Automated Essay Scoring Berbahasa Indonesia Menggunakan Hybrid Retrieval-Augmented Generation dan Contrastive In-Context Learning

Automated Essay Scoring in Indonesian Using Hybrid Retrieval-Augmented Generation and Contrastive In-Context Learning

Katon Suwida*[1], **Diana Purwitasari**[2], **Rizka Wakhidatus Sholikah***[3]
Program Studi Teknik Informatika, Fakultas Teknik Elektro dan Informatika Cerdas, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia
E-mail : 6025221039@student.its.ac.id[1], diana@if.its.ac.id[2], wakhidatus@its.ac.id*[3]*
**Corresponding author*
Received 25 June 2026; Revised 28 July 2026; Accepted 31 July 2026

Abstrak – *Automated Essay Scoring* (AES) untuk pendidikan biologi berbahasa Indonesia masih menghadapi tantangan akibat keterbatasan dataset beranotasi serta ketidakmampuan model *encoder-only* dalam mengaitkan proses penilaian dengan materi kurikulum. Meskipun *Large Language Models* (LLM) memiliki kemampuan generatif yang baik, penerapannya secara langsung masih menghadapi permasalahan halusinasi dan ketidaksesuaian dengan standar penilaian guru. Penelitian ini mengusulkan *pipeline* AES yang mengintegrasikan *Hybrid Retrieval-Augmented Generation* (RAG) dengan *Question-Anchored Contrastive In-Context Learning* (CICL). *Hybrid retriever* yang digunakan mengombinasikan *sparse retrieval* BM25, *dense retrieval* berbasis IndoBERT, serta *automated keyword boosting*. Studi ablasi terhadap 16 konfigurasi pada 312 esai biologi siswa sekolah menengah dilakukan untuk membandingkan paradigma *zero-shot* (Ollama) dan *parameter-efficient fine-tuning* (LoRA). Hasil menunjukkan bahwa penggabungan sederhana pada *Hybrid* RAG menyebabkan dilusi konteks. Kondisi ini membuat CICL berfungsi sebagai jangkar rubrik yang penting bagi model *zero-shot* (QWK = 0,7914). Temuan menarik mengungkap sebuah perbedaan perilaku: meskipun CICL meningkatkan kinerja model *zero-shot*, mekanisme ini justru bertindak sebagai informasi redundan yang menurunkan akurasi pada model LoRA (QWK = 0,8554) yang telah memiliki representasi internal. Temuan ini memberikan panduan empiris mengenai kapan penggunaan RAG dan CICL diperlukan dan kapan *fine-tuning* lebih sesuai untuk domain spesifik berbahasa Indonesia.

Kata Kunci - *Automated Essay Scoring, Retrieval-Augmented Generation, Contrastive In-Context Learning, LoRA, Hybrid Retrieval, Pendidikan Biologi*

Abstract – *Automated Essay Scoring (AES) for Indonesian biology education still faces challenges due to the lack of annotated datasets and the inability of encoder-only models to link the scoring process with curriculum materials. Although Large Language Models (LLMs) possess good generative capabilities, their direct application still faces hallucination issues and misalignment with teacher scoring standards. This study proposes an AES pipeline that integrates Hybrid Retrieval-Augmented Generation (RAG) with Question-Anchored Contrastive In-Context Learning (CICL). The hybrid retriever combines BM25 sparse retrieval, IndoBERT-based dense retrieval, and automated keyword boosting. An ablation study of 16 configurations on 312 high school student biology essays was conducted to compare the zero-shot (Ollama) and parameter-efficient fine-tuning (LoRA) paradigms. The results show that simple aggregation in Hybrid RAG causes context dilution, where CICL serves as an important rubric anchor for the zero-shot model (QWK = 0.7914). An interesting finding reveals a behavioral difference: although CICL improves the performance of the zero-shot model, this mechanism acts as redundant information that reduces accuracy in the LoRA model (QWK = 0.8554), which already possesses internal*

e-ISSN : 2356-2579 | p-ISSN : 1412-2693, DOI : 10.62411/tc.v25i3.16813

representations. These findings provide empirical guidelines on when RAG and CICL are necessary versus when fine-tuning is more appropriate for specific Indonesian language domains.

Keywords - *Automated Essay Scoring, Retrieval-Augmented Generation, Contrastive In-Context Learning, LoRA, Hybrid Retrieval, Biology Education*

1. PENDAHULUAN

Penilaian esai secara manual memakan waktu dan sering mengalami inkonsistensi penilaian antarevaluator [1], [2], [3]. AES dikembangkan untuk meningkatkan efisiensi dan konsistensi penilaian esai melalui pendekatan komputasional [2], [3]. Sistem AES telah berkembang dari pendekatan berbasis fitur statistik menuju arsitektur *deep learning* dengan kemampuan representasi semantik yang lebih kuat, termasuk model berbasis *Transformer* [1], [2], [3]. Sebuah penelitian sebelumnya yang menggunakan *semantic textual similarity* berbasis *Transformer* menunjukkan bahwa IndoBERT yang telah di-*fine-tune* mencapai MAE sebesar 0,1285 pada tugas penilaian esai bahasa Indonesia [3]. Model ini mengungguli *baseline* berbasis TF-IDF [3]. Namun, sebagian besar pendekatan AES berbasis *encoder* terutama bergantung pada kesamaan semantik antara jawaban siswa dan jawaban referensi. Ketergantungan ini membatasi kemampuan model untuk menyertakan materi pembelajaran eksternal atau pengetahuan yang selaras dengan kurikulum selama proses inferensi [1], [3].

Kemunculan *Large Language Model* (LLM) memungkinkan penilaian yang selaras dengan materi ajar. Namun, penerapan LLM secara langsung menghadapi masalah halusinasi. Model sering menghasilkan skor dan penjelasan yang tidak berdasar [4], [5]. *Retrieval-Augmented Generation* (RAG) mengatasi masalah ini dengan menyisipkan konteks buku teks ke dalam *prompt* model. Meskipun demikian, implementasi RAG standar memiliki tiga keterbatasan utama. Keterbatasan pertama adalah ketergantungan pada strategi *retrieval* bertipe tunggal. Ketergantungan ini membatasi cakupan terhadap terminologi yang beragam dan konsep spesifik domain [4]. Keterbatasan kedua adalah tidak adanya pemilihan kalimat yang terfokus. Kondisi ini membuat model terpapar pada informasi yang tidak relevan [6]. Keterbatasan ketiga adalah tidak adanya referensi batas penilaian secara eksplisit. Ketiadaan referensi ini menyulitkan model untuk membedakan karakteristik jawaban berkualitas tinggi dan rendah tanpa *fine-tuning* berskala besar [4], [6].

Integrasi *sparse retrieval*, *dense retrieval*, *automated keyword boosting* [7], [8] dan CICL [6] dalam satu *pipeline* RAG terpadu untuk penilaian spesifik domain masih belum dieksplorasi dalam literatur AES berbahasa Indonesia. Penelitian yang ada sebagian besar mengkaji peningkatan *retrieval* pada domain lain seperti aplikasi hukum [7]. Penelitian lain mengkaji strategi peningkatan *retrieval* tanpa menyertakan mekanisme kalibrasi kontrastif terstruktur [5], [6]. Penelitian sebelumnya mengenai ekstraksi kata kunci berbasis LLM menunjukkan kelayakan *zero-shot prompting* untuk pembuatan *keyphrase* secara otomatis [8]. Metode ini menawarkan potensi bagi augmentasi *retrieval* tanpa rekayasa kata kunci secara manual. Di sisi lain, pendekatan PEFT seperti LoRA menawarkan alternatif untuk menyimpan pola penilaian ke dalam bobot model. Perbandingan langsung antara pendekatan *zero-shot* RAG dan *fine-tuning* LoRA pada domain AES berbahasa Indonesia belum pernah dilakukan.

Penelitian ini mengusulkan sebuah *pipeline* RAG *hybrid* yang mengintegrasikan beberapa komponen utama. Komponen tersebut meliputi *sparse retrieval* menggunakan BM25, *dense retrieval* menggunakan *embedding* IndoBERT, *automated keyword boosting* melalui *one-shot LLM prompting*, *Cross-Encoder Reranker*, dan *Focus Mode*. Penelitian ini juga menerapkan mekanisme CICL untuk kalibrasi rubrik serta *fine-tuning* LoRA pada model Qwen3-4B. Sistem membandingkan paradigma *zero-shot* dan *fine-tuning* menggunakan *ablation study* yang terdiri dari 16 konfigurasi pada 312 sampel esai Biologi siswa sekolah menengah. Hasil evaluasi menunjukkan bahwa konfigurasi *zero-shot* terbaik mencapai skor *Quadratic Weighted Kappa* (QWK) sebesar 0,7914. Konfigurasi LoRA terbaik mencapai QWK sebesar 0,8554. Temuan ini

e-ISSN : 2356-2579 | p-ISSN : 1412-2693, DOI : 10.62411/tc.v25i3.16813

memberikan panduan empiris mengenai efektivitas kalibrasi rubrik berbasis CICL untuk model *zero-shot*, sekaligus menunjukkan batasan redundansi instruksi pada model yang telah difine-tune.

2. METODE PENELITIAN

Penelitian ini menggunakan pendekatan eksperimental kuantitatif untuk mengevaluasi kinerja *pipeline* Hybrid RAG yang terintegrasi dengan CICL dan *fine-tuning* LoRA untuk tugas AES. Tahapan penelitian terdiri atas lima fase: akuisisi data dan pembangunan basis pengetahuan, pengembangan sistem *hybrid retrieval*, perancangan mekanisme CICL dan *fine-tuning* LoRA, inferensi penilaian, serta evaluasi komparatif melalui *ablation study*. Gambar 1 menunjukkan arsitektur keseluruhan sistem yang diusulkan, yang memetakan aliran data dari masukan pertanyaan dan jawaban siswa melalui modul Hybrid Retrieval dan CICL hingga menghasilkan skor dan alasan penilaian oleh LLM.

2.1. Dataset dan Pembangunan Basis Pengetahuan

Dataset yang digunakan terdiri dari jawaban esai siswa sekolah menengah pada mata pelajaran Biologi dengan topik Sistem Pencernaan. Dataset dibagi menjadi dua subset berdasarkan stratifikasi tipe pertanyaan dan distribusi skor. Subset pelatihan mencakup 80% dari total data, yang berfungsi untuk membangun indeks CICL, mengekstraksi kata kunci, dan melatih *adapter* LoRA. Subset pengujian mencakup 20% sisanya untuk mengevaluasi kinerja model secara eksklusif.

Basis pengetahuan dibangun dari buku teks sains kurikulum nasional. Teks diekstraksi menggunakan pustaka PyMuPDF, dilanjutkan dengan mekanisme koreksi artefak PDF (seperti *kerning* dan *ligature*) untuk memperbaiki kata yang terfragmentasi akibat masalah *rendering font*. Teks yang telah diekstraksi dibersihkan dari elemen pengganggu seperti *header*, *footer*, dan nomor halaman. Selanjutnya, teks bersih dibagi ke dalam dua tingkatan segmentasi: tingkatan paragraf untuk proses RAG utama, dan tingkatan kalimat untuk *Focus Mode*. Setiap *chunk* dianotasi dengan metadata halaman untuk memfasilitasi proses *retrieval* yang terfokus.

2.2. Arsitektur Hybrid Retrieval

Pendekatan *single-retriever* memiliki keterbatasan cakupan terhadap terminologi yang beragam dan konsep spesifik domain [4], Penelitian ini mengusulkan mekanisme *Hybrid Retrieval* untuk mengatasi keterbatasan tersebut dengan mengintegrasikan tiga komponen *retrieval* yang saling melengkapi.

2.2.1. Dense Retrieval (Semantik)

Model Sentence-BERT Bahasa Indonesia [9] (varian *firqaaa/indo-sentence-bert-base*) digunakan untuk menghasilkan *embedding* semantik dari *chunk*. Model ini sebelumnya telah dievaluasi pada tugas AES berbahasa Indonesia dan menunjukkan kinerja yang kompetitif di antara model *pretrained* lainnya [1]. *Embedding* tersebut disimpan ke dalam indeks FAISS (*IndexFlatIP* untuk *Cosine Similarity*) [10]. Skor *dense* dihitung berdasarkan *cosine similarity* antara *embedding* pertanyaan dan *chunk*:

$$S_{dense} = \cos(\mathrm{q}, \mathrm{d}) = \frac{\mathrm{q.d}}{\|\mathrm{q}\|\|\mathrm{d}\|} \quad (1)$$

Variabel $q$menyatakan vektor *query* (pertanyaan) dan $d$menyatakan vektor dokumen (*chunk*).

e-ISSN : 2356-2579 | p-ISSN : 1412-2693, DOI : 10.62411/tc.v25i3.16813

2.2.2. Sparse Retrieval (Leksikal)

Algoritma Best Matching 25 (BM25) digunakan untuk menangkap terminologi biologi secara eksak [11]. Algoritma ini mengandalkan pencocokan leksikal dan frekuensi *term*, yang efektif untuk mengidentifikasi istilah ilmiah spesifik yang mungkin terlewat oleh model semantik. Skor *sparse* dihitung menggunakan rumus berikut:

$$S_{sparse} = \sum_{t \epsilon q} IDF(t). \frac{f(t,d) \cdot (k_1+1)}{f(t,d)+k_1.(1-b+b.\frac{|d|}{avgdl})} \quad (2)$$

Variabel t menyatakan *term* (kata) di dalam *query*, $f(t,d)$ menyatakan frekuensi *term* di dalam dokumen d, $|d|$ menyatakan panjang dokumen, dan $avgdl$menyatakan rata-rata panjang dokumen pada kumpulan data. Sistem menetapkan nilai parameter $k_1 = 1{,}5$ dan $b = 0{,}75$.

2.2.3. Automated Keyword Boosting

Pendekatan kurasi manual memerlukan pembaruan daftar kata kunci oleh pakar domain setiap kali kurikulum berubah [7]. Penelitian ini menggunakan *one-shot prompting* pada LLM untuk mengekstraksi istilah biologi spesifik dan bobot kepentingannya dari *chunk* materi secara otomatis [8]. Kata kunci dikategorikan ke dalam istilah utama, istilah pendukung, proses kunci, dan istilah berbobot tinggi. Skor kata kunci dihitung berdasarkan total bobot kata kunci yang muncul dalam suatu *chunk*:

$$S_{keyword} = min\left(1.0, \frac{\sum_{kw \in K_q \cap k_d} w_{kw}}{25}\right) \quad (3)$$

Variabel K_q menyatakan himpunan kata kunci dalam *query*, K_d menyatakan himpunan kata kunci dalam dokumen, dan w_{kw} menyatakan bobot kata kunci. Angka 25 berfungsi sebagai batas normalisasi agar skor akhir tidak melebihi nilai 1,0.

2.2.4. Penggabungan Skor (Score Fusion) dan Pasca-Pemrosesan

Skor dari masing-masing *retriever* dinormalisasi ke dalam rentang [0,1] menggunakan *Min-Max Scaling* sebelum proses penggabungan:

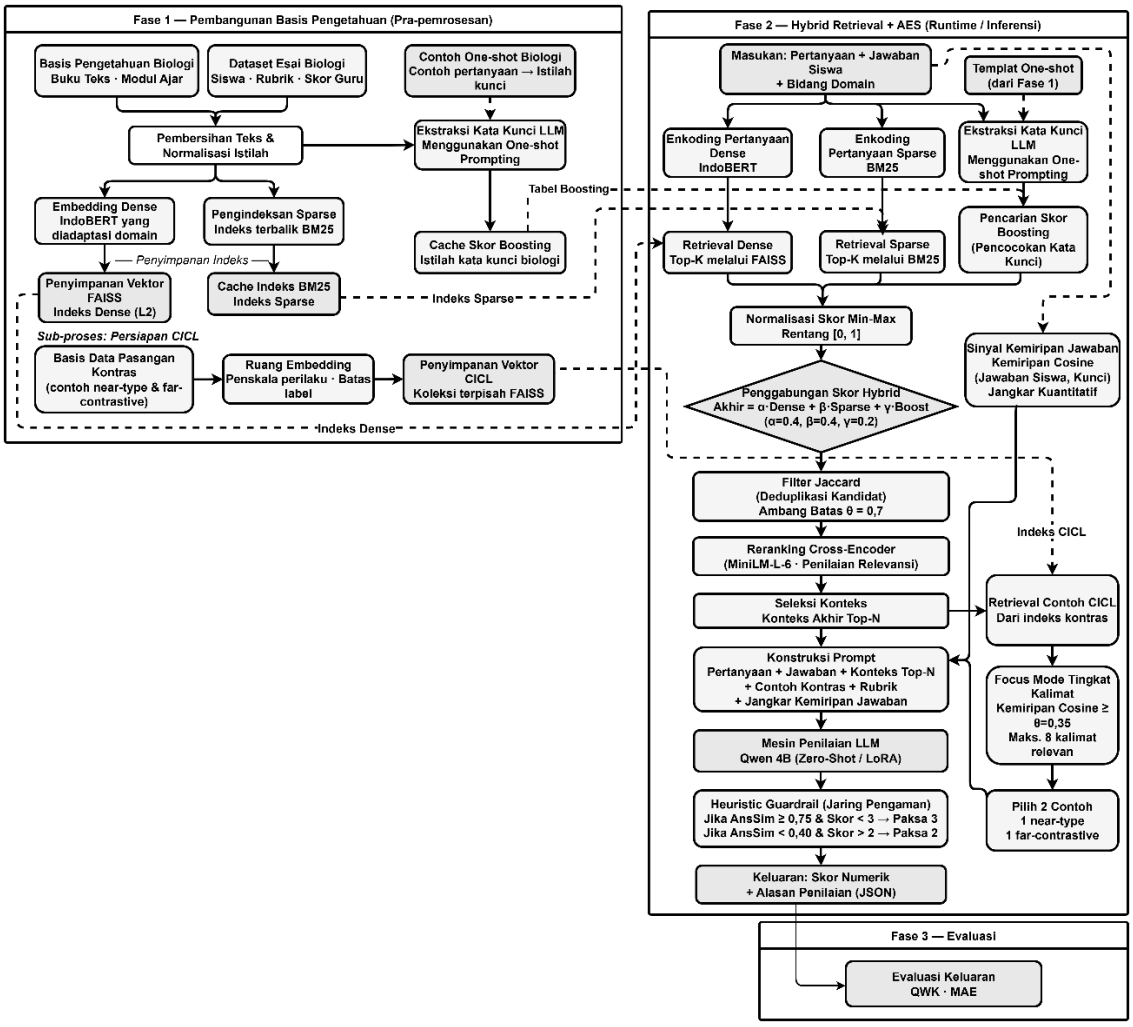
$$S_{norm} = \frac{S - S_{min}}{S_{max} - S_{min}} \quad (4)$$

Variabel S menyatakan skor awal dari *retriever*. Variabel S_{min} menyatakan skor minimum pada kumpulan dokumen. Variabel S_{max} menyatakan skor maksimum pada kumpulan dokumen. Sistem menghitung skor akhir menggunakan pembobotan linier:

$$S_{final} = w_d \cdot S_{dense} + w_s \cdot S_{sparse} + w_k \cdot S_{keyword} \quad (5)$$

Variabel S_{dense}, S_{sparse}, dan $S_{keyword}$ menyatakan skor *dense retrieval*, *sparse retrieval*, dan kata kunci yang telah dinormalisasi. Sistem menetapkan nilai bobot $w_d = 0{,}4$, $w_s = 0{,}4$, dan $w_k = 0{,}2$. Penetapan bobot ini mengacu pada konfigurasi yang digunakan dalam penelitian *hybrid retrieval* sebelumnya [7]. Selanjutnya, sistem menyaring dokumen dengan skor tertinggi menggunakan *Jaccard Similarity* dengan ambang batas sebesar 0,7 untuk menghapus *chunk* yang memiliki redundansi informasi tinggi. Ambang batas ini mengacu pada praktik standar deduplikasi teks untuk mengidentifikasi duplikasi parsial [12].

e-ISSN : 2356-2579 | p-ISSN : 1412-2693, DOI : 10.62411/tc.v25i3.16813



Gambar 1. Arsitektur Keseluruhan Sistem AES yang Diusulkan

2.2.5. Cross-Encoder Reranking

Tahap reranking dipakai supaya konteks yang masuk ke LLM lebih relevan. Kalau Bi-Encoder mengolah query dan passage masing-masing secara terpisah, Cross-Encoder memproses keduanya sekaligus lewat cross-attention. Hasilnya, skor relevansi lebih tepat, tapi biaya komputasinya juga lebih besar [9]. Di penelitian ini, kami memakai MiniLM-L-6 yang sudah di-fine-tune pada dataset MS MARCO Passage Ranking [13], [14] sebagai jalan tengah antara akurasi dan beban komputasi.

2.2.6. Focus Mode

Tahap *reranking* menghasilkan paragraf dengan skor relevansi tertinggi. Namun, menyertakan seluruh teks paragraf ke dalam *prompt* dapat memasukkan informasi yang tidak relevan. *Focus Mode* diterapkan untuk menyaring kalimat yang paling berkaitan dengan *query* dari paragraf hasil *reranking* [6]. Kalimat-kalimat tersebut dikodekan menggunakan IndoBERT, lalu dihitung kemiripan *cosine*-nya dengan *query*. Sistem hanya menyertakan kalimat dengan nilai kemiripan di atas 0,35 ke dalam konteks akhir, dengan batas maksimum 8 kalimat. Langkah ini mengurangi panjang *prompt* dan membatasi masukan LLM pada informasi yang secara langsung berkaitan dengan pertanyaan.

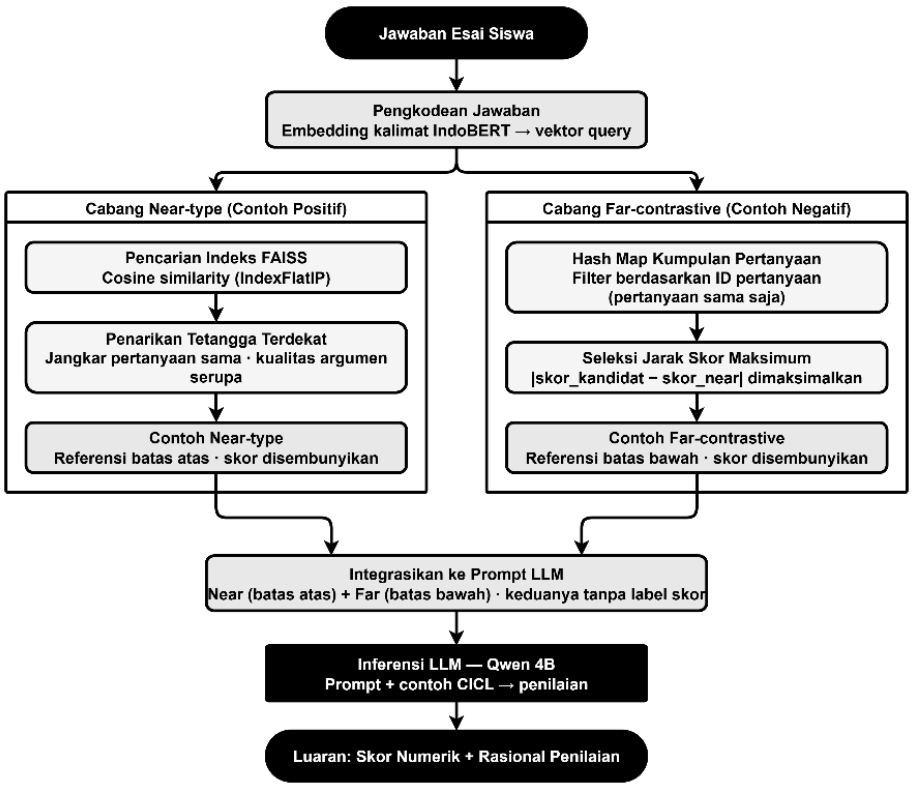
e-ISSN : 2356-2579 | p-ISSN : 1412-2693, DOI : 10.62411/tc.v25i3.16813

2.3. Contrastive In-Context Learning (CICL)

Sistem menerapkan pendekatan CICL berbasis pertanyaan (*question-anchored*) untuk mengalibrasi standar penilaian LLM [6]. Pada tahap praproses, sistem menyaring pasangan jawaban yang memiliki tingkat kemiripan semantik di atas 85% (0,85) namun memperoleh label skor yang berbeda. Penetapan ambang batas ini mengacu pada studi ablasi dalam sistem berbasis *retrieval*, yang menunjukkan bahwa nilai 0,85 memberikan keseimbangan antara mempertahankan diversitas semantik dan menghilangkan redundansi (*near-duplicate*) [15]. Dalam konteks penyusunan data CICL, penggunaan ambang ini ditujukan untuk mengurangi inkonsistensi pelabelan (*label noise*) pada data yang hampir identik, sembari mempertahankan variasi linguistik pada jawaban siswa.

Sistem mengambil contoh jawaban dari subset pelatihan selama tahap inferensi melalui dua mekanisme. Mekanisme pertama adalah *near-type*, di mana sistem mengambil contoh jawaban dengan tingkat kemiripan semantik tertinggi terhadap jawaban siswa yang sedang dinilai. Sistem menggunakan indeks FAISS yang dibangun dari *embedding* jawaban untuk menangkap kemiripan kualitas argumen. Mekanisme kedua adalah *far-contrastive*, di mana sistem mengambil contoh jawaban dengan skor yang paling berjauhan dari contoh *near-type*. Sistem mengambil contoh ini dari *Question Pools*. *Question Pools* merupakan struktur data *hash map* yang mengelompokkan jawaban berdasarkan teks pertanyaan yang sama persis, sehingga pengambilan contoh *far-contrastive* dapat dilakukan dalam waktu akses konstan selama proses inferensi.

Pendekatan ini memberikan referensi eksplisit mengenai rubrik penilaian guru manusia kepada LLM. Referensi ini membantu model membedakan jawaban berkualitas tinggi dan jawaban berkualitas rendah pada pertanyaan tertentu. Kondisi ini mengurangi masalah kalibrasi yang sering terjadi pada pendekatan *zero-shot* standar [4]. Gambar 2 mengilustrasikan mekanisme CICL tersebut. Diagram tersebut menunjukkan proses pengambilan contoh *near-type* dan *far-contrastive* berdasarkan pertanyaan dan jawaban siswa, serta proses penyisipan contoh-contoh tersebut ke dalam *prompt* LLM untuk mengkalibrasi proses penilaian.



Gambar 2. Mekanisme Pemilihan Contoh CICL Berbasis Pertanyaan

e-ISSN : 2356-2579 | p-ISSN : 1412-2693, DOI : 10.62411/tc.v25i3.16813

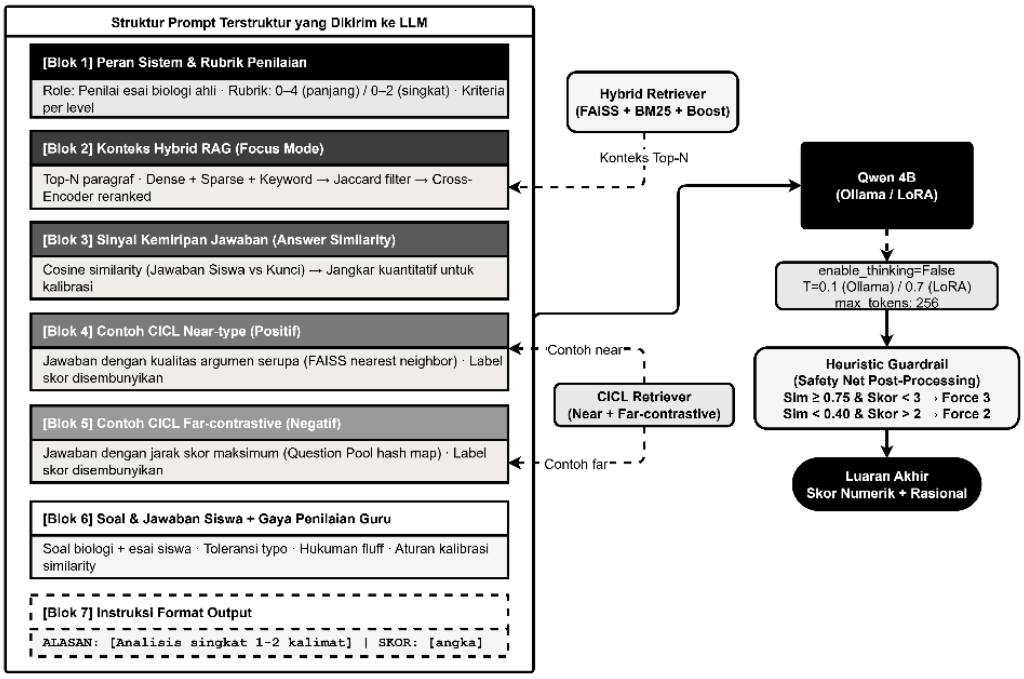
2.4. Inferensi dan Penilaian LLM

Proses penilaian menggunakan model LLM Qwen3-4B [16]. Sistem menjalankan model dasar secara *zero-shot* melalui Ollama, dan membandingkannya dengan model yang telah di-*fine-tuning* menggunakan *adapter* LoRA. Sistem menerapkan perutean berdasarkan tipe pertanyaan. Untuk esai singkat, sistem menggunakan *prompt* khusus tanpa mekanisme RAG dan CICL. Untuk esai panjang, sistem menggunakan *prompt* lengkap. Prompt lengkap ini mengintegrasikan rubrik penilaian, materi kontekstual dari Hybrid RAG, dan contoh referensi dari CICL.

```
Kamu adalah penilai jawaban soal Biologi SMP/SMA untuk materi Sistem Pencernaan.
Tugasmu adalah memberikan skor objektif berdasarkan rubrik di bawah ini.

=== RUBRIK PENILAIAN ===
{RUBRIK_PENILAIAN}

=== CONTOH REFERENSI ===
{CONTOH_CICL}

=== KONTEKS MATERI ===
{KONTEKS_MATERI}

=== SINYAL KEMIRIPAN ===
Kemiripan semantik jawaban siswa dengan jawaban referensi: {PERSEN}%
(Gunakan ini sebagai sinyal tambahan, bukan penentu tunggal)

=== SOAL DAN JAWABAN YANG DINILAI ===
Pertanyaan: {SOAL}
Jawaban Siswa: {JAWABAN_SISWA}

CATATAN PENTING & GAYA PENILAIAN GURU:
1. Abaikan kesalahan ketik (typo) minor selama konsep ilmiah yang dimaksud jelas.
2. Guru cenderung memberikan skor tinggi jika siswa menyebutkan KATA KUNCI UTAMA
dan inti jawaban. Jangan terlalu kaku pada tata bahasa.
3. HUKUMAN FLUFF: Jika siswa hanya menyebutkan daftar/definisi tanpa menjawab inti
pertanyaan, atau jawaban bertele-tele tapi salah konsep, berikan skor rendah.
4. Gunakan sinyal kemiripan semantik sebagai acuan tambahan kedekatan jawaban dengan
kunci guru.
5. ATURAN KALIBRASI SIMILARITY (PENTING):
  - Jika SINYAL KEMIRIPAN >= 75%: Siswa PASTI menangkap inti jawaban yang benar.
Berikan skor minimal 3 (atau 4). Jangan hukum karena tata bahasa.
  - Jika SINYAL KEMIRIPAN < 50%: Jawaban kemungkinan melenceng atau hanya "fluff".
Maksimal skor 2.

=== INSTRUKSI OUTPUT ===
Pikirkan kesesuaian jawaban siswa dengan rubrik secara SINGKAT (maksimal 2 kalimat),
kemudian tuliskan hasil penilaian Anda secara persis mengikuti format di bawah ini:

ALASAN: [Analisis singkat 1-2 kalimat]
SKOR: [angka]
```


e-ISSN : 2356-2579 | p-ISSN : 1412-2693, DOI : 10.62411/tc.v25i3.16813

Templat ini memuat instruksi peran, rubrik, contoh referensi, konteks materi, sinyal kemiripan, data evaluasi, aturan penilaian eksplisit, serta format keluaran. Penyusunan teks ini ditujukan untuk menjaga konsistensi inferensi dan memfasilitasi reproduksi eksperimen.



Gambar 3. *Pipeline* Konstruksi *Prompt* untuk Penilaian

Sistem merancang *prompt* dengan tiga instruksi eksplisit. Instruksi pertama mengarahkan model untuk mengabaikan kesalahan tipografi minor selama konsep ilmiahnya tetap jelas. Instruksi kedua mengarahkan model untuk mentoleransi variasi terminologi antara nama ilmiah dan istilah umum. Instruksi ketiga mengarahkan model untuk memberikan pengurangan skor pada jawaban *fluff*. Jawaban *fluff* merupakan jawaban yang bertele-tele namun kurang memiliki konten substantif.

Model menghasilkan luaran dalam format JSON terstruktur. Luaran ini berisi skor numerik dan alasan penilaian. Gambar 3 menyajikan representasi visual dari struktur *prompt*. Diagram tersebut mengilustrasikan integrasi informasi dari basis pengetahuan dan contoh CICL ke dalam satu *prompt* utuh.

2.5. Adaptasi Model dengan Fine-Tuning LoRA

Tahap *fine-tuning* bertujuan menyesuaikan LLM dengan pola penilaian guru. Proses ini menggunakan pendekatan PEFT, khususnya LoRA [17]. Metode LoRA memperbarui parameter secara efisien tanpa melatih ulang seluruh bobot model dasar. Penelitian ini menggunakan model Qwen3-4B sebagai model dasar.

Dataset *fine-tuning* mencakup pasangan soal, jawaban siswa, dan skor guru dari subset pelatihan. Dataset diformat ke dalam struktur ChatML untuk memisahkan instruksi penilaian dan respons model. Fitur penalaran (*thinking*) pada model dinonaktifkan untuk mencegah generasi token penalaran internal selama pelatihan, sehingga mengoptimalkan penggunaan memori GPU.

Proses pelatihan membekukan sebagian besar bobot model dasar dan hanya memperbarui matriks dekomposisi peringkat rendah pada lapisan *attention*. Kuantisasi 4-bit digunakan untuk memuat model dasar ke dalam memori GPU. Parameter pelatihan ditetapkan mengikuti konfigurasi standar *fine-tuning* kuantisasi rendah [18], yaitu 3 *epoch*, ukuran *batch* 1, dan akumulasi gradien 2. Laju pembelajaran (*learning rate*) ditetapkan sebesar 2e-4 dengan *warmup*

e-ISSN : 2356-2579 | p-ISSN : 1412-2693, DOI : 10.62411/tc.v25i3.16813

ratio 0,05. Model dievaluasi pada dataset validasi setiap 10 langkah pelatihan untuk memantau kinerja dan menyimpan bobot terbaik. Proses ini menghasilkan *adapter* LoRA yang merepresentasikan pola penilaian esai biologi, yang kemudian diintegrasikan ke dalam *pipeline* inferensi untuk pengujian.

2.6. Skenario Pengujian dan Metrik Evaluasi

Evaluasi dilakukan menggunakan subset pengujian yang terdiri dari 312 sampel esai. Untuk mengukur kontribusi masing-masing komponen dan membandingkan pendekatan *zero-shot* dengan *fine-tuning*, dilakukan *ablation study* yang terdiri dari 16 konfigurasi (Tabel 1). Konfigurasi ini dirancang untuk mengisolasi dampak dari metode *retrieval* (*Dense*, *Sparse*, *Keyword*, *Hybrid*), *Cross-Encoder Reranking*, *Focus Mode*, *Answer Similarity Signal*, dan CICL. Evaluasi dilakukan secara komparatif menggunakan model LLM Qwen3-4B dalam dua pendekatan: (1) *Zero-shot* melalui Ollama, dan (2) PEFT menggunakan LoRA.

Tabel 1. Matriks 16 Konfigurasi Ablation Study

Konfigurasi	*Retrieval*	*Reranker*	*Focus*	*AnsSim*	CICL	Keterangan
A	*Zero-shot*	X	X	X	✗	*Baseline*
A2	*Zero-shot*	X	X	X	✓	A + CICL
B	*Dense*	X	X	X	X	*Dense RAG*
B2	*Dense*	X	X	X	✓	B + CICL
C	*Sparse*	X	X	X	X	*Sparse RAG* (BM25)
C2	*Sparse*	X	X	X	✓	C + CICL
D	*Keyword*	X	X	X	X	*Keyword RAG*
D2	*Keyword*	X	X	X	✓	D + CICL
E	*Hybrid*	X	X	X	X	*Hybrid Retrieval*
F	*Hybrid*	X	X	X	✓	E + CICL
G0	*Hybrid*	X	X	✓	X	E + *Answer Similarity*
G	*Hybrid*	X	X	✓	✓	G0 + CICL
H0	*Hybrid*	✓	X	✓	X	G0 + *Reranker*
H	*Hybrid*	✓	X	✓	✓	H0 + CICL
I0	*Hybrid*	✓	✓	✓	X	H0 + *Focus Mode*
I	*Hybrid*	✓	✓	✓	✓	*Proposed Full*

2.6.1. Quadratic Weighted Kappa (QWK)

Metrik utama yang digunakan untuk mengukur tingkat kesepakatan antara skor prediksi dan skor guru. Metrik ini memberikan penalti yang lebih besar pada kesalahan dengan selisih ordinal yang lebih besar [19]:

$$QWK = 1 - \frac{\sum_{i,j} W_{i,j} O_{i,j}}{\sum_{i,j} W_{i,j} E_{i,j}} \qquad (6)$$

Dengan O adalah matriks observasi, E adalah matriks ekspektasi, dan W adalah bobot kuadratik yang didefinisikan sebagai:

$$W_{i,j} = \frac{(i-j)^2}{(N-1)^2} \quad (7)$$

Variabel i menyatakan skor dari penilai manusia. Variabel j menyatakan skor dari sistem. Variabel N menyatakan jumlah kategori skor yang mungkin. Matriks O mencatat jumlah

e-ISSN : 2356-2579 | p-ISSN : 1412-2693, DOI : 10.62411/tc.v25i3.16813

kesepakatan yang terjadi pada data uji. Matriks E menghitung kesepakatan yang diharapkan secara acak berdasarkan distribusi skor.

2.6.2. Mean Absolute Error (MAE)

Metrik ini mengukur rata-rata kesalahan absolut antara skor prediksi dan skor aktual:

$$MAE = \frac{1}{n}\sum_{i=1}^{n} | y_i - \hat{y}_i | \qquad (8)$$

Variabel n menyatakan jumlah seluruh sampel esai. Variabel y_i menyatakan skor aktual dari guru. Variabel $\hat{y}_i$ menyatakan skor prediksi dari sistem. Nilai MAE yang lebih kecil menunjukkan bahwa prediksi sistem lebih mendekati skor aktual.

2.6.3. Accuracy

Metrik ini mengukur proporsi ketepatan antara skor prediksi dan skor aktual:

$$Accuracy = \frac{1}{n}\sum_{i=1}^{n} I(y_i = \hat{y}_i) \qquad (9)$$

Variabel n menyatakan jumlah sampel. Variabel y_i menyatakan skor aktual. Variabel $\hat{y}_i$ menyatakan skor prediksi. Fungsi I adalah fungsi indikator yang menghasilkan nilai 1 jika prediksi sama dengan label, dan 0 jika berbeda.

3. HASIL DAN PEMBAHASAN

Bagian ini menyajikan hasil evaluasi dari 16 konfigurasi model pada 312 sampel esai biologi. Pembahasan diuraikan dalam tiga bagian. Bagian pertama membandingkan kinerja antara pendekatan *zero-shot* dan LoRA. Bagian kedua menganalisis pengaruh mekanisme CICL pada berbagai konfigurasi sistem. Bagian ketiga membahas dampak penerapan *Focus Mode* terhadap kinerja model.

3.1. Ablation Study: Zero-Shot vs LoRA

Tabel 2 merangkum hasil evaluasi dari 16 konfigurasi pengujian. Secara umum, model LoRA menunjukkan nilai QWK yang lebih tinggi dibandingkan model *zero-shot* pada hampir seluruh konfigurasi. Rata-rata peningkatan nilai QWK model LoRA mencapai +0,074. Pada konfigurasi dasar (Konfigurasi A), model LoRA mencapai QWK sebesar 0,8147 tanpa menggunakan konteks eksternal. Nilai ini lebih tinggi dibandingkan seluruh konfigurasi RAG pada model *zero-shot*. Hasil ini menunjukkan bahwa proses *fine-tuning* LoRA memungkinkan model mempelajari pola penilaian dan materi biologi dari data latih. Ketika komponen CICL ditambahkan pada konfigurasi A2, model *zero-shot* mengalami peningkatan kinerja. Sebaliknya, model LoRA mengalami sedikit penurunan kinerja karena tambahan contoh pada *prompt* tidak memberikan informasi baru bagi model yang sudah dilatih.

Selanjutnya, pengujian menggunakan satu jenis *retrieval* (Konfigurasi B, C, dan D) menunjukkan kontribusi masing-masing metode. Konfigurasi B menggunakan *Dense RAG* untuk menangkap makna semantik kalimat, sehingga meningkatkan kinerja model *zero-shot*. Konfigurasi C menggunakan *Sparse RAG* (BM25) untuk mencocokkan istilah biologi secara teks. Metode ini menghasilkan kinerja yang baik pada model LoRA (QWK = 0,8465). Konfigurasi D menggunakan *Keyword RAG* yang mengekstrak kata kunci penting dari materi. Penambahan komponen CICL pada pasangan konfigurasi ini (B2, C2, D2) membantu model *zero-shot* untuk menyesuaikan penilaian dengan rubrik guru. Namun, model LoRA menunjukkan pola yang sama, di mana penambahan CICL tidak selalu meningkatkan nilai QWK.

e-ISSN : 2356-2579 | p-ISSN : 1412-2693, DOI : 10.62411/tc.v25i3.16813

Tabel 2. Hasil *Ablation Study* pada Dataset Pengujian (n = 312)

Konfigurasi	Deskripsi Konfigurasi	QWK (*Zero-Shot*)	QWK (LoRA)	ΔQWK
A	*Baseline*	0,6451	0,8147	+0,1696
A2	A + CICL	0,7626	0,7976	+0,0350
B	*Dense RAG*	0,7392	0,8207	+0,0815
B2	B + CICL	0,7408	0,8262	+0,0854
C	*Sparse RAG* (BM25)	0,7414	0,8465	+0,1051
C2	C + CICL	0,7812	0,8160	+0,0348
D	*Keyword RAG*	0,7674	0,8153	+0,0479
D2	D + CICL	0,7864	0,8142	+0,0278
E	*Hybrid RAG* (*No* CICL)	0,7420	0,8483	+0,1063
F	E + CICL	0,7778	0,8174	+0,0396
G0	E + *Answer Similarity*	0,7460	0,8413	+0,0953
G	G0 + CICL	0,7784	0,8441	+0,0657
H0	G0 + *Reranker*	0,7284	0,8351	+0,1067
H	**H0 + CICL (*Best Zero-Shot*)**	**0,7914**	**0,8194**	**+0,0280**
I0	H0 + *Focus Mode*	0,6945	0,8155	+0,1210
I	***Proposed Full* (*Best* LoRA)**	**0,7808**	**0,8554**	**+0,0746**

Pengujian kemudian beralih ke penggabungan ketiga metode *retrieval* menjadi *Hybrid RAG* (Konfigurasi E). Pada model *zero-shot*, penggabungan ini menurunkan kinerja (QWK = 0,7420) dibandingkan Konfigurasi D (QWK = 0,7674). Penurunan ini terjadi karena *prompt* menjadi lebih panjang dan mengandung informasi yang berulang. Kondisi ini membuat model *zero-shot* kurang fokus pada kriteria penilaian utama. Penambahan CICL pada Konfigurasi F membantu mengatasi masalah tersebut dengan memberikan contoh jawaban yang kontras. Selanjutnya, Konfigurasi G0 menambahkan sinyal *Answer Similarity* sebagai acuan tambahan, dan Konfigurasi G menggabungkannya dengan CICL. Sinyal ini membantu kedua model untuk menyesuaikan skor berdasarkan tingkat kemiripan jawaban siswa dengan kunci jawaban.

Tahap berikutnya menguji penggunaan *Cross-Encoder Reranker* untuk menyaring dan mengurutkan paragraf yang paling relevan (Konfigurasi H0). Pada model *zero-shot*, penggunaan *Reranker* tanpa CICL menurunkan kinerja (QWK = 0,7284). Model *zero-shot* membutuhkan contoh tambahan untuk memahami konteks yang telah diurutkan. Ketika CICL diaktifkan pada Konfigurasi H, model *zero-shot* mencapai nilai QWK tertinggi (0,7914). CICL berfungsi sebagai acuan yang membimbing model *zero-shot* dalam mengevaluasi jawaban berdasarkan konteks yang sudah disaring oleh *Reranker*. Model LoRA pada Konfigurasi H0 dan H mempertahankan kinerja yang stabil di atas 0,81 karena tidak terlalu bergantung pada susunan *prompt*.

Terakhir, pengujian melibatkan *Focus Mode* yang menyaring kalimat paling relevan dari setiap paragraf (Konfigurasi I0). Pada model *zero-shot*, penyaringan kalimat ini menurunkan kinerja (QWK = 0,6945). Model *zero-shot* kehilangan kalimat penghubung yang dibutuhkan untuk memahami konteks secara utuh. Model LoRA tidak terpengaruh oleh penyaringan ini karena sudah mempelajari hubungan antarkonsep dari data latih. Ketika seluruh komponen digabungkan pada Konfigurasi I (*Proposed Full*), model LoRA mencapai nilai QWK tertinggi sebesar 0,8554 dan MAE sebesar 0,2917. Hasil ini menunjukkan bahwa model LoRA dapat menggunakan seluruh komponen *pipeline* tanpa mengalami gangguan dari panjangnya *prompt*.

e-ISSN : 2356-2579 | p-ISSN : 1412-2693, DOI : 10.62411/tc.v25i3.16813

3.2. Perbedaan Dampak CICL pada Model *Zero-Shot* dan LoRA

Salah satu temuan dari penelitian ini adalah perbedaan pengaruh penambahan komponen CICL terhadap kedua jenis model. Pada model *zero-shot*, penambahan CICL berfungsi sebagai panduan penilaian. Model *zero-shot* belum mempelajari standar penilaian guru secara khusus. Oleh karena itu, model *zero-shot* bergantung pada contoh eksplisit yang berisi jawaban bernilai tinggi dan jawaban bernilai rendah. Contoh eksplisit ini membantu model *zero-shot* untuk menyaring informasi yang tidak relevan dari *Hybrid RAG*. Penambahan CICL meningkatkan nilai QWK pada model *zero-shot*. Sebagai contoh, peralihan dari Konfigurasi E ke Konfigurasi F meningkatkan QWK sebesar +0,0358, dan peralihan dari Konfigurasi H0 ke Konfigurasi H meningkatkan QWK sebesar +0,0630.

Sebaliknya, kondisi yang berbeda terjadi pada model LoRA. Penambahan komponen CICL pada beberapa konfigurasi menurunkan nilai QWK. Penurunan kinerja ini tercatat pada tiga pasangan konfigurasi berikut:

1. Peralihan dari Konfigurasi E ke Konfigurasi F (*Hybrid* ± CICL) menurunkan QWK sebesar -0,0309.
2. Peralihan dari Konfigurasi C ke Konfigurasi C2 (*Sparse* ± CICL) menurunkan QWK sebesar -0,0305.
3. Peralihan dari Konfigurasi H0 ke Konfigurasi H (*Reranker* ± CICL) menurunkan QWK sebesar -0,0157.

Penurunan kinerja pada model LoRA terjadi karena proses *fine-tuning* telah mempelajari pola penilaian dari data latih. Penyertaan contoh kontras secara eksplisit ke dalam *prompt* menghasilkan informasi yang berulang. Kondisi ini mengganggu fokus model saat memproses jawaban siswa. Model LoRA yang sudah mempelajari pola penilaian mengalami penurunan akurasi ketika sistem memberikan contoh tambahan yang bersifat redundan.

3.3. Dampak Focus Mode pada Model Zero-Shot

Focus Mode dirancang untuk menyaring kalimat paling relevan dari setiap paragraf RAG menggunakan kemiripan semantik (*semantic similarity*). Sistem hanya mengambil kalimat dengan skor kemiripan di atas ambang batas tertentu (maksimum 8 kalimat per paragraf) untuk mengurangi panjang *prompt*. Namun, hasil pengujian menunjukkan penurunan kinerja pada model *zero-shot* ketika *Focus Mode* diaktifkan.

Konfigurasi I0 (Hybrid RAG + Reranker + Answer Similarity + Focus Mode, tanpa CICL) menghasilkan QWK sebesar 0,6945. Nilai ini lebih rendah dibandingkan Konfigurasi H0 (tanpa Focus Mode) yang mencapai QWK sebesar 0,7284. Penurunan ini terjadi karena penyaringan kalimat menghilangkan kalimat penghubung dan konteks pendukung yang dibutuhkan oleh model *zero-shot* untuk memahami hubungan antar konsep biologi. Model *zero-shot* tidak memiliki pengetahuan internal mengenai struktur narasi ilmiah, sehingga bergantung pada kelengkapan konteks dalam *prompt*.

Kondisi yang berbeda terjadi pada model LoRA. Konfigurasi I0 pada model LoRA menghasilkan QWK sebesar 0,8155, yang relatif stabil dibandingkan dengan konfigurasi tanpa *Focus Mode*. Model LoRA tidak terpengaruh secara signifikan oleh penyaringan kalimat karena proses *fine-tuning* telah mengodekan pola hubungan antarkonsep biologi ke dalam bobot model. Model LoRA dapat mengompensasi hilangnya informasi dalam *prompt* dengan mengandalkan representasi pengetahuan yang tersimpan dalam bobot model. Temuan ini menunjukkan bahwa *fine-tuning* meningkatkan ketahanan model terhadap pengurangan konteks, sementara model *zero-shot* lebih bergantung pada kelengkapan teks yang disediakan.

e-ISSN : 2356-2579 | p-ISSN : 1412-2693, DOI : 10.62411/tc.v25i3.16813

3.4. Analisis Kinerja Berdasarkan Tingkat Kesulitan Pertanyaan

Penelitian ini mengelompokkan hasil evaluasi berdasarkan tingkat kesulitan pertanyaan untuk menganalisis kinerja sistem pada kategori soal yang berbeda. Tabel 3 menyajikan perbandingan kinerja antara Konfigurasi H (*Zero-Shot*) dan Konfigurasi I (LoRA) pada dua kategori soal, yaitu soal singkat dan soal panjang.

Tabel 3. Hasil Evaluasi Berdasarkan Tingkat Kesulitan Pertanyaan

Tipe Soal	n	Konfigurasi	QWK	MAE	Exact Match
Singkat (Faktual)	156	Zero-Shot (Konfigurasi H)	0,8166	0,1410	92,95%
Singkat (Faktual)	156	LoRA (Konfigurasi I)	0,9349	0,0513	97,44%
Panjang (Proses)	156	Zero-Shot (Konfigurasi H)	0,6921	0,5705	53,21%
Panjang (Proses)	156	LoRA (Konfigurasi I)	0,7483	0,5321	55,13%

Kedua model menghasilkan nilai QWK yang relatif tinggi pada soal singkat. Soal jenis ini umumnya mengukur pengetahuan faktual siswa terhadap istilah biologi, seperti nama organ atau enzim. Model *Zero-Shot* (Konfigurasi H) mencapai nilai QWK sebesar 0,8166 karena sistem RAG dapat mengambil konteks faktual yang sesuai dari buku teks. Model LoRA (Konfigurasi I) mencapai nilai QWK sebesar 0,9349. Model LoRA mengkodekan hubungan antara istilah dan definisi ke dalam bobot model selama proses pelatihan, sehingga dapat memberikan skor yang lebih sesuai tanpa bergantung pada pencarian konteks eksternal.

Kinerja kedua model mengalami penurunan pada soal panjang. Soal panjang menuntut siswa untuk menjelaskan proses kausal secara bertahap, seperti mekanisme pencernaan di dalam lambung. Model *Zero-Shot* (Konfigurasi H) mencatatkan nilai QWK sebesar 0,6921. Penurunan ini terjadi karena model *Zero-Shot* cenderung bergantung pada kecocokan kata kunci eksplisit dari sistem RAG. Siswa sering kali menulis jawaban yang secara konsep benar tetapi tidak menyebutkan kata kunci spesifik, padahal guru manusia biasanya memberikan skor parsial untuk jawaban jenis ini.

Model LoRA (Konfigurasi I) mencapai nilai QWK yang lebih tinggi pada soal panjang, yaitu 0,7483. Model LoRA mengenali pola hubungan sebab-akibat dari data latih, sehingga dapat memberikan skor parsial yang lebih sesuai untuk jawaban yang tidak lengkap. Model LoRA memproses makna semantik dari jawaban siswa, sedangkan model *Zero-Shot* cenderung mencari kecocokan kata secara leksikal. Perbedaan ini menunjukkan bahwa proses *fine-tuning* meningkatkan ketepatan penilaian pada soal yang memerlukan penalaran kausal, dibandingkan dengan pendekatan *Zero-Shot* yang mengandalkan instruksi pada *prompt*.

3.5. Analisis Kualitatif dan Frekuensi Fenomena Penilaian

Penelitian ini melakukan analisis kualitatif terhadap *raw output* dan penalaran model pada konfigurasi terbaik masing-masing pendekatan. Konfigurasi terbaik untuk model *zero-shot* adalah konfigurasi H, sedangkan konfigurasi terbaik untuk model LoRA adalah konfigurasi I. Analisis ini menggunakan sampel dari dataset pengujian untuk mengamati cara kedua model menangani kesalahan tipografi, respons *fluff*, dan kesalahan fakta.

Sebagai gambaran, identifikasi frekuensi pada 312 sampel uji menunjukkan bahwa fenomena toleransi tipografi muncul pada 35 kasus (11,2%), respons *fluff* pada 36 kasus (11,5%), dan *overkoreksi* konseptual pada 10 kasus (3,2%). Tabel 4 merangkum distribusi agregat fenomena tersebut, sedangkan Tabel 5 menyajikan contoh nyata untuk membandingkan cara kedua model memproses jawaban.

e-ISSN : 2356-2579 | p-ISSN : 1412-2693, DOI : 10.62411/tc.v25i3.16813

Tabel 4. Frekuensi Agregat Fenomena Kualitatif pada Dataset Pengujian (n=312)

Fenomena Penilaian	Frekuensi (n)	Persentase	Tingkat Kesesuaian dengan Skor Guru
Toleransi Tipografi (Substansi benar, terdapat *typo*)	35	11,2%	LoRA: 62,9% (22/35) *Zero-shot*: 40,0% (14/35)
Deteksi Respons *Fluff* (Tanpa kausalitas, skor guru ≤ 2)	36	11,5%	LoRA: 77,8% (28/36) *Zero-shot*: 72,2% (26/36)
Overkoreksi Konseptual	10	3,2%	LoRA lebih mendekati pada 80% kasus

Tabel 5. Analisis Kualitatif Komparatif pada Kasus Nyata Dataset Pengujian

Fenomena	Cuplikan Jawaban Siswa (dari Dataset)	Skor Guru	Konfigurasi H (*Zero-Shot*)	Konfigurasi I (LoRA)	Mekanisme Penalaran
Toleransi *Typo* & Tata Bahasa	"Karena melarutkan vitamin C hanya membutuhkan air sedangkan vit A membutuhkan lemak dan lama untuk *dicrerna*, *mkaa* dari itu vit c *yg* dijual..."	4	4 (Berkat instruksi *prompt*)	4 (*Native*)	Ollama mengandalkan aturan eksplisit di *prompt* untuk mengabaikan *typo*. LoRA langsung memahami esensi konsep dari representasi internalnya.
Deteksi Jawaban *Fluff*	"...pencernaan makanan menjadi sedikit *bermasaalh* dan penyerapan gizi menjadi tidak maksimal"	4	2 (Berkat *Far-Contrastive* CICL)	2 (*Native*)	Zero-shot menurunkan skor karena "dicontohkan" oleh CICL bahwa jawaban baik harus menyebutkan mekanisme spesifik. LoRA menolaknya karena sudah terlatih pada pola *fluff*.
Penolakan Miskonsepsi Faktual	Soal: Enzim lipase dihasilkan oleh organ... Jawaban: "usus halus"	0	0 (Berkat RAG + *Reranker*)	0 (*Parametric Memory*)	Zero-shot memverifikasi kesalahan ini melalui *chunk* buku teks yang di-*rerank*. LoRA "tahu" fakta ini tanpa perlu konteks eksternal.

3.5.1. Toleransi Kesalahan Tipografi dan Tata Bahasa

Dari 312 sampel, teridentifikasi 35 kasus di mana jawaban siswa mengandung kesalahan tipografi namun secara substansi benar. Model LoRA berhasil menyelaraskan skor dengan guru pada 62,9% kasus, dibandingkan 40,0% pada model *zero-shot*. Sebagai contoh representatif, seorang siswa menjawab: "Karena melarutkan vitamin C hanya membutuhkan air sedangkan vit A membutuhkan lemak dan lama untuk *dicrerna*, *mkaa* dari itu vit c *yg* dijual...". Jawaban ini mengandung kesalahan ejaan (*dicrerna*, *mkaa*, *yg*), namun secara konsep benar.

1. Model *Zero-Shot* (Konfigurasi H): Memberikan skor 4 karena mengikuti instruksi eksplisit pada *prompt*: "Abaikan kesalahan ketik minor selama konsep ilmiah yang dimaksud jelas". Tanpa instruksi ini, model cenderung memberikan penalti pada tata bahasa yang tidak baku.

e-ISSN : 2356-2579 | p-ISSN : 1412-2693, DOI : 10.62411/tc.v25i3.16813

2. Model LoRA (Konfigurasi I): Memberikan skor 4 secara langsung tanpa bergantung pada instruksi *prompt*. Proses *fine-tuning* menggunakan variasi jawaban siswa yang tidak baku telah membentuk representasi internal yang memetakan makna jawaban ke konsep biologi, terlepas dari morfologi kata yang salah.

3.5.2. Deteksi Respons Fluff Berbasis Jangkar Kontras

Respons *fluff* (bertele-tele tanpa kausalitas) teridentifikasi pada 36 kasus. Kedua model cenderung memberikan skor lebih tinggi dari guru (*overscoring*) pada sebagian kasus ini (LoRA 22,2%, *zero-shot* 27,8%), namun LoRA menunjukkan tingkat deteksi yang sedikit lebih baik. Sebagai contoh, siswa menjawab pertanyaan dampak kerusakan pankreas dengan: "pencernaan makanan menjadi sedikit bermasalah dan penyerapan gizi menjadi tidak maksimal". Guru memberikan skor 2 karena tidak ada mekanisme ilmiah yang dijelaskan.

1. Model *Zero-Shot* (Konfigurasi H): Cenderung memberikan skor 3 karena adanya kecocokan leksikal. Namun, dengan adanya mekanisme CICL, model membandingkan jawaban ini dengan contoh *far-contrastive* (jawaban bernilai rendah) di dalam *prompt*, yang mengarahkan model untuk mengoreksi skor menjadi 2.
2. Model LoRA (Konfigurasi I): Memberikan skor 2 secara langsung. Model telah mempelajari pola respons *fluff* selama *fine-tuning*. Temuan ini menjelaskan mengapa penambahan CICL pada model LoRA (seperti dibahas di Sub-bab 3.2) terkadang menurunkan akurasi: contoh kontras di dalam *prompt* menjadi informasi redundan yang mengganggu fokus atensi model yang sudah memiliki standar internal.

3.5.3. Verifikasi Faktual dan Overkoreksi Konseptual

Fenomena kesalahan faktual murni sangat jarang (1 kasus dari 29 konteks serupa), namun fenomena *overkoreksi* (model memberikan skor jauh lebih rendah dari guru karena kesalahan istilah minor) teridentifikasi pada 10 kasus (3,2%). Pada 80% kasus *overkoreksi* ini, skor LoRA lebih mendekati guru dibandingkan *zero-shot*. Sebagai contoh, siswa menjawab soal organ penghasil enzim lipase dengan "usus halus" (jawaban benar: pankreas).

1. Model *Zero-Shot* (Konfigurasi H): Menolak jawaban ini (skor 0) karena mekanisme Hybrid RAG dan Cross-Encoder *Reranker* berhasil mengambil dan menempatkan paragraf buku teks yang menyebutkan getah pankreas di urutan teratas konteks.
2. Model LoRA (Konfigurasi I): Menolak jawaban ini (skor 0) dengan mengandalkan memori parametrik. Model telah mempelajari hubungan organ dan enzim dari data latih, sehingga tidak memerlukan waktu komputasi tambahan untuk pencarian konteks eksternal pada materi yang memiliki batasan kurikulum yang jelas.

Di sisi lain, pada kasus *overkoreksi* di mana siswa menulis "maltosa" alih-alih "glukosa" (guru memberi skor 4), model *zero-shot* memberikan skor 1 karena validasi konsep yang terlalu kaku. Sebaliknya, model LoRA memberikan skor 3, menunjukkan kemampuan yang lebih baik dalam mempertimbangkan substansi jawaban secara holistik, selaras dengan pertimbangan guru manusia.

4. KESIMPULAN

Penelitian ini mengevaluasi arsitektur AES berbahasa Indonesia melalui *ablation study* yang melibatkan 16 konfigurasi pada pendekatan *zero-shot* dan *fine-tuning* (LoRA). Hasil menunjukkan bahwa penggabungan sederhana pada Hybrid RAG menyebabkan *context dilution*, yang dampaknya dapat dikurangi oleh penerapan CICL pada model *zero-shot* (mencapai QWK = 0,7914). Penelitian ini juga menunjukkan perbedaan karakteristik: CICL yang berperan penting bagi model *zero-shot* justru menjadi informasi redundan yang menurunkan akurasi model LoRA (QWK = 0,8554) yang telah mempelajari pola penilaian dari data latih.

Berdasarkan hasil tersebut, tidak ada pendekatan tunggal yang paling tepat untuk semua kondisi AES pada domain spesifik. Pendekatan *zero-shot* RAG dengan CICL lebih sesuai untuk

e-ISSN : 2356-2579 | p-ISSN : 1412-2693, DOI : 10.62411/tc.v25i3.16813

sistem yang membutuhkan fleksibilitas tinggi, di mana materi baru dapat ditambahkan hanya dengan memperbarui *Knowledge Base* tanpa *retraining*. Sebaliknya, *fine-tuning* LoRA lebih direkomendasikan jika prioritas utama adalah akurasi pada penalaran kompleks dan dataset beranotasi tersedia. Penelitian selanjutnya dapat mengeksplorasi mekanisme *Soft-CICL* atau *Adaptive Prompting* untuk menentukan secara otomatis apakah model memerlukan jangkar rubrik eksplisit.

DAFTAR PUSTAKA

[1] Pulung Hendro Prastyo, Eddy Tungadi, and Shaifudin Zuhdi, "Indonesian Automated Essay Scoring: A Comparative Study of Pretrained Transformer Models," *Information Technology Education Journal*, pp. 120–130, Jun. 2025, doi: 10.59562/intec.v4i2.8069.

[2] D. Ramesh and S. K. Sanampudi, "An automated essay scoring systems: a systematic literature review," *Artif. Intell. Rev.*, vol. 55, no. 3, pp. 2495–2527, Mar. 2022, doi: 10.1007/s10462-021-10068-2.

[3] K. A. Pradani and L. H. Suadaa, "Automated Essay Scoring Menggunakan Semantic Textual Similarity Berbasis Transformer Untuk Penilaian Ujian Esai," *Jurnal Teknologi Informasi dan Ilmu Komputer*, vol. 10, no. 6, pp. 1177–1184, Dec. 2023, doi: 10.25126/jtiik.2023107338.

[4] M. Fateen, B. Wang, and T. Mine, "Beyond Scores: A Modular RAG-Based System for Automatic Short Answer Scoring With Feedback," *IEEE Access*, vol. 12, pp. 185371–185385, 2024, doi: 10.1109/ACCESS.2024.3508747.

[5] Z. Li, Z. Wang, W. Wang, K. Hung, H. Xie, and F. L. Wang, "Retrieval-augmented generation for educational application: A systematic survey," Jun. 01, 2025, *Elsevier B.V.* doi: 10.1016/j.caeai.2025.100417.

[6] S. Li, L. Stenzel, C. E. Seyed, and A. Bahrainian, "Enhancing Retrieval-Augmented Generation: A Study of Best Practices," 2025. Accessed: Jul. 30, 2026. [Online]. Available: https://github.com/ali-bahrainian/RAG_best_practices

[7] Suharyadi and I. Saputra, "Hybrid Ensemble Retrieval-Augmented Generation for Indonesian Legal Consultation with Keyword Boosting," *Journal of Novel Engineering Science and Technology*, vol. 4, no. 02, pp. 71–85, Jul. 2025, doi: 10.56741/jnest.v4i02.1042.

[8] M. Song, X. Geng, S. Yao, S. Lu, Y. Feng, and L. Jing, "Large Language Models as Zero-Shot Keyphrase Extractors: A Preliminary Empirical Study," Jan. 2024, [Online]. Available: http://arxiv.org/abs/2312.15156

[9] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks," 2019. doi: 10.18653/v1/D19-1410.

[10] J. Johnson, M. Douze, and H. Jégou, "Billion-scale similarity search with GPUs," Feb. 2017, [Online]. Available: http://arxiv.org/abs/1702.08734

[11] S. Robertson and H. Zaragoza, "The probabilistic relevance framework: BM25 and beyond," *Foundations and Trends in Information Retrieval*, vol. 3, no. 4, pp. 333–389, 2009, doi: 10.1561/1500000019.

[12] J. Kim and Y. Park, "A Survey of Text Deduplication: From Syntactic Matching to Semantic Understanding," 2026, *Institute of Electrical and Electronics Engineers Inc.* doi: 10.1109/ACCESS.2026.3658439.

[13] W. Wang, F. Wei, L. Dong, H. Bao, N. Yang, and M. Zhou, "MINILM: Deep Self-Attention Distillation for Task-Agnostic Compression of Pre-Trained Transformers." [Online]. Available: https://aka.ms/minilm.

[14] P. Bajaj *et al.*, "MS MARCO: A Human Generated MAchine Reading COmprehension Dataset," Oct. 2018, [Online]. Available: http://arxiv.org/abs/1611.09268

e-ISSN : 2356-2579 | p-ISSN : 1412-2693, DOI : 10.62411/tc.v25i3.16813

[15] R. Sheng, J. Pan, Z. Zeng, H. Chen, and W. Cao, "Game-Theoretic Multi-LLM Collaboration for Attribute-Aware Open-Vocabulary Object Detection," *Electronics (Basel).*, vol. 15, no. 13, p. 2817, Jun. 2026, doi: 10.3390/electronics15132817.
[16] Q. Team, "Qwen3 Technical Report," 2025. Accessed: Jul. 30, 2026. [Online]. Available: https://huggingface.co/Qwen
[17] E. J. Hu *et al.*, "LoRA: Low-Rank Adaptation of Large Language Models," Oct. 2021, [Online]. Available: http://arxiv.org/abs/2106.09685
[18] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, "QLoRA: Efficient Finetuning of Quantized LLMs," May 2023, [Online]. Available: http://arxiv.org/abs/2305.14314
[19] J. Cohen, "Psychological Bulletin WEIGHTED KAPPA: NOMINAL SCALE AGREEMENT WITH PROVISION FOR SCALED DISAGREEMENT OR PARTIAL CREDIT," 1968.

e-ISSN : 2356-2579 | p-ISSN : 1412-2693, DOI : 10.62411/tc.v25i3.16813