TYPE Correction
PUBLISHED 12 November 2025
DOI 10.3389/fcomp.2025.1734114

OPEN ACCESS

APPROVED BY
Frontiers Editorial Office,
Frontiers Media SA, Switzerland

*CORRESPONDENCE
Frontiers Production Office
✉ production.office@frontiersin.org

RECEIVED 28 October 2025
ACCEPTED 28 October 2025
PUBLISHED 12 November 2025

CITATION
Frontiers Production Office (2025) Correction: Optimizing architectural-feature tradeoffs in Arabic automatic short answer grading: comparative analysis of fine-tuned AraBERTv2 models. *Front. Comput. Sci.* 7:1734114.
doi: 10.3389/fcomp.2025.1734114

Correction: Optimizing architectural-feature tradeoffs in Arabic automatic short answer grading: comparative analysis of fine-tuned AraBERTv2 models

Frontiers Production Office*

Frontiers Media SA, Lausanne, Switzerland

KEYWORDS

large language model (LLMs), AraBERT, neural network, Arabic natural language processing, educational assessment, Automated Short Answer Grading (ASAG)

A Correction on
Optimizing architectural-feature tradeoffs in Arabic automatic short answer grading: comparative analysis of fine-tuned AraBERTv2 models

by Mahmood, S. A. (2025). *Front. Comput. Sci.* 7:1683272. doi: 10.3389/fcomp.2025.1683272

There was a mistake in the article as published. Tables 1–7 and Figures 1–8 were published as supplementary material when they should have been added to the main article. The corrected figures and tables appear below.

All in-text Supplementary Table and Supplementary Figure in-text citations have been changed to Table and Figure in-text citations.

The original version of this article has been updated.

Generative AI statement

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

TABLE 1 Distribution of answers by question type.

Question type	Question type (In Arabic)	Total questions	Total answers
Define the scientific term	عرف المصطلح العلمي	6	291
Explain	إشرح	21	830
What are the consequences of	ما النتائج المترتبة على	6	282
Justify or give reasons for	علل	10	465
What is the difference between	ما الفرق بين	5	217
Total	5 types	48	2,085

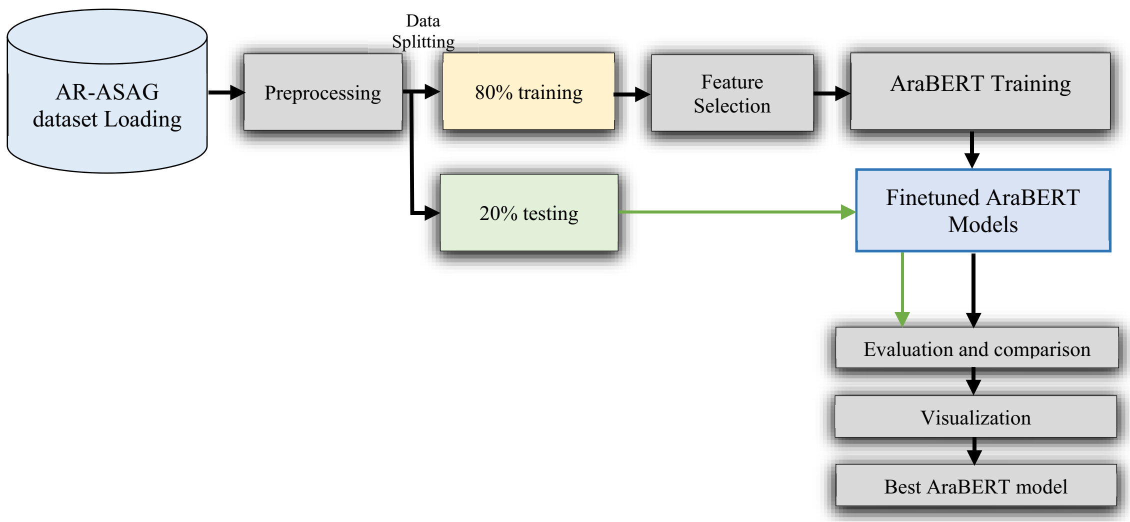

FIGURE 1
General workflow of the proposed automated Arabic short-answer grading model using AraBERTv2.

TABLE 2 Detailed distribution of randomly sampled responses across selected questions.

Q–No.	Question type	Total answers	Training answers	Test answers
1	Define the scientific term	46	36	10
26	Explain	47	37	10
28	What are the consequences of	48	38	10
35	Justify or give reasons for	51	40	11
45	What is the difference between	36	28	8

TABLE 3 Performance evaluation of AraBERTv2 with MLP model using different feature sets: training vs. testing results.

Model	Stage	No. of feature	MAE	RMSE	Pearson correlation	Spearman's correlation	Epoch 1–5
AraBERTv2 with MLP	Training	2-feature	1.14	1.51	0.847	0.85	898 → 533 → 347 → 250 → 156
		3-feature	1.2	1.58	0.818	0.816	1,026 → 614 → 263 → 185
		4-feature	0.18	0.2	0.999	0.998	713 → 34 → 13 → 9 → 7
	Testing	2-feature	1.31	1.76	0.803	0.808	
		3-feature	1.48	1.9	0.744	0.746	
		4-feature	1.77	2.22	0.691	0.689	

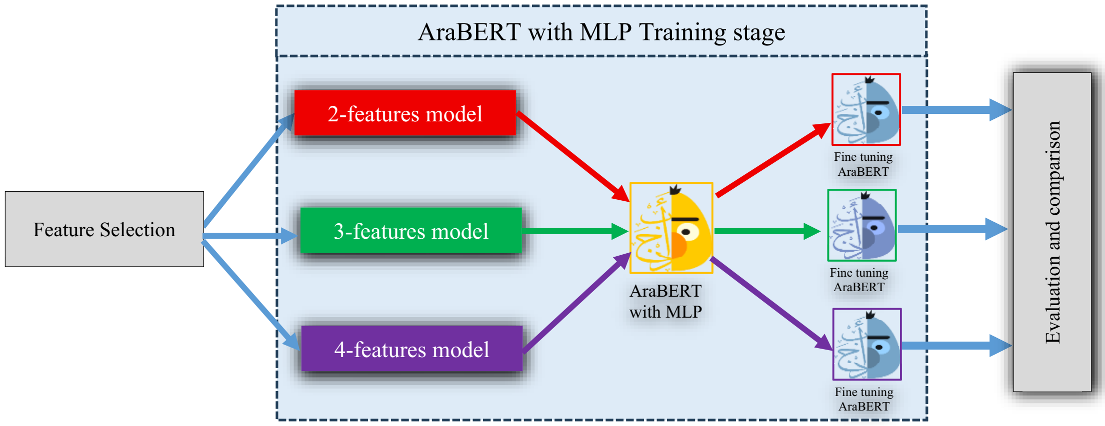

FIGURE 2
The AraBERT_MLP training methodology.

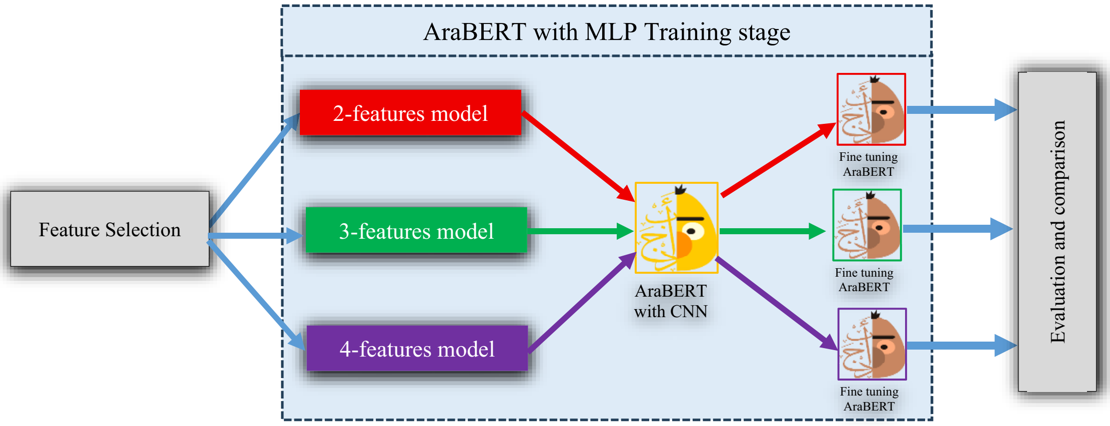

FIGURE 3
The AraBERT_CNN training methodology.

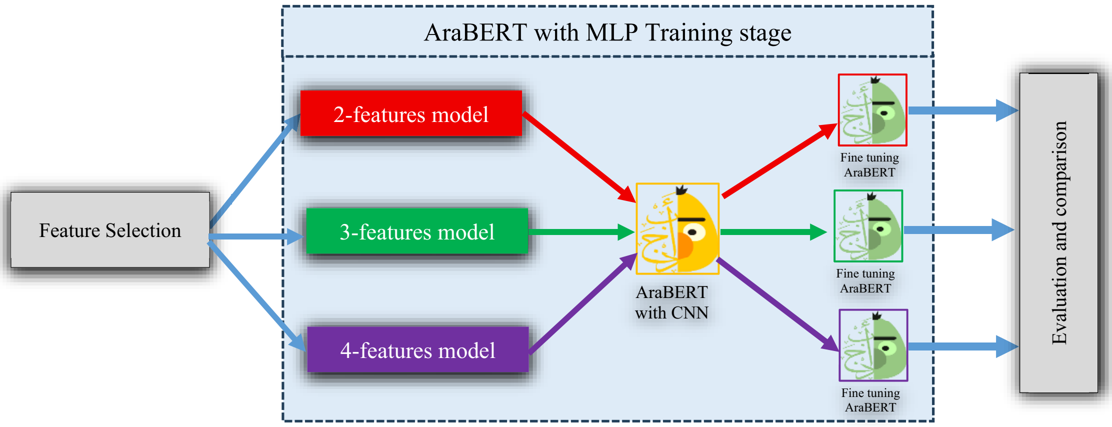

FIGURE 4
The AraBERT_LSTM training methodology.

10.3389/fcomp.2025.1734114

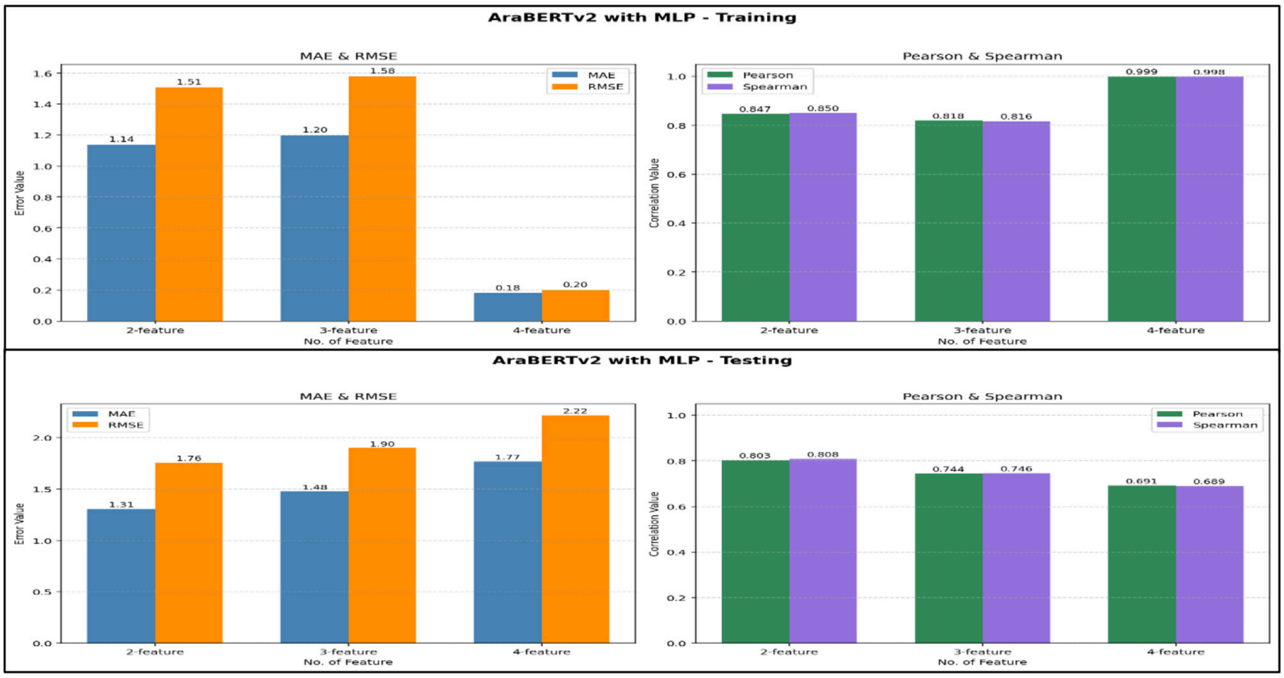

FIGURE 5
Performance evaluation of AraBERTv2 with MLP model using different feature sets: training vs. testing results.

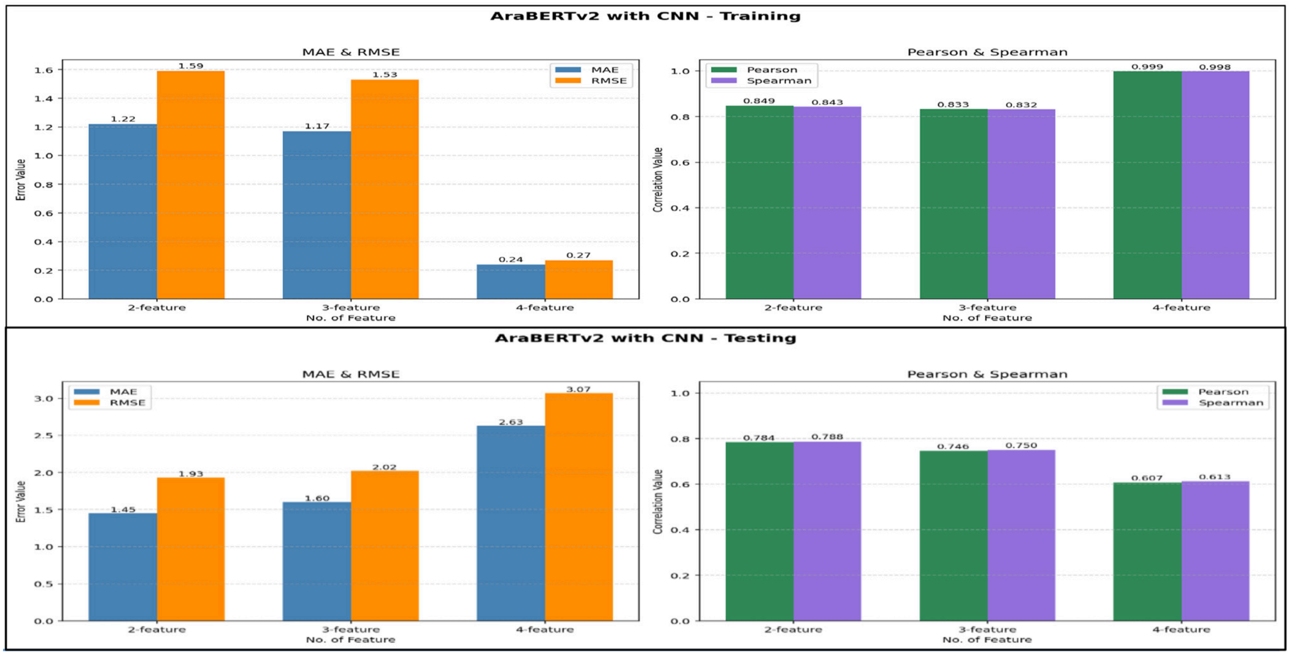

FIGURE 6
Performance evaluation of AraBERTv2 with CNN model using different feature sets: training vs. testing results.

TABLE 4 Performance evaluation of AraBERTv2 with CNN model using different feature sets: training vs. testing results.

Model	Stage	No. of features	MAE	RMSE	Pearson correlation	Spearman's correlation	Epoch 1–5
AraBERTv2 with CNN	Training	2-feature	1.22	1.59	0.849	0.843	1,092 → 610 → 427 → 306 → 227
		3-feature	1.17	1.53	0.833	0.832	1,057 → 567 → 379 → 280 → 205
		4-feature	0.24	0.27	0.999	0.998	773 → 28 → 12 → 8 → 6
	Testing	2-feature	1.45	1.93	0.784	0.788	
		3-feature	1.6	2.02	0.746	0.75	
		4-feature	2.63	3.07	0.607	0.613	

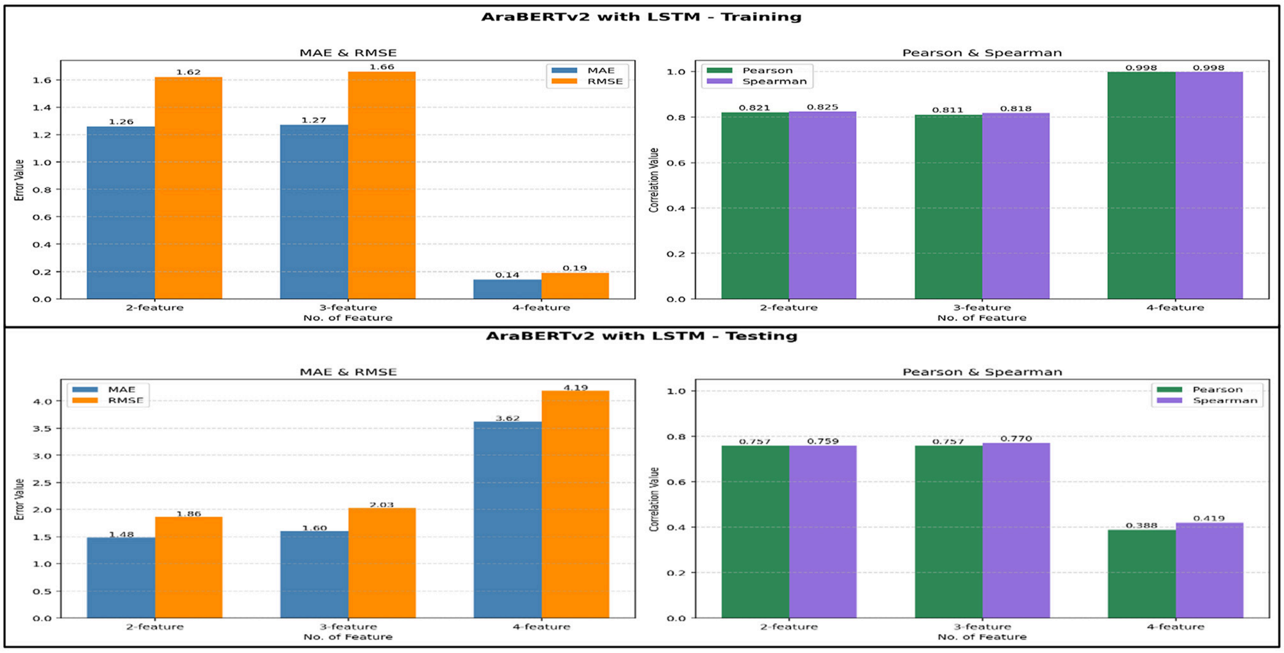

FIGURE 7
Performance evaluation of AraBERTv2 with LSTM model using different feature sets: training vs. testing results.

TABLE 5 Performance evaluation of AraBERTv2 with LSTM model using different feature sets: training vs. testing results.

Model	Stage	No. of features	MAE	RMSE	Pearson correlation	Spearman's correlation	Epoch 1–5
AraBERTv2 with LSTM	Training	2-feature	1.26	1.62	0.821	0.825	1,147 → 718 → 524 → 356 → 262
		3-feature	1.27	1.66	0.811	0.818	1,141 → 675 → 456 → 349 → 267
		4-feature	0.14	0.19	0.998	0.998	728 → 62 → 31 → 22 → 19
	Testing	2-feature	1.48	1.86	0.757	0.759	
		3-feature	1.6	2.03	0.757	0.77	
		4-feature	3.62	4.19	0.388	0.419	

TABLE 6 Performance comparison of AraBERTv2 fine-tuned models with MLP, CNN, and LSTM architectures using different feature sets.

Fine-tuned models	MAE	RMSE	Pearson correlation	Spearman's correlation
2-features-AraBERTv2 with MLP	**1.31**	**1.76**	**0.803**	**0.808**
2-features-AraBERTv2 with CNN	1.45	1.93	0.784	0.788
2-features-AraBERTv2 with LSTM	1.48	1.86	0.757	0.759
3-features-AraBERTv2 with MLP	1.48	1.9	0.744	0.746
3-features-AraBERTv2 with CNN	1.6	2.02	0.746	0.75
3-features-AraBERTv2 with LSTM	**1.6**	**2.03**	**0.757**	**0.77**
4-features-AraBERTv2 with MLP	**1.77**	**2.22**	**0.691**	**0.689**
4-features-AraBERTv2 with CNN	2.63	3.07	0.607	0.613
4-features-AraBERTv2 with LSTM	3.62	4.19	0.388	0.419

The bold values represent the optimal results obtained from our experimental analysis.

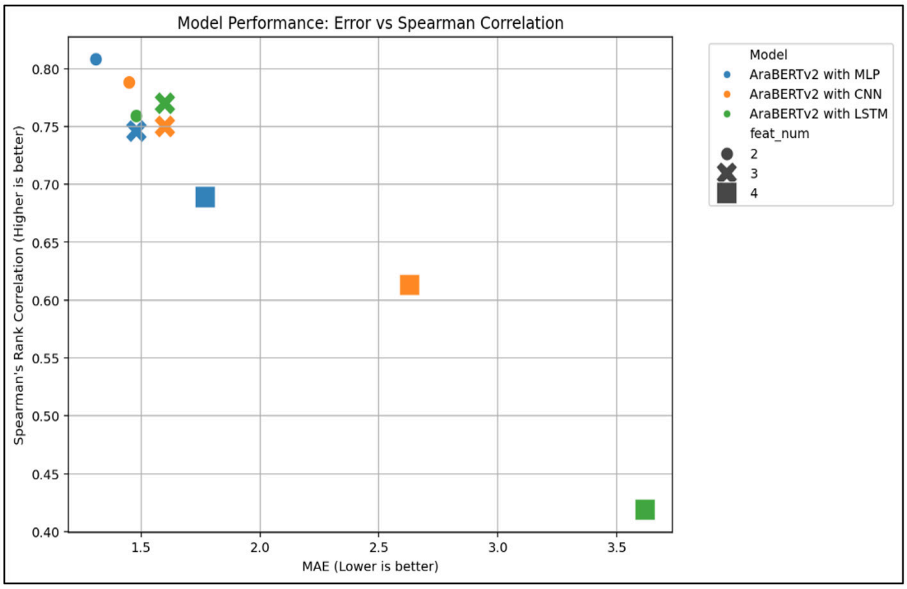

FIGURE 8
Fine-tuned models performance: MAE vs. spearman correlation.

TABLE 7 Comparative performance evaluation of Arabic Automated Short Answer Grading (ASAG) systems.

Criterion/study	Methodology	Dataset	Best RMSE	Best Pearson/Spearman	Key strength	Primary limitation
Our study (AraBERTv2)	- Fine-tuned AraBERTv2 with MLP/CNN/LSTM - Tested 2/3/4 feature configurations	AS-ARSG (2,133 answers)	1.31	- Pearson: 0.803 - Spearman: 0.808	Optimal balance between generalizability and accuracy with limited data	Performance degradation in LSTM with added features
(4)	Latent Semantic Analysis (LSA) with local/hybrid weighting	AR-ASAG (2,133 answers)	N/A	N/A	Effective semantic weighting	Limited capacity for capturing complex contextual relationships
(19)	- BERT vs. Word2Vec/AWN comparison - Intensive text preprocessing	- AR-ASAG (2,133) - Jordanian History (550)	1.00308	Pearson: 0.841902	Demonstrated BERT's superiority over traditional approaches	Heavy dependency on text normalization and stemming