HOAQ: JURNAL TEKNOLOGI INFORMASI, Volume 17 Nomor 2, Desember 2026, p292-305
p-ISSN: 2337-5280, e-ISSN: 26207427
DOI: https://doi.org/10.52972/hoaq.vol17no2

PERFORMA *LARGE LANGUAGE MODEL* DALAM OTOMASI PEMBUATAN JURNAL DAN REKONSILIASI KEUANGAN: SIMULASI PADA SISTEM INFORMASI AKUNTANSI BERBASIS *CLOUD*

Wendy[1], Octavian Hartono[2], Novita[3], Ricky Imanuel Ndaumanu[4]

[1,3] Program Studi Akuntansi, Fakultas Ekonomi dan Bisnis, Universitas Widya Dharma Pontianak
[2] Program Studi Manajemen, Fakultas Ekonomi dan Bisnis, Universitas Widya Dharma Pontianak
[4] Program Studi Informatika, Fakultas Teknologi Informasi, Universitas Widya Dharma Pontianak
Jl. Hos Cokroaminoto No. 445, Kota Pontianak - Kalimantan Barat, Indonesia.
Email: [1] wendy@widyadharma.ac.id, [2] vincente_viano@yahoo.co.id, [3] novita.uwdp@gmail.com, [4] ricky_im@widyadharma.ac.id

ABSTRAK

Transformasi digital dalam sistem informasi akuntansi membuka peluang pemanfaatan *Large Language Model* (LLM) untuk mengotomasi pembuatan jurnal dan rekonsiliasi keuangan, terutama pada KUD/UMKM pengolahan sawit yang masih menghadapi keterbatasan pencatatan manual dan semi-digital. Penelitian ini bertujuan mengevaluasi akurasi, kecepatan, dan keandalan beberapa model LLM dalam menghasilkan jurnal otomatis dan melakukan rekonsiliasi transaksi pada Sistem Informasi Akuntansi berbasis *cloud*, serta mengidentifikasi pengaruh kompleksitas transaksi dan kualitas data terhadap performanya. Penelitian ini menggunakan pendekatan kuantitatif eksperimental berbasis simulasi dengan desain faktorial yang membandingkan GPT-4, Llama 3, Mistral, dan model FinEdge-like pada dataset sintetis transaksi dan paket rekonsiliasi yang dikonstruksi berdasarkan karakteristik operasional KUD/UMKM pengolahan sawit. Evaluasi dilakukan menggunakan metrik klasifikasi, waktu pemrosesan, *error rate*, statistik deskriptif, dan ANOVA tiga-arah. Berdasarkan eksperimen terhadap dataset sintetis yang dikonstruksi dan disimulasikan sesuai desain penelitian, GPT-4 menunjukkan performa tertinggi, diikuti FinEdge-like, Llama 3, dan Mistral. Seluruh model mengalami penurunan akurasi ketika kompleksitas transaksi meningkat dan kualitas data menurun. LLM juga menunjukkan waktu pemrosesan yang jauh lebih singkat dibandingkan prosedur manual dalam skenario simulasi. Temuan ini menegaskan potensi integrasi LLM dalam SIA berbasis *cloud* sebagai mekanisme augmentasi akuntansi yang tetap memerlukan validasi manusia, *audit trail*, dan pengendalian berbasis risiko. Temuan ini merepresentasikan kinerja model dalam skenario simulasi dan belum dapat ditafsirkan sebagai bukti kinerja pada data transaksi lapangan aktual.

Kata kunci: *Large Language Model* (LLM), otomasi pembuatan jurnal, rekonsiliasi keuangan, sistem informasi akuntansi berbasis *cloud*.

ABSTRACT

Digital transformation in accounting information systems opens up opportunities to utilize Large Language Models (LLMs) to automate journal entry generation and financial reconciliation, particularly for palm oil processing cooperatives and MSMEs that still face limitations with manual and semi-digital record-keeping. This study aims to evaluate the accuracy, speed, and reliability of several LLM models in generating automated journal entries and performing transaction reconciliations in a cloud-based accounting information system, as well as to identify the impact of transaction complexity and data quality on their performance. This study employs a quantitative experimental approach based on simulation with a factorial design that compares GPT-4, Llama 3, Mistral, and FinEdge-like models on a synthetic dataset of transactions and reconciliation packages constructed based on the operational characteristics of palm oil processing KUDs/SMEs. Evaluation was conducted using classification metrics, processing time, error rate, descriptive statistics, and a three-way ANOVA. Based on experiments using the synthetic dataset constructed and simulated according to the study design, GPT-4 demonstrated the highest performance, followed by the FinEdge-like model, Llama 3, and Mistral. All models experienced a decline in accuracy as transaction complexity increased and data quality decreased. LLMs also demonstrated significantly shorter processing times compared to manual procedures in the simulation scenarios. These findings underscore the potential for integrating LLMs into cloud-based SIA as an accounting augmentation mechanism that still requires human validation, audit trails, and risk-based controls. These findings represent model performance in simulation scenarios and cannot yet be interpreted as evidence of performance on actual field transaction data.

Keywords: Large Language Model (LLM), Automated Journal Entry Generation, Financial Reconciliation, Cloud-Based Accounting Information Systems.

Copyright (c) HOAQ (High Education of Organization Archive Quality) : Jurnal Teknologi Informasi

1. PENDAHULUAN

Transformasi digital telah mengubah sistem informasi akuntansi dari mekanisme pencatatan yang dominan manual menuju ekosistem berbasis *cloud*, otomatisasi, dan kecerdasan buatan yang memungkinkan pengolahan informasi keuangan secara lebih cepat dan terintegrasi. Perkembangan literatur mutakhir menunjukkan bahwa integrasi *artificial intelligence* (AI), *robotic process automation*, dan teknologi digital dalam akuntansi berpotensi meningkatkan efisiensi, akurasi, serta reliabilitas proses akuntansi, sekaligus menggeser pekerjaan akuntan dari aktivitas transaksional menuju fungsi analitis dan pengambilan keputusan [1] . Dalam perkembangan yang lebih spesifik, *Large Language Model* (LLM) menawarkan kemampuan pemrosesan bahasa alami untuk memahami konteks, mengolah data tekstual dalam jumlah besar, melakukan penalaran finansial, serta menghasilkan keluaran berdasarkan informasi terstruktur maupun tidak terstruktur, sehingga membuka peluang pemanfaatannya dalam berbagai aktivitas keuangan yang sebelumnya membutuhkan intervensi manusia secara intensif [2]. Dalam bidang akuntansi, otomatisasi berbasis AI semakin relevan karena aktivitas rutin seperti pemasukan data dan rekonsiliasi dapat diotomatisasi untuk mengurangi risiko kesalahan manusia dan mempercepat pemrosesan informasi [3]. Perkembangan tersebut memiliki arti strategis bagi usaha mikro, kecil, dan menengah (UMKM), sebab sistem *cloud accounting* dapat meningkatkan kecepatan, akurasi, relevansi, dan kualitas informasi keuangan, tetapi manfaat tersebut bergantung pula pada kesiapan internal seperti literasi digital, infrastruktur teknologi, dan kapasitas sumber daya manusia [4] . Pada konteks Indonesia, transformasi tersebut relevan bagi Koperasi Unit Desa (KUD) dan UMKM pengolahan sawit di Kalimantan Barat yang menjadi konteks penelitian ini, karena proses pembelian tandan buah segar (TBS), penjualan produk, pembayaran kepada pemasok, penerimaan dari pembeli, serta transaksi bank menghasilkan dokumen dan sumber informasi yang harus diklasifikasikan, dijurnal, dicocokkan, dan ditelusuri secara konsisten. Dengan demikian, integrasi LLM ke dalam sistem informasi akuntansi (SIA) berbasis *cloud* berpotensi membentuk mekanisme otomasi yang tidak hanya memindahkan pencatatan ke platform digital, tetapi juga mengubah dokumen sumber menjadi informasi akuntansi dan mendukung rekonsiliasi transaksi secara otomatis.

Meskipun potensi tersebut besar, terdapat kesenjangan antara perkembangan kemampuan LLM dan bukti empiris mengenai reliabilitasnya dalam pekerjaan akuntansi yang bersifat kontekstual dan membutuhkan ketepatan tinggi. Evaluasi LLM pada pengetahuan *chartered accountancy* memperlihatkan adanya variasi performa antar-model, yang mana model seperti GPT-4 menunjukkan kemampuan yang kuat pada sejumlah tugas konseptual, tetapi tantangan tetap muncul pada komputasi numerik dan interpretasi yang membutuhkan pengetahuan domain, sehingga performa tinggi pada tugas bahasa umum tidak dapat secara otomatis dianggap setara dengan reliabilitas dalam pekerjaan akuntansi [5] . Bukti lain dari evaluasi kemampuan analisis data finansial menegaskan bahwa reliabilitas LLM dalam domain keuangan masih perlu diuji melalui dimensi kemampuan dasar, penalaran, dan keterampilan teknis pada persoalan yang menyerupai penggunaan nyata [6] . Persoalan tersebut menjadi semakin penting pada rekonsiliasi keuangan karena proses ini melibatkan pencocokan transaksi lintas sumber, identifikasi ketidaksesuaian, dan penelusuran anomaly. Pendekatan AI memang menawarkan peluang untuk mengotomasi deteksi anomali dan rekonsiliasi dalam lingkungan *cloud-native*, tetapi persoalan interpretabilitas, skalabilitas, kepatuhan, dan keandalan tetap menjadi tantangan implementasi [7]. Pada sisi adopsi teknologi, penelitian terhadap UMKM Indonesia juga menunjukkan bahwa penggunaan *cloud accounting* dipengaruhi oleh kondisi teknologi dan organisasi, termasuk kesiapan teknologi, sedangkan faktor-faktor adopsi tidak selalu menunjukkan pengaruh yang seragam pada setiap konteks [8]. Literatur Indonesia lebih lanjut menunjukkan bahwa literasi digital dan kompatibilitas teknologi berhubungan dengan persepsi kemudahan dan kegunaan *cloud accounting*, sehingga kualitas teknologi saja tidak cukup untuk menjelaskan keberhasilan penerapannya pada UMKM [9]. Kesenjangan yang menjadi dasar penelitian ini terletak pada masih terbatasnya pengujian eksperimental yang secara bersamaan membandingkan kemampuan beberapa LLM dalam menghasilkan jurnal dari dokumen sumber dan melakukan rekonsiliasi keuangan pada lingkungan SIA berbasis *cloud*, khususnya menggunakan karakteristik transaksi UMKM agroindustri sawit Indonesia. Kesenjangan tersebut menuntut evaluasi yang tidak berhenti pada klaim bahwa AI mampu meningkatkan efisiensi, tetapi mengukur secara langsung akurasi klasifikasi dan pembuatan jurnal, kesalahan rekonsiliasi, waktu pemrosesan, serta perubahan performa ketika kualitas data dan kompleksitas transaksi berbeda.

Secara konseptual, penelitian ini menempatkan performa teknis LLM dan konteks penerimaan SIA berbasis *cloud* sebagai dua lapisan yang saling berkaitan dalam otomasi akuntansi. *Technology Acceptance Model* (TAM) digunakan untuk memahami dimensi penerimaan teknologi karena bukti empiris mutakhir dalam bidang akuntansi menunjukkan bahwa *perceived usefulness* dan *perceived ease of use* memengaruhi intensi pengguna untuk menggunakan teknologi informasi akuntansi [10]. Pada UMKM Indonesia, kajian mengenai adopsi *cloud accounting* juga menunjukkan bahwa *performance expectancy*, *effort expectancy*, dan *facilitating conditions* dapat berpengaruh signifikan terhadap intensi penggunaan, sedangkan penggunaan aktual sistem berhubungan positif dengan kualitas keputusan pemilik UMKM [11]. Pada lapisan pemrosesan informasi, *Natural Language Processing* (NLP) dan *Accommodation Theory* digunakan untuk memosisikan LLM sebagai mekanisme yang mengakomodasi variasi representasi informasi pada *invoice*, kwitansi, laporan timbang, laporan harian TBS, dan sumber transaksi lainnya ke dalam representasi jurnal yang lebih terstandardisasi. Relevansi pendekatan tersebut diperkuat oleh perkembangan LLM finansial multimodal yang telah dirancang untuk menangani informasi tekstual, numerik, tabel, dan visual secara terintegrasi [12]. *Reconciliation Theory* selanjutnya menjadi dasar konseptual bagi validasi dan penyesuaian informasi antar-sumber, karena otomasi rekonsiliasi berbasis AI dapat

Copyright (c) HOAQ (High Education of Organization Archive Quality) : Jurnal Teknologi Informasi

menggabungkan ekstraksi, standardisasi, pencocokan transaksi, identifikasi anomali, dan penyediaan jejak informasi yang mendukung pemeriksaan transaksi [13]. Sementara itu, *Cloud Accounting Framework* memberikan konteks arsitektural bagi integrasi proses tersebut karena *cloud accounting* memungkinkan akses data secara lebih cepat, otomasi proses, serta peningkatan fleksibilitas pengelolaan informasi, meskipun penerapannya pada UMKM tetap berhadapan dengan hambatan seperti keamanan, biaya, dan keterbatasan infrastruktur [14]. Keempat landasan tersebut menempatkan keberhasilan otomasi bukan semata-mata sebagai persoalan kemampuan generatif LLM, melainkan sebagai hasil interaksi antara kemampuan memahami dokumen, ketepatan menghasilkan representasi akuntansi, kemampuan mencocokkan informasi lintas sumber, kualitas lingkungan teknologi, dan penerimaan pengguna.

Berdasarkan permasalahan tersebut, penelitian ini bertujuan menguji secara eksperimental performa LLM dalam otomasi pembuatan jurnal dan rekonsiliasi keuangan pada SIA berbasis *cloud* di KUD/UMKM pengolahan sawit Kalimantan Barat. Secara operasional, penelitian diarahkan untuk mengukur akurasi, presisi, *recall*, dan *F1-score* keluaran jurnal otomatis, mengevaluasi waktu pemrosesan dan tingkat kesalahan rekonsiliasi, mengidentifikasi pengaruh kualitas data dan kompleksitas transaksi terhadap performa otomasi, serta menghasilkan rekomendasi desain arsitektur SIA berbasis *cloud* yang terintegrasi LLM untuk konteks UMKM sawit. Pemilihan indikator performa yang eksplisit diperlukan karena evaluasi LLM finansial semakin menekankan penggunaan kerangka *benchmarking* yang membedakan kemampuan model pada berbagai jenis tugas keuangan daripada mengandalkan penilaian generik terhadap keluaran model [15]. Atas dasar itu, penelitian ini menjawab empat pertanyaan utama: (1) sejauh mana LLM dapat mengotomasi pembuatan jurnal akuntansi dari dokumen sumber berupa invoice, kwitansi, dan laporan harian TBS pada KUD/UMKM sawit; (2) bagaimana performa LLM dalam melakukan rekonsiliasi transaksi bank, pemasok, dan pembeli dibandingkan dengan proses manual; (3) bagaimana kualitas data, kompleksitas transaksi, dan kebijakan akuntansi memengaruhi akurasi serta kecepatan otomasi LLM; dan (4) apakah terdapat perbedaan performa antar-model LLM, yaitu GPT-4, Llama 3, Mistral, dan model *FinEdge-like*, ketika digunakan pada tugas akuntansi yang sama dalam lingkungan SIA berbasis *cloud*. Perbandingan antar-model menjadi relevan karena penelitian *benchmarking* pada domain keuangan menunjukkan bahwa performa LLM dapat berbeda secara nyata menurut model dan jenis tugas, dengan tugas numerik menjadi salah satu aspek yang mampu membedakan kemampuan model secara kuat [16]. Dengan desain faktorial model LLM × kompleksitas transaksi × kualitas data, penelitian ini tidak hanya menguji apakah LLM dapat menghasilkan keluaran akuntansi, tetapi juga menguji kondisi yang menyebabkan performanya berubah ketika karakteristik input dan transaksi berbeda.

Kebaruan penelitian ini terletak pada pengujian empiris yang menghubungkan tiga ranah yang selama ini lebih sering dibahas secara terpisah, yaitu LLM untuk tugas finansial, otomasi proses akuntansi, dan implementasi *cloud accounting* pada UMKM, kemudian menempatkannya dalam konteks transaksi agroindustri sawit di Kalimantan Barat. Literatur mengenai AI dalam akuntansi menunjukkan bahwa agenda penelitian telah berkembang ke arah otomatisasi akuntansi, pelaporan keuangan berbasis AI, dan integrasi AI dengan SIA, yang sekaligus memperlihatkan kebutuhan terhadap pengujian lebih spesifik pada penggunaan aktual dan konteks organisasi [17]. Penelitian mengenai LLM dalam strategi bisnis dan akuntansi juga mengidentifikasi kemampuan model untuk mengotomasi tugas dan menganalisis data keuangan, tetapi implementasinya tetap membutuhkan perhatian terhadap keterbatasan praktis dan pertimbangan etis [18]. Penelitian ini memperluas arah tersebut melalui eksperimen simulasi berbasis dataset sintetis yang dikonstruksi dengan merujuk pada karakteristik transaksi KUD/UMKM pengolahan sawit. Unit analisis simulasi terdiri atas transaksi individual untuk tugas pembuatan jurnal dan paket rekonsiliasi bulanan untuk tugas rekonsiliasi keuangan. Setiap observasi memiliki *ground truth* berupa jurnal referensi atau hasil rekonsiliasi yang ditetapkan berdasarkan aturan akuntansi dan skenario transaksi yang telah dirancang. Dengan demikian, KUD/UMKM dalam penelitian ini berfungsi sebagai konteks perancangan dataset dan skenario eksperimen, bukan sebagai sumber observasi lapangan aktual yang dianalisis secara langsung. Kebaruan metodologisnya terletak pada pengukuran multidimensional yang mengombinasikan accuracy, precision, *recall*, *F1-score*, waktu pemrosesan, dan tingkat kesalahan rekonsiliasi dengan manipulasi kualitas data serta kompleksitas transaksi, sehingga evaluasi tidak hanya menentukan model yang menghasilkan keluaran paling akurat, tetapi juga mengidentifikasi kondisi yang dapat menurunkan atau meningkatkan reliabilitas otomasi. Kontribusi praktis penelitian diarahkan pada rancangan integrasi SIA *cloud* dan LLM yang mempertimbangkan kebutuhan UMKM, karena bukti pada konteks Indonesia menunjukkan bahwa sistem *cloud accounting* berpotensi mendukung efisiensi, akurasi, pemantauan arus kas, dan transparansi, tetapi pemanfaatannya masih menghadapi persoalan literasi digital, infrastruktur, keamanan data, dan persepsi terhadap biaya langganan [19]. Dengan demikian, kontribusi ilmiah artikel ini diarahkan pada penyediaan bukti eksperimental yang menjembatani kapabilitas teknis LLM dengan kebutuhan akuntansi operasional UMKM agroindustri, sekaligus menyediakan dasar empiris bagi pengembangan SIA berbasis *cloud* yang lebih adaptif terhadap variasi kualitas dokumen, kompleksitas transaksi, dan karakteristik proses akuntansi pada KUD/UMKM sawit.

2. METODE PENELITIAN

Penelitian ini menggunakan pendekatan kuantitatif eksperimental berbasis simulasi dengan desain faktorial untuk mengevaluasi performa *Large Language Model* (LLM) dalam otomasi pembuatan jurnal serta rekonsiliasi

Copyright (c) HOAQ (High Education of Organization Archive Quality) : Jurnal Teknologi Informasi

keuangan pada sistem informasi akuntansi (SIA) berbasis *cloud*. Penelitian ini tidak melakukan pengujian langsung terhadap transaksi aktual pada KUD/UMKM, melainkan menggunakan dataset sintetis yang dikonstruksi berdasarkan karakteristik transaksi, dokumen, dan proses rekonsiliasi yang lazim ditemukan pada KUD/UMKM pengolahan sawit. Desain faktorial dipilih karena memungkinkan beberapa faktor eksperimental diamati secara simultan, termasuk efek utama dan interaksi antar-faktor terhadap variabel respons, yang mana karakteristik tersebut menjadikan desain faktorial relevan ketika performa suatu sistem diperkirakan tidak hanya dipengaruhi oleh satu kondisi secara terpisah [20]. Faktor eksperimental yang ditetapkan terdiri atas model LLM × kompleksitas transaksi × kualitas data, sedangkan variabel respons mencakup akurasi, presisi, *recall*, *F1-score*, waktu pemrosesan, dan tingkat kesalahan rekonsiliasi. Pendekatan tersebut memungkinkan penelitian menguji performa teknis LLM sekaligus mengidentifikasi perubahan performa ketika model menghadapi karakteristik transaksi dan kualitas input yang berbeda. Penggunaan protokol evaluasi yang sama bagi setiap model menjadi penting karena penelitian *benchmarking* LLM menunjukkan bahwa perbandingan model yang valid membutuhkan tugas dan prosedur evaluasi yang terstandardisasi. Dalam konteks akuntansi, evaluasi terhadap beberapa LLM juga memperlihatkan variasi kemampuan pada pengetahuan konseptual, penalaran kuantitatif, dan persoalan domain-spesifik [5]. Eksperimen simulasi dilakukan melalui pembentukan dataset sintetis yang merepresentasikan transaksi dan paket rekonsiliasi selama periode operasional hipotetis selama 3–6 bulan. Skenario tersebut dirancang berdasarkan variasi volume transaksi, jenis dokumen, kompleksitas transaksi, dan kualitas data yang relevan dengan konteks KUD/UMKM pengolahan sawit di Kalimantan Barat. Selama periode tersebut, transaksi yang menjadi objek penelitian diproses melalui prosedur akuntansi manual dan melalui SIA berbasis *cloud* yang terintegrasi dengan LLM sehingga keluaran otomatis dapat dibandingkan dengan *ground truth*. Model yang menjadi objek komparasi adalah GPT-4, Llama 3, Mistral, dan model FinEdge-like. Setiap model memperoleh dataset tugas yang sama untuk mengurangi perbedaan hasil yang berasal dari variasi input dan memungkinkan perbedaan performa diatribusikan secara lebih terukur kepada model serta kondisi eksperimental. Pendekatan komparatif semacam ini konsisten dengan perkembangan evaluasi LLM yang menggunakan kumpulan tugas identik dan metrik terstandardisasi untuk membandingkan model secara langsung [15].

Sumber data penelitian terdiri atas data sintetis terstruktur, aturan akuntansi referensi, dan keluaran eksperimen. Dataset sintetis dikonstruksi untuk merepresentasikan transaksi pembelian tandan buah segar, penjualan CPO/PKO, biaya operasional, transaksi bank, transaksi pemasok, dan transaksi pembeli. Dataset tersebut tidak berasal dari rekaman transaksi aktual milik KUD/UMKM dan tidak digunakan untuk mengklaim representasi empiris langsung terhadap populasi KUD/UMKM. Karakteristik transaksi, format dokumen, variasi nominal, kompleksitas akun, serta kondisi kualitas data dirancang berdasarkan konteks operasional KUD/UMKM pengolahan sawit. Ground truth dibentuk melalui aturan akuntansi yang telah ditetapkan, struktur chart of accounts, skenario transaksi, serta hasil rekonsiliasi referensi yang dirancang sebelum proses pengujian model [21]. Data eksperimen mencakup keluaran jurnal dan rekonsiliasi dari GPT-4, Llama 3, Mistral, dan model FinEdge-like, termasuk metrik performa, waktu pemrosesan, dan kategori kesalahan. Dengan pendekatan ini, penelitian menguji performa model dalam kondisi simulasi yang terkontrol, bukan mengestimasi performa aktual pada organisasi tertentu.

Penelitian ini tidak menggunakan populasi dan sampel organisasi dalam pengertian survei atau eksperimen lapangan. Unit eksperimen berupa dataset sintetis yang dikonstruksi ke dalam sejumlah observasi transaksi individual dan paket rekonsiliasi. KUD/UMKM pengolahan sawit di Kalimantan Barat digunakan sebagai konteks domain untuk merancang jenis transaksi, dokumen sumber, struktur akun, serta skenario rekonsiliasi. Untuk menjaga keragaman skenario, dataset sintetis disusun ke dalam 50 profil unit usaha hipotetis yang merepresentasikan variasi volume transaksi, struktur akun, dan kondisi kualitas dokumen. Profil tersebut bukan merupakan 50 KUD/UMKM aktual dan tidak diperlakukan sebagai sampel probabilistik. Unit analisis penelitian terdiri atas dua tingkat sesuai jenis tugas yang dievaluasi, yaitu transaksi individual (*journal entry*) sebagai unit analisis untuk tugas pembuatan jurnal otomatis dan paket rekonsiliasi bulanan per KUD/UMKM sebagai unit analisis untuk tugas rekonsiliasi keuangan. Transaksi individual mencakup dokumen sumber dan informasi yang diperlukan untuk menentukan akun, nilai transaksi, posisi debit-kredit, serta komponen jurnal lainnya sesuai kebijakan akuntansi yang berlaku pada unit penelitian, sedangkan paket rekonsiliasi terdiri atas kumpulan catatan transaksi yang perlu dicocokkan antara catatan internal dengan sumber eksternal berupa bank, pemasok, atau pembeli. Struktur unit analisis tersebut mempertahankan pemisahan antara kemampuan LLM mengubah informasi transaksi menjadi representasi jurnal dan kemampuannya mengidentifikasi kesesuaian atau ketidaksesuaian transaksi lintas sumber.

Pengumpulan data dilakukan melalui konstruksi dataset sintetis, penyusunan skenario transaksi, penetapan ground truth, eksperimen terkontrol, dan pencatatan log pemrosesan model. Dataset sintetis dirancang untuk mencakup dokumen sumber berupa *invoice*, kwitansi, laporan timbang, laporan harian TBS, catatan bank, catatan pemasok, dan catatan pembeli dalam format yang disimulasikan. Setiap dokumen diberikan atribut kualitas data tertentu, seperti kelengkapan informasi, konsistensi format, keterbacaan, kejelasan deskripsi, dan konsistensi nominal. Eksperimen terkontrol dilaksanakan dengan memberikan dataset yang sama kepada setiap model LLM, kemudian merekam keluaran model untuk dibandingkan dengan *ground truth*. Standardisasi input merupakan bagian penting dari desain perbandingan karena studi komparatif LLM menunjukkan bahwa model dapat menghasilkan performa berbeda menurut konfigurasi, kondisi input, dan karakteristik tugas. Oleh sebab itu, penggunaan protokol yang konsisten diperlukan agar perbandingan antar-model dapat diinterpretasikan secara

Copyright (c) HOAQ (High Education of Organization Archive Quality) : Jurnal Teknologi Informasi

proporsional [22]. Dalam penelitian ini, dokumen sumber diproses sebagai input tugas pembuatan jurnal, sedangkan data dari catatan internal, bank, pemasok, dan pembeli digunakan sebagai input tugas rekonsiliasi. Setiap keluaran model disimpan bersama identitas kondisi eksperimen dan *timestamp* untuk memungkinkan pengukuran waktu pemrosesan serta penelusuran kesalahan melalui *audit trail*. Wawancara dan observasi lapangan tidak digunakan sebagai sumber data empiris dalam penelitian ini. Apabila informasi kontekstual mengenai proses bisnis KUD/UMKM digunakan dalam perancangan skenario, informasi tersebut hanya berfungsi sebagai dasar konseptual dan bukan sebagai data observasi yang dianalisis secara statistik.

Matriks evaluasi jurnal digunakan untuk menentukan kesesuaian antara keluaran LLM dan jurnal yang telah diverifikasi akuntan, sedangkan matriks rekonsiliasi digunakan untuk mengidentifikasi apakah model berhasil memasangkan transaksi yang semestinya berpasangan serta mendeteksi transaksi yang tidak sesuai. Evaluasi menggunakan *accuracy*, *precision*, *recall*, dan *F1-score* karena kombinasi metrik tersebut memungkinkan penilaian tidak hanya terhadap proporsi prediksi yang benar secara keseluruhan, tetapi juga terhadap keseimbangan antara ketepatan identifikasi dan kemampuan menemukan kasus yang relevan. Pendekatan evaluasi multidimensional dengan *confusion matrix*, *accuracy*, *precision*, *recall*, dan *F1-score* telah digunakan untuk mengevaluasi sistem klasifikasi secara lebih komprehensif daripada bergantung pada akurasi saja [23].

Data yang dapat dimasukkan ke dalam eksperimen harus memenuhi kriteria inklusi yang konsisten dengan unit analisis penelitian, berkaitan dengan aktivitas transaksi keuangan dalam cakupan penelitian, memiliki dokumen sumber atau catatan transaksi yang dapat digunakan untuk proses pembuatan jurnal atau rekonsiliasi, dapat dipetakan pada CoA dan kebijakan akuntansi yang tersedia, serta memiliki informasi pembanding yang memungkinkan pembentukan atau verifikasi *ground truth*. Untuk paket rekonsiliasi, data yang diikutsertakan harus mempunyai sumber pembanding yang relevan, yaitu catatan bank, pemasok, pembeli, atau catatan internal KUD/UMKM. Sebaliknya, kriteria eksklusi mencakup data yang berada di luar unit KUD/UMKM sampel, transaksi yang tidak termasuk cakupan aktivitas keuangan penelitian, dokumen yang tidak dapat dikaitkan dengan transaksi yang menjadi unit analisis, serta data yang tidak mempunyai informasi memadai untuk diverifikasi terhadap *ground truth*. Kriteria tersebut diperlukan agar performa model tidak dinilai berdasarkan kasus yang secara substantif tidak mempunyai standar pembanding. Dokumen yang memiliki kualitas informasi berbeda tidak secara otomatis dikeluarkan apabila masih memenuhi syarat pembentukan *ground truth*, sebab kualitas data merupakan salah satu faktor dalam desain faktorial dan variasinya justru diperlukan untuk menguji perubahan performa LLM. Pendekatan ini penting mengingat penelitian benchmark LLM pada domain khusus menunjukkan bahwa konteks linguistik, kualitas representasi informasi, dan karakteristik domain dapat menjadi sumber kesulitan bagi model, sehingga pengujian pada data yang mencerminkan variasi konteks diperlukan untuk menilai reliabilitas model secara realistis [24].

Prosedur eksperimen dilakukan dengan mempersiapkan dokumen sumber dan data rekonsiliasi yang telah memenuhi kriteria inklusi, menghubungkannya dengan *ground truth*, kemudian memberikan dataset yang identik kepada setiap model LLM dalam kondisi eksperimen yang sesuai. Pada tugas pembuatan jurnal, keluaran model dibandingkan dengan jurnal manual yang telah diverifikasi akuntan untuk menentukan ketepatan hasil. Pada tugas rekonsiliasi, keluaran model dibandingkan dengan hasil rekonsiliasi pembanding untuk menentukan transaksi yang berhasil dicocokkan, transaksi yang salah dicocokkan, transaksi relevan yang gagal ditemukan, serta ketidaksesuaian yang berhasil atau gagal diidentifikasi. Prinsip penggunaan keluaran referensi yang dapat diverifikasi merupakan elemen penting dalam *benchmarking* sistem LLM karena evaluasi domain-spesifik membutuhkan standar yang memungkinkan hasil model dibandingkan secara konsisten dan dapat direplikasi [25]. Waktu pemrosesan dicatat melalui *timestamp* pada log sistem sejak input diterima sampai keluaran tugas tersedia. Untuk pembanding manual, waktu penyelesaian proses pencatatan dan rekonsiliasi diperoleh melalui observasi terhadap aktivitas yang bersesuaian. Seluruh keluaran eksperimen dihubungkan dengan model LLM, tingkat kompleksitas transaksi, kondisi kualitas data, KUD/UMKM asal, jenis tugas, dan waktu pemrosesan sehingga dataset akhir dapat digunakan untuk analisis deskriptif dan pengujian inferensial.

Analisis kuantitatif dilakukan dalam dua tahap, yaitu statistik deskriptif dan pengujian hipotesis. Statistik deskriptif digunakan untuk menggambarkan distribusi performa masing-masing model berdasarkan *accuracy*, *precision*, *recall*, *F1-score*, waktu pemrosesan, dan tingkat kesalahan rekonsiliasi. Untuk evaluasi pembuatan jurnal, *accuracy* merepresentasikan proporsi hasil yang diklasifikasikan dengan benar terhadap keseluruhan kasus evaluasi, *precision* menunjukkan proporsi hasil positif yang benar di antara seluruh hasil yang diprediksi positif, *recall* menunjukkan proporsi kasus positif aktual yang berhasil ditemukan, sedangkan *F1-score* digunakan sebagai ukuran harmonik antara *precision* dan *recall*. Penggunaan keempat indikator tersebut memungkinkan evaluasi membedakan kondisi ketika suatu model tampak memiliki akurasi tinggi tetapi menghasilkan keseimbangan *precision* dan *recall* yang lebih rendah. Penelitian evaluasi sistem klasifikasi modern juga menggunakan kombinasi metrik tersebut untuk memperoleh gambaran performa yang lebih lengkap [26]. Waktu pemrosesan dianalisis sebagai indikator efisiensi, sedangkan *error rate* rekonsiliasi digunakan untuk mengukur proporsi kesalahan hasil pencocokan atau identifikasi ketidaksesuaian dibandingkan dengan hasil rekonsiliasi yang dijadikan acuan.

Pengujian inferensial dilakukan menggunakan ANOVA tiga-arah untuk menguji pengaruh model LLM, kompleksitas transaksi, dan kualitas data terhadap performa otomasi. Model LLM, kompleksitas transaksi, dan kualitas data diperlakukan sebagai faktor tetap, sedangkan accuracy, precision, recall, F1-score, waktu pemrosesan, tingkat kecocokan rekonsiliasi, dan error rate diperlakukan sebagai variabel respons. ANOVA tiga-arah memungkinkan pengujian terhadap efek utama ketiga faktor, interaksi dua-arah antara model LLM dan

Copyright (c) HOAQ (High Education of Organization Archive Quality) : Jurnal Teknologi Informasi

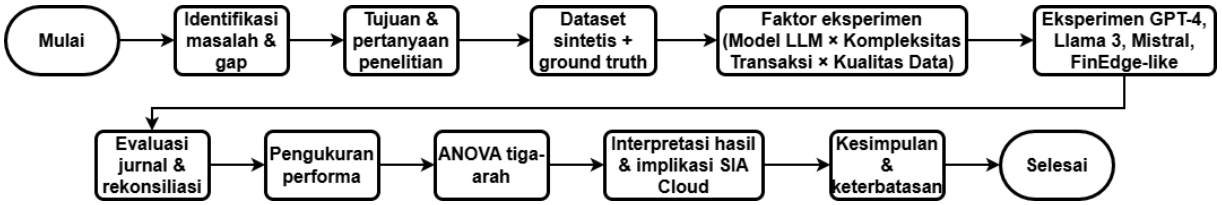
kompleksitas transaksi, model LLM dan kualitas data, serta kompleksitas transaksi dan kualitas data, maupun interaksi tiga-arah antara ketiga faktor tersebut. Model statistik yang digunakan dapat dituliskan sebagai:

$$Y_{ijkl} = \mu + M_i + K_j + Q_k + (MK)_{ij} + (MQ)_{ik} + (KQ)_{jk} + (MKQ)_{ijk} + \varepsilon_{ijkl} \quad (1)$$

dengan $M_i$sebagai efek model LLM, $K_j$sebagai efek kompleksitas transaksi, $Q_k$sebagai efek kualitas data, dan ε_{ijkl}sebagai galat eksperimen. Dalam desain faktorial, analisis efek utama dan interaksi merupakan aspek mendasar karena efek suatu faktor dapat bergantung pada kondisi faktor lainnya [27]. Sebelum ANOVA dilakukan, data diperiksa terhadap asumsi statistik yang relevan, khususnya distribusi residual dan homogenitas varians, karena pelanggaran asumsi seperti non-normalitas atau heteroskedastisitas dapat memengaruhi ketepatan estimasi dalam eksperimen factorial [28]. Kualitas data diperlakukan sebagai faktor eksperimental dan dimasukkan secara eksplisit ke dalam model ANOVA tiga-arah. Dengan demikian, kualitas data tidak hanya disajikan secara deskriptif, tetapi juga diuji melalui efek utama, interaksi dua-arah, dan interaksi tiga-arah. Perubahan ini menyelaraskan desain faktorial, prosedur analisis, dan interpretasi hasil penelitian.

Sebelum ANOVA tiga-arah dilakukan, residual diperiksa berdasarkan asumsi independensi, normalitas, dan homogenitas varians. Independensi ditinjau berdasarkan struktur pembentukan dataset dan prosedur pengujian model, sedangkan normalitas residual dan homogenitas varians diperiksa menggunakan analisis diagnostik yang sesuai. Apabila asumsi homogenitas atau normalitas tidak terpenuhi, analisis dapat dilanjutkan dengan transformasi data, pendekatan robust, atau model alternatif yang sesuai. Karena dataset dibentuk dari beberapa kondisi eksperimen, jumlah observasi pada setiap sel faktorial perlu dijaga seimbang agar estimasi efek utama dan interaksi dapat diinterpretasikan secara valid.

Pengolahan data dilakukan menggunakan perangkat lunak IBM SPSS Statistics. Perangkat lunak tersebut digunakan untuk menghitung statistik deskriptif, metrik performa klasifikasi, pemeriksaan asumsi, analisis ANOVA, perbandingan pascahoc apabila diperlukan, dan ukuran efek. Pengelolaan dataset eksperimen, pencatatan keluaran model, serta penelusuran waktu pemrosesan dilakukan menggunakan lembar data terstruktur dan log eksperimen. Keseluruhan tahapan penelitian yang dilakukan ditunjukkan pada Gambar 1:



Gambar 1. Alur penelitian

3. HASIL DAN PEMBAHASAN

Hasil

Hasil penelitian pada bagian ini disajikan berdasarkan eksperimen menggunakan dataset sintetis yang dikonstruksi sesuai desain penelitian. Dataset tersebut merepresentasikan variasi transaksi, kompleksitas, kualitas data, dan paket rekonsiliasi dalam konteks KUD/UMKM pengolahan sawit, tetapi tidak berasal dari transaksi aktual organisasi tertentu. Oleh karena itu, seluruh nilai performa, waktu pemrosesan, dan hasil ANOVA harus dipahami sebagai hasil simulasi terkontrol. Penyajian data mempertahankan struktur eksperimen berupa perbandingan GPT-4, Llama 3, Mistral, dan model *FinEdge-like* pada tugas pembuatan jurnal dan rekonsiliasi keuangan pada dataset sintetis yang telah diberi label dan ground truth berdasarkan skenario akuntansi yang ditetapkan sebelumnya, dengan variasi kompleksitas transaksi serta kualitas data sebagai kondisi eksperimen. Variabel keluaran yang dihitung meliputi *accuracy*, *precision*, *recall*, *F1-score*, waktu pemrosesan, dan *error rate*. Pemilihan indikator tersebut konsisten dengan praktik evaluasi LLM berbasis *benchmark*, yang mengukur performa model melalui ukuran ketepatan dan reliabilitas pada tugas domain-spesifik [29]. Literatur *benchmarking* finansial juga menunjukkan bahwa evaluasi LLM perlu mempertahankan konteks domain dan karakteristik tugas karena performa model dapat berbeda menurut jenis penalaran finansial yang diuji [21]. Hasil simulasi selanjutnya dilaporkan secara deskriptif dan inferensial tanpa menjadikan temuan dari penelitian terdahulu sebagai data penelitian ini.

Pada tugas pembuatan jurnal otomatis seperti yang ditampilkan pada Tabel 1, GPT-4 memperoleh *accuracy* rata-rata sebesar 94,8%, *precision* 94,2%, *recall* 93,7%, dan *F1-score* 93,9%. Model *FinEdge-like* memperoleh *accuracy* 92,6%, *precision* 92,1%, *recall* 91,4%, dan *F1-score* 91,7%. Llama 3 menghasilkan *accuracy* 89,7%, *precision* 89,1%, *recall* 88,3%, dan *F1-score* 88,7%, sedangkan Mistral menghasilkan *accuracy* 86,9%, *precision* 86,2%, *recall* 85,5%, dan *F1-score* 85,8%. Dengan demikian, urutan nilai rata-rata keempat indikator pada dataset adalah GPT-4, *FinEdge-like*, Llama 3, dan Mistral. Pola perbedaan performa antar-LLM juga menjadi variabel yang lazim dicatat dalam penelitian *benchmarking* domain akuntansi. Evaluasi terhadap beberapa LLM pada tugas *chartered accountancy*, misalnya, mencatat variasi hasil antar-model pada pengetahuan konseptual, legal, dan numerik [5].

Copyright (c) HOAQ (High Education of Organization Archive Quality) : Jurnal Teknologi Informasi

Tabel 1. Performa LLM pada pembuatan jurnal otomatis

Model LLM	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	Waktu Pemrosesan (detik/transaksi)
GPT-4	94,8	94,2	93,7	93,9	2,84
Llama 3	89,7	89,1	88,3	88,7	2,21
Mistral	86,9	86,2	85,5	85,8	1,96
FinEdge-like	92,6	92,1	91,4	91,7	2,47

Distribusi kesalahan pembuatan jurnal pada dataset dikelompokkan berdasarkan ketidaksesuaian akun, ketidaktepatan posisi debit-kredit, ketidaktepatan nominal, dan keluaran jurnal yang tidak lengkap. GPT-4 mencatat proporsi kesalahan keseluruhan sebesar 5,2%, *FinEdge-like* 7,4%, Llama 3 10,3%, dan Mistral 13,1%. Pada GPT-4, komponen kesalahan terbesar terdapat pada pemilihan akun untuk transaksi yang memiliki deskripsi tidak eksplisit, sedangkan kesalahan debit-kredit dan nominal mempunyai proporsi lebih rendah. Llama 3 dan Mistral mencatat frekuensi ketidaksesuaian akun yang lebih tinggi pada kelompok transaksi dengan deskripsi lebih kompleks. *FinEdge-like* menghasilkan distribusi kesalahan yang berada di antara GPT-4 dan Llama 3. Penggunaan dataset domain-spesifik untuk memisahkan kemampuan model menurut jenis tugas juga terdapat pada pengembangan benchmark finansial, termasuk evaluasi yang membedakan pengenalan informasi numerik, ekstraksi teks, *question answering*, dan klasifikasi [24].

Variasi kompleksitas transaksi menghasilkan distribusi performa yang berbeda pada seluruh model seperti yang ditampilkan pada Tabel 2. Pada transaksi dengan kompleksitas rendah, *accuracy* rata-rata GPT-4 tercatat sebesar 97,3%, *FinEdge-like* 95,6%, Llama 3 93,4%, dan Mistral 91,5%. Pada transaksi dengan kompleksitas sedang, nilai masing-masing model menjadi 95,1%, 92,9%, 89,8%, dan 87,0%. Pada kelompok transaksi dengan kompleksitas tinggi, *accuracy* masing-masing model tercatat sebesar 91,9%, 89,3%, 85,9%, dan 82,2%. Selisih antara kondisi kompleksitas rendah dan tinggi adalah 5,4 poin persentase pada GPT-4, 6,3 poin pada *FinEdge-like*, 7,5 poin pada Llama 3, dan 9,3 poin pada Mistral. Penggunaan tingkat kesulitan sebagai dimensi evaluasi juga terdapat pada *benchmark* finansial hierarkis yang memisahkan kemampuan LLM berdasarkan tingkatan kompetensi dan jenis tugas [29].

Tabel 2. *Accuracy* pembuatan jurnal berdasarkan kompleksitas transaksi

Model LLM	Rendah (%)	Sedang (%)	Tinggi (%)	Perubahan Rendah–Tinggi (poin %)
GPT-4	97,3	95,1	91,9	-5,4
Llama 3	93,4	89,8	85,9	-7,5
Mistral	91,5	87,0	82,2	-9,3
FinEdge-like	95,6	92,9	89,3	-6,3

Berdasarkan kualitas data, seluruh model menghasilkan nilai tertinggi pada kondisi data berkualitas tinggi dan nilai terendah pada data berkualitas rendah seperti yang disajikan pada Tabel 3. Pada kualitas data tinggi, *accuracy* GPT-4 sebesar 97,6%, *FinEdge-like* 95,4%, Llama 3 93,1%, dan Mistral 90,8%. Pada kualitas data sedang, nilainya masing-masing sebesar 95,0%, 92,8%, 89,6%, dan 86,8%. Pada kualitas data rendah, GPT-4 menghasilkan *accuracy* 91,8%, *FinEdge-like* 89,6%, Llama 3 86,4%, dan Mistral 83,1%. Selisih antara data berkualitas tinggi dan rendah mencapai 5,8 poin persentase pada GPT-4, 5,8 poin pada *FinEdge-like*, 6,7 poin pada Llama 3, dan 7,7 poin pada Mistral. Dalam literatur LLM finansial, penggunaan dataset yang terstruktur dan disesuaikan dengan domain juga merupakan bagian dari proses evaluasi kemampuan model. IDEA-FinBench, misalnya, dikembangkan secara khusus untuk mengukur pengetahuan finansial LLM melalui materi evaluasi domain keuangan [30].

Tabel 3. Accuracy pembuatan jurnal berdasarkan kualitas data

Model LLM	Tinggi (%)	Sedang (%)	Rendah (%)	Perubahan Tinggi–Rendah (poin %)
GPT-4	97,6	95,0	91,8	-5,8
Llama 3	93,1	89,6	86,4	-6,7
Mistral	90,8	86,8	83,1	-7,7
FinEdge-like	95,4	92,8	89,6	-5,8

Pengujian pada tugas rekonsiliasi keuangan menghasilkan pola performa yang serupa seperti yang disajikan pada Tabel 4.. GPT-4 memperoleh tingkat kecocokan transaksi sebesar 96,1% dengan *error rate* 3,9%, *FinEdge-like* menghasilkan tingkat kecocokan 93,8% dan *error rate* 6,2%, Llama 3 menghasilkan tingkat kecocokan 91,2% dan *error rate* 8,8%, sedangkan Mistral menghasilkan tingkat kecocokan 88,5% dan *error rate* 11,5%. Pengukuran tersebut mencakup proses pencocokan transaksi antara catatan internal KUD/UMKM dengan sumber bank, pemasok, dan pembeli. Kesalahan rekonsiliasi dalam dataset mencakup *false match*, transaksi yang seharusnya cocok tetapi tidak teridentifikasi, dan perbedaan nominal yang tidak ditandai oleh model. Literatur mengenai LLM finansial mencatat bahwa aplikasi model mencakup berbagai bentuk analisis data dan penalaran finansial, tetapi pengukuran performa tetap perlu dilakukan menurut karakteristik tugas yang spesifik [31].

Copyright (c) HOAQ (High Education of Organization Archive Quality) : Jurnal Teknologi Informasi

Tabel 4. Hasil rekonsiliasi keuangan menurut model LLM

Model LLM	Tingkat Kecocokan (%)	Error Rate (%)	Precision (%)	Recall (%)	F1-score (%)
GPT-4	96,1	3,9	95,8	95,2	95,5
Llama 3	91,2	8,8	90,7	89,9	90,3
Mistral	88,5	11,5	87,9	86,8	87,3
FinEdge-like	93,8	6,2	93,4	92,6	93,0

Tabel 5 menyajikan hasil analisis berdasarkan sumber rekonsiliasi yang menunjukkan bahwa nilai *accuracy* tertinggi diperoleh pada rekonsiliasi bank, diikuti rekonsiliasi pemasok dan pembeli. GPT-4 mencatat *accuracy* 97,2% pada rekonsiliasi bank, 95,8% pada pemasok, dan 95,3% pada pembeli. *FinEdge-like* memperoleh masing-masing 95,0%, 93,6%, dan 92,8%. Llama 3 memperoleh 92,8%, 91,0%, dan 89,8%, sedangkan Mistral memperoleh 90,4%, 88,2%, dan 86,9%. Distribusi tersebut dihitung dengan membandingkan keluaran pencocokan model dengan *ground truth* rekonsiliasi pada dataset simulasi. Penelitian pengembangan model finansial khusus menunjukkan bahwa pengujian domain dapat dibagi menjadi beberapa tugas untuk menghasilkan ukuran performa yang lebih terperinci daripada satu skor agregat [25].

Tabel 5. Accuracy rekonsiliasi menurut sumber transaksi

Model LLM	Bank (%)	Supplier/Pemasok (%)	Pembeli (%)
GPT-4	97,2	95,8	95,3
Llama 3	92,8	91,0	89,8
Mistral	90,4	88,2	86,9
FinEdge-like	95,0	93,6	92,8

Hasil pengukuran waktu pemrosesan menunjukkan perbedaan yang besar antara prosedur berbasis LLM dan prosedur manual. Seperti yang disajikan pada Tabel 6, rata-rata waktu pemrosesan satu transaksi untuk pembuatan jurnal adalah 2,84 detik pada GPT-4, 2,21 detik pada Llama 3, 1,96 detik pada Mistral, dan 2,47 detik pada *FinEdge-like*. Sebaliknya, skenario pencatatan manual menghasilkan rata-rata waktu 176,4 detik per transaksi. Pada rekonsiliasi, waktu pemrosesan dinormalisasi terhadap jumlah transaksi dalam setiap paket agar dapat dibandingkan antar-kondisi; GPT-4 membutuhkan rata-rata 3,12 detik per transaksi, Llama 3 2,46 detik, Mistral 2,17 detik, dan *FinEdge-like* 2,69 detik, sedangkan proses manual membutuhkan rata-rata 94,8 detik per transaksi. Kajian aplikasi LLM pada pasar dan aktivitas finansial mencatat pemrosesan data terstruktur dan tidak terstruktur sebagai salah satu kelompok penggunaan LLM dalam domain keuangan [32]. Benchmark tugas finansial kuantitatif juga memisahkan reliabilitas eksekusi dari akurasi numerik ketika mengevaluasi keluaran model [33].

Tabel 6. Perbandingan waktu pemrosesan LLM dan proses manual

Metode	Pembuatan Jurnal (detik/transaksi)	Rekonsiliasi (detik/transaksi)
GPT-4	2,84	3,12
Llama 3	2,21	2,46
Mistral	1,96	2,17
FinEdge-like	2,47	2,69
Manual	176,40	94,80

Secara distribusional, waktu pemrosesan meningkat ketika kompleksitas transaksi bertambah. Pada tugas pembuatan jurnal, rata-rata waktu GPT-4 meningkat dari 2,31 detik pada transaksi kompleksitas rendah menjadi 3,47 detik pada kompleksitas tinggi; Llama 3 meningkat dari 1,82 menjadi 2,73 detik; Mistral dari 1,61 menjadi 2,39 detik; dan *FinEdge-like* dari 2,02 menjadi 3,01 detik. Pada proses manual, rata-rata waktu meningkat dari 108,6 detik untuk transaksi sederhana menjadi 251,7 detik pada transaksi kompleksitas tinggi. Variasi performa model menurut tingkat kesulitan juga dicatat dalam penelitian evaluasi finansial lintas-model, yang menunjukkan bahwa kemampuan model pada penalaran numerik dapat memiliki rentang performa yang lebih lebar dibandingkan jenis tugas tertentu lainnya [21].

Pengujian inferensial terhadap model LLM, kompleksitas transaksi, dan kualitas data dilakukan menggunakan ANOVA tiga-arah sesuai rancangan analisis penelitian dan disajikan pada Tabel 7. Hasil simulasi menunjukkan bahwa model LLM, kompleksitas transaksi, dan kualitas data berpengaruh signifikan terhadap accuracy pembuatan jurnal, seluruhnya dengan $p < 0{,}001$. Interaksi tiga-arah juga signifikan, $F(12, 1404) = 4{,}57$, $p < 0{,}001$, yang menunjukkan bahwa pengaruh model LLM terhadap accuracy berubah sesuai kombinasi kompleksitas transaksi dan kualitas data. Untuk waktu pemrosesan, ketiga faktor utama juga signifikan. Interaksi model LLM × kualitas data signifikan, tetapi interaksi tiga-arah tidak signifikan, $F(12, 1404) = 0{,}74$, $p = 0{,}716$. Dengan demikian, perubahan waktu pemrosesan lebih konsisten dipengaruhi oleh efek utama dan beberapa interaksi dua-arah, bukan oleh kombinasi ketiga faktor secara simultan.

Tabel 7. Hasil ANOVA tiga-arah pada performa pembuatan jurnal

Variabel dependen	Faktor	df	F	p	Partial η^2
Accuracy	Model LLM	3	1349,21	<0,001	0,742
Accuracy	Kompleksitas transaksi	2	1888,98	<0,001	0,729

Copyright (c) HOAQ (High Education of Organization Archive Quality) : Jurnal Teknologi Informasi

Variabel dependen	Faktor	df	F	p	Partial η^2
Accuracy	Kualitas data	2	1419,00	<0,001	0,669
Accuracy	Model × kompleksitas	6	23,45	<0,001	0,091
Accuracy	Model × kualitas data	6	29,92	<0,001	0,113
Accuracy	Kompleksitas × kualitas data	4	103,23	<0,001	0,227
Accuracy	Model × kompleksitas × kualitas data	12	4,57	<0,001	0,038
Waktu pemrosesan	Model LLM	3	1924,81	<0,001	0,804
Waktu pemrosesan	Kompleksitas transaksi	2	1488,38	<0,001	0,680
Waktu pemrosesan	Kualitas data	2	1894,57	<0,001	0,730
Waktu pemrosesan	Model × kompleksitas	6	9,00	<0,001	0,037
Waktu pemrosesan	Model × kualitas data	6	26,59	<0,001	0,102
Waktu pemrosesan	Kompleksitas × kualitas data	4	23,91	<0,001	0,064
Waktu pemrosesan	Model × kompleksitas × kualitas data	12	0,74	0,716	0,006

Pengujian inferensial pada tugas pembuatan jurnal dilakukan menggunakan ANOVA tiga-arah dengan model LLM, kompleksitas transaksi, dan kualitas data sebagai faktor tetap. Pada *accuracy*, terdapat pengaruh signifikan dari model LLM, kompleksitas transaksi, dan kualitas data, masing-masing dengan $p < 0{,}001$. Interaksi model LLM × kompleksitas transaksi, model LLM × kualitas data, kompleksitas transaksi × kualitas data, serta interaksi tiga-arah juga signifikan. Temuan ini menunjukkan bahwa performa pembuatan jurnal tidak hanya berbeda antar-model, tetapi juga berubah berdasarkan kombinasi kompleksitas transaksi dan kualitas data. Pada waktu pemrosesan, ketiga faktor utama dan seluruh interaksi dua-arah signifikan, sedangkan interaksi tiga-arah tidak signifikan ($p = 0{,}716$).

Pada tingkat kecocokan rekonsiliasi yang disajikan pada Tabel 8, model LLM, kompleksitas transaksi, dan kualitas data menunjukkan pengaruh yang signifikan dengan $p < 0{,}001$. Interaksi tiga-arah juga signifikan, $F(12, 1404) = 2{,}87$, $p < 0{,}001$, sehingga perbedaan performa model perlu ditafsirkan berdasarkan kombinasi kompleksitas transaksi dan kualitas data. Pada waktu rekonsiliasi, model LLM, kompleksitas transaksi, dan kualitas data berpengaruh signifikan. Interaksi kompleksitas transaksi × kualitas data signifikan pada tingkat $p = 0{,}034$, tetapi interaksi tiga-arah tidak signifikan, $F(12, 1404) = 1{,}56$, $p = 0{,}096$. Artinya, kombinasi ketiga faktor belum menunjukkan pengaruh simultan yang cukup kuat terhadap waktu rekonsiliasi pada simulasi ini. Perbedaan performa antar-model pada domain keuangan juga merupakan objek pengukuran dalam benchmark kontemporer karena model yang mempunyai performa tinggi pada satu jenis tugas tidak selalu mempunyai tingkat ketepatan yang sama pada tugas finansial lainnya [34].

Tabel 8. Hasil ANOVA tiga-arah pada rekonsiliasi keuangan

Variabel dependen	Faktor	df	F	p	Partial η^2
Tingkat kecocokan	Model LLM	3	1126,23	<0,001	0,706
Tingkat kecocokan	Kompleksitas transaksi	2	1929,86	<0,001	0,733
Tingkat kecocokan	Kualitas data	2	1509,37	<0,001	0,683
Tingkat kecocokan	Model × kompleksitas	6	16,03	<0,001	0,064
Tingkat kecocokan	Model × kualitas data	6	8,70	<0,001	0,036
Tingkat kecocokan	Kompleksitas × kualitas data	4	69,80	<0,001	0,166
Tingkat kecocokan	Model × kompleksitas × kualitas data	12	2,87	<0,001	0,024
Waktu rekonsiliasi	Model LLM	3	1471,88	<0,001	0,759
Waktu rekonsiliasi	Kompleksitas transaksi	2	628,38	<0,001	0,472
Waktu rekonsiliasi	Kualitas data	2	1288,84	<0,001	0,647
Waktu rekonsiliasi	Model × kompleksitas	6	9,03	<0,001	0,037
Waktu rekonsiliasi	Model × kualitas data	6	12,94	<0,001	0,052
Waktu rekonsiliasi	Kompleksitas × kualitas data	4	2,60	0,034	0,007
Waktu rekonsiliasi	Model × kompleksitas × kualitas data	12	1,56	0,096	0,013

Pada tugas rekonsiliasi keuangan, ANOVA tiga-arah menunjukkan bahwa model LLM, kompleksitas transaksi, dan kualitas data berpengaruh signifikan terhadap tingkat kecocokan transaksi ($p < 0{,}001$). Seluruh interaksi, termasuk interaksi tiga-arah antara model LLM, kompleksitas transaksi, dan kualitas data, juga signifikan. Hasil tersebut menunjukkan bahwa efektivitas model dalam mencocokkan transaksi bergantung pada kondisi input yang dihadapi. Untuk waktu rekonsiliasi, ketiga faktor utama serta interaksi model × kompleksitas, model × kualitas data, dan kompleksitas × kualitas data menunjukkan pengaruh signifikan. Namun, interaksi tiga-arah tidak signifikan ($p = 0{,}096$).

Tabulasi silang antara kompleksitas transaksi dan kualitas data menunjukkan nilai performa tertinggi pada kombinasi transaksi berkompleksitas rendah dengan data berkualitas tinggi dan nilai terendah pada kombinasi transaksi berkompleksitas tinggi dengan data berkualitas rendah seperti tampak pada Tabel 9. Pada GPT-4, *accuracy* pembuatan jurnal pada kedua kondisi tersebut masing-masing sebesar 98,4% dan 88,9%. Pada *FinEdge-like*, nilainya sebesar 97,1% dan 86,2%; Llama 3 sebesar 95,4% dan 81,8%; sedangkan Mistral sebesar 93,7% dan 77,9%. Pada kombinasi kompleksitas sedang dan kualitas data sedang, *accuracy* masing-masing model adalah

Copyright (c) HOAQ (High Education of Organization Archive Quality) : Jurnal Teknologi Informasi

95,2%, 93,0%, 89,7%, dan 87,1%. Evaluasi LLM pada teks finansial juga menunjukkan penggunaan dataset berlabel dan pengukuran terpisah menurut kondisi tugas sebagai prosedur pengujian performa model [35].

Tabel 9. Accuracy pada kombinasi kondisi eksperimen terpilih

Model	Kompleksitas Rendah + Kualitas Tinggi (%)	Kompleksitas Sedang + Kualitas Sedang (%)	Kompleksitas Tinggi + Kualitas Rendah (%)
GPT-4	98,4	95,2	88,9
Llama 3	95,4	89,7	81,8
Mistral	93,7	87,1	77,9
FinEdge-like	97,1	93,0	86,2

Hasil kategorisasi kesalahan menunjukkan empat kategori utama pada pembuatan jurnal, yaitu kesalahan pemilihan akun, kesalahan debit-kredit, kesalahan nominal, dan ketidaklengkapan elemen jurnal. Dari keseluruhan kesalahan yang tercatat pada dataset, kesalahan pemilihan akun mencakup 43,6%, ketidaktepatan debit-kredit 21,8%, ketidaktepatan nominal 18,7%, dan ketidaklengkapan jurnal 15,9%. Pada transaksi kompleksitas tinggi, proporsi kesalahan pemilihan akun meningkat menjadi 49,2% dari keseluruhan kesalahan pada kelompok tersebut. Pada data berkualitas rendah, kesalahan yang berkaitan dengan identifikasi nominal dan informasi transaksi mencakup 27,4% dari seluruh kesalahan pada kelompok kualitas data rendah. Pengukuran kesalahan secara terpisah menurut jenis tugas digunakan pula dalam evaluasi sistem finansial berbasis LLM karena evaluasi tunggal berbasis akurasi tidak mencakup seluruh dimensi keluaran finansial [36].

Pada rekonsiliasi, kategorisasi kesalahan menghasilkan tiga kelompok, yaitu transaksi tidak terpasangkan (*unmatched transaction*), pencocokan yang salah (*false match*), dan perbedaan nominal atau atribut transaksi yang tidak terdeteksi. *Unmatched transaction* mencakup 46,3% dari keseluruhan kesalahan rekonsiliasi, *false match* 31,5%, dan ketidakterdeteksian perbedaan nominal atau atribut 22,2%. GPT-4 mencatat *false match rate* sebesar 1,4%, *FinEdge-like* 2,1%, Llama 3 3,2%, dan Mistral 4,5%. Proporsi transaksi relevan yang gagal dipasangkan masing-masing tercatat 1,7%, 2,8%, 4,1%, dan 5,3%. Literatur mengenai audit laporan keuangan berbasis LLM juga mengidentifikasi deteksi kesalahan, pemeriksaan kepatuhan, dan penyaringan anomali sebagai kelompok tugas yang dievaluasi dalam penerapan LLM pada informasi akuntansi [37].

Hasil tabulasi akhir berdasarkan seluruh indikator kuantitatif memperlihatkan bahwa GPT-4 mempunyai nilai *accuracy* jurnal 94,8%, *F1-score* jurnal 93,9%, tingkat kecocokan rekonsiliasi 96,1%, dan *error rate* rekonsiliasi 3,9%. *FinEdge-like* mencatat masing-masing 92,6%, 91,7%, 93,8%, dan 6,2%. Llama 3 menghasilkan 89,7%, 88,7%, 91,2%, dan 8,8%, sedangkan Mistral menghasilkan 86,9%, 85,8%, 88,5%, dan 11,5%. Literatur evaluasi LLM menunjukkan bahwa perbandingan antar-model dengan ukuran multidimensional diperlukan karena keluaran model pada tugas domain-spesifik dapat berbeda secara substansial [38]. Penelitian lainnya juga mendokumentasikan perubahan performa LLM ketika konteks tugas berubah, dengan penurunan akurasi pada kondisi yang membutuhkan perubahan penilaian berdasarkan konteks [39].

Pembahasan

Hasil penelitian menunjukkan bahwa performa Large Language Model (LLM) dalam otomasi akuntansi dipengaruhi secara simultan oleh karakteristik model, kompleksitas transaksi, dan kualitas data. Pada tugas pembuatan jurnal, GPT-4 menghasilkan accuracy sebesar 94,8% dan F1-score 93,9%, diikuti FinEdge-like sebesar 92,6% dan 91,7%, Llama 3 sebesar 89,7% dan 88,7%, serta Mistral sebesar 86,9% dan 85,8%. Pola yang sama terlihat pada rekonsiliasi, dengan tingkat kecocokan masing-masing 96,1%, 93,8%, 91,2%, dan 88,5%. Temuan ini menunjukkan bahwa kemampuan generatif tidak menghasilkan performa yang seragam antar-model. Perbedaan tersebut relevan dengan penelitian benchmarking LLM finansial yang menekankan bahwa performa model sangat bergantung pada jenis tugas, kebutuhan penalaran, serta karakteristik domain keuangan [5], [15], [40]. Dengan demikian, pemilihan model untuk otomasi akuntansi perlu didasarkan pada tugas dan kondisi penggunaannya, bukan hanya pada kemampuan bahasa umum model.

Performa pembuatan jurnal yang relatif tinggi menunjukkan bahwa LLM memiliki kemampuan untuk mentransformasikan informasi dari dokumen sumber menjadi representasi akuntansi yang lebih terstruktur. Dokumen seperti invoice, kwitansi, laporan timbang, dan laporan harian TBS mengandung informasi mengenai tanggal, pihak transaksi, nominal, barang atau jasa, serta deskripsi aktivitas ekonomi. Informasi tersebut dapat diekstraksi dan dipetakan ke struktur jurnal melalui kemampuan pemrosesan bahasa alami. Temuan ini sejalan dengan penelitian Li, Wu, dan Vasarhelyi yang menunjukkan bahwa LLM mampu mengekstraksi informasi finansial dari sumber tidak terstruktur dengan tingkat akurasi tinggi dalam skenario pengujian tertentu [41]. Namun, hasil penelitian ini juga menunjukkan bahwa kesalahan paling banyak berasal dari pemilihan akun. Hal tersebut mengindikasikan bahwa tantangan utama bukan hanya mengekstraksi informasi, tetapi menentukan makna ekonomi transaksi dan menghubungkannya dengan perlakuan akuntansi yang sesuai. Dengan kata lain, kemampuan menghasilkan teks jurnal tidak identik dengan kemampuan menentukan keputusan akuntansi.

Penurunan performa pada transaksi dengan kompleksitas lebih tinggi memperkuat argumentasi tersebut. Accuracy GPT-4 menurun dari 97,3% pada transaksi berkompleksitas rendah menjadi 91,9% pada transaksi berkompleksitas tinggi, sedangkan penurunan Mistral mencapai 9,3 poin persentase. Hasil ANOVA menunjukkan bahwa kompleksitas transaksi berpengaruh signifikan terhadap accuracy dan terdapat interaksi antara model dan

Copyright (c) HOAQ (High Education of Organization Archive Quality) : Jurnal Teknologi Informasi

kompleksitas. Kondisi ini menunjukkan bahwa peningkatan kompleksitas bukan sekadar meningkatkan jumlah informasi yang harus diproses, tetapi juga meningkatkan kebutuhan model untuk mengintegrasikan beberapa informasi, mempertimbangkan hubungan antar-akun, serta menangani deskripsi transaksi yang ambigu. Temuan tersebut sejalan dengan penelitian mengenai evaluasi kemampuan LLM dalam mendeteksi kesalahan pada dokumen keuangan yang menunjukkan bahwa tugas dengan kebutuhan penalaran lebih kompleks dapat menjadi sumber penurunan reliabilitas model [42]. Oleh karena itu, transaksi sederhana dan berulang lebih sesuai untuk otomasi langsung, sedangkan transaksi multi-akun, transaksi penyesuaian, dan transaksi yang memerlukan interpretasi kebijakan akuntansi memerlukan tingkat pengawasan yang lebih tinggi.

Kualitas data merupakan faktor lain yang secara konsisten memengaruhi performa. *Accuracy* GPT-4 turun dari 97,6% pada data berkualitas tinggi menjadi 91,8% pada data berkualitas rendah, sedangkan Mistral turun dari 90,8% menjadi 83,1%. Pada kombinasi paling menantang, yaitu kompleksitas tinggi dan kualitas data rendah, accuracy kembali mengalami penurunan pada seluruh model. Hasil ANOVA juga memperlihatkan pengaruh signifikan kualitas data terhadap accuracy serta interaksi kualitas data dengan faktor lainnya. Temuan ini memperlihatkan bahwa keberhasilan otomasi LLM tidak dapat dipisahkan dari kualitas input yang diberikan kepada sistem. Hal tersebut penting karena karakter probabilistik LLM memungkinkan sistem tetap menghasilkan keluaran meskipun informasi yang diterima tidak lengkap atau ambigu. Literatur mengenai hallucination LLM menunjukkan bahwa keluaran yang tampak koheren tidak selalu sepenuhnya didukung oleh informasi yang tersedia [43]. Dalam konteks SIA, kondisi tersebut dapat meningkatkan risiko apabila keluaran model langsung dimasukkan ke dalam catatan akuntansi tanpa proses validasi. Karena itu, pengendalian kualitas dokumen perlu ditempatkan sebagai bagian dari arsitektur otomasi, bukan sebagai aktivitas tambahan setelah keluaran LLM dihasilkan.

Hasil rekonsiliasi memperlihatkan bahwa kemampuan LLM juga dapat digunakan untuk mencocokkan transaksi dari beberapa sumber. Tingkat kecocokan berada pada rentang 88,5% sampai 96,1%, sedangkan sumber bank menghasilkan accuracy yang lebih tinggi dibandingkan pemasok dan pembeli. Kemampuan tersebut dapat dijelaskan karena rekonsiliasi tidak selalu hanya membutuhkan kecocokan nominal, tetapi juga mempertimbangkan tanggal, pihak transaksi, deskripsi, dan referensi transaksi. LLM memiliki kemampuan untuk memanfaatkan hubungan semantik antar-elemen tersebut sehingga dapat menghasilkan kandidat kecocokan pada kondisi ketika aturan pencocokan sederhana mungkin tidak memadai. Namun, munculnya false match dan unmatched transaction menunjukkan bahwa pendekatan probabilistik tidak dapat sepenuhnya menggantikan pemeriksaan deterministik. Literatur mengenai LLM finansial juga menekankan bahwa aplikasi pada tugas keuangan perlu mempertimbangkan kombinasi pemahaman konteks, penalaran, dan validasi terhadap informasi sumber [2], [44]. Dengan demikian, penggunaan LLM untuk rekonsiliasi lebih tepat ditempatkan sebagai mekanisme penghasil kandidat kecocokan dan deteksi anomali, sementara keputusan akhir untuk kasus berisiko tetap membutuhkan validasi.

Temuan mengenai waktu pemrosesan memperlihatkan manfaat efisiensi yang cukup besar dalam skenario simulasi. Pembuatan jurnal membutuhkan rata-rata 176,4 detik per transaksi melalui prosedur manual, sedangkan LLM berada pada rentang 1,96–2,84 detik. Pada rekonsiliasi, proses manual membutuhkan 94,8 detik per transaksi dibandingkan 2,17–3,12 detik menggunakan LLM. Perbedaan ini menunjukkan bahwa nilai utama otomasi bukan hanya peningkatan kecepatan sebuah transaksi, tetapi potensi pengurangan beban pekerjaan berulang ketika volume transaksi meningkat. Temuan tersebut konsisten dengan literatur cloud accounting yang mengidentifikasi efisiensi proses dan percepatan pengolahan informasi sebagai salah satu manfaat digitalisasi akuntansi [4]. Namun, waktu inferensi model tidak seharusnya dipahami sebagai waktu penyelesaian proses akuntansi secara keseluruhan. Dalam implementasi nyata masih terdapat proses validasi, koreksi, persetujuan, logging, serta penanganan transaksi pengecualian. Oleh karena itu, penelitian lanjutan perlu mengukur waktu total dari input sampai transaksi memperoleh persetujuan final, bukan hanya waktu pemrosesan model.

Dari perspektif reliabilitas, hasil penelitian memperlihatkan pentingnya membedakan antara performa teknis dan risiko akuntansi. *Accuracy*, precision, recall, dan F1-score memberikan ukuran kuantitatif mengenai kemampuan model, tetapi setiap kesalahan tidak memiliki konsekuensi yang sama. Kesalahan pemilihan akun pada transaksi bernilai kecil, misalnya, memiliki implikasi yang berbeda dibandingkan kesalahan pada transaksi material. Demikian pula, false match pada rekonsiliasi dapat lebih berisiko daripada transaksi yang secara eksplisit ditandai sebagai unmatched karena false match dapat membuat transaksi bermasalah terlihat telah sesuai. Oleh karena itu, hasil penelitian memberikan dasar bahwa evaluasi LLM pada SIA perlu dikembangkan dari sekadar pengukuran akurasi menuju evaluasi berbasis risiko. Pendekatan tersebut sejalan dengan perkembangan penelitian LLM finansial yang semakin menekankan evaluasi berdasarkan karakteristik tugas dan kondisi penggunaan [38].

Implikasi praktis dari temuan tersebut adalah perlunya arsitektur SIA berbasis cloud yang menerapkan mekanisme *human-in-the-loop*. Transaksi rutin, berulang, memiliki dokumen lengkap, dan memiliki pola akun yang stabil dapat diarahkan ke pemrosesan otomatis. Sebaliknya, transaksi dengan confidence rendah, dokumen tidak lengkap, transaksi multi-akun, transaksi penyesuaian, atau transaksi yang melewati batas materialitas perlu diarahkan kepada akuntan untuk pemeriksaan. Pendekatan ini konsisten dengan gagasan bahwa LLM dalam akuntansi lebih tepat digunakan sebagai mekanisme augmentasi daripada sebagai pengganti penuh proses profesional. Literatur mengenai penerapan LLM dalam akuntansi juga menekankan pentingnya kualitas data, adaptasi domain, keamanan, dan tata kelola dalam penerapan teknologi tersebut [17].

Copyright (c) HOAQ (High Education of Organization Archive Quality) : Jurnal Teknologi Informasi

Temuan penelitian juga memiliki implikasi terhadap desain sistem. Integrasi LLM sebaiknya tidak ditempatkan sebagai satu-satunya mekanisme pengambilan keputusan, tetapi sebagai salah satu lapisan dalam arsitektur SIA. LLM dapat digunakan untuk ekstraksi dokumen, penyusunan kandidat jurnal, pencocokan transaksi, dan identifikasi ketidaksesuaian. Setelah itu, aturan deterministik dapat digunakan untuk memeriksa keseimbangan debit-kredit, validitas akun terhadap *chart of accounts*, duplikasi transaksi, konsistensi nominal, dan batas materialitas. Transaksi yang gagal melewati pemeriksaan tersebut dapat dikembalikan kepada pengguna untuk dikoreksi. Pendekatan hibrida semacam ini memberikan pemisahan antara kemampuan interpretasi semantik LLM dan pengendalian akuntansi yang bersifat deterministik [45]. Selain meningkatkan reliabilitas, arsitektur tersebut juga mendukung audit trail karena setiap keluaran dapat ditelusuri kembali ke dokumen sumber, model, parameter pengujian, serta keputusan pengguna.

Aspek penerimaan pengguna juga perlu diperhatikan. Hasil eksperimen menunjukkan bahwa sistem dapat memproses transaksi dengan cepat dan relatif akurat, tetapi manfaat teknis tersebut belum otomatis menjamin penggunaan aktual pada KUD/UMKM. Penelitian mengenai cloud accounting di Indonesia menunjukkan bahwa literasi digital, kompatibilitas teknologi, dan kondisi organisasi berkaitan dengan penerimaan dan penggunaan sistem [46], [47]. Dalam konteks LLM, pengguna tidak hanya membutuhkan kemampuan menjalankan sistem, tetapi juga kemampuan memahami alasan keluaran, memeriksa sumber informasi, mengenali tingkat ketidakpastian, dan melakukan koreksi. Oleh karena itu, antarmuka SIA terintegrasi LLM perlu menampilkan dokumen sumber, akun yang dipilih, nominal, status rekonsiliasi, confidence level, serta alasan atau dasar pencocokan yang tersedia. Transparansi tersebut dapat membantu mengurangi kesenjangan antara performa teknis model dan kepercayaan pengguna terhadap hasil otomasi.

Secara keseluruhan, kontribusi utama penelitian ini bukan sekadar menunjukkan bahwa salah satu model menghasilkan nilai rata-rata lebih tinggi daripada model lainnya. Kontribusi yang lebih penting adalah menunjukkan bahwa performa LLM merupakan hasil interaksi antara model, kompleksitas transaksi, dan kualitas data. GPT-4 menunjukkan performa tertinggi dalam skenario eksperimen ini, tetapi hasil tersebut tidak dapat diperlakukan sebagai peringkat universal seluruh LLM karena dataset yang digunakan bersifat sintetis dan dikonstruksi berdasarkan karakteristik KUD/UMKM pengolahan sawit. Dengan demikian, pertanyaan yang lebih relevan bagi implementasi bukan hanya model mana yang menghasilkan skor tertinggi, tetapi pada kondisi transaksi dan kualitas data seperti apa suatu model mampu memenuhi tingkat reliabilitas yang dibutuhkan. Pendekatan tersebut memperkuat kebutuhan benchmarking berbasis tugas dan kondisi penggunaan sebagaimana ditekankan dalam penelitian LLM finansial [40].

Akhirnya, keterbatasan dataset sintetis harus menjadi pertimbangan utama dalam membaca hasil penelitian. Dataset memang dirancang berdasarkan karakteristik transaksi KUD/UMKM pengolahan sawit dan dilengkapi ground truth, tetapi belum menangkap seluruh variasi, ketidakteraturan, serta dinamika transaksi aktual. Artikel sendiri menegaskan bahwa hasil eksperimen belum dapat ditafsirkan sebagai bukti kinerja LLM pada transaksi lapangan aktual. Oleh karena itu, penelitian berikutnya perlu melakukan validasi eksternal menggunakan data transaksi nyata yang telah dianonimkan, melibatkan lebih dari satu validator akuntansi, serta menambahkan ukuran berbasis risiko seperti material error rate, human correction rate, false reconciliation rate, dan proporsi transaksi yang dapat diselesaikan tanpa intervensi. Dengan pengembangan tersebut, hasil simulasi dapat menjadi dasar menuju evaluasi implementasi LLM yang lebih realistis pada SIA berbasis cloud.

4. SIMPULAN

Eksperimen simulasi ini menunjukkan bahwa *Large Language Model* memiliki potensi untuk mendukung otomasi pembuatan jurnal dan rekonsiliasi keuangan dalam skenario SIA berbasis cloud yang merepresentasikan konteks KUD/UMKM pengolahan sawit. Berdasarkan hasil eksperimen terhadap dataset sintetis yang dikonstruksi secara terkontrol, GPT-4 menghasilkan performa paling tinggi pada indikator akurasi, *precision*, *recall*, *F1-score*, dan tingkat kecocokan rekonsiliasi, diikuti model *FinEdge-like*, Llama 3, dan Mistral. Seluruh model mengalami penurunan performa ketika kompleksitas transaksi meningkat dan kualitas data menurun, sementara waktu pemrosesan berbasis LLM tetap jauh lebih singkat dibandingkan proses manual dalam skenario simulasi. Hasil tersebut menjawab bahwa LLM dapat digunakan untuk mengotomasi aktivitas pencatatan dan pencocokan transaksi, tetapi penggunaannya tidak dapat dilepaskan dari kualitas dokumen sumber, struktur transaksi, kebijakan akuntansi, dan mekanisme verifikasi terhadap keluaran model.

Secara teoretis dan konseptual, penelitian ini memperluas pemahaman mengenai otomasi akuntansi dengan menempatkan performa LLM sebagai hasil interaksi antara kemampuan pemrosesan bahasa, kompleksitas informasi, kualitas data, dan konteks penggunaan SIA berbasis *cloud*. Penelitian ini juga memberikan kontribusi praktis berupa dasar konseptual bagi pengembangan arsitektur SIA yang mengintegrasikan LLM sebagai lapisan pemrosesan dokumen, penyusunan kandidat jurnal, pencocokan transaksi, dan deteksi ketidaksesuaian, tetapi tetap mempertahankan *human-in-the-loop* untuk transaksi yang ambigu, material, atau berisiko tinggi. Dengan demikian, integrasi LLM lebih tepat diarahkan pada model augmentasi akuntansi daripada otonomi penuh, sehingga efisiensi teknologi tetap berjalan bersama prinsip keterlacakan, akuntabilitas, dan pengendalian internal.

Namun, temuan tersebut tidak dapat diperlakukan sebagai estimasi kinerja pada data lapangan aktual. Penelitian lanjutan perlu melakukan validasi eksternal menggunakan data transaksi nyata yang telah dianonimkan dan memperoleh persetujuan etis serta izin organisasi terkait.

Copyright (c) HOAQ (High Education of Organization Archive Quality) : Jurnal Teknologi Informasi

DAFTAR PUSTAKA

[1] C. Sampaio and R. Silva, "Digital Transformation in Accounting: An Assessment of Automation and AI Integration," *International Journal of Financial Studies*, vol. 13, no. 4, p. 206, Nov. 2025, doi: 10.3390/ijfs13040206.

[2] Y. Nie *et al.*, "A Survey of Large Language Models for Financial Applications: Progress, Prospects and Challenges," Jun. 2024.

[3] S. Oviya, N. Sharadha, T. Bhuvaneswari, S. Vijayalakshmi, and S. Sushma, "The Impact of Automation and Ai in Revolutionising Traditional Accounting Methods," *Journal of Informatics Education and Research*, 2024, doi: 10.52783/jier.v4i2.1001.

[4] F. R. P. Faiz, S. A. D. Safina, C. A. Citra, A. Angelia, I. N. Iqlima, and C. D. H. N. Christyan, "Systematic Literatur Review: Pengaruh Cloud Accounting Terhadap Efektivitas Sistem Pelaporan Keuangan Di UMKM," *Jurnal Ekonomika Dan Bisnis (JEBS)*, vol. 5, no. 3, pp. 527–533, May 2025, doi: 10.47233/jebs.v5i3.2769.

[5] J. Gupta *et al.*, "Large Language Models Acing Chartered Accountancy," *SN Comput. Sci.*, vol. 6, no. 8, p. 957, Nov. 2025, doi: 10.1007/s42979-025-04497-x.

[6] S. Liu *et al.*, "FinDABench: Benchmarking Financial Data Analysis Ability of Large Language Models," Jun. 2024.

[7] K. Akshat, "AI-Driven Reconciliation Agents for Financial Accuracy and Compliance in Cloud-Native Data Pipelines," *International Research Journal on Advanced Engineering Hub (IRJAEH)*, vol. 3, no. 09, pp. 3502–3508, Sep. 2025, doi: 10.47392/IRJAEH.2025.0514.

[8] D. Permatasari, N. F. Mohammed, N. A. Shafie, and M.-E. Sánchez-del-Río-Vázquez, "Technological-Organisational-Environmental Determinants of Cloud Accounting System Usage in Indonesian SMEs," *Journal of Nusantara Studies (JONUS)*, vol. 10, no. 2, pp. 566–589, Jul. 2025, doi: 10.24200/jonus.vol10iss2pp566-589.

[9] A. M. Musyaffi *et al.*, "Exploring Technological Factors and Cloud Accounting Adoption in MSMEs: A Comprehensive TAM Framework," *International Review of Management and Marketing*, vol. 15, no. 1, pp. 283–292, Dec. 2024, doi: 10.32479/irmm.17451.

[10] J. Slezák, "Modeling Technology Acceptance in Accounting Area," *Engineering Economics*, vol. 36, no. 3, pp. 291–302, Jun. 2025, doi: 10.5755/j01.ee.36.3.33645.

[11] D. Permatasari, N. Farizal, and N. Aima, "Exploring Factors Influencing the Adoption of Cloud Accounting Systems in Indonesian Micro Small and Medium Enterprises: A Unified Theory of Acceptance and Use of Technology Based Analysis," *Management and Accounting Review*, vol. 23, no. 1, Apr. 2024, doi: 10.24191/MAR.V23i01-07.

[12] G. Bhatia, E. M. B. Nagoudi, H. Cavusoglu, and M. Abdul-Mageed, "FinTral: A Family of GPT-4 Level Multimodal Financial Large Language Models," Jun. 2024.

[13] N. Vali, S. Saha, S. Koora, and Md. A. Vali, "Automating Financial Workflows: AI-Powered Revenue Reconciliation for Real Estate Enterprises," *International Journal of Financial Management and Economics*, vol. 8, no. 2, pp. 273–283, Jan. 2025, doi: 10.33545/26179210.2025.v8.i2.595.

[14] D. Qoriah, "Adoption of Cloud Accounting: Opportunities, Barriers, and Implications for SMEs," *Summa : Journal of Accounting and Tax*, vol. 3, no. 3, pp. 191–204, Jul. 2025, doi: 10.61978/summa.v3i3.979.

[15] G. Matlin, M. Okamoto, H. Pardawala, Y. Yang, and S. Chava, "Financial Language Model Evaluation (FLaME)," in *Findings of the Association for Computational Linguistics: ACL 2025*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2025, pp. 22633–22679. doi: 10.18653/v1/2025.findings-acl.1164.

[16] R. S. Pall, "IndiaFinBench: An Evaluation Benchmark for Large Language Model Performance on Indian Financial Regulatory Text," May 2026.

[17] M. Taqi, "AI Application in Accounting Studies," *Review on Islamic Accounting*, vol. 4, no. 2, Feb. 2025, doi: 10.58968/ria.v4i2.598.

[18] Z. Li and H. Qin, "Large Language Models (LLMs) in Business Strategies and Accounting: Opportunities and Challenges," *Management Journal for Advanced Research*, vol. 4, no. 6, pp. 1–7, Dec. 2024, doi: 10.54741/mjar.4.6.1-7.

[19] A. P. Oktavia, A. Nuraliati, and L. Ulfiana, "The Role of Cloud Accounting in Improving MSME Performance in The Digital Economy Era," *Oikonomia : Journal of Management Economics and Accounting*, vol. 3, no. 1, pp. 1–10, Dec. 2025, doi: 10.61942/oikonomia.v3i1.494.

[20] P. H. Mainardi and E. D. Bidoia, "Fundamental Concepts And Recent Applications of Factorial Statistical Designs," *Brazilian Journal of Biometrics*, vol. 40, no. 1, Mar. 2022, doi: 10.28951/bjb.v40i1.552.

[21] R. S. Pall, "IndiaFinBench: An Evaluation Benchmark for Large Language Model Performance on Indian Financial Regulatory Text," May 2026.

[22] S. Y. A. Safavi-Naini *et al.*, "Benchmarking Proprietary and Open-Source Language and Vision-Language Models for Gastroenterology Clinical Reasoning," *NPJ Digit. Med.*, vol. 8, no. 1, p. 797, Nov. 2025, doi: 10.1038/s41746-025-02174-0.

Copyright (c) HOAQ (High Education of Organization Archive Quality) : Jurnal Teknologi Informasi

[23] T.-T.-H. Le, A. A. Adiputra, J. Yun, and H. Kim, "Anomaly Detection in Industrial Machine Sounds Using High-Frequency Features and Gate Recurrent Unit Networks," *IEEE Access*, vol. 13, pp. 77165–77186, 2025, doi: 10.1109/ACCESS.2025.3565812.
[24] X. Peng *et al.*, "Plutus: Benchmarking Large Language Models in Low-Resource Greek Finance," Feb. 2025.
[25] S. Noels, J. De Blaere, and T. De Bie, "A Dutch Financial Large Language Model," in *Proceedings of the 5th ACM International Conference on AI in Finance*, New York, NY, USA: ACM, Nov. 2024, pp. 283–291. doi: 10.1145/3677052.3698628.
[26] M. M. Bishal, Md. R. H. Chowdory, A. Das, and M. A. Kabir, "COVIDHealth: A novel Labeled Dataset and Machine Learning-Based Web Application for Classifying COVID-19 discourses on Twitter," *Heliyon*, vol. 10, no. 14, p. e34103, Jul. 2024, doi: 10.1016/j.heliyon.2024.e34103.
[27] T. J. Nasser and B. K. Mohammed, "Estimate Factorial Experimental Design by Using Some Methods With an Application," *International Journal of Statistics and Applied Mathematics*, vol. 10, no. 8S, pp. 43–52, Aug. 2025, doi: 10.22271/maths.2025.v10.i8Sa.2149.
[28] R. Ramírez-Tapia, A. J. Ríos-Lira, Y. V. Pantoja-Pacheco, J. A. Vázquez-López, and E. A. Ruelas-Santoyo, "Quantitative Analysis of the Balance Property in Factorial Experimental Designs 24 to 28," *Mathematics*, vol. 10, no. 20, p. 3812, Oct. 2022, doi: 10.3390/math10203812.
[29] D. Stanishevskii *et al.*, "FINESSE-Bench: A Hierarchical Benchmark Suite for Financial Domain Knowledge and Technical Analysis in Large Language Models," May 2026.
[30] C. Yang, C. Xu, and Y. Qi, "Financial Knowledge Large Language Model," Jun. 2024.
[31] Y. Nie *et al.*, "A Survey of Large Language Models for Financial Applications: Progress, Prospects and Challenges," Jun. 2024.
[32] A. Jadhav and V. Mirza, "Large Language Models in Equity Markets: Applications, Techniques, and Insights," *Front. Artif. Intell.*, vol. 8, Aug. 2025, doi: 10.3389/frai.2025.1608365.
[33] A. Srivastava, S. Jung, and S. Mateega, "Market-Bench: Evaluating Large Language Models on Introductory Quantitative Trading and Market Dynamics," Jan. 2026.
[34] X. Peng *et al.*, "Ebisu: Benchmarking Large Language Models in Japanese Finance," Feb. 2026.
[35] M. Birti, A. Maurino, and F. Osborne, "Optimizing Large Language Models for ESG Activity Detection in Financial Texts," in *Proceedings of the 6th ACM International Conference on AI in Finance*, New York, NY, USA: ACM, Nov. 2025, pp. 856–863. doi: 10.1145/3768292.3770371.
[36] D. Shu, H. Yuan, Y. Wang, Y. Liu, H. Zhang, and M. Du, "FinChart-Bench: Benchmarking Financial Chart Comprehension in Vision-Language Models," in *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2026, pp. 13447–13466. doi: 10.18653/v1/2026.acl-long.615.
[37] K. N. Sumiari, I. K. Parnata, I. G. A. A. Pramitari, and I. M. A. Putrayasa, "The Emergence of Large Language Models in Financial Report Auditing: Opportunities, Benchmarks, Risks, and The Road Ahead," *Greenation International Journal of Economics and Accounting*, vol. 4, no. 2, pp. 364–373, Jun. 2026, doi: 10.38035/gijea.v4i2.901.
[38] Z. Liu, X. Zhang, K. Yang, Q. Xie, J. Huang, and S. Ananiadou, "FMDLlama: Financial Misinformation Detection Based on Large Language Models," in *Companion Proceedings of the ACM on Web Conference 2025*, New York, NY, USA: ACM, May 2025, pp. 1153–1157. doi: 10.1145/3701716.3715599.
[39] D. Lee, H. Yun, M. Cha, S. Park, S. Park, and J. Kim, "EconCausal: A Context-Aware Economic Reasoning Benchmark for Large Language Models," May 2026.
[40] S. Lin *et al.*, "Evaluation and Benchmarking Suite for Financial Large Language Models and Agents," Feb. 2026.
[41] D. Chandra, "Applications of Large Language Model Reasoning in Feature Generation," Mar. 2025.
[42] Y. He *et al.*, "Are Large Language Models Reliable Reviewers? A Benchmark for Error Detection in Financial Documents," in *Findings of the Association for Computational Linguistics: ACL 2026*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2026, pp. 29625–29643. doi: 10.18653/v1/2026.findings-acl.1481.
[43] C. Woesle, L. Fischer-Brandies, and R. Buettner, "A Systematic Literature Review of Hallucinations in Large Language Models," *IEEE Access*, vol. 13, pp. 148231–148253, 2025, doi: 10.1109/ACCESS.2025.3601206.
[44] Z. Xie *et al.*, "The CLEF-2026 FinMMEval Lab: Multilingual and Multimodal Evaluation of Financial AI Systems," Feb. 2026.
[45] K. Veeramani, A. Joseph N, and P. M, "Neuro-Symbolic Reasoning Engine for Tax Optimisation," *Front. Artif. Intell.*, vol. 9, May 2026, doi: 10.3389/frai.2026.1802755.
[46] A. Hamzah, D. Suhendar, and A. Z. Arifin, "Factors Affecting Cloud Accounting Adoption In SMEs," *Jurnal Akuntansi*, vol. 27, no. 3, pp. 442–464, Sep. 2023, doi: 10.24912/ja.v27i3.1520.
[47] E. Novianti and F. Aligarh, "MSME Adoption of Cloud Accounting: A TAM and UTAUT Perspective," *Proceedings of Economics Business Innovation & Creativity*, vol. 2, no. 1, pp. 442–463, Jun. 2025, doi: 10.32493/ebic.v2i1.51050.

Copyright (c) HOAQ (High Education of Organization Archive Quality) : Jurnal Teknologi Informasi