Journal of Artificial Intelligence and Technology Information
Website : https://ejournal.techcart-press.com/index.php/jaiti
E-ISSN 2985-6396
P-ISSN 2985-5306
Volume 4, Number 3, September 2026
https://doi.org/10.58602/jaiti.v4i3.317
Page 495-505

Analisis Groundedness Model LLM dalam Pembentukan Soal Cerita Matematika dari Teks Bacaan Berbahasa Indonesia

Feri Andriawan Saputra[1*], Agung Prasetya[2], Mohamad Khoirul Ansor[3]
[1,2,3]Program Studi Informatika, Fakultas Sains dan Teknologi, Universitas Bhinneka PGRI, Tulungagung, Indonesia
Email: [1*]feris4296@gmail.com, [2]agung@ubhi.ac.id, [3]ansorulkhoir@gmail.com

Abstrak

Soal cerita matematika (*Math Word Problem*/MWP) merupakan instrumen evaluasi penting yang mengintegrasikan kemampuan literasi dan numerasi. Perkembangan *Large Language Model* (LLM) membuka peluang otomatisasi pembuatan MWP dari teks bacaan, namun terhambat oleh fenomena *hallucination* yang menurunkan kualitas pedagogis soal. Penelitian ini bertujuan: (1) mengembangkan kerangka evaluasi *groundedness* berbasis enam aspek sebagai produk utama, dan (2) mengevaluasi secara empiris tingkat *groundedness* MWP yang dihasilkan model gemma3:1b dari teks bacaan BSE SD Kurikulum Merdeka. Pendekatan *Research and Development* (R&D) dengan model ADDIE digunakan. Enam aspek dikembangkan: *Entity Consistency* (EC), *Event Consistency* (EV), *Context Alignment* (CA), *Semantic Alignment* (SA), *Numerical Groundedness* (QN), dan *Question Relevance* (QR), masing-masing dinilai dengan skala biner 0–1. Evaluasi dilakukan terhadap 460 soal valid dari empat teks bacaan dengan lima konfigurasi parameter (E1–E5). Hasil menunjukkan gemma3:1b *grounded* pada teks sederhana–menengah (Teks 1: 73,25%; Teks 3: 75,91%) namun gagal pada teks kompleks (Teks 4: 47,22%). Konfigurasi E3 (temperature=0,3; top_p=0,9) memberikan keseimbangan terbaik. Kerangka evaluasi reliabel dengan Cohen's Kappa 0,68–0,81.

Kata Kunci: Groundedness; Large Language Model; Math Word Problem; hallucination; evaluasi NLP;

Abstract

Math Word Problems (MWP) are essential evaluation instruments integrating literacy and numeracy competencies. Large Language Model (LLM) development enables automated MWP generation from reading texts, yet the hallucination phenomenon significantly reduces pedagogical quality. This study aims to: (1) develop a six-aspect groundedness evaluation framework, and (2) empirically evaluate MWP groundedness generated by gemma3:1b from Indonesian Kurikulum Merdeka reading texts. A Research and Development (R&D) approach using ADDIE was employed. Six aspects were developed: Entity Consistency (EC), Event Consistency (EV), Context Alignment (CA), Semantic Alignment (SA), Numerical Groundedness (QN), and Question Relevance (QR), each rated on a binary scale of 0–1. Evaluation covered 460 valid MWPs from four reading texts across five parameter configurations (E1–E5). Results show gemma3:1b achieved adequate groundedness on simple-to-medium texts (Text 1: 73.25%; Text 3: 75.91%) but failed on complex texts (Text 4: 47.22%). Configuration E3 (temperature=0.3, top_p=0.9) provided optimal balance. The framework demonstrated reliable inter-rater agreement (Cohen's Kappa 0.68–0.81).

Keywords: Groundedness; Large Language Model; Math Word Problem; hallucination; NLP evaluation;

***Corresponding Author:**
Feri Andriawan Saputra, Universitas Bhinneka PGRI, Indonesia
Email: feris4296@gmail.com

Journal of Artificial Intelligence and Technology Information
Website : https://ejournal.techcart-press.com/index.php/jaiti
E-ISSN 2985-6396
P-ISSN 2985-5306
Volume 4, Number 3, September 2026
https://doi.org/10.58602/jaiti.v4i3.317
Page 495-505

1. PENDAHULUAN

Kemampuan matematika merupakan kompetensi dasar yang menjadi tolok ukur mutu pendidikan global. Hasil PISA 2022 menempatkan Indonesia di posisi ke-70 dari 81 negara dengan skor rata-rata 366 poin, jauh di bawah rata-rata OECD sebesar 472 poin [1]. Hanya 1% siswa Indonesia yang mencapai level mahir, sementara 82% tidak mencapai kompetensi dasar (Level 2). Kondisi ini mengindikasikan kesenjangan mendasar antara kemampuan komputasi numerik yang diajarkan di kelas dan kemampuan penalaran matematis kontekstual yang dibutuhkan dalam kehidupan nyata.

Salah satu instrumen evaluasi yang dirancang menjembatani kesenjangan tersebut adalah soal cerita matematika (*Math Word Problem*/MWP). MWP menyajikan permasalahan dalam narasi yang merepresentasikan situasi nyata, menuntut siswa memahami teks, menafsirkan konteks, dan memetakan informasi verbal ke representasi matematis [2]. Kemampuan menyelesaikan MWP melibatkan pemahaman teks, pembangunan model situasi, translasi ke model matematis, dan interpretasi hasil secara terpadu [2]. Dalam Kurikulum Merdeka, MWP menjadi instrumen strategis yang mengintegrasikan penguatan literasi dan numerasi secara bersamaan [3].

Perkembangan *Large Language Model* (LLM) seperti GPT-4 [4] dan model instruksi [16] membuka peluang baru dalam otomatisasi pembuatan MWP berbasis konteks bacaan. Namun, pemanfaatan LLM dalam generasi MWP menghadapi tantangan mendasar berupa fenomena *hallucination*. Ji et al. [5] mengklasifikasikan *hallucination* ke dalam dua tipe: (1) *intrinsic hallucination*, yaitu keluaran yang bertentangan langsung dengan teks sumber; dan (2) *extrinsic hallucination*, yaitu penambahan informasi yang tidak dapat diverifikasi dari teks. Dalam konteks MWP, *hallucination* dapat muncul sebagai penambahan tokoh, peristiwa, relasi kuantitatif, atau perubahan konteks yang tidak ada dalam teks sumber, sehingga menurunkan kualitas pedagogis soal yang dihasilkan.

Konsep yang relevan untuk mengukur fenomena ini adalah *groundedness*: sejauh mana teks yang dihasilkan dapat ditelusuri langsung ke sumber informasi yang diberikan [6]. Maynez et al. [6] membedakan *groundedness* dari *factuality* (kesesuaian fakta dunia nyata) dan *faithfulness* (kesetiaan terhadap makna sumber) [9]. Penelitian *groundedness* telah banyak dilakukan dalam domain *abstractive summarization* [6][7] dan *dialogue generation* [8], namun penerapannya pada evaluasi MWP berbahasa Indonesia belum dieksplorasi. Nie et al. [13] secara khusus mengkaji *numerical hallucination* dalam LLM — kategori yang sangat relevan bagi pembuatan MWP yang menuntut akurasi kuantitatif. Pada tataran klasifikasi MWP, Prasetya & Aprianto [17] telah mengembangkan pendekatan berbasis klaster dan Transformer untuk identifikasi operasi aritmatika dalam teks soal, yang memperkuat urgensi kualitas teks MWP sebagai fondasi sistem tersebut.

Penelitian ini menggunakan model gemma3:1b dari keluarga Gemma 3 (Google DeepMind) yang diakses lokal melalui Ollama — model *compact* 1 miliar parameter yang dapat dijalankan pada GPU kelas konsumen tanpa biaya API. Pemilihan model ini didasarkan pada representativitas model open-source berparameter kecil yang relevan untuk penerapan di lingkungan pendidikan dengan sumber daya terbatas.

Berbeda dengan penelitian *groundedness* sebelumnya yang umumnya berfokus pada evaluasi *factual consistency* secara tunggal dalam domain *summarization* atau *dialogue generation* [6]–[8], penelitian ini tidak hanya mengevaluasi tingkat groundedness keluaran LLM, tetapi juga mengembangkan enam indikator *groundedness* yang mempertimbangkan aspek konsistensi entitas, peristiwa, konteks, semantik, numerik, serta relevansi pertanyaan secara bersamaan dan spesifik untuk domain *Math Word Problem* (MWP) berbahasa Indonesia. Pemilihan keenam aspek tersebut didasarkan pada karakteristik unik MWP yang menuntut konsistensi naratif sekaligus akurasi kuantitatif secara simultan — karakteristik yang belum sepenuhnya tercakup oleh kerangka groundedness maupun *faithfulness* yang telah ada, yang pada umumnya hanya menilai kesesuaian makna secara umum tanpa memisahkan dimensi numerik dan relevansi pertanyaan sebagai indikator yang independen [6], [9], [13]. Dengan demikian, kebaruan penelitian ini terletak pada pengembangan instrumen evaluasi multi-aspek yang lebih granular dan operasional, bukan sekadar penerapan konsep groundedness yang sudah ada pada domain baru.

Journal of Artificial Intelligence and Technology Information
Website : https://ejournal.techcart-press.com/index.php/jaiti
E-ISSN 2985-6396
P-ISSN 2985-5306
Volume 4, Number 3, September 2026
https://doi.org/10.58602/jaiti.v4i3.317
Page 495-505

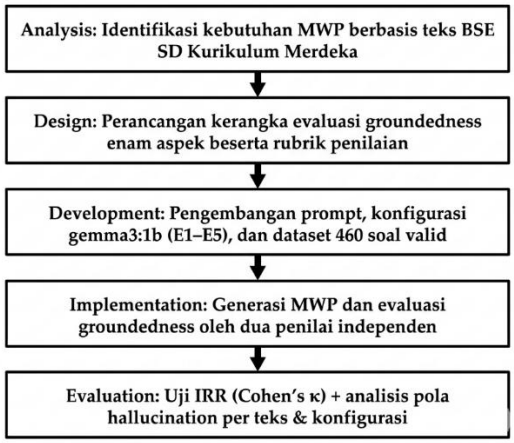
Secara lebih eksplisit, perbedaan kerangka evaluasi yang dikembangkan pada penelitian ini dengan framework *groundedness/faithfulness* sebelumnya dapat dijelaskan sebagai berikut. Maynez et al. [6] dan SummaC [7] mengukur *groundedness/faithfulness* sebagai satu skor tunggal pada level kalimat atau dokumen dalam konteks ringkasan, tanpa memisahkan sumber pelanggaran berdasarkan jenis informasi (entitas, peristiwa, angka, atau relasi semantik). Dziri et al. [9] mengusulkan taksonomi *faithfulness* yang bersifat konseptual-deskriptif, namun belum menyediakan instrumen penilaian operasional berskala biner yang dapat direplikasi secara konsisten antar penilai. Nie et al. [13] berfokus secara eksklusif pada *numerical hallucination* tanpa mengaitkannya dengan aspek naratif lain seperti konsistensi entitas atau relevansi pertanyaan. Kerangka yang dikembangkan pada penelitian ini berbeda dari ketiganya dalam tiga hal mendasar: (1) memecah *groundedness* menjadi enam indikator independen (EC, EV, CA, SA, QN, QR) yang masing-masing dinilai dan dilaporkan secara terpisah, bukan sebagai satu skor gabungan, sehingga pola pelanggaran per aspek dapat diidentifikasi secara spesifik; (2) secara eksplisit memasukkan *Question Relevance* (QR) sebagai indikator baru yang mengukur validitas logis pertanyaan matematis terhadap konten bacaan — aspek yang tidak dijumpai pada framework groundedness manapun karena task summarization dan dialogue generation tidak memiliki komponen pertanyaan matematis; dan (3) mengintegrasikan dimensi numerik (QN) sebagai indikator setara dengan dimensi semantik dan kontekstual, bukan sebagai kategori hallucination yang berdiri sendiri seperti pada Nie et al. [13]. Dengan struktur granular tersebut, kerangka ini mampu mengidentifikasi pola pelanggaran groundedness yang spesifik per aspek dan per konfigurasi parameter — sesuatu yang tidak dapat diperoleh dari skor groundedness/faithfulness tunggal pada penelitian-penelitian sebelumnya.

Berdasarkan latar belakang tersebut, penelitian ini memiliki dua tujuan spesifik: (1) mengembangkan kerangka evaluasi *groundedness* berbasis enam aspek (EC, EV, CA, SA, QN, QR) yang operasional, berindikator terukur, dan reliabel secara inter-rater untuk domain MWP berbahasa Indonesia sebagai produk utama penelitian; dan (2) mengevaluasi secara empiris tingkat *groundedness* MWP yang dihasilkan model gemma3:1b dari empat teks bacaan BSE SD Kurikulum Merdeka menggunakan lima konfigurasi parameter, guna mengidentifikasi batas kemampuan dan pola pelanggaran *groundedness* model tersebut.

2. METODE PENELITIAN

Penelitian ini menggunakan pendekatan *Research and Development* (R&D) dengan model prosedural ADDIE (*Analysis, Design, Development, Implementation, Evaluation*) [10]. Model ADDIE dipilih karena penelitian ini menghasilkan produk berupa kerangka evaluasi groundedness yang memerlukan tahapan pengembangan sistematis, mulai dari analisis kebutuhan domain, perancangan indikator dan rubrik penilaian, pengembangan instrumen serta dataset, implementasi pada model LLM, hingga evaluasi reliabilitas antar penilai; alur bertahap tersebut sesuai dengan karakteristik penelitian R&D yang berorientasi pada produk evaluatif, berbeda dengan model R&D lain yang lebih berorientasi pada pengembangan produk fisik atau media pembelajaran. Penelitian menghasilkan dua komponen produk: (1) kerangka evaluasi *groundedness* MWP berbasis enam aspek dengan rubrik penilaian berskala biner 0–1 beserta protokol inter-rater reliability (IRR); dan (2) dataset 460 soal cerita matematika valid beranotasi dari empat teks bacaan BSE SD Kurikulum Merdeka. Alur tahapan penelitian disajikan pada Gambar 1.

Journal of Artificial Intelligence and Technology Information
Website : https://ejournal.techcart-press.com/index.php/jaiti
E-ISSN 2985-6396
P-ISSN 2985-5306
Volume 4, Number 3, September 2026
https://doi.org/10.58602/jaiti.v4i3.317
Page 495-505



Gambar 1. Tahapan penelitian menggunakan model ADDIE.

Penelitian ini menggunakan model ADDIE (Analysis, Design, Development, Implementation, Evaluation) sebagai kerangka pengembangan penelitian. Pada tahap Analysis, dilakukan identifikasi kebutuhan pengembangan *Math Word Problem* (MWP) berbasis teks bacaan BSE SD Kurikulum Merdeka. Tahap Design berfokus pada perancangan kerangka evaluasi *groundedness* yang terdiri atas enam aspek beserta rubrik penilaian sebagai instrumen evaluasi. Selanjutnya, pada tahap Development, dikembangkan *prompt*, konfigurasi model gemma3:1b (E1–E5), serta penyusunan dataset yang menghasilkan 460 soal valid. Tahap Implementation dilakukan dengan menghasilkan MWP menggunakan model LLM, kemudian mengevaluasi *groundedness* setiap soal oleh dua penilai independen. Terakhir, pada tahap Evaluation, dilakukan pengujian Inter-rater Reliability (IRR) menggunakan Cohen's Kappa serta analisis pola *hallucination* berdasarkan teks bacaan dan konfigurasi parameter yang digunakan.

2.1. Kerangka Evaluasi Enam Aspek Groundedness

Enam aspek *groundedness* dikembangkan berdasarkan kajian literatur dan analisis kebutuhan domain MWP berbahasa Indonesia: (1) *Entity Consistency* (EC) — kesesuaian entitas (tokoh, objek, lokasi) dengan teks sumber; (2) *Event Consistency* (EV) — kesesuaian peristiwa dan alur kejadian; (3) *Context Alignment* (CA) — keselarasan latar dan domain aktivitas; (4) *Semantic Alignment* (SA) — kesetiaan makna dan relasi semantik antar entitas; (5) *Numerical Groundedness* (QN) — kewajaran dan kontekstualisasi angka dalam teks; dan (6) *Question Relevance* (QR) — relevansi pertanyaan matematis sebagai kelanjutan logis bacaan. Setiap aspek dinilai menggunakan skala biner 0–1 (1 = terpenuhi, 0 = tidak terpenuhi). Skor groundedness total (Gi) dihitung menggunakan persamaan $Gi = \Sigma Si(j)$, dengan $j \in \{EC, EV, CA, SA, QN, QR\}$. Persentase groundedness dihitung menggunakan rumus $(Gi/6) \times 100\%$. Suatu soal dikategorikan Grounded (G) apabila memperoleh skor minimal 66,67% (setidaknya empat dari enam aspek terpenuhi), sedangkan skor di bawah 66,67% dikategorikan sebagai Tidak Grounded (TG).

2.2. Dataset, Model, dan Konfigurasi Parameter

Empat teks bacaan dipilih secara purposif dari BSE SD Kurikulum Merdeka mewakili spektrum kompleksitas narasi: Teks 1 (*Kiki dan Cici*, sederhana/20 kalimat), Teks 2 (*Darman dan Darmin*, multi-tokoh/28 kalimat), Teks 3 (*Anike dan Bibit TOGA*, angka eksplisit/32 kalimat), dan Teks 4 (*Ditukar dengan Apa?/Hutan Kelayau*, kompleks-panjang/45 kalimat). Setiap kombinasi teks × operasi (penjumlahan, pengurangan, perkalian, pembagian) × konfigurasi (E1–E5) menghasilkan 6 soal (target 480 soal; valid: 460 soal). Lima konfigurasi parameter disajikan pada Tabel 1.

Pemilihan keempat teks bacaan didasarkan pada tiga kriteria utama, yaitu (1) kesesuaian dengan jenjang kelas III–VI SD sesuai Kurikulum Merdeka, (2) variasi jumlah kata dan tokoh untuk

Journal of Artificial Intelligence and Technology Information
Website : https://ejournal.techcart-press.com/index.php/jaiti
E-ISSN 2985-6396
P-ISSN 2985-5306
Volume 4, Number 3, September 2026
https://doi.org/10.58602/jaiti.v4i3.317
Page 495-505

merepresentasikan gradasi kompleksitas naratif, dan (3) ketersediaan konteks yang memungkinkan pembentukan empat jenis operasi aritmatika dasar (penjumlahan, pengurangan, perkalian, dan pembagian). Secara rinci, Teks 1 (*Kiki dan Cici*) memuat ±180 kata dengan 2 tokoh utama dan alur linear; Teks 2 (*Darman dan Darmin*) memuat ±260 kata dengan 4 tokoh; Teks 3 (*Anike dan Bibit TOGA*) memuat ±310 kata dengan 3 tokoh dan mengandung angka eksplisit pada narasinya; sedangkan Teks 4 (*Hutan Kelayau*) memuat ±430 kata dengan 6 tokoh dan struktur naratif berlapis (sistem barter bertingkat). Dari target 480 soal (4 teks × 4 operasi × 5 konfigurasi × 6 pengulangan), sebanyak 20 soal (4,17%) dinyatakan tidak valid dan dikeluarkan dari dataset karena format keluaran tidak sesuai templat terstruktur atau tidak memuat operasi matematis yang diminta secara eksplisit, sehingga dataset akhir yang dianalisis berjumlah 460 soal. Proses pembuatan soal dilakukan dengan mengeksekusi setiap kombinasi teks, operasi, dan konfigurasi parameter melalui model gemma3:1b yang dijalankan secara lokal melalui Ollama pada perangkat dengan GPU kelas konsumen, tanpa melibatkan proses *fine-tuning* tambahan terhadap model.

Tabel 1. Konfigurasi Parameter Model gemma3:1b

Kode	Temp.	Top-P	Rep.Pen.	Num_ctx	Rasional
E1	0,2	0,8	1,1	4096	Output paling konservatif dan deterministik terhadap teks sumber
E2	0,3	0,8	1,1	4096	Output stabil, variasi minimal
E3*	0,3	0,9	1,1	4096	Konfigurasi utama/baseline — keseimbangan terbaik (direkomendasikan)
E4	0,4	0,9	1,1	4096	Uji dampak peningkatan temperatur moderat
E5	0,5	0,9	1,1	4096	Uji dampak randomness tinggi terhadap hallucination

Tabel 1 menunjukkan lima konfigurasi parameter yang digunakan untuk mengevaluasi pengaruh variasi temperature, top-p, dan repeat penalty terhadap kualitas *groundedness* soal cerita matematika yang dihasilkan model gemma3:1b. Konfigurasi disusun dari tingkat kreativitas rendah (E1) hingga lebih tinggi (E5), sedangkan nilai **num_ctx** dipertahankan konstan sebesar 4096 agar kapasitas konteks tidak menjadi variabel yang memengaruhi hasil. Konfigurasi E3 dipilih sebagai *baseline* karena memberikan keseimbangan antara variasi keluaran dan konsistensi terhadap teks sumber.

2.3. Desain Prompt dan Prosedur Penilaian

Prompt dirancang berdasarkan prinsip *groundedness-constrained generation* [11]: model secara eksplisit diperintahkan membatasi generasi pada informasi yang tersedia dalam teks bacaan, dengan larangan eksplisit penambahan entitas, peristiwa, atau angka yang tidak kontekstual. Setiap prompt memuat: (1) instruksi peran sebagai guru matematika SD; (2) teks bacaan sumber secara lengkap; (3) instruksi operasi matematika spesifik; (4) larangan penambahan informasi; dan (5) format output terstruktur.

2.4. Uji Inter-rater Reliability

Reliabilitas antar penilai diuji menggunakan Cohen's Kappa (κ) [14][15] dengan target threshold $\kappa \geq 0{,}61$ (kategori *Substantial* menurut Landis & Koch [14]). Dua penilai independen melakukan penilaian terhadap subset 80 soal menggunakan *stratified sampling* dari total 460 soal, memastikan representasi proporsional dari setiap teks dan konfigurasi. Aspek dengan $\kappa < 0{,}61$ dikalibrasi melalui diskusi berpedoman contoh bersama hingga threshold tercapai.

Journal of Artificial Intelligence and Technology Information
Website : https://ejournal.techcart-press.com/index.php/jaiti
E-ISSN 2985-6396
P-ISSN 2985-5306
Volume 4, Number 3, September 2026
https://doi.org/10.58602/jaiti.v4i3.317
Page 495-505

3. HASIL DAN PEMBAHASAN

3.1. Groundedness per Teks Bacaan

Evaluasi dilakukan terhadap 460 soal cerita matematika valid yang dihasilkan dari empat teks bacaan, sedangkan sebanyak 80 soal dipilih secara stratified sampling untuk dianalisis secara mendalam pada setiap aspek groundedness. Skor setiap aspek kemudian dinormalisasi ke dalam persentase (0–100%) untuk memudahkan perbandingan antar teks bacaan. Tabel 2 menyajikan hasil evaluasi groundedness pada masing-masing teks bacaan.

Tabel 2. Perbandingan Skor Groundedness Antar Teks Bacaan (%)

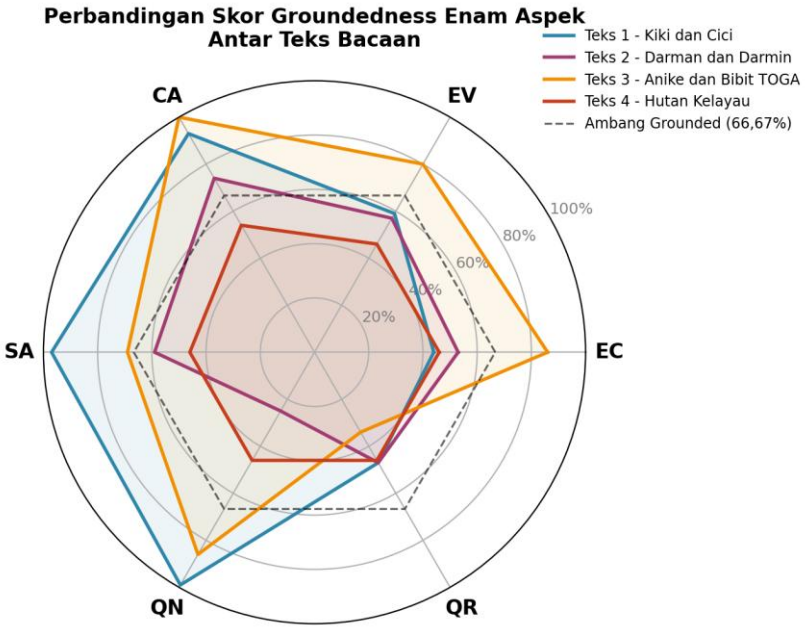
Teks Bacaan	EC	EV	CA	SA	QN	QR	Rata-rata
Teks 1 – Kiki dan Cici	44%	59%	93%	97%	99%	47%	73,25% (G)
Teks 2 – Darman dan Darmin	53%	57%	74%	59%	25%	47%	52,64% (TG)
Teks 3 – Anike dan Bibit TOGA	86%	80%	100%	69%	86%	34%	75,91% (G)
Teks 4 – Hutan Kelayau	46%	46%	54%	46%	46%	46%	47,22% (TG)

Keterangan: G = Grounded (rata-rata ≥66,67%); TG = Tidak Grounded (rata-rata <66,67%).

Teks 1 (*Kiki dan Cici*, sederhana) memperoleh skor 73,25% karena alur narasi yang ringkas dan linear. Kelemahan utama muncul pada EC (44%) akibat penambahan objek tematik yang tidak ada di bacaan, dan QR (47%) akibat kegagalan membangun relasi satu-ke-banyak untuk operasi perkalian. Teks 2 (*Darman dan Darmin*, multi-tokoh) mengalami penurunan ke 52,64% (Tidak Grounded). Penyebab utama adalah pencampuran atribut antar tokoh (EC: 53%) dan kegagalan total aspek QN (25%) yang menghasilkan angka fiktif tidak wajar.

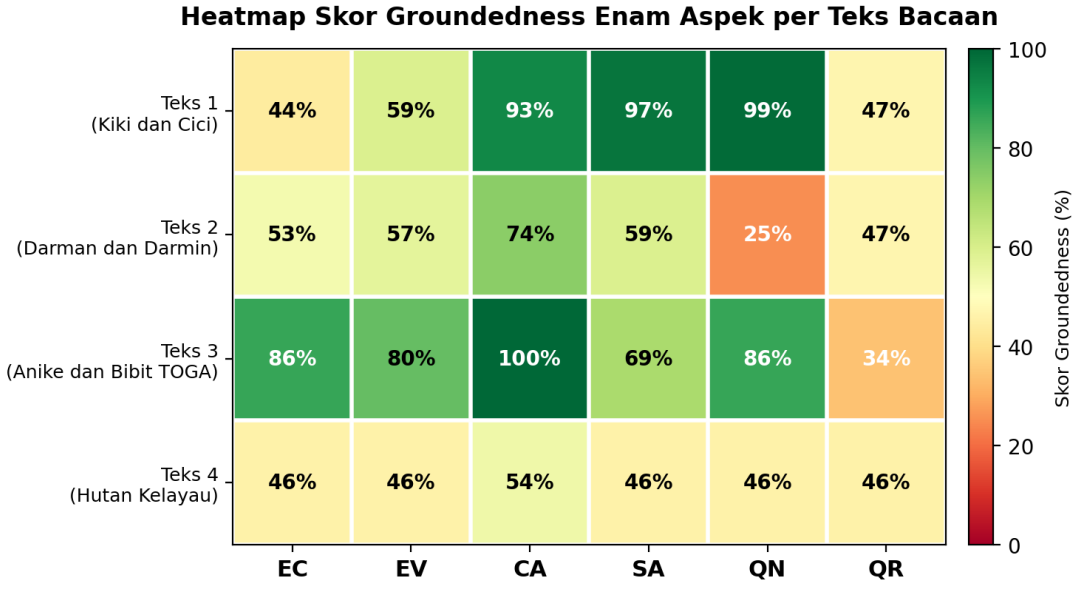
Teks 3 (*Anike dan Bibit TOGA*, angka eksplisit) mencapai skor tertinggi 75,91%. Namun, keberadaan angka eksplisit justru memicu pola baru *relational numerical hallucination* — penggunaan angka faktual dari bacaan dalam relasi semantik yang tidak valid — yang menekan skor SA dan QR. Teks 4 (Hutan Kelayau, kompleks-panjang) memperoleh skor terendah sebesar 47,22% dan dikategorikan sebagai Tidak Grounded. Penurunan performa ini diduga berkaitan dengan meningkatnya kompleksitas narasi, banyaknya tokoh, alur barter bertingkat, serta keterbatasan model dalam mempertahankan konsistensi konteks. Kondisi tersebut menyebabkan munculnya berbagai bentuk pelanggaran groundedness, seperti perubahan konteks (*context shift*), inkonsistensi entitas, dan relasi numerik yang tidak sesuai, sehingga menurunkan skor pada hampir seluruh aspek evaluasi. Temuan tersebut menunjukkan bahwa peningkatan kompleksitas narasi berpengaruh terhadap kemampuan model dalam mempertahankan groundedness. Hasil ini sejalan dengan Ji et al. [5] yang menyatakan bahwa model LLM berukuran kecil lebih rentan mengalami *hallucination* ketika memproses informasi dengan hubungan semantik dan numerik yang kompleks. Dengan demikian, semakin kompleks struktur bacaan, semakin besar peluang model menghasilkan keluaran yang tidak sepenuhnya sesuai dengan teks sumber.

Journal of Artificial Intelligence and Technology Information
Website : https://ejournal.techcart-press.com/index.php/jaiti
E-ISSN 2985-6396
P-ISSN 2985-5306
Volume 4, Number 3, September 2026
https://doi.org/10.58602/jaiti.v4i3.317
Page 495-505



Gambar 2. Radar chart perbandingan skor groundedness enam aspek (EC, EV, CA, SA, QN, QR) antar empat teks bacaan, dengan garis putus-putus menandai ambang batas Grounded (66,67%).

Gambar 2 memvisualisasikan pola perbedaan skor keenam aspek *groundedness* secara simultan antar teks bacaan. Visualisasi ini memperjelas temuan pada Tabel 2, yaitu bahwa Teks 4 secara konsisten berada di bawah ambang batas Grounded pada hampir seluruh aspek (bentuk radar yang relatif menyempit dan merata di sekitar 46–54%), sedangkan Teks 1 dan Teks 3 menunjukkan profil yang timpang — sangat kuat pada aspek CA, SA, dan QN, namun lemah pada EC dan QR. Pola ini menegaskan bahwa kegagalan groundedness pada model gemma3:1b tidak bersifat seragam pada seluruh aspek, melainkan spesifik pada kombinasi aspek dan karakteristik teks tertentu, sehingga evaluasi berbasis rata-rata tunggal berisiko menyamarkan pola pelanggaran yang sesungguhnya terjadi.



Gambar 3. Heatmap skor groundedness enam aspek (EC, EV, CA, SA, QN, QR) untuk masing-masing teks bacaan, sebagai visualisasi komplementer dari Gambar 2.

Journal of Artificial Intelligence and Technology Information
Website : https://ejournal.techcart-press.com/index.php/jaiti
E-ISSN 2985-6396
P-ISSN 2985-5306
Volume 4, Number 3, September 2026
https://doi.org/10.58602/jaiti.v4i3.317
Page 495-505

Gambar 3 menyajikan representasi *heatmap* dari data yang sama pada Tabel 2 dan Gambar 2, dengan gradasi warna dari merah (skor rendah) hingga hijau (skor tinggi) untuk mempermudah identifikasi sel bermasalah secara sekilas. Pola yang tampak konsisten dengan radar chart: kolom QN pada Teks 2 (25%) dan kolom EC pada Teks 1 (44%) tampil sebagai sel merah gelap yang menandakan titik pelanggaran groundedness paling kritis, sedangkan baris Teks 4 menunjukkan gradasi warna oranye yang relatif homogen di seluruh kolom — mengonfirmasi bahwa kegagalan model pada teks kompleks bersifat menyeluruh dan tidak terkonsentrasi pada aspek tertentu. Kombinasi radar chart dan heatmap ini memberikan dua sudut pandang saling melengkapi: radar chart menonjolkan profil relatif antar aspek dalam satu teks, sedangkan heatmap memudahkan perbandingan lintas teks pada aspek yang sama.

3.2. Hasil Uji Inter-rater Reliability

Tabel 3 menunjukkan bahwa seluruh aspek groundedness memperoleh nilai Cohen's Kappa (κ) di atas 0,61, sehingga memenuhi kategori Substantial Agreement hingga Almost Perfect Agreement berdasarkan kriteria Landis dan Koch [14]. Hasil ini menunjukkan bahwa rubrik penilaian yang dikembangkan memiliki tingkat konsistensi yang baik antar penilai dan layak digunakan sebagai instrumen evaluasi groundedness.

Tabel 3. Hasil Uji Inter-rater Reliability (Cohen's Kappa)

Aspek Groundedness	Nilai κ	Interpretasi	Status
EC – Entity Consistency	0,79	Substantial	Diterima
EV – Event Consistency	0,74	Substantial	Diterima
CA – Context Alignment	0,81	Almost Perfect	Diterima
SA – Semantic Alignment	0,68	Substantial	Diterima
QN – Numerical Groundedness	0,75	Substantial	Diterima
QR – Question Relevance	0,72	Substantial	Diterima

Seluruh nilai Cohen's Kappa berada di atas ambang batas 0,61 sehingga instrumen dinyatakan memiliki reliabilitas yang memadai dan dapat digunakan pada proses evaluasi groundedness. Aspek Context Alignment (CA) memperoleh nilai κ tertinggi sebesar 0,81 (Almost Perfect Agreement). Tingginya kesepakatan ini menunjukkan bahwa indikator kesesuaian konteks, seperti latar, objek, dan domain aktivitas, relatif mudah diidentifikasi secara konsisten oleh kedua penilai. Sebaliknya, Semantic Alignment (SA) memperoleh nilai κ terendah sebesar 0,68 (Substantial Agreement). Perbedaan ini mengindikasikan bahwa penilaian terhadap kesesuaian makna dan relasi semantik memerlukan interpretasi yang lebih mendalam sehingga berpotensi menimbulkan variasi penilaian antar evaluator. Secara keseluruhan, hasil uji reliabilitas mengonfirmasi bahwa kerangka evaluasi groundedness enam aspek yang dikembangkan memiliki konsistensi penilaian yang baik. Temuan ini mendukung penggunaan kerangka tersebut sebagai instrumen evaluasi untuk mengidentifikasi berbagai bentuk pelanggaran groundedness pada soal cerita matematika yang dihasilkan oleh LLM, termasuk pelanggaran yang tidak dapat diukur menggunakan metrik otomatis seperti BLEU maupun BERTScore [6], [7].

3.3. Analisis Pola Hallucination

Analisis kualitatif mengidentifikasi empat pola utama pelanggaran *groundedness*. Pertama, *entity hallucination*: pada teks sederhana bersifat *extrinsic* (penambahan objek tematik yang tidak ada di bacaan); pada teks kompleks bersifat *intrinsic* (penggantian tokoh yang sudah ditetapkan dalam narasi) [5]. Kedua, *numerical hallucination* dalam dua sub-pola: angka fiktif tidak wajar (*extrinsic*, dominan Teks

Journal of Artificial Intelligence and Technology Information
Website : https://ejournal.techcart-press.com/index.php/jaiti
E-ISSN 2985-6396
P-ISSN 2985-5306
Volume 4, Number 3, September 2026
https://doi.org/10.58602/jaiti.v4i3.317
Page 495-505

4, misal "1.000 batang kangkung") dan *relational numerical hallucination* (*intrinsic*, dominan Teks 3) — temuan baru yang memperluas kategorisasi Nie et al. [13] dengan menambahkan dimensi relasional pada *hallucination* numerik.

Ketiga, kegagalan sistematis QR pada operasi perkalian di seluruh konfigurasi (E1–E5) dan seluruh teks bacaan. Model gemma3:1b tidak mampu mengonstruksi relasi satu-ke-banyak (*one-to-many*) dari narasi nonkuantitatif — kegagalan intrinsik yang hanya dapat diatasi melalui *few-shot prompting* dengan template relasi eksplisit [11]. Keempat, *context shift*: pada Teks 4 bersifat *intrinsic* (sistem barter dikonversi menjadi transaksi konvensional) akibat keterbatasan *effective context window* model compact [5].

3.4. Analisis Konfigurasi Parameter

Pada teks sederhana–menengah (Teks 1–3), perbedaan skor antar konfigurasi kecil (<5%), menunjukkan stabilitas *groundedness* dalam rentang parameter yang diuji. Pada Teks 4, konfigurasi E1 (temperature=0,2) menghasilkan skor 3–5% lebih tinggi dibanding E5 (temperature=0,5), mengkonfirmasi bahwa temperatur lebih rendah membantu mengurangi *hallucination* pada beban kontekstual tinggi. Konfigurasi E3 (temperature=0,3; top_p=0,9; repeat_penalty=1,1) secara konsisten memberikan keseimbangan terbaik antara variasi leksikal dan konsistensi *groundedness* di semua kondisi pengujian. Temuan ini menunjukkan bahwa pengaturan parameter decoding berpengaruh terhadap kemampuan model dalam mempertahankan groundedness, terutama ketika memproses teks dengan kompleksitas narasi yang tinggi. Temperatur yang lebih rendah menghasilkan keluaran yang lebih konservatif sehingga mampu mengurangi munculnya hallucination, sedangkan peningkatan temperatur cenderung meningkatkan variasi keluaran namun berisiko menurunkan konsistensi terhadap teks sumber. Hasil ini sejalan dengan Wang et al. [12], yang menyatakan bahwa penalaran matematis bertahap (multi-step reasoning) masih menjadi tantangan utama bagi model bahasa, khususnya model berukuran kecil yang memiliki keterbatasan dalam mempertahankan relasi numerik dan semantik secara bersamaan. Dengan demikian, pemilihan konfigurasi parameter menjadi faktor penting dalam meminimalkan hallucination ketika LLM digunakan untuk menghasilkan soal cerita matematika berbasis teks bacaan.

Secara mekanistis, nilai *top-p* 0,9 pada konfigurasi E3 memberikan ruang variasi leksikal yang cukup luas untuk menghasilkan kalimat soal yang natural, namun tetap membatasi pemilihan token pada kumpulan probabilitas kumulatif tertinggi sehingga token dengan probabilitas sangat rendah — yang berpotensi memunculkan entitas atau angka tidak relevan — cenderung tidak terpilih. Nilai *top-p* yang lebih rendah (0,8) pada E1 dan E2 menghasilkan keluaran yang lebih kaku dan repetitif, sedangkan *top-p* yang sama (0,9) pada E4 dan E5 apabila dikombinasikan dengan *temperature* lebih tinggi terbukti meningkatkan risiko *hallucination* karena distribusi probabilitas token menjadi lebih rata, sehingga token-token di luar konteks bacaan memiliki peluang lebih besar untuk terpilih. Adapun *temperature* 0,3 pada E3 dinilai optimum karena berada pada titik keseimbangan antara determinisme — *temperature* rendah seperti pada E1 (0,2) menghasilkan variasi kalimat yang terlalu terbatas dan cenderung mengulang pola templat — dan keacakan berlebih — *temperature* tinggi seperti pada E5 (0,5) meningkatkan kecenderungan model "berimprovisasi" di luar informasi teks sumber. Secara umum, hubungan antara *temperature* dan *hallucination* bersifat tidak linear namun cenderung positif pada beban kontekstual tinggi: peningkatan *temperature* memperbesar entropi distribusi token, yang pada teks dengan struktur naratif kompleks (Teks 4) meningkatkan peluang model memilih token yang secara statistik masuk akal tetapi tidak *grounded* pada teks sumber, sebagaimana ditunjukkan oleh selisih skor 3–5% antara E1 dan E5 pada Teks 4. Sebaliknya, pada teks sederhana–menengah (Teks 1–3), pengaruh *temperature* terhadap *hallucination* relatif kecil karena ruang kemungkinan token yang valid secara kontekstual masih cukup terbatas sehingga model tetap terarah meskipun *temperature* dinaikkan.

Perbandingan langsung antar konfigurasi pada Tabel 1 memperkuat argumen di atas. E1 dan E2 menggunakan *top-p* 0,8 sedangkan E3, E4, dan E5 menggunakan *top-p* 0,9; dengan *temperature* ditahan konstan pada 0,3 antara E2 dan E3, satu-satunya variabel yang berubah adalah *top-p*. Kenaikan *top-p*

Journal of Artificial Intelligence and Technology Information
Website : https://ejournal.techcart-press.com/index.php/jaiti
E-ISSN 2985-6396
P-ISSN 2985-5306
Volume 4, Number 3, September 2026
https://doi.org/10.58602/jaiti.v4i3.317
Page 495-505

dari 0,8 ke 0,9 pada perbandingan E2 vs E3 memperluas kumpulan token kandidat dari kelompok probabilitas kumulatif yang lebih sempit menjadi lebih luas, sehingga model memperoleh sedikit ruang tambahan untuk memilih padanan kata atau struktur kalimat yang lebih variatif tanpa harus keluar dari token-token berprobabilitas tinggi yang umumnya selaras dengan konteks bacaan. Secara empiris, perluasan terbatas ini tidak menurunkan konsistensi terhadap teks sumber karena probabilitas kumulatif 0,9 masih mengecualikan ekor distribusi token yang berprobabilitas sangat rendah — wilayah yang paling sering menjadi sumber *hallucination* entitas maupun angka tidak relevan. Nilai *top-p* yang lebih tinggi lagi (mendekati 1,0) tidak diujikan karena berisiko menghapus batas atas probabilitas kumulatif tersebut, sehingga token berprobabilitas rendah berpeluang lebih besar untuk terpilih terutama ketika dikombinasikan dengan *temperature* di atas 0,3. Kombinasi *temperature*=0,3 dan *top-p*=0,9 pada E3 dinilai sebagai titik optimal karena keduanya bekerja saling melengkapi: *temperature* 0,3 mengendalikan bentuk keseluruhan distribusi probabilitas agar tidak terlalu datar (mencegah pemilihan token berisiko tinggi), sementara *top-p* 0,9 membatasi wilayah pengambilan sampel pada rentang probabilitas kumulatif tertinggi (mencegah ekor distribusi ikut terpilih). Ketika salah satu parameter dinaikkan tanpa yang lain — misalnya *temperature* dinaikkan pada E4 dan E5 dengan *top-p* tetap 0,9 — efek pengendali dari *top-p* tidak lagi cukup untuk mengimbangi pelebaran distribusi akibat *temperature* yang lebih tinggi, sebagaimana ditunjukkan oleh penurunan skor pada Teks 4 dari E3 ke E5. Dengan demikian, keunggulan E3 bukan berasal dari nilai *top-p* atau *temperature* secara terpisah, melainkan dari interaksi keduanya yang bersama-sama membatasi ruang token kandidat pada wilayah yang secara statistik paling konsisten dengan teks sumber.

4. KESIMPULAN

Penelitian ini menghasilkan dua kontribusi utama. Pertama, kerangka evaluasi *groundedness* MWP berbasis enam aspek (EC, EV, CA, SA, QN, QR) terbukti operasional dan reliabel dengan Cohen's Kappa 0,68–0,81 (*Substantial* hingga *Almost Perfect*). Kerangka ini mampu menangkap nuansa pelanggaran *groundedness* yang tidak terdeteksi oleh metrik evaluasi otomatis. Kedua, model gemma3:1b memadai untuk teks sederhana–menengah (Teks 1: 73,25%; Teks 3: 75,91%) namun tidak dapat diandalkan untuk teks kompleks (Teks 4: 47,22%), dengan rata-rata keseluruhan 62,25%. Terdapat korelasi negatif yang kuat antara kompleksitas narasi dan tingkat *groundedness*.

Secara teoretis, penelitian ini memberikan kontribusi berupa perluasan konsep *groundedness*, yang sebelumnya banyak diterapkan secara tunggal pada tugas *summarization* dan *dialogue generation*, menjadi kerangka multi-dimensi yang memisahkan aspek konsistensi entitas, peristiwa, konteks, semantik, numerik, dan relevansi pertanyaan sebagai indikator yang independen namun saling melengkapi. Kontribusi teoretis lain adalah diperkenalkannya kategori *relational numerical hallucination* sebagai perluasan taksonomi *hallucination* numerik yang sebelumnya dikemukakan oleh Nie et al. [13], yang menunjukkan bahwa *hallucination* pada domain MWP tidak hanya terjadi pada level penyebutan angka (angka fiktif), tetapi juga pada level relasi semantik-numerik antar entitas dalam teks. Temuan ini memperkaya pemahaman teoretis mengenai bagaimana model bahasa berukuran kecil (*compact LLM*) gagal mempertahankan keterhubungan informasi kuantitatif dan naratif secara simultan, sehingga dapat menjadi landasan bagi pengembangan taksonomi *hallucination* yang lebih komprehensif pada tugas generasi teks berbasis konteks di luar domain MWP.

Empat pola pelanggaran teridentifikasi: *entity hallucination*, *relational numerical hallucination* (temuan baru yang memperluas kategorisasi), kegagalan sistematis QR pada operasi perkalian, dan *context shift*. Konfigurasi E3 direkomendasikan sebagai parameter optimal. Praktisi pendidikan disarankan menggunakan model LLM *compact* hanya untuk teks naratif sederhana–menengah (≤300 kata, ≤3 tokoh utama) dengan verifikasi manual wajib pada aspek SA, QN, dan QR. Penelitian selanjutnya dapat memperluas evaluasi dengan membandingkan performa beberapa model Large Language Model (LLM) open-source yang memiliki ukuran parameter berbeda, menerapkan teknik prompting yang lebih kompleks seperti few-shot prompting atau chain-of-thought prompting, serta menguji kerangka groundedness yang dikembangkan pada dataset yang lebih beragam. Pengembangan tersebut diharapkan mampu meningkatkan kualitas generasi soal cerita matematika sekaligus memperluas

Journal of Artificial Intelligence and Technology Information
Website : https://ejournal.techcart-press.com/index.php/jaiti
E-ISSN 2985-6396
P-ISSN 2985-5306
Volume 4, Number 3, September 2026
https://doi.org/10.58602/jaiti.v4i3.317
Page 495-505

penerapan kerangka evaluasi groundedness pada domain pendidikan. Dengan demikian, kerangka evaluasi groundedness yang dikembangkan tidak hanya dapat digunakan untuk mengevaluasi kualitas soal cerita matematika yang dihasilkan LLM, tetapi juga berpotensi menjadi acuan evaluasi pada tugas generasi teks pendidikan lainnya yang berbasis konteks.

REFERENCES

[1] OECD, PISA 2022 Results (Volume I): The State of Learning and Equity in Education, OECD Publishing, 2023. https://doi.org/10.1787/53f23881-en

[2] L. Verschaffel, W. Van Dooren, B. Greer, and S. Mukhopadhyay, "Reconceptualising word problems as exercises in mathematical modelling," J. für Mathematik-Didaktik, vol. 31, no. 1, pp. 9–29, 2010. https://doi.org/10.1007/s13138-010-0007-x

[3] Kementerian Pendidikan, Kebudayaan, Riset, dan Teknologi, Panduan Pembelajaran dan Asesmen Kurikulum Merdeka, Kemendikbudristek, 2022. https://kurikulum.kemdikbud.go.id

[4] T. B. Brown et al., "Language models are few-shot learners," Adv. Neural Inf. Process. Syst., vol. 33, pp. 1877–1901, 2020. https://arxiv.org/abs/2005.14165

[5] Z. Ji et al., "Survey of hallucination in natural language generation," ACM Comput. Surv., vol. 55, no. 12, pp. 1–38, 2023. https://doi.org/10.1145/3571730

[6] J. Maynez, S. Narayan, B. Bohnet, and R. McDonald, "On faithfulness and factuality in abstractive summarization," in Proc. 58th ACL, 2020, pp. 1906–1919. https://doi.org/10.18653/v1/2020.acl-main.173

[7] P. Laban, T. Schneider, C.-S. Wu, and M. A. Hearst, "SummaC: Re-visiting NLI-based models for inconsistency detection in summarization," Trans. Assoc. Comput. Linguist., vol. 10, pp. 163–177, 2022. https://doi.org/10.1162/tacl_a_00453

[8] H. Rashkin, E. M. Smith, M. Li, and Y.-L. Boureau, "Increasing faithfulness in knowledge-grounded dialogue with controllable features," in Proc. 59th ACL-IJCNLP, 2021, pp. 763–777. https://doi.org/10.18653/v1/2021.acl-long.58

[9] N. Dziri, H. Rashkin, T. Shi, and K. McKeown, "Faithfulness in natural language generation: A systematic survey," arXiv, 2022. https://arxiv.org/abs/2203.05227

[10] R. M. Branch, Instructional Design: The ADDIE Approach, Springer, 2009. https://doi.org/10.1007/978-0-387-09506-6

[11] S. Mishra, D. Khashabi, C. Baral, Y. Choi, and H. Hajishirzi, "Reframing instructional prompts to GPTk's language," in Findings ACL 2022, pp. 589–612. https://doi.org/10.18653/v1/2022.findings-acl.50

[12] P. Wang, L. Pang, M. Lan, J. Shen, S. Cheng, and X. Cheng, "Math-NLP: Mathematical reasoning in large language models," arXiv, 2023. https://arxiv.org/abs/2304.01399

[13] Y. Nie, Y. Tian, M. Bansal, and R. Rudinger, "Do large language models know what they don't know? A survey on numerical reasoning," arXiv, 2023. https://arxiv.org/abs/2305.09898

[14] J. R. Landis and G. G. Koch, "The measurement of observer agreement for categorical data," Biometrics, vol. 33, no. 1, pp. 159–174, 1977. https://doi.org/10.2307/2529310

[15] M. L. McHugh, "Interrater reliability: The kappa statistic," Biochemia Medica, vol. 22, no. 3, pp. 276–282, 2012. https://doi.org/10.11613/BM.2012.031

[16] L. Ouyang et al., "Training language models to follow instructions with human feedback," Adv. Neural Inf. Process. Syst., vol. 35, pp. 27730–27744, 2022. https://arxiv.org/abs/2203.02155

[17] A. Prasetya and M. D. Aprianto, "Structural classification of Indonesian arithmetic word problems using hierarchical agglomerative clustering," JUKTISI, 2026. https://doi.org/10.62712/juktisi.v5i1.1026