Al Matla : Journal of Arabic Language Education
e-ISSN: 3125-0505
https://doi.org/00.0000/ Al Matla.x0x0.000
Volume 1, Issue. 2, 2026, Page 54-67
journal.staimh.ac.id

PERKEMBANGAN AUTOMATED ESSAY SCORING UNTUK BAHASA ARAB: SYSTEMATIC LITERATURE REVIEW DARI MACHINE LEARNING HINGGA LARGE LANGUAGE MODELS

Imam Safri[*1], Asep Hasan Sulaeman[*2], Firmansah Setia Budi[*3]

[123] Sekolah Tinggi Agama Islam Minhajul Haq Purwakarta Indonesia
e-mail: safri.rifqi20@gmail.com , ahsanlegendkarawang@gmail.com , firmansahsetiabudi212@gmail.com

Submitted: 09-07-2026	Revised : 25-08-2026	Accepted: 27-09-2026	Published: 31-10-2026

ABSTRACT *Student writing assessment is an important component of the educational process. However, manual essay assessment is time-consuming and labor-intensive and may be influenced by evaluator subjectivity. Automated Essay Scoring (AES) has been developed as a technology-based solution to support faster and more consistent writing assessment. Nevertheless, the implementation of AES for Arabic still faces various challenges compared with English. These challenges are related to the characteristics of Arabic, including its root-and-pattern morphology, complex word formation processes, and variations in language use and dialects.This article presents a Systematic Literature Review (SLR) examining the development of AES methodologies for Arabic. The review traces the evolution of approaches from manual feature extraction using traditional Machine Learning, through Deep Learning based on Transformer architectures, to recent developments utilizing generative Large Language Models (LLMs). The findings indicate that Transformer-based models demonstrate strong capabilities in understanding relationships between different parts of a text within a broader context. Meanwhile, LLMs such as Gemini, DeepSeek, and Jais offer new opportunities for more comprehensive writing assessment, including vocabulary, writing style, and mechanics.The review also identifies several important developments, including the use of synthetic data through error injection techniques to address the limited availability of annotated Arabic corpora. In addition, prompt relevance modeling can be used to ensure that assessed essays are genuinely aligned with the assigned topic or question. Overall, the development of Arabic AES demonstrates a shift from systems that merely generate numerical scores toward systems capable of functioning as artificial intelligence-based tutors.Such systems have the potential to provide more informative and transparent feedback while adapting to the characteristics of learners' writing.*

Keywords: *Automated Essay Scoring, Arabic Language, Machine Learning, Large Language Models, Systematic Literature Review, Natural Language Processing.*

ABSTRAK Evaluasi tulisan siswa merupakan bagian penting dalam proses pendidikan. Namun, penilaian esai secara manual membutuhkan banyak waktu dan tenaga serta berpotensi dipengaruhi oleh subjektivitas penilai. *Automated Essay Scoring* (AES) dikembangkan sebagai solusi berbasis teknologi untuk membantu proses penilaian tulisan secara lebih cepat dan konsisten. Meskipun demikian, penerapan AES untuk bahasa Arab masih menghadapi berbagai tantangan dibandingkan bahasa Inggris. Hal tersebut berkaitan dengan karakteristik bahasa Arab yang memiliki morfologi akar-dan-pola, proses pembentukan kata yang kompleks, serta variasi penggunaan bahasa dan dialek. Artikel ini menyajikan *Systematic Literature Review* (SLR) yang membahas perkembangan metodologi AES untuk bahasa Arab. Kajian ini menelusuri perubahan pendekatan mulai dari penggunaan ekstraksi fitur secara manual dengan *Machine Learning* tradisional, kemudian berkembang menuju *Deep Learning* berbasis *Transformer*, hingga perkembangan terbaru yang memanfaatkan *Large Language Models* (LLMs) generatif. Hasil kajian menunjukkan bahwa model berbasis *Transformer* memiliki kemampuan yang baik dalam memahami hubungan antarbagian teks dalam konteks yang panjang. Sementara itu, LLM seperti Gemini, Deepseek, dan Jais memberikan peluang baru dalam melakukan penilaian tulisan secara lebih luas, termasuk aspek kosa kata, gaya bahasa, dan mekanika penulisan. Kajian ini juga menemukan beberapa perkembangan penting, seperti penggunaan data sintetis melalui teknik *error injection* untuk mengatasi keterbatasan korpus bahasa Arab yang telah diberi anotasi. Selain itu, pemodelan *prompt relevance* dapat digunakan untuk memastikan bahwa tulisan yang dinilai benar-benar sesuai dengan topik atau pertanyaan yang diberikan. Secara umum, perkembangan AES bahasa Arab menunjukkan perubahan dari sistem yang hanya menghasilkan skor numerik menuju sistem

© 2026 by the authors; This is an Open Access article distributed under the terms of the Creative Commons Attribution-ShareAlike 4.0 International License. (https://creativecommons.org/licenses/by-sa/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

yang dapat berperan sebagai tutor kecerdasan buatan. Sistem tersebut berpotensi memberikan umpan balik yang lebih informatif, transparan, dan dapat menyesuaikan diri dengan karakteristik tulisan pemelajar.

Kata Kunci: *Automated Essay Scoring*, Bahasa Arab, *Machine Learning*, *Large Language Models*, *Systematic Literature Review*, *Natural Language Processing*.

1. PENDAHULUAN

Evaluasi dan penilaian tulisan merupakan bagian penting dalam pendidikan karena dapat memberikan informasi kepada pendidik mengenai tingkat pemahaman, kemampuan berpikir kritis, dan kemampuan peserta didik dalam menyampaikan gagasan melalui bahasa (Ghazawi, 2025). Namun, penilaian esai secara manual memiliki beberapa kendala. Proses tersebut membutuhkan waktu dan tenaga yang cukup besar, terutama ketika jumlah tulisan yang harus dinilai sangat banyak. Selain itu, hasil penilaian juga dapat dipengaruhi oleh subjektivitas penilai (Barakat et al., 2021; Alqahtani & Alsaif, 2019).

Kondisi tersebut mendorong berkembangnya teknologi *Automated Essay Scoring* (AES) atau sistem penilaian esai otomatis. AES merupakan teknologi berbasis komputasi yang menggunakan Pemrosesan Bahasa Alami (*Natural Language Processing*/NLP) untuk membantu menilai tulisan peserta didik. Sistem ini dapat digunakan untuk menganalisis berbagai aspek tulisan, seperti tata bahasa, koherensi, dan kesesuaian isi dengan topik yang diberikan. Proses penilaian yang dapat dilakukan secara otomatis juga memungkinkan hasil diperoleh dengan lebih cepat dan dapat membantu mengurangi beban kerja pendidik (Fagbohun et al., 2024; Ghazawi & Simpson, 2025).

Perkembangan AES telah lebih dahulu terlihat pada bahasa yang memiliki sumber daya digital cukup besar, khususnya bahasa Inggris. Teknologi tersebut bahkan telah digunakan dalam berbagai bentuk aplikasi penilaian tulisan (Ramesh & Sanampudi, 2022). Sementara itu, pengembangan *Arabic Automated Essay Scoring* (AAES) masih membutuhkan penelitian lebih lanjut. Kondisi ini berkaitan dengan karakteristik bahasa Arab yang memiliki struktur morfologi yang kompleks, keterbatasan dataset berukuran besar yang telah dilengkapi anotasi berdasarkan aspek atau *trait* tertentu, serta adanya variasi dialek yang cukup luas (Ghazawi, 2025; Qwaider et al., 2025).

Pengembangan AES untuk bahasa Arab menjadi penting karena bahasa tersebut digunakan secara luas dalam pendidikan di kawasan Timur Tengah dan Afrika Utara, termasuk dalam berbagai konteks pendidikan Islam. Sistem penilaian otomatis dapat membantu menyediakan proses evaluasi yang lebih objektif, konsisten, dan dapat diterapkan pada jumlah tulisan yang besar. Oleh karena itu, pengembangan teknologi AES yang mampu memahami karakteristik bahasa Arab menjadi kebutuhan yang semakin penting (Putri et al., 2025).

Tinjauan literatur sistematis ini membahas perkembangan AES untuk bahasa Arab dengan melihat perubahan teknologi dari waktu ke waktu. Pembahasan dimulai dari penggunaan ekstraksi fitur secara manual dengan pendekatan *Machine Learning* tradisional, kemudian berlanjut pada penggunaan *Deep Learning*, hingga perkembangan model berbasis *Transformer* dan *Large Language Models* (LLMs) (Ghazawi & Simpson, 2025; Fagbohun et al., 2024).

Kajian ini juga membahas bagaimana perkembangan teknologi NLP dan kecerdasan buatan digunakan untuk menghadapi berbagai karakteristik bahasa Arab. Dengan demikian, pembahasan tidak hanya berfokus pada kemampuan teknis model dalam menghasilkan skor, tetapi juga pada manfaatnya dalam proses pembelajaran. Pendekatan ini diharapkan dapat

memberikan gambaran mengenai perkembangan AES bahasa Arab sekaligus peluang pemanfaatannya sebagai teknologi yang dapat mendukung evaluasi dan pembelajaran bahasa Arab.

2. METODE PENELITIAN

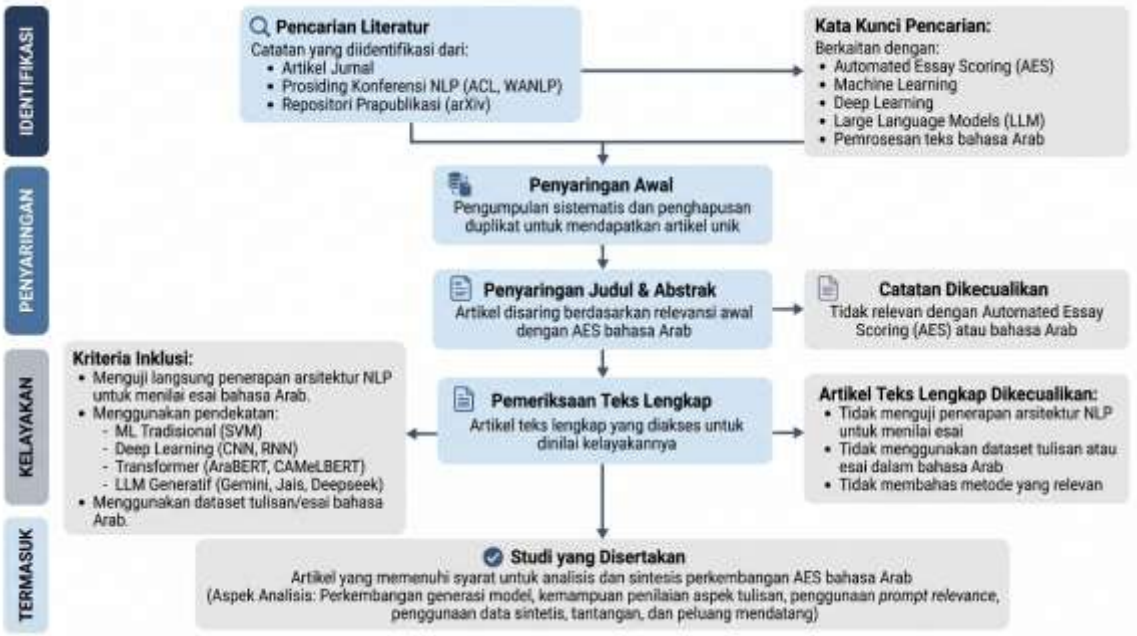
Penelitian ini menggunakan metode *Systematic Literature Review* (SLR) untuk mengidentifikasi, menyeleksi, dan menganalisis berbagai penelitian yang membahas perkembangan *Automated Essay Scoring* untuk bahasa Arab (Putri et al., 2025). Metode SLR digunakan agar berbagai hasil penelitian yang relevan dapat dikumpulkan secara sistematis dan kemudian dibandingkan untuk melihat perkembangan pendekatan AES dari waktu ke waktu.

Proses pencarian literatur mencakup berbagai publikasi ilmiah, seperti artikel jurnal, prosiding konferensi di bidang *Natural Language Processing* (NLP), termasuk ACL dan WANLP, serta artikel prapublikasi yang tersedia melalui repositori seperti arXiv. Pencarian dilakukan dengan menggunakan kata kunci yang berkaitan dengan AES, *Machine Learning*, *Deep Learning*, *Large Language Models*, dan pemrosesan teks bahasa Arab.

Literatur yang dipilih mencakup penelitian yang secara langsung menguji penerapan berbagai arsitektur NLP untuk menilai esai berbahasa Arab. Pendekatan yang dikaji mencakup *Machine Learning* tradisional seperti *Support Vector Machine*, model *Deep Learning* seperti CNN dan RNN, model berbasis *Transformer* seperti AraBERT dan CAMeLBERT, hingga LLM generatif seperti Gemini, Jais, dan Deepseek. Fokus utama diberikan pada penelitian yang menggunakan dataset tulisan atau esai dalam bahasa Arab.

Setelah literatur terkumpul, hasil penelitian kemudian dikelompokkan berdasarkan perkembangan teknologi dan kemampuan yang ditawarkan oleh masing-masing pendekatan. Pengelompokan tersebut mencakup beberapa aspek, yaitu perkembangan teknologi dari setiap generasi model, kemampuan melakukan penilaian terhadap berbagai aspek tulisan, penggunaan *prompt relevance* untuk memastikan kesesuaian tulisan dengan topik yang diberikan, serta penggunaan data sintetis untuk mengatasi keterbatasan data beranotasi.

Melalui tahapan tersebut, penelitian ini berupaya memberikan gambaran yang sistematis mengenai perkembangan AES bahasa Arab, mulai dari sistem berbasis fitur sederhana hingga pemanfaatan LLM yang lebih kompleks. Hasil sintesis kemudian digunakan untuk mengidentifikasi perkembangan utama, tantangan yang masih dihadapi, serta peluang pengembangan AES bahasa Arab pada penelitian selanjutnya.



3. HASIL DAN DISKUSI

3.1 Anatomi Kompleksitas Linguistik Bahasa Arab dalam Komputasi

Sebelum membahas perkembangan metode dalam *Automated Essay Scoring* (AES), penting untuk memahami terlebih dahulu alasan bahasa Arab menjadi salah satu bahasa yang cukup sulit diproses oleh komputer. Kesulitan tersebut tidak hanya berkaitan dengan keterbatasan teknologi, tetapi juga dengan karakteristik bahasa Arab yang memiliki struktur linguistik yang kompleks. Kondisi ini menyebabkan metode NLP konvensional tidak selalu mampu memahami dan menilai tulisan bahasa Arab secara akurat, sehingga diperlukan pendekatan dan pemodelan yang lebih sesuai dengan karakteristik bahasa tersebut (Ghazawi, 2025; Ghazawi & Simpson, 2025).

Salah satu tantangan utama terdapat pada sistem morfologi bahasa Arab yang bersifat non-konkatenatif atau dikenal sebagai sistem akar-dan-pola (*root-and-pattern*). Pada bahasa Inggris, pembentukan kata umumnya dilakukan dengan menambahkan prefiks atau sufiks pada kata dasar. Sebaliknya, bahasa Arab dapat membentuk banyak kata turunan dari satu akar konsonan dengan mengubah pola dan susunan vokal di dalam kata tersebut (Ghazawi & Simpson, 2025). Kondisi ini membuat proses identifikasi kata dasar dan ekstraksi fitur menjadi lebih kompleks.

Selain sistem akar-dan-pola, bahasa Arab juga memiliki karakteristik aglutinatif. Dalam sistem ini, unsur bahasa seperti preposisi, konjungsi, dan kata ganti dapat melekat langsung pada kata utama sehingga membentuk satu rangkaian kata (Ghazawi & Simpson, 2025). Kondisi tersebut dapat menjadi masalah bagi tokenizer standar yang digunakan dalam berbagai model bahasa, termasuk tokenizer berbasis *Byte-Pair Encoding*. Proses tokenisasi dapat menghasilkan pemecahan kata yang terlalu banyak atau *over-segmentation*. Akibatnya, satu kata dapat dipecah menjadi beberapa bagian kecil yang tidak selalu mempertahankan

makna linguistiknya secara utuh. Selain dapat mengurangi kualitas pemahaman teks, kondisi tersebut juga dapat meningkatkan beban komputasi model (Ghazawi, 2025).

Tantangan lainnya berkaitan dengan penggunaan tanda baca dan diakritik atau harakat. Dalam tulisan siswa, tanda baca tidak selalu digunakan secara konsisten, sedangkan harakat umumnya tidak dituliskan. Kondisi ini dapat menimbulkan ambiguitas dalam memahami makna dan fungsi suatu kata dalam kalimat. Oleh karena itu, sistem AES membutuhkan kemampuan memahami konteks yang kuat agar dapat menentukan makna kata secara tepat (Ghazawi & Simpson, 2025; Agrawal, 2025).

Selain faktor struktural, bahasa Arab juga memiliki karakteristik sosiolinguistik berupa diglosia. Dalam praktiknya, siswa dapat menggunakan bahasa Arab Standar Modern (*Modern Standard Arabic*/MSA) bersamaan dengan dialek regional yang biasa digunakan dalam komunikasi sehari-hari. Perpindahan atau pencampuran penggunaan bahasa tersebut dikenal sebagai *code-switching* (Ghazawi, 2025; Alhafni, 2025). Kondisi ini dapat menambah kesulitan bagi sistem AES karena model harus mampu membedakan variasi bahasa formal dan informal sekaligus memahami konteks penggunaannya.

Berbagai karakteristik tersebut menunjukkan bahwa pengembangan AES untuk bahasa Arab tidak cukup hanya dengan menerapkan metode yang telah berhasil digunakan pada bahasa lain. Sistem perlu mempertimbangkan karakteristik morfologi, tokenisasi, tanda baca, harakat, serta variasi penggunaan bahasa Arab. Ringkasan berbagai hambatan tersebut disajikan pada tabel berikut untuk memperlihatkan pengaruh masing-masing karakteristik terhadap pengembangan dan rekayasa sistem AES :

Tabel 1. Karakteristik Linguistik Bahasa Arab dan Implikasinya terhadap Sistem Evaluasi Esai Otomatis (Automated Essay Scoring/AES) Berbasis Kecerdasan Buatan

Karakteristik Linguistik	Deskripsi Teknis	Dampak pada Sistem Evaluasi AES
Morfologi Akar-dan-Pola	Sistem derivasi makna non-konkatenatif dari akar konsonan tunggal.	Pencocokan kata kunci statis gagal; membutuhkan *stemmer* komputasional tingkat lanjut.
Aglutinasi	Afiks (preposisi/konjungsi) menempel langsung pada kata utama.	Menyebabkan *over-segmentasi* pada LLM standar, memperpanjang batas sekuens memori.
Ambiguitas Ortografis	Penulisan tanpa diakritik (harakat) dan inkonsistensi tanda baca.	Mengacaukan deteksi batas kalimat dan memicu kesalahan pemahaman *i'rab* (tata bahasa).
Diglosia / Dialek	Perbedaan radikal antara MSA tertulis dan variasi lisan regional.	Model berisiko memberikan penalti tidak adil (*bias*) pada esai yang mengandung campur kode dialek.

3.2 Evolusi Paradigmatis Sistem Penilaian Esai Bahasa Arab

Berbagai kompleksitas dalam bahasa Arab mendorong para peneliti untuk mengembangkan metode penilaian esai yang semakin sesuai dengan karakteristik bahasa tersebut. Berdasarkan hasil tinjauan literatur, perkembangan AES bahasa Arab dapat dikelompokkan ke dalam empat generasi utama. Setiap generasi menunjukkan perubahan dalam cara sistem memahami, menganalisis, dan memberikan penilaian terhadap tulisan siswa (Ghazawi & Simpson, 2025).

Generasi Pertama (Pendekatan Heuristik dan *Machine Learning*)

Pada tahap awal, sistem AES banyak menggunakan fitur yang dirancang dan diekstraksi secara manual. Algoritma *Machine Learning* klasik seperti *Support Vector Machines* (SVM) dan pohon keputusan digunakan untuk mengolah fitur-fitur tersebut (Alqahtani & Alsaif, 2019). Fitur yang digunakan umumnya masih sederhana, misalnya jumlah kata, jarak *Levenshtein*, serta karakteristik leksikal tertentu. Pendekatan tersebut kemudian dikombinasikan dengan metode *stemming* hibrida dan sumber leksikal seperti *Arabic WordNet* (AWN) untuk meningkatkan kemampuan sistem dalam mengenali karakteristik bahasa Arab (Barakat et al., 2021).

Pendekatan generasi pertama cukup membantu dalam mendeteksi kesalahan tertentu, terutama kesalahan pada tingkat ejaan. Namun, sistem masih memiliki keterbatasan dalam memahami hubungan antarkalimat dan konteks wacana secara keseluruhan. Sistem juga relatif mudah dipengaruhi oleh perubahan bentuk kalimat atau parafrase karena penilaiannya sangat bergantung pada fitur yang telah ditentukan sebelumnya (Ghazawi, 2025; Mahmoud et al., 2024).

Generasi Kedua (*Deep Learning*)

Perkembangan berikutnya ditandai dengan penggunaan *Deep Learning*. Pada tahap ini, sistem mulai menggunakan *word embeddings* dan jaringan saraf seperti LSTM, GRU, serta kombinasi CNN-RNN. Pendekatan tersebut memungkinkan sistem memahami hubungan antarkata dengan lebih baik dan mengenali kesamaan makna, termasuk hubungan antara kata-kata yang memiliki makna serupa (Ghazawi & Simpson, 2025; Machhout et al., 2021).

Meskipun memberikan peningkatan dibandingkan metode sebelumnya, model *Deep Learning* masih memiliki keterbatasan ketika digunakan untuk esai yang panjang. Salah satu masalahnya adalah *vanishing gradient*, yaitu kondisi ketika informasi dari bagian awal teks semakin sulit dipertahankan ketika model memproses bagian teks yang lebih panjang. Akibatnya, ketika sistem sampai pada paragraf akhir sebuah esai argumentatif, informasi penting dari paragraf awal dapat tidak lagi dipertimbangkan secara optimal.

Generasi Ketiga (Model *Transformer* Khusus Bahasa Arab)

Perkembangan selanjutnya terjadi dengan hadirnya arsitektur *Transformer*. Teknologi ini menggunakan mekanisme *self-attention* yang memungkinkan model memperhatikan hubungan antara berbagai bagian teks secara lebih efektif. Dengan pendekatan tersebut, masalah dalam memahami hubungan atau dependensi yang panjang dapat dikurangi.

Model bahasa Arab yang telah melalui proses pra-pelatihan, seperti AraBERT, ARBERT, dan CAMeLBERT, kemudian banyak digunakan dalam berbagai tugas pemrosesan bahasa Arab. Model-model tersebut mampu menganalisis hubungan antarkata dalam konteks yang lebih luas dan menjadi salah satu pendekatan utama dalam pengembangan AES bahasa Arab (Fagbohun et al., 2024; Ghazawi, 2025).

Dalam beberapa eksperimen, pendekatan *Cross-encoder* berbasis BERT mampu menghasilkan tingkat akurasi yang tinggi, dengan nilai *Quadratic Weighted Kappa* (QWK) mencapai 0,88. Hasil tersebut menunjukkan kemampuan model dalam menghasilkan penilaian yang mendekati penilaian manusia. Namun, model ini masih memiliki keterbatasan dalam menjelaskan alasan di balik skor yang diberikan. Dengan kata lain, sistem dapat menghasilkan nilai yang akurat, tetapi belum tentu mampu memberikan penjelasan atau umpan balik yang mudah dipahami dan

bermanfaat bagi siswa untuk memperbaiki tulisannya (Ghazawi & Simpson, 2025; Alqahtani & Alsaif, 2019).

Generasi Keempat (Kognisi Generatif Berbasis LLM)

Generasi terbaru ditandai dengan penggunaan *Large Language Models* (LLMs) seperti GPT-4, Gemini, Deepseek, dan Jais. Perkembangan ini membawa perubahan yang lebih besar terhadap konsep AES. Sistem tidak lagi hanya digunakan untuk memberikan skor akhir, tetapi mulai diarahkan untuk berfungsi sebagai tutor berbasis kecerdasan buatan yang dapat memberikan umpan balik secara langsung dan lebih terperinci.

LLM memiliki kemampuan untuk menganalisis berbagai aspek tulisan, seperti penggunaan bahasa, gaya penulisan, struktur dan abstraksi wacana, serta memberikan saran perbaikan. Berbeda dengan pendekatan sebelumnya yang sering membutuhkan *fine-tuning* menggunakan dataset khusus, LLM dapat dimanfaatkan melalui teknik *zero-shot* atau *few-shot prompting*. Dengan cara tersebut, model dapat diarahkan untuk melakukan tugas penilaian tertentu tanpa harus selalu melalui proses pelatihan ulang yang ekstensif (Ghazawi & Simpson, 2025; Fagbohun et al., 2024).

Perubahan ini menunjukkan bahwa perkembangan AES bahasa Arab tidak hanya berkaitan dengan peningkatan akurasi skor. Fokusnya mulai bergeser menuju sistem yang mampu memahami kualitas tulisan secara lebih menyeluruh dan memberikan umpan balik yang dapat digunakan siswa dalam proses belajar. Dengan demikian, AES berkembang dari sekadar alat evaluasi menjadi bagian dari ekosistem pembelajaran yang lebih bersifat formatif dan interaktif.

Tabel 2. Perkembangan Generasi Arsitektur Sistem Evaluasi Esai Otomatis (Automated Essay Scoring/AES) dan Kapasitas Evaluatifnya dalam Konteks Bahasa Arab

Generasi	Fokus Arsitektural	Logika Evaluasi Inti	Kapasitas Kontekstual	Hasil Output Formatif
Gen 1	*Rule-based*, SVM, *Stemmer* Hibrida, AWN	Deteksi sintaks permukaan, perhitungan panjang kata.	Terisolasi (terbatas tingkat kata tunggal atau klausa).	Sangat Buruk (hanya skor numerik).
Gen 2	Jaringan Saraf (*LSTM, CNN-RNN*)	*Word embeddings*, urutan sekuensial memori.	Moderat (korelasi makna dalam satu kalimat).	Buruk (skor holistik akhir).
Gen 3	*Encoder Transformer* (*AraBERT, CAMeLBERT*)	*Self-attention* paralel, perhitungan *cross-encoder*.	Kuat (menguasai koherensi antar paragraf).	Menengah (skor QWK tinggi, tanpa narasi).
Gen 4	*Generative LLMs* (*Jais, Gemini, GPT-4, Deepseek*)	*Zero-shot / Few-shot prompting*, penalaran kausal.	Sangat Kuat (pemahaman retorika tingkat wacana).	Sangat Baik (skor multidimensi dengan umpan balik).

3.3 Kinerja Komparatif Model LLM Mutakhir dalam Penilaian Multidimensi

Hasil penelitian menunjukkan bahwa tidak ada satu LLM yang selalu memberikan hasil terbaik pada seluruh aspek penilaian AES (Alqahtani & Alsaif, 2019; Ghazawi & Simpson, 2025). Oleh karena itu, penilaian esai perlu dilakukan dengan melihat beberapa aspek atau *trait* secara terpisah. Pendekatan ini memungkinkan sistem untuk mengetahui kelebihan dan keterbatasan masing-masing model dalam menilai tulisan.

Dalam beberapa penelitian, **Gemini** menunjukkan kemampuan yang kuat dalam menilai aspek yang membutuhkan pemahaman dan penyusunan gagasan secara lebih abstrak, terutama dalam

hal pengembangan ide dan penggunaan kosakata. Penggunaan mekanisme yang memungkinkan model melakukan proses "revisi internal" juga dilaporkan dapat meningkatkan kinerja Gemini hingga 70% dalam penilaian aspek kosakata (Ghazawi & Simpson, 2025).

Sementara itu, **Deepseek** menunjukkan kemampuan yang kuat dalam menilai aspek mekanika penulisan. Model ini dapat mengidentifikasi kesalahan pada tingkat yang lebih rinci, termasuk berbagai kesalahan tata bahasa yang terdapat dalam teks (Ghazawi & Simpson, 2025). Perbedaan tersebut menunjukkan bahwa kemampuan setiap LLM dapat berbeda bergantung pada aspek tulisan yang dinilai.

Model yang dikembangkan secara khusus dengan mempertimbangkan karakteristik bahasa Arab, seperti **Jais**, juga menunjukkan keunggulan pada proses tokenisasi. Jais menggunakan *tokenizer* berbasis BPE yang dirancang untuk menangani bahasa Arab dan bahasa lainnya. Model ini dilatih menggunakan sekitar 10 miliar token bahasa Arab dan bahasa pemrograman. Pendekatan tersebut membantu mengurangi masalah yang dapat muncul ketika tokenizer yang dikembangkan terutama untuk bahasa Inggris digunakan pada teks Arab, misalnya pemotongan preposisi Arab menjadi bagian-bagian yang kurang sesuai (Ghazawi & Simpson, 2025; Habash et al., 2025).

Selain menggunakan satu model, penelitian juga menunjukkan potensi penggunaan metode *Ensemble*. Pendekatan ini menggabungkan hasil dari beberapa LLM, misalnya dengan menggunakan MLP untuk menggabungkan keputusan dari beberapa model. Hasilnya menunjukkan bahwa kombinasi beberapa model dapat memberikan hasil yang lebih baik dalam menilai aspek gaya dan organisasi tulisan dibandingkan penggunaan satu model saja. Hal ini dimungkinkan karena setiap model dapat memberikan sudut pandang yang berbeda dan saling melengkapi dalam proses evaluasi (Ghazawi & Simpson, 2025).

Menariknya, cara memberikan instruksi kepada model juga memengaruhi hasil penilaian. Penggunaan *bilingual prompt*, yaitu instruksi atau rubrik diberikan dalam bahasa Inggris sementara esai tetap ditulis dalam bahasa Arab, dilaporkan dapat menghasilkan penilaian yang lebih stabil dibandingkan beberapa pendekatan lainnya (Ghazawi, 2025). Temuan ini menunjukkan bahwa tidak hanya jenis model yang penting, tetapi juga cara model tersebut diarahkan dalam melakukan proses penilaian.

3.4 Isu Krusial: Mengevaluasi Relevansi Topik (*Prompt Relevance*)

Salah satu masalah penting dalam sistem AES adalah kemampuan untuk memastikan bahwa isi esai benar-benar menjawab topik atau pertanyaan yang diberikan. Sebuah esai dapat memiliki tata bahasa yang baik dan struktur yang rapi, tetapi isinya dapat saja tidak berkaitan dengan pertanyaan yang diberikan. Jika kondisi tersebut tidak terdeteksi, sistem dapat memberikan skor tinggi kepada tulisan yang sebenarnya tidak menjawab tugas yang diberikan (Habash et al., 2025).

Masalah ini menjadi semakin penting pada penilaian esai bahasa Arab karena pertanyaan atau *prompt* yang diberikan sering kali cukup singkat. Prompt yang pendek dapat membuat sistem lebih sulit menentukan batas dan inti topik yang harus dibahas oleh siswa. Akibatnya, sistem dapat mengalami *topic drift*, yaitu kondisi ketika penilaian bergeser dari topik yang sebenarnya diminta (Habash et al., 2025; Alqahtani & Alsaif, 2019).

Untuk mengatasi masalah tersebut, beberapa penelitian mengembangkan sistem penilaian yang dilakukan melalui beberapa tahap atau lapisan (Habash et al., 2025). Pendekatan tersebut antara lain mencakup:

1. **Ekspansi Prompt Semantik**

   Tahap pertama dilakukan dengan memperluas informasi yang terdapat dalam *prompt*. Model seperti *AraVec Wikipedia-SkipGram* dapat digunakan untuk menemukan kata-kata yang memiliki hubungan atau kemiripan makna dengan kata-kata dalam pertanyaan. Dengan demikian, sistem tidak hanya mencari kesamaan kata secara langsung, tetapi juga mempertimbangkan sinonim dan hubungan makna sebelum membandingkan *prompt* dengan isi esai (Putri et al., 2025).

2. **Pemodelan Similaritas dengan *Cross-lingual Embeddings***

   Tahap berikutnya adalah mengukur tingkat kemiripan makna antara pertanyaan dan jawaban siswa. Model seperti *paraphrase-multilingual-MiniLM* dapat digunakan untuk menghitung jarak atau kedekatan semantik antara gagasan dalam esai dengan inti pertanyaan yang diberikan oleh guru. Pendekatan ini membantu sistem menentukan apakah isi tulisan masih berada dalam topik yang seharusnya dibahas (Putri et al., 2025).

3. **Kategorisasi Menggunakan LLM**

   LLM berukuran relatif ringan juga dapat digunakan secara khusus sebagai pengklasifikasi. Dalam pendekatan ini, model tidak langsung digunakan untuk memberikan skor esai, tetapi terlebih dahulu menentukan apakah jawaban siswa relevan dengan pertanyaan. Hasil klasifikasi sederhana, misalnya berupa "na'am" atau "laa", kemudian digunakan untuk menentukan apakah tulisan dapat dilanjutkan ke tahap penilaian tata bahasa dan aspek lainnya (Putri et al., 2025).

Pendekatan bertahap tersebut menunjukkan bahwa penilaian esai tidak cukup hanya dengan mengukur kualitas bahasa. Sistem juga perlu memastikan bahwa tulisan benar-benar menjawab tugas yang diberikan. Dengan demikian, *prompt relevance* menjadi salah satu komponen penting untuk meningkatkan keandalan dan integritas sistem AES.

3.5 Demistifikasi Korpus Data dan Inovasi Rekayasa Data Sintetis

Selain persoalan model dan metode penilaian, ketersediaan data juga menjadi salah satu tantangan utama dalam pengembangan AES bahasa Arab. Sistem kecerdasan buatan membutuhkan data dalam jumlah dan kualitas yang memadai untuk dapat belajar melakukan penilaian secara akurat. Namun, korpus bahasa Arab yang telah diverifikasi dan dilengkapi dengan anotasi untuk kebutuhan penilaian esai masih relatif terbatas (Qwaider et al., 2025).

Keterbatasan tersebut menjadi persoalan penting karena model AES membutuhkan contoh tulisan yang beragam, lengkap dengan informasi mengenai kualitas dan karakteristik masing-masing tulisan. Data yang sedikit dapat membatasi kemampuan model untuk memahami berbagai jenis kesalahan, gaya penulisan, tingkat kemampuan siswa, dan variasi bahasa yang muncul dalam tulisan.

Untuk mengatasi keterbatasan tersebut, para peneliti memanfaatkan dan mengembangkan berbagai dataset yang tersedia. Selain menggunakan data asli, perkembangan terbaru juga mulai mengarah pada pemanfaatan **data sintetis**, yaitu data yang dibuat atau dimodifikasi secara

terkontrol untuk menghasilkan contoh tulisan dengan karakteristik tertentu. Pendekatan ini dapat membantu memperluas jumlah data pelatihan tanpa sepenuhnya bergantung pada pengumpulan dan anotasi tulisan siswa secara manual.

Dengan demikian, pengembangan AES bahasa Arab tidak hanya bergantung pada peningkatan kemampuan model, tetapi juga pada ketersediaan dan kualitas data yang digunakan. Rekayasa data sintetis menjadi salah satu pendekatan yang berpotensi membantu mengatasi keterbatasan korpus, terutama ketika data tulisan siswa yang telah diberi anotasi masih sulit diperoleh.

Tabel 3. Perbandingan Korpus dan Dataset Evaluasi Esai Bahasa Arab Berdasarkan Karakteristik, Cakupan, dan Keterbatasannya

Nama Korpus Dataset	Skala dan Karakteristik Data	Cakupan dan Keunggulan	Kekurangan dan Keterbatasan
ZAEBUC	Korpus esai kurikulum sarjana dwibahasa (Arab-Inggris).	Disertai label kemahiran standar CEFR Eropa (A1-C2) yang sangat komprehensif (Ghazawi, 2025; Qwaider et al., 2025)[5].	Variasi topik sangat sempit dan skor aslinya privat untuk peneliti.
QAES	Diturunkan dari repositori *Qatari Corpus of Argumentative Writing*.	Repositori publik murni dengan dimensi berlapis dan label relevansi soal eksplisit (Putri et al., 2025)[9].	Hanya memuat dua pertanyaan esai spesifik (risiko *overfitting*) (Qwaider et al., 2025; Habash et al., 2025)[7].
AR-AES	Koleksi 12 ragam pertanyaan ujian terbuka (argumen/deskriptif).	Paling representatif mereplikasi ujian nyata; metrik absolut *benchmark* AraBERT (Ghazawi, 2025)[1].	Rentan kesalahan segmentasi jika diproses tanpa *tokenizer* modern.
AR-ASAG	Koleksi penulisan tingkat pendek.	Optimal untuk akurasi klausa tunggal (Ghazawi & Simpson, 2025)[2].	Tidak berguna untuk evaluasi koherensi antar-paragraf.

Sumber: Diolah penulis berdasarkan Ghazawi dan Simpson (2024), Bashendy et al. (2024), serta dokumentasi korpus ZAEBUC.

Untuk mengatasi keterbatasan jumlah dan variasi data, komunitas NLP mulai mengembangkan pendekatan **data sintetis bersyarat melalui injeksi kesalahan yang ditargetkan** (*targeted error injection*) (Qwaider et al., 2025). Dalam pendekatan ini, peneliti menggunakan GPT-4o untuk menghasilkan ribuan esai dengan tingkat kemampuan bahasa berdasarkan skala CEFR, mulai dari A1 hingga C2. Namun, tulisan yang dihasilkan oleh model bahasa umumnya terlalu baik dan tidak sepenuhnya menggambarkan kesalahan yang biasa ditemukan dalam tulisan siswa.

Oleh karena itu, data sintetis tersebut kemudian dimodifikasi secara sengaja dengan menggunakan algoritma *Maximum Likelihood Estimation* (MLE) dan klasifikasi kesalahan berdasarkan taksonomi *ARETA*. Proses ini memungkinkan peneliti memasukkan jenis kesalahan tertentu sesuai dengan tingkat kemampuan bahasa. Misalnya, kesalahan penggunaan *gender* gramatikal dapat diberikan pada teks tingkat A1, sedangkan kesalahan penggunaan preposisi dapat diterapkan pada teks tingkat B1 (Qwaider et al., 2025; Belkebir & Habash, 2021). Esai yang telah diberi kesalahan secara terkontrol tersebut kemudian digunakan sebagai data

pelatihan untuk membangun model klasifikasi yang lebih mampu mengenali berbagai kesalahan yang menyerupai kesalahan alami dalam tulisan siswa (Qwaider et al., 2025).

3.6 Implikasi Pedagogis, Pelacakan Formatif, dan Mitigasi Bias LLM

Perkembangan AES menunjukkan bahwa keberhasilan sistem tidak cukup diukur hanya berdasarkan nilai akurasi atau ukuran statistik lainnya. Sistem juga perlu memberikan manfaat nyata bagi proses pembelajaran. Dalam perkembangan terbaru, AES mulai diarahkan untuk tidak hanya bertindak sebagai alat pemberi nilai, tetapi juga sebagai sistem yang dapat membantu siswa memahami kesalahan dan memperbaiki kemampuan menulisnya (Alqahtani & Alsaif, 2019; Fagbohun et al., 2024).

Salah satu contoh perkembangan tersebut dapat dilihat pada arsitektur komputasi seperti *Arwi* (Alhafni et al., 2025). Platform ini menonaktifkan fitur koreksi otomatis pada tahap tertentu agar tulisan asli siswa tetap dapat dipertahankan. Selanjutnya, sistem menggunakan modul *Grammar Error Detection & Correction* (GED/C) yang mengacu pada taksonomi *ARETA*. Modul tersebut tidak hanya mendeteksi kesalahan, tetapi juga memberikan penjelasan mengenai alasan suatu bentuk ortografi atau morfologi dianggap tidak sesuai (Alhafni et al., 2025; Grammar correction, 2025).

Hasil penilaian tersebut kemudian dapat diintegrasikan dengan sistem AES untuk melihat perkembangan kemampuan siswa dari waktu ke waktu. Dengan menyimpan riwayat hasil penilaian, sistem dapat membantu menggambarkan *learning trajectory* atau perkembangan belajar siswa secara lebih terstruktur. Pendekatan ini membuat hasil AES tidak berhenti pada satu nilai, tetapi dapat digunakan sebagai informasi untuk memantau perkembangan kemampuan menulis siswa.

Meskipun demikian, penggunaan LLM dalam penilaian otomatis masih memiliki beberapa risiko. Salah satunya adalah *LLM hallucination*, yaitu ketika model memberikan informasi atau koreksi yang sebenarnya tidak tepat. Dalam konteks penilaian bahasa Arab, kondisi ini dapat terjadi ketika sistem menganggap suatu bentuk *i'rab* sebagai kesalahan, padahal bentuk tersebut secara akademis masih benar (Grammar correction, 2025; Agrawal, 2025).

Persoalan lain berkaitan dengan kemungkinan munculnya bias algoritmik. Model yang dikembangkan berdasarkan data tertentu dapat memberikan penilaian yang kurang sesuai terhadap siswa yang menggunakan bahasa Arab sebagai bahasa kedua atau siswa yang menggunakan dialek yang kurang banyak terwakili dalam data pelatihan, seperti dialek Levantine, Teluk, atau Palestina (Fagbohun et al., 2024; Alhafni, 2025). Kondisi tersebut menunjukkan pentingnya memastikan bahwa sistem AES dapat mengenali keragaman latar belakang linguistik siswa.

Salah satu pendekatan yang dapat digunakan untuk mengurangi masalah tersebut adalah memasukkan informasi mengenai latar belakang bahasa atau *user profiling* ke dalam *prompt*. Informasi tersebut dapat membantu model mempertimbangkan latar belakang vernakular siswa ketika melakukan evaluasi, sehingga penilaian tidak hanya didasarkan pada satu standar bahasa secara kaku. Pendekatan ini diharapkan dapat membantu menghasilkan penilaian yang lebih adil dan sesuai dengan konteks pemelajar (Alhafni et al., 2025; IALab UC, 2025).

4. KESIMPULAN

Perkembangan *Natural Language Processing* (NLP) dalam bidang *Automated Essay Scoring* (AES) bahasa Arab menunjukkan perubahan yang cukup besar. Sistem yang pada awalnya menghadapi berbagai keterbatasan dalam memahami struktur morfologi dan sintaksis bahasa Arab kini telah berkembang menuju model yang mampu memahami teks dalam konteks yang lebih luas.

Perkembangan tersebut dapat dilihat melalui empat tahapan utama. Generasi pertama menggunakan pendekatan berbasis *stemming*, ekstraksi fitur, dan *Machine Learning* tradisional. Generasi kedua mulai memanfaatkan jaringan saraf dan model yang mampu memahami hubungan antarkata secara berurutan. Generasi ketiga berkembang melalui penggunaan model *Transformer* yang memiliki kemampuan lebih baik dalam memahami hubungan antarbagian teks. Selanjutnya, generasi keempat memanfaatkan *Large Language Models* (LLMs) yang memungkinkan sistem melakukan analisis dan memberikan umpan balik dengan kemampuan memahami konteks yang lebih kompleks.

Meskipun pengembangan AES bahasa Arab masih menghadapi masalah keterbatasan korpus yang memiliki anotasi multidimensi, berbagai penelitian telah mengembangkan beberapa solusi. Masalah kesesuaian isi esai dengan topik atau *prompt relevance* dapat ditangani melalui penggunaan *cross-lingual embeddings* dan metode pengukuran kesamaan semantik. Sementara itu, keterbatasan jumlah data dapat dikurangi melalui penggunaan data sintetis yang dihasilkan dengan model seperti GPT-4o dan kemudian diberi kesalahan secara terkontrol melalui teknik *error injection*.

Perkembangan sistem seperti platform formatif yang menggabungkan deteksi dan koreksi kesalahan tata bahasa (GED/C), penilaian AES, serta pelacakan perkembangan siswa menunjukkan adanya perubahan fungsi AES. Sistem tidak lagi hanya digunakan untuk menghasilkan skor, tetapi juga mulai digunakan untuk memberikan informasi diagnostik dan membantu siswa memahami kesalahan dalam tulisannya.

Namun, penerapan LLM dalam AES tetap memerlukan perhatian terhadap akurasi, *hallucination*, dan potensi bias terhadap pemelajar dengan latar belakang linguistik yang beragam. Oleh karena itu, pengembangan AES bahasa Arab ke depan perlu tidak hanya berorientasi pada peningkatan akurasi model, tetapi juga pada transparansi penilaian, kualitas umpan balik, keberagaman data, dan keadilan bagi berbagai kelompok pemelajar. Dengan pendekatan tersebut, AES dapat berkembang menjadi teknologi pendukung pembelajaran yang lebih adaptif dan bermanfaat bagi proses pengembangan kemampuan menulis bahasa Arab.

DAFTAR PUSTAKA

Agrawal, T. (2025). *Graphing the truth: Structured visualizations for automated hallucination detection in LLMs*. arXiv preprint arXiv:2512.00673.

Abdurahman, I., Salim, A., Wangsadanureja, M., Alkhotiri, I., & Alpiansyah, R. (2026). Modernization of Human Resource Management in Islamic Boarding Schools: An Analysis of the Outlines of Personnel and Work Systems (GBKSK) Minhajul Haq. *Labina : Journal of Islamic Education Management*, *1*(1), 29-39. https://doi.org/10.66891/8wxyjc16

Alhafni, B. (2025). Benchmarking Arabic authorship attribution and style transfer with large language models. *Proceedings of The Third Arabic Natural Language Processing Conference*.

Alpiansyah, R., Affandy, A., & Dzulfikri. (2026). Branding Strategies For Islamic Educational Institutions: The Use Of Digital Marketing And Influencer Marketing To Attract Gen Z Prospective Students To Islamic-Based Schools. *Labina : Journal of Islamic Education Management*, *1*(2), 01-14. https://doi.org/10.66891/5brb6v48

Alpiansyah, R., Abdurahman, I., Alkhotiri, I., Wangsadanureja, M., & Dzulfikri. (2026). MANAJEMEN KURIKULUM DAN EKSTRAKURIKULER BERBASIS KARAKTER: INTEGRASI PENDIDIKAN JASMANI, PEMBIASAAN SPIRITUAL, DAN NEUROSAINS DALAM PEMBENTUKAN ADAB ISLAMI. *Labina : Journal of Islamic Education Management*, *1*(1), 62-90. https://doi.org/10.66891/v4dpda25

Alqahtani, A., & Alsaif, A. (2019). Automatic evaluation for Arabic essays: A rule-based system. *2019 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)*, 1–7.

Barakat, H. M., Ismail, M. A., & Ravana, S. D. (2021). Automated Arabic essay scoring based on hybrid stemming with WordNet. *Malaysian Journal of Computer Science*, 55–67.

Belkebir, R., & Habash, N. (2021). ARETA: Arabic error taxonomy and annotation guidelines. *Proceedings of the WANLP*.

Fagbohun, O. E., et al. (2024). Beyond scoring: Automated text scoring systems and language learning. *arXiv preprint arXiv:2606.09830v2*.

Ghazawi, R. N. (2025). *Automatic essay scoring in Arabic: Development, evaluation, and advanced techniques* [Doctoral thesis, University of Bristol].

Ghazawi, R. N., & Simpson, E. (2025). Advancing Arabic automated essay scoring through cross-encoder BERT models and interpretable explanations. *Data, 10*(9), 148.

Ghazawi, R. N., & Simpson, E. (2025). How well can LLMs grade essays in Arabic? *arXiv preprint arXiv:2501.16516*.

Habash, N., et al. (2025). Evaluating prompt relevance in Arabic automatic essay scoring: Insights from synthetic and real-world data. *Proceedings of The Third Arabic Natural Language Processing Conference*.

Machhout, R. A., Zribi, C. B. O., & Bouzid, S. M. (2021). Arabic automatic essay scoring systems: An overview study. *International Conference on Intelligent Systems Design and Applications*, 1164–1176.

Mahmoud, S., Nabil, E., & Torki, M. (2024). Automatic scoring of Arabic essays: A parameter-efficient approach for grammatical assessment. *IEEE Access, 12*, 142555–142568.

Putri, R. K. A., Kusaeri, K., & Suparto, S. (2025). Deep learning in automated essay scoring for Islamic education: A systematic review. *Online Learning in Educational Research (OLER), 5*(2), 319–338.

Qwaider, C., Chirkunov, K., Alhafni, B., Habash, N., & Briscoe, T. (2025). Enhancing Arabic automated essay scoring with synthetic data and error injection. *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications*, 549–563.

Ramesh, D., & Sanampudi, S. K. (2022). An automated essay scoring systems: A systematic literature review. *Artificial Intelligence Review, 55*, 2495–2527.