HIJM
HORIZON: Indonesian Journal of Multidisciplinary

p-ISSN: 2987-7776
e-ISSN: 2987-7180
Vol. 4, No. 4, 2026

GENERATOR SOAL CERITA KURIKULUM MATEMATIKA DASAR BERBASIS LARGE LANGUAGE MODEL

Jalmo Rohadi[1], Agung Prasetya[2], Mohamad Khoirul Ansor[3]
[1, 2, 3] Universitas Bhinneka PGRI, Jl. Mayor Sujadi Timur No.7, Tulungagung, Jawa Timur, Indonesia
Email: jalmo.student@ubhi.ac.id

Article History

Received: 10-08-2026

Revision: 24-08-2026

Accepted: 28-08-2026

Published: 31-08-2026

Abstract. Math word problems (MWPs) are essential for developing the reasoning skills of elementary school students; however, manual creation is time-consuming and results in limited problem variety. This study aims to design, implement, and evaluate a Large Language Model (LLM)-based math word problem generator that aligns with the *Kurikulum Merdeka* (Independent Curriculum) and produces valid, solvable problems in Indonesian. The system was developed using LLaMA 3.2 3B Instruct via QLoRA fine-tuning with 4-bit NormalFloat quantization, trained on a dataset of 3,185 entries covering 19 topics and 6 grade levels. The system generated 385 problems across 77 grade-topic combinations and was evaluated against a sample of 95 problems through assessments by three elementary school teachers and the G-Eval metric. Evaluation results indicate that 93 out of 95 problems met validity criteria (97.89%), with an average quality score of 4.95 out of 5.00. A comparison of the two evaluation methods revealed a Pearson correlation of 0.94 and a Cohen's Kappa of 1.00. These findings demonstrate the system's potential as a tool to assist teachers in generating math word problems that align with the *Kurikulum Merdeka*.

Keywords: Large Language Model, Fine-tuning, Math Word Problem, Merdeka Curriculum, PostValidator.

Abstrak. Soal cerita matematika (*Math Word Problem*/MWP) penting untuk melatih kemampuan penalaran siswa Sekolah Dasar, tetapi penyusunannya secara manual membutuhkan waktu dan variasi soal yang terbatas. Penelitian ini bertujuan merancang, mengimplementasikan, dan mengevaluasi generator soal cerita matematika berbasis *Large Language Model* (LLM) yang selaras dengan Kurikulum Merdeka serta menghasilkan soal yang valid dan dapat diselesaikan (*solvable*) dalam Bahasa Indonesia. Sistem dikembangkan menggunakan LLaMA 3.2 3B Instruct melalui *fine-tuning* QLoRA dengan kuantisasi 4-bit NormalFloat pada 3.185 data yang mencakup 19 topik dan 6 tingkat kelas. Sistem menghasilkan 385 soal dari 77 kombinasi kelas-topik dan dievaluasi terhadap 95 soal sampel melalui penilaian tiga guru Sekolah Dasar dan G-Eval. Hasil evaluasi menunjukkan 93 dari 95 soal memenuhi kriteria validitas (97,89%) dengan rata-rata skor kualitas 4,95 dari 5,00. Hasil perbandingan kedua metode evaluasi menunjukkan korelasi Pearson sebesar 0,94 dan Cohen's Kappa sebesar 1,00. Temuan ini menunjukkan bahwa sistem berpotensi menjadi alat bantu guru dalam menghasilkan soal cerita matematika yang sesuai dengan Kurikulum Merdeka.

Kata Kunci: Large Language Model, Fine-tuning, Math Word Problem, Kurikulum Merdeka, PostValidator.

How to Cite: Rohadi, J., Prasetya, A., & Ansor, M. K. (2026). Generator Soal Cerita Kurikulum Matematika Dasar Berbasis *Large Language Model*. *HORIZON: Indonesian Journal of Multidisciplinary*, 4 (4), 6072-6083. http://doi.org/10.54373/hijm.v4i4.7369

PENDAHULUAN

Soal cerita matematika atau *Math Word Problem* (MWP) merupakan bentuk evaluasi yang menyajikan permasalahan matematis dalam narasi bahasa alami. Penyelesaian MWP tidak hanya menuntut kemampuan melakukan operasi hitung, tetapi juga kemampuan memahami konteks, mengidentifikasi informasi yang relevan, menerjemahkan situasi ke dalam model matematika, dan menentukan strategi penyelesaian. Oleh karena itu, kualitas MWP yang digunakan dalam pembelajaran perlu memperhatikan aspek matematis sekaligus aspek kebahasaan dan kesesuaian dengan karakteristik peserta didik.

Perkembangan *Large Language Model* (LLM) membuka peluang untuk mengotomatisasi pembangkitan MWP dengan menghasilkan soal dalam jumlah besar dan dengan variasi konteks yang lebih beragam dibandingkan penyusunan secara manual. Beberapa penelitian telah memanfaatkan model LLaMA-2 dan Mistral melalui pendekatan *few-shot prompting* untuk menghasilkan soal matematika. Pendekatan tersebut memiliki keunggulan dalam menghasilkan soal secara cepat tanpa memerlukan proses pelatihan model secara khusus, tetapi kualitas keluaran sangat bergantung pada rancangan *prompt* dan kemampuan model yang digunakan. Dengan demikian, pendekatan *prompting* saja belum sepenuhnya menjamin konsistensi struktur soal, kesesuaian tingkat kesulitan, dan kebenaran matematis dari soal yang dihasilkan.

Penelitian lain melalui proyek MATHWELL menekankan pentingnya keterlibatan guru dalam proses pengembangan dan anotasi data. Pendekatan tersebut memiliki keunggulan dari sisi kesesuaian pedagogis karena penilaian guru dapat digunakan untuk mengarahkan kualitas soal yang dihasilkan. Namun, ketergantungan pada anotasi manusia juga dapat meningkatkan kebutuhan waktu dan sumber daya ketika jumlah soal yang dihasilkan semakin besar. Penelitian mengenai *Chain-of-Thought* (CoT) *prompting* menunjukkan bahwa pemberian langkah penalaran dapat membantu model menghasilkan proses penyelesaian yang lebih terstruktur. Pendekatan ini relevan untuk MWP karena hubungan antara narasi, model matematika, dan jawaban perlu dipastikan secara logis. Akan tetapi, CoT tidak dengan sendirinya menjamin bahwa setiap soal yang dihasilkan valid dan dapat diselesaikan. Pendekatan *self-consistency* dapat mengurangi ketidakstabilan keluaran dengan membandingkan beberapa jalur penalaran, tetapi mekanisme tersebut lebih berfokus pada konsistensi proses jawaban daripada pemeriksaan menyeluruh terhadap struktur soal, konteks, dan solvabilitasnya.

Selain penelitian yang berfokus pada MWP, pengembangan sistem digital dengan pendekatan yang sistematis juga telah diterapkan pada penelitian informatika di Indonesia. Salaam dan Iskandar (2024), misalnya, mengembangkan sistem informasi berbasis *website* menggunakan model ADDIE. Kelebihan pendekatan tersebut terletak pada proses pengembangan yang dilakukan secara bertahap berdasarkan kebutuhan pengguna dan evaluasi pada setiap tahap. Meskipun objek penelitiannya berbeda dengan generator soal matematika, prinsip pengembangan secara modular dan evaluatif relevan dengan penelitian ini, khususnya dalam merancang sistem yang tidak hanya menghasilkan keluaran, tetapi juga melakukan pemeriksaan sebelum keluaran diberikan kepada pengguna.

Berdasarkan kajian tersebut, terdapat tiga persoalan yang belum sepenuhnya teratasi. Pertama, sebagian penelitian pembangkitan MWP menggunakan model bahasa umum dengan pendekatan *prompting*, sehingga belum secara khusus dioptimalkan untuk karakteristik soal cerita matematika berbahasa Indonesia. Kedua, keterlibatan guru dalam penelitian terdahulu dapat meningkatkan kualitas pedagogis, tetapi belum tentu praktis untuk proses generasi dalam jumlah besar. Ketiga, teknik penalaran seperti CoT dan *self-consistency* membantu meningkatkan kualitas proses penyelesaian, tetapi belum memberikan mekanisme validasi berlapis untuk memastikan bahwa soal yang dihasilkan memiliki struktur yang benar, konteks yang logis, dan solusi yang dapat diperoleh secara matematis. Selain itu, sebagian besar kajian terdahulu berorientasi pada lingkungan bahasa Inggris, sehingga karakteristik linguistik dan konteks pembelajaran matematika di Indonesia belum menjadi fokus utama.

Penelitian ini mengisi celah tersebut dengan mengembangkan generator MWP berbahasa Indonesia yang disesuaikan dengan Kurikulum Merdeka menggunakan LLaMA 3.2 3B Instruct melalui *fine-tuning* QLoRA. Pembeda utama penelitian ini terletak pada integrasi model yang telah diadaptasi dengan *prompt* terstruktur dan mekanisme validasi berlapis untuk memeriksa kualitas keluaran, termasuk solvabilitas dan potensi kesalahan sebelum soal disajikan. Dengan pendekatan tersebut, penelitian tidak hanya berfokus pada kemampuan LLM dalam menghasilkan soal, tetapi juga pada mekanisme pengendalian kualitas keluaran yang diperlukan agar sistem lebih sesuai untuk mendukung penyusunan MWP pada jenjang Sekolah Dasar.

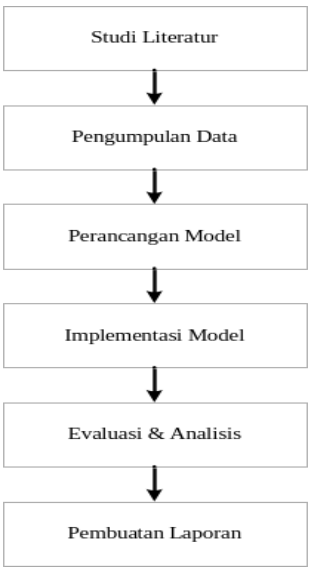
Berdasarkan *research gap* tersebut, penelitian ini berfokus pada pengembangan dan evaluasi generator soal cerita matematika berbasis LLM yang sesuai dengan Kurikulum Merdeka dan karakteristik Bahasa Indonesia. Penelitian mencakup pengembangan dataset, perancangan sistem generasi dan validasi, penerapan *fine-tuning* QLoRA, serta evaluasi kualitas soal melalui penilaian guru dan G-Eval. Dengan demikian, kontribusi utama penelitian

ini adalah menghasilkan pendekatan generasi MWP yang menggabungkan adaptasi model, struktur *prompt*, dan validasi keluaran dalam satu sistem.

METODE

Penelitian ini menggunakan metode *Research and Development* (R&D) untuk mengembangkan dan mengevaluasi sistem generator soal cerita matematika berbasis *Large Language Model* (LLM). Pengembangan sistem dilakukan melalui enam tahapan utama, yaitu (1) studi literatur, (2) pengumpulan dan penyusunan dataset, (3) perancangan model dan arsitektur sistem, (4) implementasi model, (5) evaluasi dan analisis hasil, serta (6) penyusunan laporan penelitian. Pendekatan pengembangan digunakan untuk menghasilkan sistem yang mampu membangkitkan soal cerita matematika berbahasa Indonesia sesuai Kurikulum Merdeka sekaligus melakukan pemeriksaan terhadap kualitas dan solvabilitas soal sebelum soal diberikan sebagai keluaran akhir.

Secara umum, proses penelitian dimulai dengan studi literatur sebagai dasar dalam menentukan karakteristik dataset, metode adaptasi model, arsitektur generator, mekanisme validasi, dan metode evaluasi. Tahap berikutnya adalah penyusunan dataset dan anotasi berdasarkan tingkat kelas serta topik matematika. Dataset tersebut digunakan untuk melakukan *fine-tuning* model LLaMA 3.2 3B Instruct menggunakan metode QLoRA. Model yang telah diadaptasi kemudian diintegrasikan ke dalam sistem generator yang dilengkapi *prompt* terstruktur, *PostValidator*, pemeriksaan solvabilitas, dan *ConstraintFallback*. Sistem selanjutnya diuji dengan menghasilkan soal pada berbagai kombinasi kelas dan topik, kemudian kualitas keluarannya dievaluasi menggunakan *Human Evaluation* dan G-Eval.



Gambar 1. Alur penelitian pengembangan generator soal cerita matematika berbasis LLM.

Tahapan Penelitian

Tahapan penelitian terdiri atas enam tahap yang saling berkaitan. Studi literatur dilakukan untuk mengidentifikasi metode pembentukan dataset, arsitektur generator, teknik *fine-tuning*, mekanisme validasi, dan pendekatan evaluasi kualitas soal cerita matematika yang relevan. Hasil studi literatur digunakan sebagai dasar dalam merancang sistem. Tahap kedua adalah pengumpulan dan penyusunan dataset. Tahap ketiga meliputi perancangan model dan arsitektur sistem, sedangkan tahap keempat merupakan implementasi model dan integrasi seluruh modul. Tahap kelima dilakukan untuk mengevaluasi kualitas soal yang dihasilkan dan menganalisis hasil pengujian. Tahap terakhir adalah penyusunan laporan penelitian berdasarkan seluruh hasil pengembangan dan evaluasi.

Pengumpulan Dataset

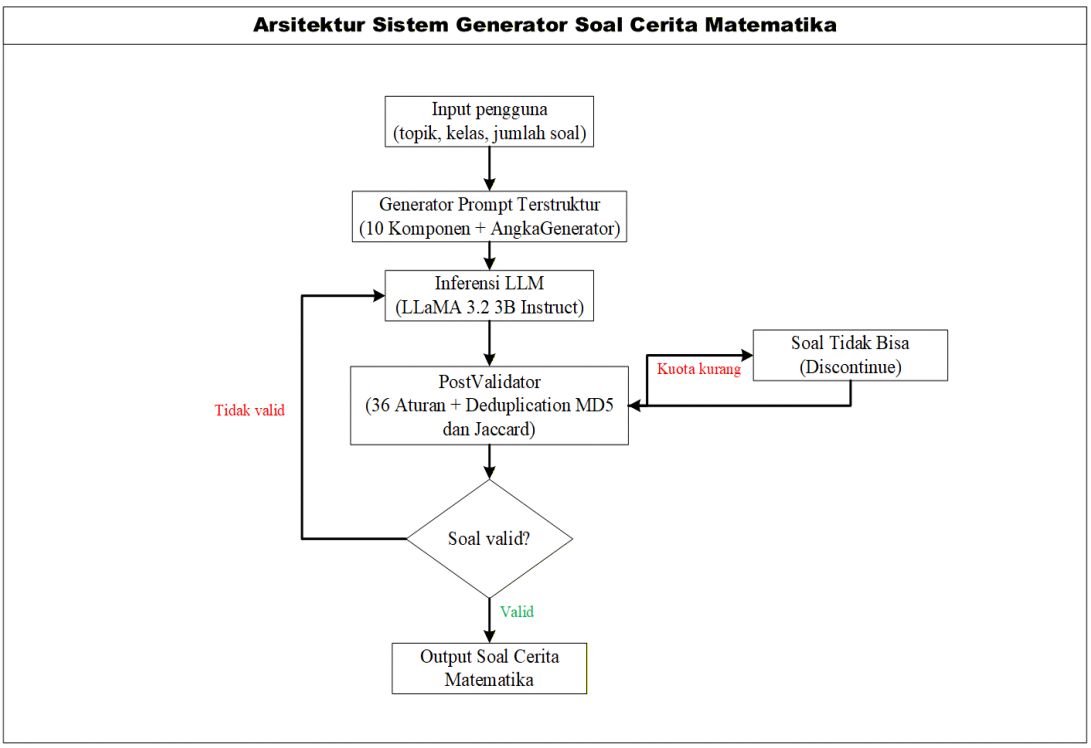
Dataset disusun secara manual tanpa bantuan sistem AI generatif guna menjaga keaslian data dan meminimalkan bias model, dengan tiga jenis dataset utama: Initial Dataset, Annotated Dataset, dan Evaluation Dataset, seluruhnya versi Bahasa Indonesia. Initial Dataset dibentuk dari 493 soal cerita matematika tingkat SD yang dikumpulkan dari buku sekolah dan Buku Sekolah Elektronik (BSE), kemudian diperluas melalui augmentasi menjadi 3.185 data yang mencakup 19 topik dan 6 tingkat kelas. Annotated Dataset dilengkapi label kelas dan ruang lingkup materi sesuai Capaian Pembelajaran Kurikulum Merdeka Fase A-C, dengan validasi anotasi oleh minimal dua penilai manusia. Evaluation Dataset disusun independen dari data pelatihan agar hasil evaluasi bersifat objektif.

Arsitektur Model Generator

Sistem generator dirancang menggunakan LLaMA 3.2 3B Instruct yang diadaptasi melalui metode *Parameter-Efficient Fine-Tuning* (PEFT), yaitu QLoRA dengan kuantisasi 4-bit NormalFloat. Proses *fine-tuning* menggunakan 3.185 pasangan *prompt-output* yang diperoleh dari dataset yang telah disusun. Arsitektur sistem terdiri atas enam modul fungsional yang bekerja secara berurutan, yaitu:

- Modul Input Pengguna, yang mengenali parameter berupa topik, kelas, dan jumlah soal menggunakan kamus alias dan *fuzzy matching*.
- Modul Generator Prompt Terstruktur, yang menyusun *prompt* dinamis berdasarkan sepuluh komponen dan menggunakan AngkaGenerator untuk mengendalikan nilai numerik sesuai rentang kemampuan kelas.

- Modul Inferensi LLM, yang menggunakan LLaMA 3.2 3B Instruct untuk menghasilkan narasi soal.
- Modul PostValidator, yang menerapkan 36 aturan validasi untuk memeriksa struktur dan isi soal, termasuk aturan khusus untuk setiap topik. Modul ini juga dilengkapi mekanisme *deduplication* menggunakan hash MD5 dan *Jaccard similarity*.
- Modul Output Soal Final, yang menampilkan soal yang telah memenuhi seluruh kriteria validasi.
- Modul ConstraintFallback, yang berfungsi sebagai mekanisme cadangan deterministik. Modul ini menghasilkan soal menggunakan templat khusus setiap topik apabila jumlah soal yang memenuhi kriteria belum terpenuhi setelah beberapa iterasi generasi.



Gambar 2. Arsitektur sistem generator soal cerita matematika yang terdiri atas enam modul fungsional.

Model bahasa diadaptasi melalui fine-tuning Parameter-Efficient Fine-Tuning (PEFT), khususnya QLoRA dengan kuantisasi 4-bit NormalFloat, menggunakan 3.185 pasangan prompt-output dari dataset yang telah disusun. Pemeriksaan solvability dilakukan melalui empat modul yaitu ekstraksi informasi numerik, pembentukan model matematika, penyelesaian persamaan menggunakan symbolic solver berbasis aturan aritmatika, dan pemeriksaan konsistensi untuk memastikan persamaan memiliki satu solusi unik tanpa konflik data.

Metode Evaluasi

Evaluasi dilakukan setelah sistem selesai diimplementasikan. Sistem digunakan untuk menghasilkan lima soal pada setiap kombinasi kelas dan topik yang valid. Dari 77 kombinasi kelas-topik diperoleh 385 soal. Selanjutnya, 95 soal dipilih sebagai sampel evaluasi menggunakan *proportional random sampling*, dengan proporsi lima soal untuk setiap 19 topik yang tersedia. Evaluasi dilakukan menggunakan dua pendekatan, yaitu *Human Evaluation* dan G-Eval. *Human Evaluation* dilakukan oleh tiga guru Sekolah Dasar yang aktif mengajar. Para penilai memeriksa validitas soal berdasarkan dua belas kategori kesalahan dan memberikan skor 1-5 pada empat aspek kualitas, yaitu kesesuaian tingkat kelas, ketepatan operasi matematika, kejelasan bahasa, dan kesesuaian konteks.

G-Eval digunakan sebagai metode evaluasi otomatis berbasis model bahasa dengan menggunakan Claude 3.5 Sonnet. Kriteria penilaian G-Eval disusun agar setara dengan kriteria yang digunakan pada *Human Evaluation*. Hasil kedua metode kemudian dibandingkan untuk mengetahui tingkat kesesuaian penilaian. Korelasi skor numerik dianalisis menggunakan koefisien Pearson, sedangkan kesesuaian klasifikasi valid/error dianalisis menggunakan Cohen's Kappa.

Teknik Analisis Data

Data hasil evaluasi dianalisis secara kuantitatif. Persentase validitas dihitung berdasarkan perbandingan jumlah soal yang memenuhi kriteria valid dengan jumlah seluruh soal yang dievaluasi. Skor kualitas dihitung berdasarkan rata-rata penilaian pada empat aspek yang telah ditentukan. Selanjutnya, hasil *Human Evaluation* dan G-Eval dibandingkan menggunakan koefisien korelasi Pearson untuk mengetahui hubungan antara skor yang diberikan oleh kedua metode. Cohen's Kappa digunakan untuk mengukur tingkat kesepakatan klasifikasi valid/error antara penilaian manusia dan evaluasi otomatis. Hasil analisis digunakan untuk menilai kualitas keluaran sistem serta konsistensi G-Eval terhadap penilaian guru.

HASIL

Hasil Generasi Soal

Sistem generator berhasil menghasilkan lima soal pada setiap kombinasi kelas dan topik yang ditetapkan. Dari 77 kombinasi kelas-topik, seluruh kombinasi berhasil menghasilkan soal sehingga diperoleh 385 soal. Evaluasi dilakukan terhadap 95 soal yang dipilih secara *proportional random sampling* dari 19 topik, dengan lima soal pada setiap topik. Dengan demikian, sampel evaluasi mencakup 24,7% dari keseluruhan soal yang dihasilkan. Perlu

dicatat bahwa jumlah soal pada setiap kelas tidak sepenuhnya seimbang karena kombinasi kelas-topik yang tersedia dalam Kurikulum Merdeka berbeda antar fase. Oleh karena itu, perbandingan kualitas berdasarkan kelas perlu diinterpretasikan secara hati-hati karena jumlah kombinasi dan representasi data pada setiap kelas tidak sama. Analisis berdasarkan topik dilakukan untuk melihat apakah kesalahan dan skor kualitas terkonsentrasi pada materi tertentu.

Hasil Human Evaluation

Evaluasi oleh tiga guru Sekolah Dasar menunjukkan bahwa kualitas soal yang dihasilkan tergolong sangat baik. Rata-rata skor keseluruhan mencapai 4,95 dari skala 5. Penilaian mencakup kesesuaian tingkat kelas, ketepatan operasi matematika, kejelasan bahasa, dan kesesuaian konteks.

Tabel 1. Rata-rata skor kualitas soal berdasarkan aspek penilaian *human evaluation*

Aspek Penilaian	Rata-rata Skor (1-5)
Kesesuaian Tingkat Kelas	4,96
Ketepatan Operasi Matematika	4,95
Kejelasan Bahasa	4,95
Kesesuaian Konteks	4,93
Rata-rata Keseluruhan	4,95

Berdasarkan Tabel 1, seluruh aspek memperoleh skor di atas 4,90. Skor tertinggi terdapat pada aspek kesesuaian tingkat kelas (4,96), sedangkan skor terendah terdapat pada kesesuaian konteks (4,93). Selisih skor yang relatif kecil, yaitu 0,03, menunjukkan bahwa kualitas soal relatif konsisten pada keempat aspek yang dinilai. Dari 95 soal yang dievaluasi, sebanyak 93 soal (97,89%) dinyatakan valid dan 2 soal (2,11%) dikategorikan sebagai *error*. Kesalahan pertama ditemukan pada soal pembagian kelas 4 yang tidak *solvable* karena operasi 205 ÷ 8 menghasilkan bilangan tidak bulat. Kesalahan kedua terdapat pada soal Bilangan Bulat kelas 6 yang dinilai tidak realistis karena menggunakan frasa "naik -2 lantai" yang kontradiktif secara logis. Kedua temuan tersebut menunjukkan bahwa sistem telah mampu mendeteksi sebagian besar kesalahan struktural dan matematis, tetapi masih memiliki keterbatasan dalam menangkap kesalahan yang bersifat semantik dan kontekstual.

Jika ditinjau berdasarkan topik, kedua soal *error* tersebut berasal dari topik Pembagian kelas 4 dan Bilangan Bulat kelas 6. Sementara itu, 17 topik lainnya tidak menunjukkan *error* pada sampel evaluasi. Dengan demikian, kesalahan tidak tersebar secara merata pada seluruh topik, tetapi terkonsentrasi pada kasus tertentu yang berkaitan dengan solvabilitas operasi dan

kewajaran konteks. Temuan ini menunjukkan bahwa kualitas keluaran secara umum relatif konsisten, meskipun beberapa topik masih memerlukan pemeriksaan semantik yang lebih kuat.

Hasil G-Eval dan Korelasi dengan Human Evaluation

Evaluasi otomatis menggunakan G-Eval menghasilkan rata-rata skor kualitas sebesar 4,94 dari skala 5. Sebanyak 93 dari 95 soal dinyatakan valid dan dua soal dikategorikan sebagai *error*. Dengan demikian, tingkat validitas berdasarkan G-Eval juga sebesar 97,89%. G-Eval mengidentifikasi dua soal yang sama dengan *Human Evaluation*, yaitu soal Pembagian kelas 4 dan Bilangan Bulat kelas 6.

Tabel 2. Perbandingan hasil G-Eval dan *human evaluation*

Metrik	Nilai
Korelasi Pearson (skor kualitas)	0,94
Cohen's Kappa (klasifikasi valid/error)	1,00
Akurasi G-Eval terhadap Human Evaluation	100%
Presisi deteksi kesalahan	100%
Recall deteksi kesalahan	100%

Perbedaan rata-rata skor antara *Human Evaluation* dan G-Eval hanya sebesar 0,01, sehingga secara deskriptif hasil penilaian kedua metode menunjukkan pola yang sangat serupa. Korelasi Pearson sebesar 0,94 menunjukkan hubungan linier yang sangat kuat antara skor yang diberikan oleh kedua metode. Artinya, soal yang memperoleh skor tinggi dari evaluator manusia cenderung memperoleh skor tinggi pula dari G-Eval.

Cohen’s Kappa sebesar 1,00 menunjukkan adanya kesepakatan sempurna pada sampel penelitian dalam klasifikasi valid dan *error*. Namun, nilai tersebut tidak dapat secara langsung diartikan bahwa G-Eval sepenuhnya setara dengan penilaian manusia. Interpretasi tersebut perlu dibatasi pada data dan kategori yang diuji. Jumlah soal *error* dalam sampel relatif sedikit, yaitu hanya dua dari 95 soal, sehingga distribusi kelas valid dan *error* tidak seimbang. Kondisi ini dapat memengaruhi interpretasi ukuran kesepakatan dan menyebabkan nilai Kappa terlihat sangat tinggi. Oleh karena itu, Pearson, Cohen’s Kappa, serta metrik akurasi, presisi, dan *recall* perlu dipandang secara bersama-sama, bukan menggunakan satu metrik sebagai satu-satunya bukti kesetaraan G-Eval dengan evaluator manusia.

Hasil tersebut menunjukkan bahwa G-Eval memiliki potensi sebagai alat bantu evaluasi awal dalam proses pengembangan generator soal, khususnya untuk penyaringan kualitas secara berulang. Meskipun demikian, evaluasi manusia tetap diperlukan, terutama untuk menilai aspek semantik, kewajaran konteks, dan kesesuaian pedagogis yang belum seluruhnya dapat ditangkap oleh mekanisme otomatis.

DISKUSI

Hasil penelitian menunjukkan bahwa sistem generator soal cerita matematika yang dikembangkan menghasilkan tingkat validitas sebesar 97,89% dengan skor kualitas rata-rata 4,95 dari 5,00. Temuan ini menunjukkan bahwa integrasi *prompt* terstruktur sepuluh komponen, *fine-tuning* QLoRA, serta mekanisme validasi berlapis melalui *PostValidator* dan *ConstraintFallback* mampu menghasilkan soal yang relatif konsisten dari sisi tingkat kelas, ketepatan operasi matematika, kejelasan bahasa, dan kesesuaian konteks. Temuan tersebut menjadi penting karena penelitian terdahulu mengenai pembangkitan *Math Word Problem* (MWP) berbasis *Large Language Model* (LLM) lebih banyak menggunakan pendekatan *few-shot prompting* dan berfokus pada kemampuan model menghasilkan variasi soal, sementara aspek validasi keluaran dan jaminan *solvability* belum menjadi bagian utama dari arsitektur sistem.

Temuan penelitian ini sekaligus menjawab *research gap* terkait kebutuhan pengembangan generator MWP yang secara khusus disesuaikan dengan karakteristik Bahasa Indonesia dan konteks Kurikulum Merdeka. Sistem tidak hanya menggunakan model LLM untuk menghasilkan narasi soal, tetapi juga memasukkan parameter kelas, topik, rentang angka, serta struktur soal ke dalam *prompt* dan proses validasi. Dengan demikian, pembangkitan soal tidak sepenuhnya bergantung pada kemampuan generatif model, tetapi dikendalikan melalui aturan yang disesuaikan dengan kebutuhan pembelajaran matematika sekolah dasar di Indonesia. Pendekatan ini menjadi pembeda dari penelitian terdahulu yang umumnya menggunakan model dan konteks berbahasa Inggris serta belum secara eksplisit mengintegrasikan karakteristik kurikulum dan konteks pembelajaran Indonesia.

Dari sisi *solvability*, penggunaan mekanisme pemeriksaan yang mencakup ekstraksi informasi numerik, pembentukan model matematika, penyelesaian persamaan, dan pemeriksaan konsistensi memberikan lapisan pengamanan tambahan terhadap keluaran LLM. Mekanisme tersebut mampu mencegah sebagian besar kesalahan struktural, seperti ketidaksesuaian operasi dan data numerik yang tidak dapat menghasilkan solusi. Namun, ditemukannya dua soal *error* pada evaluasi manusia menunjukkan bahwa validasi berbasis aturan belum sepenuhnya mampu memahami aspek semantik dan kewajaran konteks. Salah satu soal menghasilkan operasi pembagian yang tidak menghasilkan bilangan bulat, sedangkan soal lainnya mengandung konteks bilangan bulat yang secara logis tidak sesuai. Dengan demikian, tingginya tingkat validitas tidak berarti sistem telah bebas dari kesalahan, tetapi menunjukkan bahwa mekanisme validasi yang dikembangkan efektif dalam mengurangi sebagian besar kesalahan yang dapat dideteksi secara struktural.

Temuan tersebut juga memperjelas keterbatasan pendekatan berbasis aturan dibandingkan kemampuan evaluasi manusia. *PostValidator* efektif untuk memeriksa pola yang telah didefinisikan, tetapi masih terbatas dalam menilai hubungan makna antarkomponen dalam narasi. Hal ini menunjukkan bahwa pengembangan generator MWP selanjutnya perlu mengombinasikan validasi deterministik dengan pemeriksaan semantik berbasis model bahasa. Pendekatan hibrida tersebut berpotensi melengkapi kelemahan masing-masing metode, yaitu aturan deterministik untuk kesalahan yang terstruktur dan model bahasa untuk mendeteksi ketidakwajaran konteks, kontradiksi, serta koherensi narasi.

Hasil evaluasi G-Eval yang menunjukkan korelasi Pearson sebesar 0,94 dan kesepakatan klasifikasi Cohen's Kappa sebesar 1,00 juga memberikan indikasi bahwa evaluasi otomatis memiliki kesesuaian yang tinggi dengan penilaian manusia pada sampel penelitian ini. Namun, nilai tersebut perlu dipahami secara proporsional karena evaluasi hanya dilakukan terhadap 95 soal dan melibatkan dua kategori klasifikasi, yaitu valid dan *error*. Dengan demikian, hasil tersebut belum dapat dijadikan dasar untuk menyatakan bahwa G-Eval sepenuhnya menggantikan evaluasi manusia. G-Eval lebih tepat diposisikan sebagai alat bantu untuk penyaringan awal dan evaluasi berskala besar, sedangkan evaluasi manusia tetap diperlukan untuk memeriksa aspek pedagogis dan semantik yang lebih kompleks.

Penelitian ini memberikan kontribusi dengan mengembangkan generator MWP yang tidak hanya berorientasi pada kemampuan menghasilkan soal, tetapi juga mengintegrasikan penyesuaian terhadap konteks pembelajaran matematika di Indonesia, pemeriksaan *solvability*, dan validasi keluaran secara berlapis. Dengan demikian, penelitian ini mengisi celah penelitian terdahulu yang cenderung menitikberatkan pada kemampuan generatif LLM tanpa memberikan mekanisme pengendalian keluaran yang komprehensif. Meskipun demikian, keterbatasan dalam mendeteksi kesalahan semantik menunjukkan bahwa sistem masih memerlukan pengembangan lebih lanjut sebelum digunakan sebagai pengganti penuh proses penyusunan dan penelaahan soal oleh guru.

KESIMPULAN

Penelitian ini berhasil merancang, mengimplementasikan, dan mengevaluasi sistem generator soal cerita matematika berbasis LLM yang selaras dengan Kurikulum Merdeka. Sistem dengan enam modul fungsional dan model LLaMA 3.2 3B Instruct yang di-fine-tune menggunakan QLoRA berhasil menghasilkan 385 soal secara konsisten untuk seluruh 77 kombinasi kelas-topik. Evaluasi terhadap 95 soal sampel menunjukkan tingkat validitas 97,89% dengan skor kualitas rata-rata 4,95 dari 5,00 pada Human Evaluation, serta korelasi

Pearson 0,94 dan Cohen's Kappa sempurna ($\kappa = 1{,}00$) terhadap G-Eval. Temuan ini mengonfirmasi bahwa pendekatan prompt terstruktur yang dikombinasikan dengan validasi solvabilitas berlapis mampu menghasilkan soal cerita matematika yang valid, solvable, dan sesuai konteks Bahasa Indonesia, sehingga berpotensi menjadi alat bantu yang andal bagi guru Sekolah Dasar. Penelitian lanjutan disarankan untuk mengembangkan validator berbasis model bahasa guna menangkap kesalahan semantik, memperluas cakupan ke soal multi-langkah, serta menguji model LLM berukuran lebih besar untuk meningkatkan kualitas narasi soal.

REFERENSI

Christ, B. R. (2024). MATHWELL: Generating Educational Math Word Problems Using Teacher Annotations.

Dettmers, T., & May, L. G. (n.d.). QLoRA: Efficient Finetuning of Quantized LLMs.

Hought, T. (2023). Self-Consistency Improves Chain of Thought Reasoning. 1-24.

Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., & Liu, T. (2024). A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. ACM Transactions on Information Systems, 1(1), 1-58. https://doi.org/10.1145/3703155

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y., Chen, D., Dai, W., Chan, H., & Madotto, A. (2024). Survey of Hallucination in Natural Language Generation. 1(1).

Koto, F., & Li, H. (2023). Large Language Models Only Pass Primary School Exams in Indonesia. 4.

Liu, Y., Iter, D., & Xu, Y. (n.d.). G-Eval: NLG Evaluation using GPT-4 with Better Human Alignment.

Mulla, N., & Gharpure, P. (2023). Automatic question generation: a review of methodologies, datasets, evaluation metrics, and applications. Progress in Artificial Intelligence, 12(1), 1-32. https://doi.org/10.1007/s13748-023-00295-9

Salaam, P. A., & Iskandar, J. (2024). Pengembangan Sistem Informasi Digital Berbasis Website Menggunakan Pendekatan ADDIE di Desa Cikalong Sukahaji - Majalengka. JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika), 9(2), 1022-1030. https://doi.org/10.29100/jipi.v9i2.5535

Wei, J., Schuurmans, D., Chi, E. H., Le, Q. V., & Zhou, D. (2022). Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. NeurIPS, 1-43.

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, J., Dong, Z., Du, Y., Yang, C., Chen, Y., Chen, Z., & Jiang, J. (2023). A Survey of Large Language Models. 1-144.