/4/14

Dahou, A., Dahou, A. H., Cheragui, M. A., Abdedaiem, A., Al-Qaness, M. A. A., Elaziz, M. A., Ewees, A. A., & Zheng, Z. (2025). A survey on dialect Arabic processing and analysis: Recent advances and future trends. *ACM Transactions on Asian and Low-Resource Language Information Processing, 24*(8). https://doi.org/10.1145/3747290

David, I., & Gelbard, R. (2025). Using machine learning for systematic literature review: Case in point, agile software development. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 15*(1), Article e1569. https://doi.org/10.1002/widm.1569

Essameldin, O. A., Elbeih, A. O., Gomaa, W. H., & Elsersy, W. F. (2025). Arabic dialect classification using RNNs, transformers, and large language models: A comparative analysis. In *2025 3rd International Conference on Intelligent Methods, Systems, and Applications (IMSA)* (pp. 472–477). IEEE. https://doi.org/10.1109/IMSA65733.2025.11167859

Ferroud, C., Maghfour, M., & Elouardighi, A. (2026). A comparative study of lexicon-based, machine learning, deep learning, and LLM methods for sentiment analysis on standard and dialectal Arabic texts. *Lecture Notes in Networks and Systems, 1640*, 23–35. https://doi.org/10.1007/978-3-032-07785-1_3

Galhom, A., Abobakr, A., Noseer, M., Abdalgwad, S., Nour, R., & Fares, A. (2025). AASTE: Arabic aspect sentiment triplet extraction. In *2025 7th Novel Intelligent and Leading Emerging Sciences Conference (NILES)* (pp. 453–456). IEEE. https://doi.org/10.1109/NILES68063.2025.11232333

Guessoum, A., Berkani, L., Hadj Ameur, M. S., & Aouichat, A. (2026). Artificial intelligence and large language models for a new education and higher education paradigm in the Arab world. In *Higher education in the Arab world: Artificial intelligence* (pp. 373–397). Springer. https://doi.org/10.1007/978-3-031-99068-7_15

Hamed, I., Sabty, C., Abdennadher, S., Vu, N. T., Solorio, T., & Habash, N. (2025). A survey of code-switched Arabic NLP: Progress, challenges, and future directions. In *Proceedings of the International Conference on Computational Linguistics* (pp. 4561–4585).

Hasanaath, A., Alansari, A., Ashraf, A., Salmane, C., Luqman, H., & Ezzini, S. (2025). AraReasoner: Evaluating reasoning-based LLMs for Arabic NLP. In *Findings of the Association for Computational Linguistics: EMNLP 2025* (pp. 18898–18914). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.findings-emnlp.1028

Hithnawi, R. I., Hamarsheh, M. M. N., & Maree, M. (2025). AraBERT for Arabic cyberbullying detection in Facebook comments. *Journal of Cybersecurity, 11*(1), Article tyaf030. https://doi.org/10.1093/cybsec/tyaf030

Ibrahim, M., Gervás, P., & Méndez, G. (2026). AI-Cinema: A hybrid framework for Arabic movie scenario generation with traditional storytelling and cultural dialogs. *Complexity, 2026*(1), Article 9978799. https://doi.org/10.1155/cplx/9978799

Khader, K. A., Hussein, M. S., & Abu-Issa, A. S. (2025). Adapting large language models for Arabic: Comparative evaluation, fine-tuning, and ethical deployment. *IEEE Access, 13*, 182621–182632. https://doi.org/10.1109/ACCESS.2025.3623796

Kitchenham, B., Pearl Brereton, O., Budgen, D., Turner, M., Bailey, J., & Linkman, S. (2009). Systematic literature reviews in software engineering: A systematic literature review. *Information and Software Technology, 51*(1), 7–15. https://doi.org/10.1016/j.infsof.2008.09.009

Mohawesh, R., AlQarni, A. A., Alkhushayni, S. M., Daradkeh, T., & Bany Salameh, H. (2025). A new multilingual framework for fake reviews detection based on a large language model. *The Journal of Supercomputing, 81*(10). https://doi.org/10.1007/s11227-025-07636-6

Mubarak, H., Mohamed, A., & Hawasly, M. (2025). AraSafe: Benchmarking safety in Arabic LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2025* (pp. 9976–9992). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.findings-emnlp.529

Nashir, W. A., Mohsen, A. M., Al-Shargabi, A. A., Nour, M. K., & Al-Onazi, B. B. (2025). A complete, multi-layered Quranic treebank dataset with hybrid syntactic annotations for classical Arabic processing. *Data in Brief, 62*, Article 111940. https://doi.org/10.1016/j.dib.2025.111940

Ouali, S., El Garouani, S., & Chajia, M. (2025). Integrating artificial intelligence into the Arabic medical domain: A review of current progress, challenges, and future directions. In *2025 International Conference on Circuit, Systems, and Communication (ICCSC)*. IEEE. https://doi.org/10.1109/ICCSC66714.2025.11135224

Page, M. J., Moher, D., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., … McKenzie, J. E. (2021). PRISMA 2020

explanation and elaboration: Updated guidance and exemplars for reporting systematic reviews. *BMJ, 372*, Article n160. https://doi.org/10.1136/bmj.n160

Saiful Bari, M., Alnumay, Y., Alzahrani, N. A., Alotaibi, N. M., Alyahya, H. A., AlRashed, S., Mirza, F. A., Alsubaie, S. Z., Alahmed, H. A., Alabduljabbar, G., Alkhathran, R., Almushayqih, Y., Alnajim, R., Alsubaihi, S., Al Mansour, M., Alrubaian, M., Alammari, A., Alawami, Z., Al-Thubaity, A., … Khan, H. (2025). ALLAM: Large language models for Arabic and English. In *The 13th International Conference on Learning Representations (ICLR 2025)* (pp. 59235–59270).

Shang, G., Abdine, H., Khoubrane, Y., Mohamed, A., Abbahaddou, Y., Ennadir, S., Momayiz, I., Ren, X., Moulines, E., Nakov, P., Vazirgiannis, M., & Xing, E. (2025). Atlas-Chat: Adapting large language models for low-resource Moroccan Arabic dialect. In *Proceedings of the International Conference on Computational Linguistics* (pp. 9–30).

Shi, Z., & Agrawal, R. (2025). A comprehensive survey of contemporary Arabic sentiment analysis: Methods, challenges, and future directions. In *Findings of the Association for Computational Linguistics: NAACL 2025* (pp. 3760–3772). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.findings-naacl.208

Snyder, H. (2019). Literature review as a research methodology: An overview and guidelines. *Journal of Business Research, 104*, 333–339. https://doi.org/10.1016/j.jbusres.2019.07.039

Tahtah, O., Zinedine, A., & Fardousse, K. (2025). Arabic legal text classification using pre-trained transformers and deep learning. In *2025 International Conference on Circuit, Systems, and Communication (ICCSC)*. IEEE. https://doi.org/10.1109/ICCSC66714.2025.11135374

Thomas, J., & Harden, A. (2008). Methods for the thematic synthesis of qualitative research in systematic reviews. *BMC Medical Research Methodology, 8*, Article 45. https://doi.org/10.1186/1471-2288-8-45