Sistemasi: Jurnal Sistem Informasi
Volume 15, Nomor 5, 2026: 1764-1776
ISSN:2302-8149
e-ISSN:2540-9719

Implementasi *Zero-Shot Learning* pada *Edge AI* untuk Interpretasi Deterministik Metrik Kelistrikan menggunakan Model Bahasa Besar Terkuantisasi

Implementation of Zero-Shot Learning on Edge AI for Deterministic Interpretation of Electrical Metrics using Quantized Large Language Models

[1]Salman Al Majali*, [2]Ganjar M Faisal, [3]Novi Rukhviayanti
[1,2,3]Program Studi Teknik Informatika, Fakultas Teknik, Sekolah Tinggi Manajemen Informatika dan Komputer Indonesia Mandiri
[1,2,3]Jl. Belitung No. 7, Merdeka, Kec. Sumur Bandung, Kota Bandung, Jawa Barat 40113, Indonesia
*e-mail: undermod007@gmail.com

(***received*:** 24 April 2026, ***revised*:** 2 May 2026, ***accepted*:** 4 May 2026)

Abstrak

Interpretasi otomatis data metrik listrik bangunan sangat penting untuk menilai keandalan dan kecocokan sistem listrik. Kemunculan Model Bahasa Besar (LLMs) membuka peluang baru untuk mengotomatisasi inspeksi dan interpretasi deterministik nilai metrik ini tanpa input manual. Penelitian ini mengevaluasi kinerja komputasi lokal (Edge AI) dalam menafsirkan dan mengklasifikasikan status listrik menggunakan pendekatan Zero-Shot Learning tanpa perlu melatih ulang model. Aturan interpretasi didasarkan pada standar PUIL 2020 dan mencakup parameter seperti deviasi tegangan, frekuensi, beban arus, ketidakseimbangan, dan faktor daya. Pengujian perbandingan melibatkan dua model kuantisasi 8-bit: Meta Llama 3.1 (8B) dan Alibaba Qwen 2.5 (7B), yang dievaluasi menggunakan 200 sampel data historis panel listrik bangunan (100 normal, 100 anomali). Penilaian meliputi metrik kinerja LLM (akurasi sintaksis dan semantik), klasifikasi deteksi anomali, dan efisiensi sumber daya perangkat keras. Hasil menunjukkan bahwa Alibaba Qwen 2.5 (7B) mengungguli Meta Llama 3.1 (8B) dalam penalaran matematis, dengan akurasi 91,50% dan presisi 95,60%, dengan false positive minimal, serta menyelesaikan analisis 42 menit lebih cepat, dengan penggunaan RAM puncak sebesar 8,9 GB. Sebaliknya, Llama 3.1 menunjukkan kepekaan berlebihan, menghasilkan akurasi 57,50%, presisi 54,19%, dan penggunaan memori yang lebih tinggi (11,9 GB). Temuan ini menunjukkan bahwa efektivitas Zero-Shot Learning dalam LLM untuk tugas logika lebih bergantung pada bias pelatihan model daripada jumlah parameter. Model yang dilatih khusus untuk pemrograman dan matematika (seperti Qwen 2.5) lebih andal, konsisten, dan efisien dalam menafsirkan metrik listrik dibandingkan dengan model percakapan umum.

Kata kunci: edge AI, interpretasi metrik listrik, anomali kelistrikan, model bahasa besar, PUIL 2020, zero-shot learning, Llama 3.1, Qwen 2.5.

Abstract

Automatic interpretation of building electrical metric data is essential for assessing the reliability and suitability of electrical systems. The emergence of Large Language Models (LLMs) has created new opportunities to automate the inspection and deterministic interpretation of these metric values without requiring manual input. This study evaluates the performance of local computing (Edge AI) in interpreting and classifying electrical system status using a Zero-Shot Learning approach without the need for model retraining. The interpretation rules were based on the PUIL 2020 standard and included parameters such as voltage deviation, frequency, current load, imbalance, and power factor. The comparative evaluation involved two 8-bit quantized models: Llama 3.1 (8B) and Qwen 2.5 (7B), tested using 200 historical building electrical panel data samples (100 normal and 100 anomalous).
The assessment covered LLM performance metrics (syntactic and semantic accuracy), anomaly detection classification, and hardware resource efficiency. The results show that Qwen 2.5 (7B) outperformed Llama 3.1 (8B) in mathematical reasoning tasks, achieving an accuracy of 91.50% and a pre-

 ISSN:2302-8149
e-ISSN:2540-9719

cision of 95.60%, with minimal false positives. In addition, Qwen completed the analysis 42 minutes faster while using a peak RAM consumption of 8.9 GB. In contrast, Llama 3.1 demonstrated excessive sensitivity, resulting in an accuracy of 57.50%, a precision of 54.19%, and higher memory usage (11.9 GB). These findings indicate that the effectiveness of Zero-Shot Learning in LLMs for logical reasoning tasks depends more on the model's training bias than on the number of parameters. Models specifically trained for programming and mathematical reasoning, such as Qwen 2.5, are more reliable, consistent, and efficient in interpreting electrical metrics compared to general conversational models.

Keywords: *edge AI, electrical metric interpretation, electrical anomalies, large language model, PUIL 2020, zero-shot learning, Llama 3.1, Qwen 2.5.*

1 Pendahuluan

Di era digital, sangat penting bagi sebuah organisasi untuk menyelaraskan strategi bisnisnya dengan sistem informasi guna mendapatkan keunggulan kompetitif [1]. Banyak studi sebelumnya telah menciptakan sistem berbasis web untuk manajemen inventaris [2], pemesanan makanan [3], dan informasi pariwisata [4], dengan memanfaatkan pendekatan pengembangan seperti *Agile Scrum* dan *Waterfall*. Namun, sistem ini biasanya bergantung pada aturan tetap atau logika deterministik yang tidak secara otomatis diperbarui sebagai respons terhadap data baru. Sebagai alternatif, pendekatan sistem pakar berbasis *forward-chaining* dapat digunakan [5], yang mampu memberikan saran berbasis aturan, tetapi memerlukan pembaruan manual terhadap basis pengetahuan. Masalah ini muncul karena sistem aplikasi bergantung pada aturan statis atau logika deterministik yang membatasi kemampuannya untuk memproses dan menampilkan data secara efektif. Akibatnya, ada kebutuhan yang semakin besar untuk mengadopsi teknologi yang lebih adaptif dan fleksibel seperti kecerdasan buatan atau *Artificial Intelligence* (AI).

AI adalah teknologi yang memungkinkan komputer dan mesin untuk meniru kecerdasan manusia [6]. AI didasarkan pada algoritma dan model matematika untuk memproses data, mengenali pola, dan membuat keputusan cerdas [7]. AI berupaya meningkatkan kemampuan teknologi dengan meniru pembelajaran, pemahaman, pemecahan masalah, pengambilan keputusan, kreativitas, dan tindakan secara mandiri pada manusia [8]. Contoh nyata dari upaya meniru kecerdasan manusia dapat ditemukan pada model bahasa besar atau *Large Language Model* (LLM).

LLMs adalah sistem pembelajaran mendalam (*deep learning*) yang dilatih pada sejumlah besar data dan dengan banyak parameter untuk memahami dan menghasilkan bahasa alami serta konten mirip buatan manusia lainnya. Pelatihan ekstensif ini memungkinkannya untuk memprediksi dan menghasilkan teks berdasarkan input, sehingga dapat melakukan percakapan, menjawab pertanyaan, atau bahkan menulis kode. Model-model ini biasanya dilatih dengan sumber data dunia nyata yang luas (misalnya, publikasi online, buku, artikel berita, media sosial, dan konten berbasis web lainnya), yang membantu memahami konsep-konsep kompleks dan melakukan generalisasi lebih efektif terhadap tugas-tugas baru dengan pelatihan spesifik minimal [9].

Penerapan model bahasa besar (LLM) pada data metrik untuk deteksi anomali masih kurang dieksplorasi, sehingga menimbulkan pertanyaan tentang efektivitas potensialnya. Namun, memanfaatkan LLM untuk mengidentifikasi anomali dan menginterpretasikan data metrik dapat secara signifikan mengurangi ketergantungan pada interpretasi manusia, yang dapat dibatasi oleh kemampuan operator dan proses pengambilan keputusan yang tidak konsisten. Hal ini membuka jalan bagi sistem otomatisasi yang lebih canggih, memastikan bahwa pengguna tetap mendapat informasi yang baik tentang apa yang sedang terjadi dan dapat bertindak cepat untuk mencegah bencana. Meskipun demikian, pendekatan yang menjanjikan ini juga menghadirkan tantangan baru yang perlu dipertimbangkan dengan matang.

Dalam lingkungan berisiko tinggi, seperti pada sistem kelistrikan, di mana informasi yang salah dapat menyebabkan bencana atau kegagalan layanan, mengandalkan LLM juga menjadi masalah karena pengetahuannya terbatas pada data pelatihan, sehingga menimbulkan pertanyaan tentang keandalan dan akurasinya. Selain itu, pelatihan ulang LLM memakan waktu dan sumber daya yang besar. Karena LLM mampu memahami struktur dan hubungan yang rumit, mereka menjadi solusi potensial untuk masalah umum dalam deteksi anomali. Penggunaan pembelajaran zero-shot dapat

 ISSN:2302-8149
e-ISSN:2540-9719

membantu mengatasi keterbatasan ini dengan membimbing model untuk mengikuti instruksi dan tugas tertentu [10].

Untuk mengatasi masalah ini, makalah ini menyarankan pendekatan di mana LLM berperan sebagai filter interpretatif utama melalui panduan prompt. Model ini dibatasi untuk menganalisis data metrik listrik dalam batas operasional yang sudah ditetapkan agar mengurangi risiko halusinasi AI. Untuk itu, kami mengintegrasikan "Persyaratan Umum Instalasi Listrik (PUIL) 2020" sebagai batasan yang harus dipatuhi. Dengan mengubah angka metrik yang kompleks menjadi narasi yang mudah dipahami manusia, LLM membantu operator menafsirkan status sistem dengan lebih jelas. Pendekatan ini memastikan bahwa LLM memberikan wawasan yang cepat dan relevan kepada operator manusia, yang tetap memegang otoritas pengambilan keputusan akhir, sehingga menyeimbangkan pembelajaran zero-shot yang cepat dengan kebutuhan pemeriksaan keselamatan.

2 Tinjauan Literatur

Kemajuan terbaru dalam LLM telah secara signifikan memperluas penggunaannya di luar pemrosesan bahasa alami tradisional, mencakup tugas penalaran kompleks dan deteksi anomali dalam berbagai sistem siber-fisik. Kemampuan alami LLM dalam memahami sintaks terstruktur dan logika kontekstual tanpa memerlukan penyetelan khusus telah terbukti dalam lingkungan pengembangan terpadu, menunjukkan kesadaran kontekstual mereka [11]. Kemampuan penalaran ini telah berhasil diperluas ke bidang fisik dan waktu, di mana LLM digunakan untuk deteksi anomali secara *real-time* dan perencanaan reaktif dalam sistem robotik dengan menggabungkan penalaran generatif yang lambat dan pengklasifikasi anomali yang cepat [12]. Begitu pula, kemampuan AI generatif dalam memproses, memisahkan, dan mengelompokkan data sinyal berurutan untuk mendeteksi anomali telah terbukti efektif dalam aplikasi medis yang kompleks. Hal ini menunjukkan bahwa LLM memiliki kemampuan penalaran zero-shot yang kuat yang dapat diterapkan baik pada data terstruktur maupun data deret waktu [13].

Dalam konteks khusus infrastruktur kritis dan data tabular, penerapan LLM mulai menunjukkan hasil yang sangat menjanjikan dibandingkan dengan model pembelajaran mesin tradisional. Sebuah kerangka kerja berbasis AI generatif terbaru menunjukkan efektivitas sistem dialog berorientasi tugas dalam mendeteksi anomali pada pesan multicast (seperti GOOSE dan SV) dalam komunikasi smart grid, membuktikan bahwa LLM dapat beradaptasi terhadap ancaman siber baru jauh lebih cepat daripada teknik pembelajaran mesin konvensional yang melibatkan manusia secara langsung atau metode statis [14]. Selain itu, penggunaan model sumber terbuka dengan 7 miliar parameter pada data keamanan siber berbentuk tabel menunjukkan keberhasilan yang signifikan. Dengan rekayasa *prompt* yang strategis seperti penalaran berantai (*chain-of-thought reasoning*) dan pengenalan sampel data, model LLM yang lebih kecil mampu mengungguli model yang jauh lebih besar [15]. Ini menyoroti potensi praktis dan efisiensi komputasi dalam lingkungan terbatas, mengubah paradigma bahwa deteksi anomali yang sangat akurat secara ketat memerlukan arsitektur dengan banyak parameter.

Peralihan dari AI berbasis *cloud* ke *Edge AI* yang lokal merupakan langkah penting bagi sistem yang membutuhkan privasi data tinggi, latensi rendah, dan operasi tanpa koneksi terus-menerus. Penerapan LLM secara offline baru-baru ini dibuktikan melalui pengembangan sistem interaktif yang menggunakan model kuantisasi dengan 7 miliar parameter, yang mampu melakukan interaksi bahasa alami secara langsung dan pengolahan data tanpa bergantung pada layanan *cloud* [16]. Ini sejalan dengan meningkatnya kebutuhan akan penyebaran data tabular sensitif secara privat [15]. Menjalankan LLM kuantisasi di perangkat keras lokal mengurangi latensi dan risiko keamanan terkait panggilan *API cloud*, sehingga cocok untuk pemantauan industri secara terus-menerus dan aplikasi pengelolaan fasilitas.

Walaupun telah ada kemajuan besar, masih ada kekurangan penting dalam interpretasi deterministik terhadap metrik listrik fisik. Literatur saat ini lebih sering memakai LLM untuk deteksi ancaman keamanan siber, analisis paket jaringan, atau klasifikasi anomali secara umum, di mana batas-batas anomali dipelajari atau dievaluasi secara probabilistik berdasarkan pola abstrak [14], [15]. Tidak ada penelitian khusus yang mengevaluasi kemampuan *zero-shot* dari *Large Language Models* (LLM) yang dikonversi ke bentuk kuantisasi dalam menginterpretasi parameter fisik tabular mentah, seperti deviasi tegangan, frekuensi, dan ketidakseimbangan arus, terhadap standar teknik mutlak dan deterministik seperti PUIL 2020. Selain itu, meskipun kemampuan semantik LLM sudah banyak

 ISSN:2302-8149
e-ISSN:2540-9719

terdokumentasi, halusinasi komputasi bawaan dari berbagai arsitektur model saat melakukan evaluasi matematis ketat terhadap ambang desimal belum banyak dieksplorasi. Artikel ini bertujuan mengatasi kekurangan tersebut dengan mengimplementasikan dan mengevaluasi kerangka *zero-shot learning* pada *Edge AI* menggunakan LLM yang dikonversi ke kuantisasi, khususnya Llama 3.1 dan Qwen 2.5. Dengan menganalisis relevansi kontekstual dan secara ketat mengukur tingkat halusinasi matematis melalui agen penilai AI, penelitian ini menawarkan wawasan komparatif tentang efektivitas LLM sumber terbuka yang dilokalisasi dalam menggantikan pemrograman berbasis aturan yang kaku untuk interpretasi anomali listrik secara ketat dan deterministik.

3 Metode Penelitian

Makalah ini menguraikan pendekatan sistematis dalam menerapkan LLM untuk interpretasi dan deteksi anomali data metrik listrik. Pendekatan tersebut terdiri dari empat tahap utama: pengumpulan data, pemrosesan data, desain *prompt*, dan penilaian metrik.

Pengumpulan Data

Data kelistrikan dikumpulkan melalui pembacaan sensor dan disimpan dalam basis data pemantauan. Pencatatan dilakukan setiap 15 menit untuk memantau perubahan parameter sistem tiga fasa. Parameter yang direkam mencakup 11 atribut utama, termasuk:

Tabel 1 Atribut Parameter

Parameter	Atribut
Tanggal dan Waktu	DateLog
Tegangan	VoltageR, VoltageS, VoltageT
Arus	CurrentR, CurrentS, CurrentT
Faktor Daya	PowerFacR, PowerFacS, PowerFacT
Frekuensi	Freq

Tabel 1 memperlihatkan empat parameter utama dalam sistem kelistrikan yang mencerminkan kondisi atau kesehatan, beban, dan efisiensi secara langsung. Atribut-atribut ini menyimpan data mentah di basis data, menunjukkan nilai fisik yang diambil oleh sensor. Tabel 2 menyajikan contoh nilai yang secara alami menandakan adanya anomali dan nilai yang sesuai standar dasar. Nilai yang mencurigai diindikasikan dengan huruf tebal.

Tabel 2 Contoh representatif data sensor mentah yang belum dinormalisasi

ID	Tegangan R, S, T (V)	Arus R, S, T (A)	Kapasitas Arus (A)	Frekuensi (Hz)	Faktor Daya (R, S, T)
1	227.64, 224.88, 225.76	153.768, 161.07, 153.079	400	50.033	0.939, 0.949, 0.953
2	229.7, 226.93, 227.27,	138.597, 141.693, 135.616	400	50.002	0.942, 0.954, 0.952
3	**231.53**, 229.47, 229.8	73.974, 77.208, 72.477	400	50.018	0.922, 0.928, 0.923
4	**232.73**, 230.59, **231.01**	54.302, 65.694, 63.327	400	49.985	**0.845**, 0.886, 0.924

 ISSN:2302-8149
e-ISSN:2540-9719

Tabel 2 menampilkan empat contoh sampel dari metrik sensor fisik mentah sebelum dinormalisasi. Dalam sistem kami, data mentah ini tidak langsung digunakan dalam model bahasa. Sebagai gantinya, data melewati tahap praproses di mana data dinormalisasi menjadi nilai persentase seperti deviasi tegangan, beban arus, dan ketidakseimbangan arus agar sesuai dengan standar PUIL 2020.

Pemrosesan Data

Pemrosesan data mencakup mengubah data mentah menjadi format yang lebih terstruktur, konsisten, dan kontekstual agar siap digunakan untuk inferensi. Tidak seperti manusia, model AI seperti LLM tidak menafsirkan informasi dengan cara yang sama; mereka bergantung pada pengenalan pola [17]. Memberikan urutan angka sensor mentah langsung ke LLM dalam format zero-shot sering kali menyebabkan penalaran yang kurang akurat. Untuk membantu LLM membuat inferensi yang lebih tepat, penting untuk melakukan pra-pemrosesan data agar pola terstruktur dapat ditangkap dan nilai numerik dinormalisasi.

Makalah ini membahas dua teknik utama dalam pengolahan data: restrukturisasi format dan normalisasi nilai. Awalnya, data mentah yang disimpan sebagai *array* dalam PHP diubah ke format *JavaScript Object Notation* (JSON). Penggunaan struktur JSON membantu LLM untuk lebih memahami konteks data melalui pasangan kunci-nilai yang secara eksplisit mendefinisikan skema informasi. Format data seperti JSON dapat mengurangi ambiguitas saat interpretasi dan meningkatkan kinerja LLM dalam analisis. Selain itu, data numerik mentah dalam JSON diubah menjadi bentuk deskriptif atau metrik deviasi, seperti persentase penurunan tegangan relatif terhadap batas nominal. Penelitian menunjukkan bahwa penggunaan struktur data yang terdefinisi dengan baik seperti JSON dengan skema, bersama dengan normalisasi format numerik, dapat secara signifikan mengurangi "*hallucinations*", meminimalkan ambiguitas, dan meningkatkan akurasi LLM dalam deteksi anomali di sistem siber-fisik dan pengelolaan fasilitas listrik [18], [19].

Tabel 3 menampilkan hasil normalisasi data sensor, di mana pengukuran absolut seperti tegangan dan arus diubah menjadi persentase dan deviasi relatif. Pendekatan ini memastikan dataset sesuai dengan batasan deterministik yang ditetapkan dalam standar PUIL 2020.

Tabel 3 Contoh representatif data sensor yang dinormalisasi

ID	Deviasi Tegangan (%)	Arus Rata-Rata (A)	Ketidakseimbangan Arus (%)	Beban Arus (%)	Deviasi Frekuensi (Hz)	Faktor Daya (fasa terkecil)
1	3.47%, 2.22%, 2.62%	155.97	3.12% (from 161.07 A)	38.99%	+0.033,	0.939 (R)
2	4.41%, 3.15%, 3.30%	138.64	2.27% (from 141.69 A)	34.66%	+0.002,	0.942 (R),
3	**5.24%**, 4.30%, 4.45%	74.55	3.17% (from 77.21 A)	18.64%	+0.018	0.922 (R)
4	**5.79%**, 4.81%, **5.00%**	61.11	9.31% (from 65.69 A)	15.28%	-0.015	**0.845** (R)

Langkah pemrosesan ini berfungsi sebagai jembatan penting dalam proses kognitif LLM. Contohnya, pada Sampel ID 3, tegangan fase R mentah diubah menjadi deviasi sebesar 5,24%. Karena batas standar maksimal deviasi positif adalah +5%, nilai ini dinyatakan sebagai persentase yang dinormalisasi, sehingga memudahkan menentukan apakah ambang batas terlampaui. Pendekatan

 ISSN:2302-8149
e-ISSN:2540-9719

ini mengurangi kebutuhan LLM untuk melakukan perhitungan kompleks, seperti menghitung selisih persentase antara 231,5 V dan 220 V, yang berisiko menyebabkan kesalahan atau halusinasi.

Demikian pula, Sampel ID 4 menunjukkan anomali multi-metrik, dengan deviasi tegangan sebesar 5,79% dan penurunan faktor daya kritis menjadi 0,845, sedikit di bawah batas 0,85. Dengan memasukkan metrik yang telah dihitung dan dinormalisasi ini dalam format JSON yang terstruktur ke dalam LLM, sistem mengurangi beban aritmatika AI. Hal ini memungkinkan LLM untuk memusatkan seluruh jendela konteks dan kapasitas pemrosesannya pada kekuatan utamanya: deduksi logis, penalaran deterministik, dan klasifikasi anomali yang akurat sesuai aturan yang berlaku.

Desain Prompt

Model Bahasa Besar (LLMs) sangat efektif dalam Pembelajaran Tanpa Pengamatan, yang berarti model mampu menyelesaikan tugas tanpa data pelatihan sebelumnya untuk tugas tersebut [20]. Namun, meskipun memiliki kemampuan tersebut, efektivitas Model Bahasa Besar (LLM) sangat bergantung pada kejelasan dan struktur instruksi yang diberikan [21]. Studi ini dirancang secara modular untuk mengurangi kesalahan persepsi dan memastikan bahwa model bahasa besar (LLM) mengikuti batasan operasional sistem listrik, terinspirasi dari metode pengembangan sistem berurutan seperti Waterfall, yang membagi tahapan secara berurutan [22], [23].

Struktur prompt dibagi menjadi lima modul utama, yang dieksekusi secara berurutan oleh model:

1. Spesifikasi Peran
   Modul ini menjadikan LLM sebagai pengawas fasilitas kelistrikan yang sangat ahli, dengan fokus pada pendekatan yang ketat dan analitis dalam penalaran.
2. Kesadaran Skema
   LLM dapat memberikan query atau kesimpulan yang salah jika salah menafsirkan struktur data. Modul ini secara jelas menjelaskan apa yang diwakili oleh setiap kunci JSON dalam data input, misalnya, VoltageR menunjukkan tegangan pada fasa R.
3. Injeksi Aturan – PUIL 2020
   Modul ini membentuk inti dari penalaran berbasis aturan sistem. Modul ini mengubah parameter operasional aman yang diambil dari Persyaratan Umum Instalasi Listrik 2020 (PUIL) menjadi kendala logika mutlak. Model ini bergantung sepenuhnya pada aturan-aturan tersebut sebagai dasar utama untuk menghasilkan wawasan.
4. Blok Data Masukan
   Modul ini berisi data metrik listrik harian yang dikumpulkan setiap 15 menit dari sensor Acuview, kemudian diproses dan disusun dalam format JSON yang telah dinormalisasi.
5. Zero-Shot CoT dan Skema Output
   Modul ini menerapkan metode *Zero-Shot Chain-of-Thought* dengan instruksi: “Mari kita pikirkan dan evaluasi langkah demi langkah.” Selanjutnya, model dipaksa merangkum proses tersebut ke dalam format JSON standar yang dapat dibaca otomatis oleh sistem lain.

Metrik Evaluasi

Klasifikasi anomali listrik menggunakan LLM adalah salah satu bentuk pemodelan data yang sebanding dengan algoritma pohon keputusan (*decision tree*) dalam Python, yang bergantung pada data terstruktur dan kriteria objektif [24]. Studi ini mengevaluasi efektivitas model menggunakan dua pendekatan: format khusus LLM yang dipadukan dengan evaluasi semantik serta analisis klasifikasi statistik.

1. Metrik Pengukuran Performa LLM (Format dan Semantik)
   Dalam pengukuran kinerja, sebuah sistem tidak hanya dinilai dari efisiensi teknis, tetapi juga harus mempertimbangkan keseimbangan antara akurasi, ketelitian, dan konsistensi hasil [25]. Evaluasi output LLM dilakukan melalui metode penilaian proses informasi yang terstruktur, menggunakan empat kriteria utama.
   a. Keabsahan Eksekusi
      Untuk menjamin keandalan sistem, semua output LLM divalidasi atas keabsahan eksekusinya dengan memeriksa format JSON, mirip dengan pendekatan black-box dalam pengembangan sistem informasi yang memastikan setiap fungsi beroperasi sesuai spesifikasi tanpa menginspeksi kode internal [4].

 ISSN:2302-8149
e-ISSN:2540-9719

Proses ini mengevaluasi apakah sistem menghasilkan output yang sintaksis benar dan dapat dieksekusi. Dalam studi ini, validitas diukur berdasarkan kemampuan model untuk menghasilkan objek JSON yang bersih dan dapat diparse oleh sistem PHP tanpa kesalahan, serta tanpa elemen non-JSON seperti teks pengantar atau anotasi *markdown*.

b. Kebenaran
Menilai apakah akurasi faktual sesuai dengan kebenaran dasar yang sebenarnya. Metode ini memeriksa apakah klasifikasi akhir ("Normal" atau "Anomali") yang dihasilkan oleh LLM cocok dengan penilaian manual oleh ahli listrik sesuai dengan PUIL 2020.

c. Relevansi Kontekstual
Menilai sejauh mana niat di balik query input sesuai dengan output dari model. Relevansi bergantung pada kekuatan argumen dalam bagian '*Step_By_Step_Reasoning*'. AI dianggap relevan jika perhitungan deviasinya menggunakan parameter sensor yang tepat dan mengutip artikel yang relevan.

d. Tingkat Halusinasi
Evaluasi sejauh mana model memperkenalkan informasi palsu atau tidak relevan dalam responsnya. Sebuah LLM dianggap halusinasi jika ia mengada-adakan nilai numerik sensor yang tidak ada dalam data input asli atau menyusun aturan operasional fiktif yang tidak disebutkan dalam draf PUIL 2020.

2. Metrik Klasifikasi Deteksi Anomali
Untuk menilai keandalan sistem, klasifikasi dari LLM disajikan dalam matriks kebingungan (*confusion matrix)* yang menunjukkan empat kemungkinan hasil.

a. True Positive (TP)
Sistem LLM mendeteksi adanya anomali, memastikan bahwa kondisi yang sebenarnya memang seperti itu.

b. False Positive (FP)
Sistem LLM memberikan peringatan meskipun kondisi normal, menandakan alarm palsu.

c. True Negative (TN)
Sistem LLM mengidentifikasi kondisi normal dan memastikan bahwa situasi tersebut memang dalam keadaan normal.

d. False Negative (FN)
Sistem LLM menandai sebuah kondisi sebagai normal, walaupun sebenarnya itu adalah anomali (kasus berbahaya).

Berdasarkan matriks, evaluasi dilaksanakan dengan menggunakan perhitungan metrik statistik berikut:

a. Akurasi (*accuracy)*
Akurasi berfungsi untuk menunjukkan berapa banyak prediksi yang benar dari keseluruhan prediksi yang dilakukan oleh model. Matriks ini memberi gambaran bagaimana kinerja secara keseluruhan, tetapi dapat menyesatkan apabila satu kelas lebih dominan. Misalnya, model yang sering memprediksi data dengan status anomali lebih banyak dengan benar mungkin memiliki akurasi tinggi, tetapi gagal menangkap aspek penting dari data dengan status normal.

$$Accuracy = (TP + TN) / (TP + TN + FP + FN) \quad (1)$$

Rumus (1) menghitung persentase prediksi yang benar secara keseluruhan, termasuk kasus normal dan anomali, dari semua data yang tersedia.

b. Presisi (*precision)*
Matriks ini berfokus pada kualitas prediksi positif dari model dengan menunjukkan berapa banyak dari prediksi 'positif' yang sebenarnya benar. Hal ini penting dalam situasi di mana *false positives* perlu diminimalkan, seperti dalam mendeteksi email spam atau penipuan.

 ISSN:2302-8149
e-ISSN:2540-9719

$$Precision = TP / (TP + FP) \quad (2)$$

Rumus (2) mengukur tingkat ketepatan model dalam mendeteksi anomali yang sesungguhnya, di mana prediksi *false positive* dapat menimbulkan informasi yang keliru dan menyebabkan kegagalan sistem.

c. Sensitivitas (*recall)*
Mengukur seberapa baik model dalam memprediksi kasus positif dengan menunjukkan proporsi dari kasus *true positive* yang terdeteksi dari seluruh kasus positif yang sebenarnya. *Recall* yang tinggi sangat penting ketika kegagalan mendeteksi kasus positif memiliki konsekuensi yang signifikan.

$$Recall = TP / (TP + FN) \quad (3)$$

Rumus (3) menggambarkan kemampuan sistem dalam mendeteksi semua anomali yang sebenarnya ada. Dalam sistem listrik, hal ini merupakan metrik terpenting yang perlu ditingkatkan, karena satu anomali (FN) dapat berakibat pada kerusakan fatal pada peralatan sistem kelistrikan.

d. Skor F1
Karena sangat sulit mencapai nilai *precision* dan *recall* keduanya 100%, *F1-score* muncul sebagai solusi yang seimbang dengan menggabungkan hasil dari rumus (2) dan rumus (3) menjadi satu metrik untuk menyeimbangkan pertukaran (*trade-off*) keduanya. Metrik ini memberikan gambaran yang lebih baik tentang kinerja keseluruhan sebuah model, terutama untuk *dataset* yang tidak seimbang. Metrik ini berguna ketika baik *false positives* maupun *false negatives* sama pentingnya.

4 Hasil dan Pembahasan

Sistem deteksi anomali listrik berbasis Model Bahasa Besar (LLM) diuji menggunakan pendekatan Zero-Shot Learning, di mana model tidak diberi data pelatihan atau jawaban dalam prompt. Sebagai gantinya, model mendapatkan instruksi yang tegas tentang batas parameter listrik sesuai standar PUIL 2020, termasuk Deviansi Tegangan, Frekuensi, Beban Arus, Ketidakseimbangan Arus, dan Faktor Daya.

Evaluasi dilakukan dengan menggunakan 200 sampel historis dari panel distribusi sekunder, terdiri dari 100 data normal dan 100 data anomali atau gangguan. Untuk menilai apakah penerapan komputasi lokal (*Edge AI*) layak, dilakukan pengujian perbandingan dengan dua model dasar canggih menggunakan mesin Ollama. Kedua model tersebut diatur dengan kuantisasi 8-bit (Q8_0) dan jendela konteks sebanyak 128.000 token agar perbandingan adil. Model yang dievaluasi adalah:

1. Meta Llama 3.1 (8 Miliar / 8B Parameters)
2. Alibaba Qwen 2.5 (7 Miliar / 7B Parameters)

Kinerja kedua model dievaluasi berdasarkan dua metrik utama dari metodologi penelitian: metrik bawaan LLM (Format dan Semantik) serta metrik klasifikasi untuk deteksi anomali.

Evaluasi Metrik Kinerja Model Bahasa Besar (Format dan Semantik)

Kualitas output teks dari LLM dievaluasi melalui dua pendekatan, yaitu Validitas Eksekusi dan Kebenaran, yang diterapkan pada 200 dataset.

Untuk menilai relevansi kontekstual dan tingkat halusinasi secara akurat, dipilih satu set pilot berisi 10 sampel representatif dari setiap model (total 20 sampel validasi) dari log inferensi. Untuk menghindari bias evaluasi sirkular, di mana model lebih cenderung memilih keluaran dari dirinya sendiri, digunakan model pihak ketiga independen, DeepSeek, sebagai agen penilai AI (LLM sebagai hakim).

Evaluator DeepSeek dikalibrasi secara cermat sebelum dilakukan penilaian, menggunakan prompt sistem yang ketat untuk mendefinisikan batas deterministik standar PUIL 2020 secara akurat. Untuk menjaga objektivitas, proses validasi yang melibatkan manusia diterapkan pada set pilot. Seorang ahli teknik elektro secara manual memverifikasi skor biner yang diberikan oleh DeepSeek pada 20 sampel pilot gabungan untuk kedua metrik. Hasilnya dirangkum dalam Tabel 4 di bawah ini.

 ISSN:2302-8149
e-ISSN:2540-9719

Tabel 4 Hasil evaluasi metode pengukuran kinerja LLM

Kriteria Evaluasi	Meta Llama 3.1 (8B)	Alibaba Qwen 2.5 (7B)
Keabsahan Pelaksanaan	200 / 200 (**100%**)	200 / 200 (**100%**)
Kebenaran	115 / 200 (**57.50%**)	183 / 200 (**91.50%**)
Relevansi Konteks*	40 / 40 (**100%**)	40 / 40 (**100%**)
Tingkat Halusinasi*	29 / 40 (**72.50%**)	3 / 40 (**7.50%**)

**Dievaluasi menggunakan pengambilan sampel acak (n=40).*

Data kuantitatif dari tabel digunakan sebagai dasar untuk analisis perbandingan di bawah ini antara kedua model.

1. Keabsahan Pelaksanaan
   Meta Llama 3.1 (8B) dan Alibaba Qwen 2.5 (7B) keduanya mencapai tingkat akurasi 100% dalam pelaksanaan (200 dari 200 data poin). Kedua model tidak menghasilkan teks naratif (obrolan) maupun sintaks Markdown, yang berpotensi mengganggu struktur JSON. Secara empiris, ini menunjukkan bahwa desain instruksi *zero-shot* yang digunakan dalam studi ini sangat efektif dalam memastikan model bahasa mematuhi aturan format yang ketat.
2. Kebenaran
   Meskipun tampilannya sempurna, Meta Llama 3.1 (8B) hanya mencapai tingkat ketepatan faktual sebesar 57,50%, dengan 115 prediksi yang akurat. Sebaliknya, Alibaba Qwen 2.5 (7B) menunjukkan hasil yang jauh lebih baik, dengan tingkat kebenaran faktual sebesar 91,50% dan 183 prediksi yang benar.
3. Relevansi Konteks
   Hasil pengambilan sampel menunjukkan fenomena menarik dalam pemrosesan bahasa alami (*Natural Language Processing*), di mana kedua model mencapai tingkat relevansi kontekstual yang sempurna (100%). Ini mengindikasikan bahwa secara struktural, baik Llama maupun Qwen mampu menyusun argumen teks kondisional yang sangat kohesif tanpa *disconnection logic* (keterputusan penalaran).
4. Tingkat Halusinasi
   Meskipun memiliki alasan yang tampaknya relevan, Meta Llama 3.1 menunjukkan tingkat halusinasi matematis yang fatal sebesar 72,50%. Llama 3.1 berulang kali melakukan perhitungan deviasi fiktif, seperti menyimpulkan bahwa deviasi sebesar -4,41% berada dalam kisaran normal -4% hingga 5%. Ini menjadi penyebab utama menurunnya metrik keakuratan faktual Llama. Sebaliknya, Alibaba Qwen 2.5 terbukti sangat stabil, dengan tingkat halusinasi yang berkurang sebesar 7,50%. Model ini dapat secara deterministik merepresentasikan logika operasional matematis dalam bahasa alami tanpa perlu rekayasa numerik secara artifisial.
   Penilaian keandalan antarpenilai menggunakan *Cohen's Kappa* untuk mengukur konsistensi dalam mengidentifikasi kesalahan matematika. Kalibrasi awal mencapai skor *Cohen's Kappa* sebesar 0,90, menunjukkan kesepakatan yang hampir sempurna. Secara statistik, ini mendukung bahwa skor semantik dan tingkat halusinasi matematis yang diberikan oleh agen DeepSeek sangat objektif, dapat diandalkan, dan layak digeneralisasi ke dataset yang lebih besar. Hasil tersebut disajikan dalam Tabel 5 di bawah.

Tabel 5 Kesepakatan antara penilai deepseek (AI) dan para ahli manusia pada sampel percobaan berjumlah 20 sampel

Metrik	Jumlah Sampel	Kesepakatan yang Diamati	Kesepakatan yang Diharapkan	Cohen's Kappa	Interpretasi
Tingkat Halusinasi (Skor Tanpa Halusinasi)	20	0.95	0.50	0.90	Hampir sepenuhnya setuju

 ISSN:2302-8149
e-ISSN:2540-9719

Evaluasi Metrik Klasifikasi Deteksi Anomali

Metode ini mengevaluasi kinerja sistem seperti model pembelajaran mesin biner tradisional, dengan memanfaatkan matriks kebingungan untuk mengklasifikasikan data sensor listrik.

Tabel 6 Perbandingan metrik klasifikasi deteksi anomali

Metrik Klasifikasi	Meta Llama 3.1 (8B)	Alibaba Qwen 2.5 (7B)
Jumlah Prediksi "Anomali"	179 / 200 (89.5%)	91 / 200 (45.5%)
False Positive Percentage	82.0%	4.0%
True Positive (TP)	97	87
True Negative (TN)	18	96
False Positive (FP)	82	4
False Negative (FN)	3	13
Accuracy	57.50%	91.50%
Precision	54.19%	95.60%
Recall	97.00%	87.00%
F1-Score	69.53%	91.10%

Berdasarkan metrik pada tabel 6, analisis komparatif tentang kinerja klasifikasi antara kedua model adalah sebagai berikut:

1. Akurasi

   Pengukuran akurasi mengindikasikan tingkat keberhasilan keseluruhan dari prediksi model. Terdapat perbedaan yang cukup nyata antara kedua arsitektur model tersebut. Alibaba Qwen 2.5 (7B) meraih akurasi sebesar 91,50%, menunjukkan bahwa model ini sangat cocok digunakan untuk otomatisasi inspeksi. Sebaliknya, Meta Llama 3.1 (8B) hanya mendapatkan akurasi 57,50%, yang secara statistik hanya sedikit lebih baik dari tebakan acak.

   Uji two-proportion z-test dilakukan untuk menentukan apakah perbedaan akurasi antara kedua model tersebut secara statistik signifikan. Hipotesis nol (H_0) menyatakan bahwa kedua model memiliki tingkat akurasi yang sama. Pengujian ini menggunakan rumus (4) berikut:

$$z=\frac{p_1\text{-}p_2}{\sqrt{p(1\text{-}p)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}=\frac{0.575\text{-}0.915}{\sqrt{0.745\times0.255\times\left(\frac{1}{200}+\frac{1}{200}\right)}}=\text{-}7.80 \quad (4)$$

   Rumus (4) menghitung nilai statistik z dengan hasil 7,80 ($p < 0,0001$). Karena nilai p jauh lebih kecil dari 0,05, hipotesis nol ditolak. Ini menunjukkan bahwa perbedaan akurasi antara model-model tersebut signifikan secara statistik, dengan Qwen 2.5 jelas mengungguli Llama 3.1 dalam tugas deteksi anomali listrik ini.

2. Presisi

   Meta Llama 3.1 (8B) menunjukkan skor presisi yang sangat rendah, yaitu 54,19%, yang disebabkan oleh peningkatan jumlah *false positive* (FP), di mana model secara keliru menandai 83 data listrik normal sebagai "anomali". Tingginya jumlah FP ini berkorelasi langsung dengan tingkat halusinasi yang tinggi (seperti yang diukur pada metrik sebelumnya), karena model mengalami masalah saat menghitung rentang desimal. Sebaliknya, Alibaba Qwen 2.5 (7B) menunjukkan keandalan klasifikasi yang lebih baik dengan presisi sebesar 95,60%, menghasilkan hanya 4 *false positive* dari 100 data normal. Hal ini menunjukkan bahwa meskipun Qwen 2.5 mengklasifikasikan status "anomali", hasil tersebut dapat dianggap sangat terpercaya.

3. Sensitivitas

   Meta Llama 3.1 (8B) mencapai *recall* sebesar 97%, berhasil mendeteksi 97 dari 100 anomali nyata dan hanya mengalami 3 *false negatives*. Meskipun mengesankan, tingkat *recall* tinggi ini harus dianalisis bersama dengan perilaku klasifikasinya. Berdasarkan matriks kebingungan, Llama 3.1 mengklasifikasi 179 dari 200 sampel (89,5%) sebagai "anomali",

 ISSN:2302-8149
e-ISSN:2540-9719

menunjukkan alarm yang tinggi. Labeling yang agresif ini menangkap sebagian besar anomali nyata, tetapi juga menyebabkan 82 *false positives* dari 100 sampel normal, sehingga *False Positive Rate* (FPR) mencapai 82% dan *precision* sebesar 54,19%. Oleh karena itu, tingginya *recall* tidak otomatis menunjukkan model yang memiliki penalaran lebih baik, melainkan cenderung memprediksi "anomali", terutama dalam kasus tidak pasti. Hal ini juga berkorelasi dengan tingkat halusinasi yang tinggi sebanyak 72,50%, di mana nilai normal sering disalahartikan sebagai pelanggaran.

Alibaba Qwen 2.5 (7B) mencapai tingkat *recall* 87%, berhasil mengidentifikasi 87 dari 100 anomali tanpa mengurangi presisi. Model ini memberi label "anomali" pada 45,5% sampel, menggunakan pendekatan berhati-hati berbasis aturan yang hanya menandai pelanggaran eksplisit terhadap batas PUIL 2020. Pendekatan logis ini menghasilkan lebih sedikit *false positive*, meskipun 13 anomali tetap terlewatkan.

4. Skor F1

   Kelemahan utama Llama 3.1 dalam menjaga akurasi menyebabkan skor F1-nya turun tajam menjadi 69,53%. Sebaliknya, Alibaba Qwen 2.5 berhasil menyeimbangkan akurasi tinggi dan *recall* yang cukup, sehingga menghasilkan skor F1 yang sangat baik sebesar 91,10%.

Hasil Evaluasi Penggunaan Sumber Daya Komputasi

Rekaman penggunaan sumber daya perangkat keras selama 5 menit pertama proses inferensi dilakukan bersamaan dengan total waktu proses untuk 200 data uji.

1. Meta Llama 3.1 (8B)

   Dibutuhkan total waktu 3 jam, 21 menit, dan 41 detik. Penggunaan sumber daya mencapai puncaknya di 11,9 GB RAM dan 49,8% pemanfaatan CPU.
2. Alibaba Qwen 2.5 (7B)

   Dibutuhkan total 2 jam, 39 menit, dan 15 detik. Penggunaan sumber daya mencapai puncaknya di 8,9 GB RAM dan 49,2% pemanfaatan CPU.

Analisis Kepatuhan Zero-Shot Learning Berdasarkan Metrik Kinerja

Berdasarkan parameter evaluasi di Tabel 6, terdapat korelasi yang sangat kuat antara kinerja semantik LLM dan kualitas klasifikasi untuk deteksi anomali. Perbedaan kinerja kedua model ini terkait dengan tujuan pelatihan yang berbeda.

Meta Llama 3.1 menunjukkan kelemahan dalam metrik semantik, seperti kesulitan mempertahankan logika matematis di berbagai konteks. Hal ini menyebabkan skor presisi yang rendah sebesar 54,19%. Jumlah *false positives* yang tinggi (83 data) menunjukkan bahwa model sering kali mengeluarkan alarm palsu. Model ini mengalami kesulitan dalam penalaran deduktif saat mengevaluasi ambang listrik menggunakan pendekatan *zero-shot*, sehingga sering mengambil keputusan *fallback* yang salah menafsirkan anomali sebagai normal. Meskipun *recall* Llama 3.1 terlihat sangat tinggi di angka 97%, angka ini menyesatkan karena mencerminkan model yang terlalu sensitif dan menandai hampir semua data sebagai *noise*.

Sebaliknya, kemampuan Alibaba Qwen 2.5 dalam mempertahankan konsistensi semantik sangat bergantung pada tingkat presisi sebesar 95,60% dan akurasi 91,50% yang sangat tinggi. Jumlah kesalahan deteksi sangat kecil, hanya 4 *false positive* dari 100 data normal. Tingginya presisi dan skor F1 (91,10%) secara empiris menunjukkan bahwa model Qwen, yang dilatih secara mendalam pada data pemrograman (pengkodean), jauh lebih andal dalam mengikuti aturan logika ketat dari PUIL 2020.

Keterbatasan

Meskipun Alibaba Qwen 2.5 (7B) menunjukkan hasil yang menjanjikan dalam deteksi anomali listrik, penting untuk memperhatikan beberapa keterbatasan yang memengaruhi kemampuan generalisasi temuan tersebut.

Awalnya, dataset dikumpulkan dari satu panel listrik (220 V, MCB 400 A), yang mungkin tidak mewakili tingkat tegangan lainnya seperti 380 V atau 20 kV, atau jenis panel yang berbeda. Selain itu, hasil ini bersifat spesifik terhadap standar PUIL 2020; kerangka regulasi alternatif seperti IEEE atau IEC dapat menghasilkan kinerja yang berbeda untuk LLM. Selanjutnya, sampel sebanyak 200 data dan penggunaan *preprocessing* harian mungkin tidak cukup untuk mendeteksi anomali transien atau pola kerusakan langka. Kuantisasi 8-bit dapat menurunkan akurasi penalaran dibandingkan dengan

 ISSN:2302-8149
e-ISSN:2540-9719

model presisi penuh. Oleh karena itu, penting untuk menilai pendekatan ini pada berbagai infrastruktur listrik dan standar sebelum melakukan generalisasi. Akhirnya, studi ini merupakan simulasi *Edge AI*, bukan penerapan langsung pada perangkat keras *edge* nyata.

5 Kesimpulan

Pendekatan *zero-shot learning* telah terbukti efektif dalam metrik listrik tanpa memerlukan penyesuaian model tambahan. Namun, keberhasilannya dalam mendeteksi anomali matematis dan memastikan kepatuhan terhadap format sintaks JSON sangat bergantung pada bias pelatihan yang melekat pada arsitektur model dasar, bukan sekadar jumlah parameternya. Model Alibaba Qwen 2.5 (7B) tampil jauh lebih baik daripada Meta Llama 3.1 (8B) dalam tugas-tugas penalaran deterministik. Qwen 2.5 mendapatkan akurasi 91,50% dan presisi 95,60%, menunjukkan pemahaman semantik yang andal dalam kondisi logika ketat tanpa menghasilkan banyak *false positive*. Sebaliknya, Meta Llama 3.1 cenderung bereaksi berlebihan, menyebabkan tingginya *false positive* dan hanya mencapai presisi 54,19%. Temuan ini juga berpotensi meningkatkan efisiensi operasional melalui percepatan interpretasi data dan deteksi anomali, yang selanjutnya dapat meningkatkan produktivitas, kepercayaan pengguna, dan kepuasan terhadap sistem [26]. Jumlah parameter LLM tidak selalu berkorelasi langsung dengan efisiensi komputasi lokal (*Edge Computing*). Pengujian menunjukkan bahwa Qwen 2.5 (7B) memproses data metrik 200 kali lebih cepat, selesai 42 menit lebih awal, dan memerlukan RAM 3 GB lebih sedikit dibandingkan dengan Llama 3.1 (8B). Keunggulan efisiensi ini sebagian besar karena ukuran kosakata *tokenizer* Qwen yang lebih efektif dalam mengompresi bahasa Indonesia dan struktur JSON.

Referensi

[1] S. J. Safitri, G. A. Ramdhaniawan, A. Asro, and N. Rukhviyanti, "Analisis Literatur Review Perencanaan Strategi Sistem Informasi menggunakan Metode Metode Five Competitive Force Pada CV. Bio Chitosan Indonesia," *Bridge : Jurnal publikasi Sistem Informasi dan Telekomunikasi*, Vol. 2, No. 4, pp. 319–327, Sep. 2024, DOI: 10.62951/bridge.v2i4.263.

[2] N. Widaningsih, N. Windiyanti, and N. Rukhviyanti, "*Web-based Inventory Information System using Agile Scrum Method at CV Tunggal Putra Jaya*," *Sistemasi: Jurnal Sistem Informasi*, Vol. 14, No. 3, pp. 1471–1488, May 2025, DOI: 10.32520/STMSI.V14I3.5253.

[3] A. Fadhil, K. Al Jufri, S. A. Paskalis, and N. Rukhviyanti, "*Design of A Web-based Regional Food Ordering Information System at Seribu Rasa Restaurant*," *Jurnal Inovtek Polbeng - Seri Informatika*, Vol. 10, No. 1, p. 2025, Mar. 2025, DOI: 10.35314/mb5xe359.

[4] P. Adinda, D. Eviliana, and N. Rukhviyanti, "*Website-based Baduy Tourism Information System using the Software Development Life Cycle Method*," *INOVTEK Polbeng - Seri Informatika*, Vol. 10, No. 1, pp. 538–549, Mar. 2025, DOI: 10.35314/V8VTVT27.

[5] D. Zatnika and N. Rukhviyanti, "Penerapan Metode *Forward Chaining* pada Sistem Pakar Rekomendasi Mobil *Second* dari Aspek Penghasilan Kerja," *Jurnal Penelitian Inovatif*, Vol. 4, No. 4, pp. 2463–2476, Dec. 2024, DOI: 10.54082/jupin.759.

[6] A. N. Fatyandri, B. Guo, and E. S. Muchsinati, "*Impact of Artificial Intelligence and Human Resource Management on Leadership Organization Performance*," *Jurnal Teknologi dan Manajemen Informatika*, Vol. 10, No. 2, pp. 123–132, Dec. 2024, DOI: 10.26905/jtmi.v10i2.14060.

[7] E. S. Eriana and A. Zein, *Artificial Intelligence (AI)*. Eureka Media Aksara, 2023.

[8] B. Bidang *et al.*, "*Artificial Intelligence*," Jan. 2026.

[9] A. Mumuni and F. Mumuni, "*Large Language Models for Artificial General Intelligence (AGI): A Survey of Foundational Principles and Approaches*," Jan. 2025, Accessed: Mar. 30, 2026. [Online]. Available: http://arxiv.org/abs/2501.03151

[10] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa, "*Large Language Models are Zero-Shot Reasoners*," *Adv. Neural Inf. Process. Syst.*, Vol. 35, May 2022, Accessed: Nov. 03, 2025. [Online]. Available: https://arxiv.org/pdf/2205.11916

[11] D. Nam, A. MacVean, V. Hellendoorn, B. Vasilescu, and B. Myers, "*Using an LLM to Help with Code Understanding*," *Proceedings - International Conference on Software Engineering*, Vol. 13, pp. 1184–1196, May 2024, DOI: 10.1145/3597503.3639187;ISSUE:ISSUE:DOI.

[12] R. Sinha, A. Elhafsi, C. Agia, M. Foutter, E. Schmerling, and M. Pavone, "*Real-Time Anomaly Detection and Reactive Planning with Large Language Models*," Jul. 2024, Accessed: Mar. 31, 2026. [Online]. Available: http://arxiv.org/abs/2407.08735

[13] Y. Torabi, "*AI-Driven Cardiorespiratory Signal Processing: Separation, Clustering, and Anomaly Detection*," Feb. 2026, Accessed: Mar. 31, 2026. [Online]. Available: http://arxiv.org/abs/2602.09210

[14] A. Zaboli, S. L. Choi, T.-J. Song, and J. Hong, "*A Novel Generative AI-based Framework for Anomaly Detection in Multicast Messages in Smart Grid Communications*," Jun. 2024, Accessed: Mar. 31, 2026. [Online]. Available: http://arxiv.org/abs/2406.05472

[15] X. Zhao, X. Leng, L. Wang, N. Wang, and Y. Liu, "*Efficient Anomaly Detection in Tabular Cybersecurity Data using Large Language Models*," *SCI. Rep.*, Vol. 15, No. 1, Dec. 2025, DOI: 10.1038/s41598-025-88050-z.

[16] S. Shafian, "*Development of an Offline Interactive Chemistry Tutor using Generative AI and Large Language Model*," *Journal of Advanced Research in Computing and Applications Journal homepage*, Vol. 40, pp. 63–75, 2025, DOI: 10.37934/arca.40.1.6375.

[17] K. Mahowald, A. A. Ivanova, I. A. Blank, N. Kanwisher, J. B. Tenenbaum, and E. Fedorenko, "*Dissociating Language and Thought in Large Language Models*," *Trends Cogn. SCI.*, Vol. 28, No. 6, pp. 517–540, Jan. 2023, DOI: 10.1016/j.tics.2024.01.011.

[18] K. Buga, R. Tesic, E. Koyuncu, and T. Hanne, "*Large Language Models for Structured Information Processing in Construction and Facility Management*," *Electronics 2025, Vol. 14, Page 4106*, Vol. 14, No. 20, p. 4106, Oct. 2025, DOI: 10.3390/ELECTRONICS14204106.

[19] Y. Liu, H. Wu, and B. Liu, "*A Rule-Aware Prompt Framework for Structured Numeric Reasoning in Cyber-Physical Systems*," Dec. 2025, Accessed: Mar. 31, 2026. [Online]. Available: http://arxiv.org/abs/2512.12794

[20] C. Shorten *et al.*, "*StructuredRAG: JSON Response Formatting with Large Language Models*," Aug. 2024, Accessed: Nov. 03, 2025. [Online]. Available: https://arxiv.org/pdf/2408.11061

[21] A. Soffiyun Nada and S. Farisa Chaerul Haviana, "Implementasi *Zero-Shot Learning* untuk Prediksi Solusi dari Kalimat Masalah pada Artikel Ilmiah menggunakan *Large Language Models (LLM)*," *Jurnal Transistor Elektro dan Informatika (TRANSISTOR EI*, Vol. 7, No. 1, p. 2025, 2025, DOI: http://dx.doi.org/10.30659/ei.7.1.%25p.

[22] A. Muharom and N. Rukhviyanti, "*Development of Web-based Multimedia Learning for Grade 3 Elementary School Mathematics*," *Jurnal Inovtek Polbeng - Seri Informatika*, Vol. 10, No. 2, pp. 1142–1152, Jul. 2025, DOI: 10.35314/sj1qng08.

[23] D. N. Pryatama and N. Rukhviyanti, "Rancang Bangun Aplikasi Stok Barang dengan *QRcode* menggunakan Metode *Waterfall* dan *Framwork Laravel* pada Konveksi Sfgiandra," *Jurnal Kridatama Sains dan Teknologi*, Vol. 7, No. 01, pp. 71–89, Feb. 2025, DOI: 10.53863/KST.V7I01.1488.

[24] K. V. Benedict and N. Rukhviyanti, "*Analysis of the Classification of Data on the Launch of Apple Mobile Phone Prices in China and Pakistan using the Decision Tree Algorithm in Python Programming*," *Eduvest-Journal of Universal Studies*, Vol. 5, No. 9, pp. 10534–10546, Sep. 2025, DOI: 10.59188/eduvest.v5i9.51409.

[25] N. Rukhviyanti, "*Balanced Scorecard: Performance Measurement of University Mediated By Student Loyalty*," *Jurnal Manajemen*, Vol. 29, No. 2, pp. 400–420, Jun. 2025, DOI: 10.24912/JM.V29I2.2694.

[26] B. Wiguna Nugraha and N. Rukhviyanti, "*The Effect of Work Engagement, Work-Life Balance, and Work Overload on Employee Productivity: The Role of Job Satisfaction as Mediating Variable at BRI Employees In Bandung City*," *Indonesian Interdisciplinary Journal of Sharia Economics (IIJSE)*, Vol. 7, No. 2, 2024, DOI: 10.31538/iijse.v7i2.5235.