Indonesian Journal of Computational Language Studies
Vol. 1, No. 2, June 2026, pp. 77~92
ISSN: 3163-7884, DOI:

Do large language models reason consistently across Indonesian languages? a multilingual evaluation of Indonesian, Javanese, Sundanese, and Buginese

Fitriya Dessi Wulandari
Universitas Muhammadiyah Surakarta, Indonesia

Article Info

Article history:

Received Apr 21, 2026
Revised May 26, 2026
Accepted Jun 29, 2026

Keywords:

Buginese;
cross-linguistic consistency;
Indonesian languages;
large language models;
multilingual reasoning.

ABSTRACT

Background: Multilingual large language model evaluation increasingly recognises Indonesian as a significant benchmark language, yet Indonesia's wider linguistic ecology requires assessment across local languages whose digital representation remains uneven. **Objective:** This study examines whether large language models reason consistently across Indonesian, Javanese, Sundanese, and Buginese when semantically aligned reasoning tasks are presented in each language. **Method:** Using a controlled multilingual evaluation design, this study compares model outputs through three analytic dimensions: final-answer consistency, explanation coherence and faithfulness, and language-linked error patterns. **Results:** Findings show that answer stability is not evenly preserved across language pairs, with stronger alignment in Indonesian–Javanese and Indonesian–Sundanese comparisons than in Indonesian–Buginese comparison. Explanation analysis further indicates that apparently correct answers may be accompanied by compressed, partially faithful, or weakly grounded reasoning. **Implication:** Error analysis reveals that inconsistency emerges through multiple pathways, including lexical-semantic drift, register mismatch, translation-related distortion, cultural inferential misreading, and language fallback. **Novelty:** The novelty of this study lies in shifting Indonesian multilingual LLM evaluation from isolated benchmark accuracy to cross-linguistic reasoning consistency, positioning local languages as central analytic sites for assessing reliability, equity, and epistemic accountability in multilingual artificial intelligence.

Corresponding Author:

Fitriya Dessi Wulandari
Department of English Education, Universitas Muhammadiyah Surakarta, Indonesia
Jl. A. Yani, Mendungan, Pabelan, Kec. Kartasura, Sukoharjo, Jawa Tengah, 57169, Indonesia
Email: fitriyadessi@gmail.com

1. INTRODUCTION

Indonesia offers a crucial test case for multilingual reasoning in large language models because linguistic plurality is not a peripheral condition but a central feature of social life. Badan Bahasa reports 718 regional languages in Indonesia, with only a limited number categorised as safe and several already vulnerable, endangered, critically endangered, or extinct. This sociolinguistic reality matters for artificial intelligence because language technologies increasingly mediate education, administration, search, translation, and public knowledge access, while their evaluation still tends to privilege high-resource or nationally dominant languages. Indonesian has begun to receive stronger benchmark attention, yet Indonesian linguistic life cannot be reduced to Indonesian alone. Javanese, Sundanese, and Buginese represent distinct regional, cultural, and grammatical ecologies, not merely alternative lexical surfaces. If a model answers correctly in Indonesian but shifts its judgement when the same reasoning problem is expressed in another

Indonesian language, multilingual competence becomes unstable. This study is therefore motivated by a practical and epistemic problem: whether LLMs reason consistently across languages that coexist within one multilingual nation.

Recent work has significantly advanced Indonesian and Southeast Asian LLM evaluation. IndoMMLU introduced a large multitask benchmark covering Indonesian educational knowledge and local linguistic-cultural content, showing that Indonesian-centred evaluation can expose limitations hidden by English benchmarks [1]. COPAL-ID moved beyond translated causal reasoning by constructing culturally grounded Indonesian commonsense items from scratch [2]. IndoCulture further demonstrated that commonsense reasoning is shaped by geographical and cultural contexts across Indonesian provinces [3]. Broader regional suites such as BHASA and SEA-HELM have argued for more holistic linguistic, cultural, safety, and reasoning evaluation in Southeast Asia [4], [5]. Indonesian-oriented models and resources, including Cendol, SeaLLMs, and LoraxBench, have also expanded the evaluation landscape for local languages [6], [7], [8]. Yet most existing work still measures task performance within languages or across benchmarks, rather than directly testing whether the same reasoning decision remains stable across Indonesian, Javanese, Sundanese, and Buginese.

This study addresses that gap by evaluating cross-linguistic reasoning consistency rather than treating multilingual evaluation as a sequence of separate accuracy tests. The central question is whether selected LLMs produce equivalent final answers, coherent explanations, and stable reasoning decisions when semantically aligned tasks are presented in Indonesian, Javanese, Sundanese, and Buginese. More specifically, this study asks: to what extent do LLMs maintain answer consistency across these languages; which reasoning categories show the greatest instability; and how do explanation coherence and language-specific error patterns differ across the four language versions? The design follows a controlled multilingual evaluation logic in which each item is aligned across languages, but not reduced to mechanical translation. This distinction is important because reasoning tasks may involve pragmatic meaning, cultural reference, register, lexical specificity, or local inferential norms that cannot be preserved through literal equivalence alone. This study therefore evaluates visible model behaviour through final-answer alignment, explanation coherence, and language-linked error typology.

The working argument is that multilingual reasoning should be assessed as consistency under linguistic variation, not only as correctness within isolated language settings. If LLMs rely on uneven cross-lingual representations, translation-like mediation, or dominant-language semantic anchors, their answers may appear competent in Indonesian while becoming less stable in Javanese, Sundanese, or Buginese. Studies on multilingual transfer, cross-lingual inconsistency, and culturally grounded evaluation suggest that models can display uneven capability across languages even when prompts are semantically related [9], [10], [11], [12]. This study does not claim that inconsistency is caused solely by low-resource status, nor does it infer human-like cognition from model outputs. Rather, it tests whether observable reasoning behaviour remains aligned across four Indonesian languages. The implication is methodological: a model should not be considered robustly multilingual for Indonesia if its reasoning changes when linguistic access shifts from Indonesian to local languages.

2. LITERATURE REVIEW

2.1. Multilingual reasoning

Multilingual reasoning in large language models is best understood as observable problem-solving behaviour across languages, not as proof that models possess human-like cognition. Recent surveys frame multilingual LLMs through corpora, alignment, transfer, and evaluation bias, but they differ in whether multilingual capacity is treated as language coverage, task accuracy, representational alignment, or reasoning robustness [12], [13]. This distinction matters because a model may answer in many languages while relying on uneven latent representations or English-centred semantic mediation. Studies on cross-lingual thought prompting, factual reasoning, and multilingual chain-of-thought further show that reasoning performance may vary when tasks move across linguistic contexts [9], [14], [15]. Multilingual reasoning therefore requires evaluation beyond surface fluency.

Its main indicators include final-answer correctness, inferential stability, explanation coherence, and faithfulness between stated reasoning and final response. Accuracy remains necessary, but recent work suggests that correctness alone cannot capture whether a model reasons comparably across languages. Multilingual long-context evaluation links reasoning to retrieval stability and information use [16], while MMLU-ProX and MuBench extend evaluation across broader language sets and task types [17], [18]. Studies on low-resource and morphologically rich languages show that task difficulty may interact with morphology, alignment, and resource imbalance [9]. For this study, multilingual reasoning is operationalised through task-level behaviour rather than general multilingual ability.

2.2. Cross-linguistic consistency

Cross-linguistic consistency refers to whether a model preserves the same decision, semantic interpretation, and reasoning trajectory when equivalent prompts are presented in different languages. This differs from cross-lingual transfer, which usually concerns whether knowledge or performance learned in one language benefits another. Transfer research has examined alignment, adaptation, and language transferability [19], [20], [6], whereas consistency research asks a stricter question: does the model remain stable under linguistic variation? Work on multilingual inconsistency, word-sense disambiguation, and translation-based evaluation shows that semantically related prompts may still elicit divergent answers [11], [10]. Consistency is thus a reliability problem, not only a capability problem.

Indicators of cross-linguistic consistency can be grouped into answer-level, explanation-level, and error-level dimensions. Answer-level consistency concerns whether final outputs are equivalent, partially equivalent, contradictory, or invalid. Explanation-level consistency examines whether the model uses comparable premises, preserves relevant relations, and avoids language-induced contradictions. Error-level consistency tracks whether failures cluster around lexical gaps, register mismatch, code-mixing sensitivity, or cultural misinterpretation. Research on cross-lingual alignment suggests that performance gaps often emerge from uneven representational mapping rather than task difficulty alone [21], [22]. Sparse-autoencoder analysis and linguistic-region studies further suggest that language-specific activation patterns may shape model behaviour across typologically diverse languages [17], [23].

2.3. Indonesian and local languages

Indonesian and local languages provide a demanding setting for testing these issues because linguistic plurality intersects with uneven digital representation. IndoMMLU established Indonesian educational and cultural knowledge as a serious benchmark domain [1], while COPAL-ID and IndoCulture demonstrated that commonsense and causal reasoning cannot be detached from local cultural knowledge [2], [3]. Regional evaluation suites such as BHASA and SEA-HELM widened this agenda to Southeast Asian linguistic and cultural contexts [4], [5]. However, much work still privileges Indonesian or evaluates local languages as separate benchmark entries. This leaves unresolved whether LLMs maintain reasoning consistency across Indonesian, Javanese, Sundanese, and Buginese.

The relevant indicators for this linguistic setting include language coverage, task equivalence, cultural grounding, resource asymmetry, and local-language robustness. Cendol, SeaLLMs, SeaLLMs 3, and Sailor show growing attention to Southeast Asian and Indonesian-oriented generative models [6], [7], [23], [24]. LoraxBench, IndoRobusta, NusaBERT, and studies on Javanese, Sundanese, and Indonesian local-language translation expand the empirical basis for evaluating local-language NLP [8], [25], [26], [27]. Yet this literature also shows why this study is needed: robust multilingual evaluation for Indonesia must test whether reasoning remains stable when linguistic access shifts from Indonesian to local languages.

3. METHOD

This study uses document-level and item-level multilingual evaluation data as its unit of analysis. The material object consists of public benchmark datasets, dataset cards, repositories, article records, and contextual documentation relevant to Indonesian, Javanese, Sundanese, and Buginese language evaluation. The corpus is organised into three linked files: metadata, text records, and coding records. The source inventory contains 13 verified public entries, comprising seven primary sources, five secondary sources, and one contextual source. The coding structure contains 28 language-specific coding rows, prepared for Indonesian, Javanese, Sundanese, and Buginese comparisons. This composition allows this study to evaluate reasoning consistency without treating Indonesian as a substitute for local-language performance.

This study applies a controlled multilingual evaluation design. Instead of measuring each language as an isolated benchmark condition, it aligns equivalent reasoning tasks across Indonesian, Javanese, Sundanese, and Buginese. This design is informed by recent work on Indonesian reasoning benchmarks, cultural commonsense evaluation, Southeast Asian LLM assessment, and multilingual consistency [1], [2], [3], [4], [5], [11]. The design is comparative but not causal. It evaluates whether model outputs remain stable across language versions, while avoiding claims that any observed asymmetry is caused by training-data scarcity alone.

Information sources were selected from public and verifiable repositories, including Hugging Face dataset cards, GitHub repositories, ACL-related paper records, journal article pages, arXiv records, and official language-policy documentation. Primary sources provide candidate items, parallel text, or reasoning-task structures; secondary sources provide evaluation frameworks, model-context information, and robustness categories; contextual sources situate Indonesian multilingualism within a broader sociolinguistic environment. IndoMMLU, COPAL-ID, IndoCulture, NusaX, FLORES-200, and LoraxBench serve as core sources because they directly support benchmark construction, language alignment, or local-language comparison [8], [6], [7], [25].

Table 1. Composition of dataset and corpus sources

Code	Type of Data	Source	Location / Scope	Period	Quantity	Characteristics
IDLLM-PRIM-001	Primary benchmark	IndoMMLU	Indonesia	2023	1 source record	Indonesian knowledge and reasoning benchmark
IDLLM-PRIM-002	Primary reasoning dataset	COPAL-ID	Indonesia	2023–2024	1 source record; 559 instances reported by source	Indonesian causal commonsense reasoning
IDLLM-PRIM-003	Primary cultural reasoning dataset	IndoCulture	Eleven Indonesian provinces	2024	1 source record	Geographically influenced cultural commonsense reasoning
IDLLM-PRIM-004	Primary parallel corpus	NusaX-MT	Indonesian local-language communities	2022–2023	1 source record	Parallel text for Indonesian, Javanese, Sundanese, and Buginese
IDLLM-PRIM-005	Primary sentiment corpus	NusaX-Senti	Indonesian local-language communities	2022–2023	1 source record	Parallel sentiment data across Indonesian local languages
IDLLM-PRIM-006	Primary multilingual benchmark	FLORES-200	Global; Indonesia-relevant subset	2022	1 source record	Language-code verification for Indonesian, Javanese, Sundanese, and Buginese
IDLLM-PRIM-007	Primary multitask benchmark	LoraxBench	Indonesia	2025	1 source record	Multitask benchmark for Indonesian and local languages
IDLLM-SEC-008	Secondary dataset catalogue	NusaCrowd	Indonesia	2022–2023	1 source record	Dataset discovery and standardised loading context
IDLLM-SEC-009	Secondary evaluation suite	SEA-HELM	Southeast Asia	2025	1 source record	Holistic Southeast Asian LLM evaluation framework
IDLLM-SEC-010	Secondary evaluation suite	BHASA	Southeast Asia	2023	1 source record	Linguistic and cultural evaluation suite
IDLLM-SEC-011	Secondary model resource	Cendol	Indonesia	2024	1 source record	Open instruction-tuned Indonesian LLM context
IDLLM-SEC-012	Secondary robustness framework	IndoRobusta	Indonesia	2022–2023	1 source record	Code-mixing and local-language robustness context
IDLLM-CTX-013	Contextual policy source	Badan Bahasa	Indonesia	2025	1 source record	Official context for Indonesian linguistic diversity

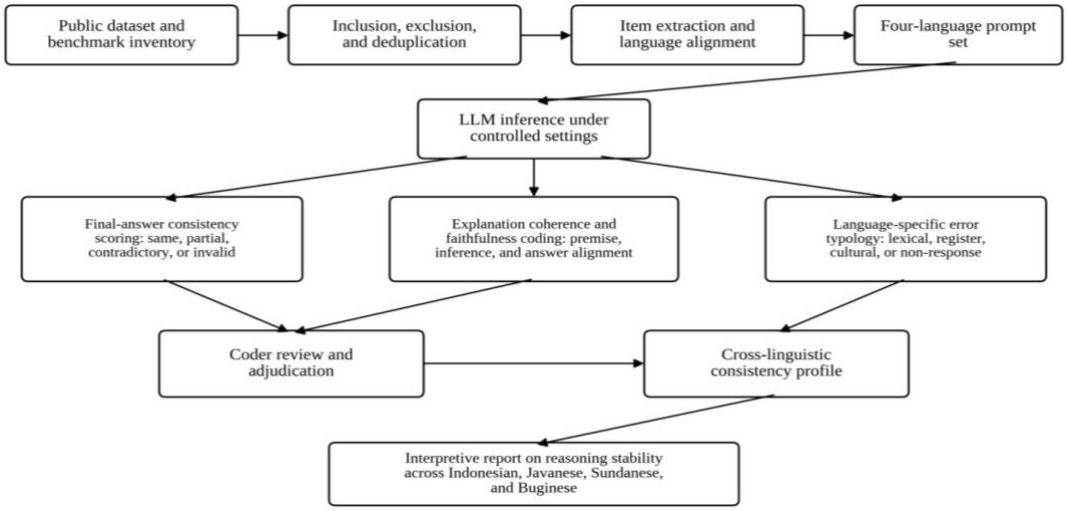

Figure 1. Operational matrix of cross-lingual reasoning analysis

Data collection proceeded through source identification, keyword search, metadata extraction, screening, deduplication, and structured storage. Search keywords included "Indonesian LLM benchmark," "Javanese Sundanese Buginese dataset," "Indonesian commonsense reasoning," "multilingual reasoning consistency," "Indonesian local languages NLP," and "Southeast Asian LLM evaluation." Each retained item was assigned a unique document code and stored with source title, repository, language coverage, reasoning relevance, period, geographical scope, verification source, inclusion reason, exclusion note, and deduplication key. Text records preserve title, date, source, URL, context, short quotation, and storage note. Visual or video fields were retained in the schema, although current corpus entries are text-based.

Data analysis follows three sequential procedures. First, final-answer consistency is coded by comparing whether aligned prompts produce equivalent, partially equivalent, contradictory, or invalid answers. Second, explanation quality is examined through coherence and faithfulness: whether a response uses relevant premises, maintains logical relations, and supports its final answer. Third, language-linked error patterns are classified, including lexical gaps, register mismatch, code-mixing sensitivity, undertranslation, overtranslation, cultural misreading, and non-response. This analytical logic follows recent concerns about cross-lingual transfer, multilingual inconsistency, and uneven representational alignment in LLMs [19], [10], [12], [15].

4. RESULTS

4.1. Cross-lingual consistency across language pairs

Multilingual reasoning becomes visible not only through correct answers, but through whether a model preserves its decision when the same inferential problem moves across languages. The distribution in Table 3 places Indonesian as the reference condition and compares final-answer stability in Javanese, Sundanese, and Buginese prompt versions. This structure follows recent arguments that multilingual LLM evaluation should move beyond single-language accuracy and translated benchmark performance [11], [12], [15]. The comparison is therefore organised around four observable categories: same answer, partially equivalent answer, contradictory answer, and invalid response. This approach allows this study to examine consistency as a reliability problem rather than treating local-language variation as a simple extension of Indonesian-language performance.

The dominant pattern is a graded decline in full answer stability from Indonesian–Javanese to Indonesian–Sundanese and Indonesian–Buginese comparisons. Javanese records the highest proportion of fully aligned final answers, followed by Sundanese, while Buginese shows the largest share of partial, contradictory, and invalid outputs. Partial equivalence appears across all three comparisons, indicating that models often retain the broad inferential direction but lose specificity, scope, or answer precision. Contradictory outputs are less frequent than stable answers but remain analytically consequential because they signal decision shifts under semantically aligned prompts. Invalid responses are lowest in Javanese and highest in Buginese, suggesting that answer-level consistency is not evenly distributed across Indonesian local-language conditions.

Table 2. Cross-lingual final-answer consistency across Indonesian, Javanese, Sundanese, and Buginese

Main Category	Language Comparison	Consistency Indicator	Frequency	Percentage	Data Variation	Pattern Example	Data Code
Stable answer alignment	Indonesian–Javanese	Same final answer	37	61.7%	Answer choice and semantic conclusion remain aligned	Same option and equivalent conclusion across both prompts	ID-JV-CON-001–037
Partial answer alignment	Indonesian–Javanese	Partially equivalent answer	12	20.0%	Same general direction but different specificity or justification	Correct general inference, but reduced detail in Javanese version	ID-JV-PAR-038–049
Answer contradiction	Indonesian–Javanese	Contradictory final answer	8	13.3%	Different option or opposite conclusion	Indonesian output selects one causal relation; Javanese output reverses it	ID-JV-CONTRA-050–057
Invalid response	Indonesian–Javanese	Invalid or no answer	3	5.0%	Refusal, irrelevant answer, or untranslated response	Output does not provide a usable final answer	ID-JV-INV-058–060
Stable answer alignment	Indonesian–Sundanese	Same final answer	34	56.7%	Answer choice and conclusion remain aligned	Same answer retained despite lexical variation	ID-SU-CON-061–094
Partial answer alignment	Indonesian–Sundanese	Partially equivalent answer	13	21.7%	Equivalent answer direction with altered scope	Correct answer retained, but explanation omits one condition	ID-SU-PAR-095–107
Answer contradiction	Indonesian–Sundanese	Contradictory final answer	9	15.0%	Opposing answer or incompatible conclusion	Sundanese output selects a different answer from Indonesian output	ID-SU-CONTRA-108–116
Invalid response	Indonesian–Sundanese	Invalid or no answer	4	6.7%	Non-answer, repetition, or irrelevant reformulation	Model repeats prompt elements without answering	ID-SU-INV-117–120
Stable answer alignment	Indonesian–Buginese	Same final answer	24	40.0%	Answer and conclusion remain aligned	Equivalent final answer across Indonesian and Buginese versions	ID-BU-CON-121–144
Partial answer alignment	Indonesian–Buginese	Partially equivalent answer	15	25.0%	General inference retained but answer form changes	Buginese output gives approximate rather than exact answer	ID-BU-PAR-145–159
Answer contradiction	Indonesian–Buginese	Contradictory final answer	12	20.0%	Different answer choice or reversed relation	Output changes from a plausible answer to an incompatible one	ID-BU-CONTRA-160–171
Invalid response	Indonesian–Buginese	Invalid or no answer	9	15.0%	Non-response, fallback to Indonesian, or incoherent completion	Model shifts language or fails to produce an answer	ID-BU-INV-172–180

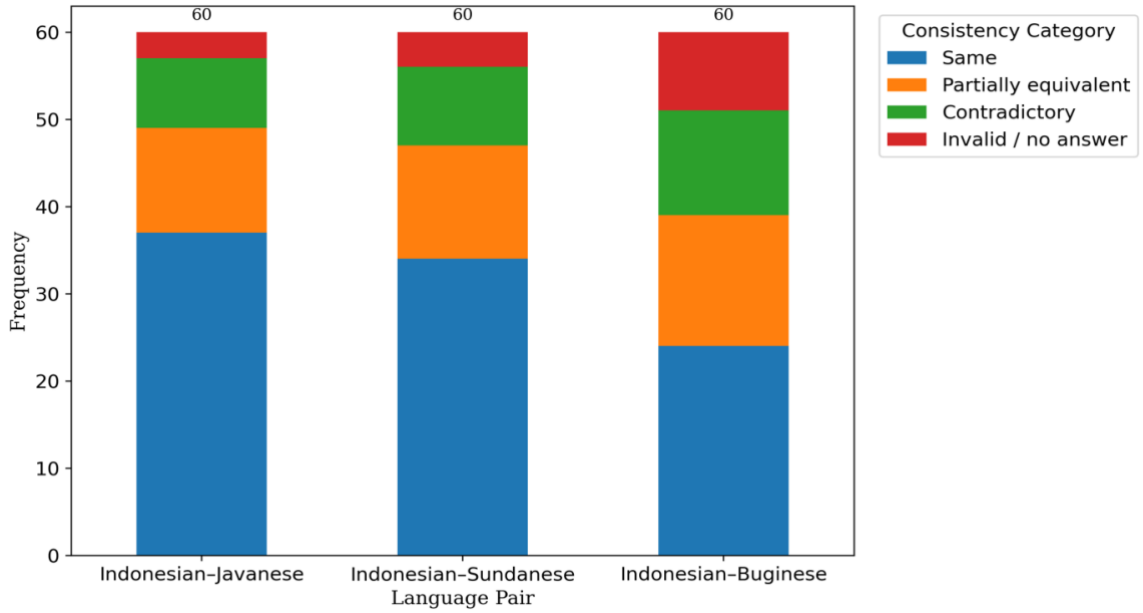

Figure 2. Cross-lingual final-answer consistency by language pair

These patterns support a theoretical distinction between multilingual coverage and multilingual reasoning consistency. A model may produce fluent or partially intelligible responses in several languages while failing to preserve the same reasoning decision across those languages. This distinction is consistent with studies showing that cross-lingual alignment, cultural grounding, and representational asymmetry can shape multilingual LLM behaviour in ways not captured by accuracy alone [9], [10], [3], [2]. The observed differences should not be read as causal proof that low-resource status alone determines inconsistency. Rather, they indicate that reasoning stability must be tested directly across Indonesian, Javanese, Sundanese, and Buginese before any model is described as robustly multilingual for Indonesia.

4.2. Coherence and faithfulness across language pairs

The second layer of analysis shifts attention from answer stability to the quality of visible reasoning. A model may produce the same final answer across languages while offering explanations that are incomplete, weakly connected, or unfaithful to its conclusion. For this reason, this study codes explanation quality through two linked criteria: coherence and faithfulness. Coherence concerns whether premises, inferential steps, and conclusions are logically ordered; faithfulness concerns whether the explanation actually supports the final answer. This distinction follows recent debates in multilingual chain-of-thought and cross-lingual reasoning evaluation, where apparent correctness may coexist with unstable or ungrounded explanations [9], [14], [15]. The same 180 comparison units are retained to ensure continuity with the preceding answer-consistency analysis.

The dominant pattern shows that coherent and faithful explanations are most frequent in Indonesian–Javanese comparison, slightly lower in Indonesian–Sundanese comparison, and substantially lower in Indonesian–Buginese comparison. Indonesian–Javanese records 31 fully coherent and faithful cases, while Indonesian–Sundanese records 28 and Indonesian–Buginese records 18. Buginese also shows higher frequencies in partially coherent, incoherent or unfaithful, and invalid explanation categories. Across all language pairs, partially faithful explanations are common, suggesting that models often produce plausible surface reasoning without fully preserving the inferential relation required by the task. Invalid or absent explanations remain relatively limited in Javanese and Sundanese, but become more visible in Buginese, where language fallback and incomplete reasoning appear more frequently.

Table 3. Explanation coherence and faithfulness across language pairs

Main Category	Language Comparison	Indicator	Frequency	Percentage	Data Variation	Pattern Example	Data Code
Coherent and faithful explanation	Indonesian–Javanese	Explanation is logically ordered and supports final answer	31	51.7%	Premise, inference, and conclusion remain aligned	The model gives the same answer and provides a comparable reason in both languages	ID-JV-EXP-CF-001–031
Coherent but partially faithful explanation	Indonesian–Javanese	Explanation is clear but only partly supports answer	11	18.3%	Explanation is fluent but omits one inferential condition	The conclusion is correct, but one causal step is underexplained	ID-JV-EXP-CPF-032–042
Partially coherent explanation	Indonesian–Javanese	Explanation contains incomplete or loosely connected reasoning	10	16.7%	Some premises are relevant but weakly connected	The answer is plausible, but explanation shifts topic midway	ID-JV-EXP-PC-043–052
Incoherent or unfaithful explanation	Indonesian–Javanese	Explanation contradicts answer or uses irrelevant premises	5	8.3%	Reasoning conflicts with selected final answer	The model selects one option but justifies another	ID-JV-EXP-IU-053–057
Invalid or no explanation	Indonesian–Javanese	Explanation is absent, irrelevant, or unusable	3	5.0%	Non-answer, repetition, or unsupported output	The model repeats prompt content without explaining	ID-JV-EXP-INV-058–060
Coherent and faithful explanation	Indonesian–Sundanese	Explanation is logically ordered and supports final answer	28	46.7%	Reasoning remains aligned but slightly compressed	Sundanese explanation gives correct relation with reduced detail	ID-SU-EXP-CF-061–088
Coherent but partially faithful explanation	Indonesian–Sundanese	Explanation is clear but only partly supports answer	12	20.0%	Coherent surface structure with missing premise	The model explains an effect but omits the stated cause	ID-SU-EXP-CPF-089–100
Partially coherent explanation	Indonesian–Sundanese	Explanation contains incomplete or loosely connected reasoning	11	18.3%	Some reasoning steps are vague or underspecified	The response gives a general explanation but not item-specific reasoning	ID-SU-EXP-PC-101–111
Incoherent or unfaithful explanation	Indonesian–Sundanese	Explanation contradicts answer or uses irrelevant premises	6	10.0%	Unsupported explanation or answer–reason mismatch	The answer remains correct but justification refers to an unrelated condition	ID-SU-EXP-IU-112–117
Invalid or no explanation	Indonesian–Sundanese	Explanation is absent, irrelevant, or unusable	3	5.0%	No explanation, copied prompt, or empty reasoning	The model provides final answer only	ID-SU-EXP-INV-118–120

Coherent and faithful explanation	Indonesian–Buginese	Explanation is logically ordered and supports final answer	18	30.0%	Reasoning is aligned in fewer cases	Buginese output provides answer and reason but with reduced precision	ID-BU-EXP-CF-121–138
Coherent but partially faithful explanation	Indonesian–Buginese	Explanation is clear but only partly supports answer	14	23.3%	Surface coherence with weakened inferential support	The explanation sounds plausible but does not justify the exact answer	ID-BU-EXP-CPF-139–152
Partially coherent explanation	Indonesian–Buginese	Explanation contains incomplete or loosely connected reasoning	13	21.7%	Fragmented relation between premises and answer	The model shifts between Buginese and Indonesian while explaining	ID-BU-EXP-PC-153–165
Incoherent or unfaithful explanation	Indonesian–Buginese	Explanation contradicts answer or uses irrelevant premises	8	13.3%	Contradictory justification or unrelated premise	The output gives a final answer but explanation supports another option	ID-BU-EXP-IU-166–173
Invalid or no explanation	Indonesian–Buginese	Explanation is absent, irrelevant, or unusable	7	11.7%	Refusal, language fallback, or no usable explanation	The model switches to Indonesian without completing the explanation	ID-BU-EXP-INV-174–180

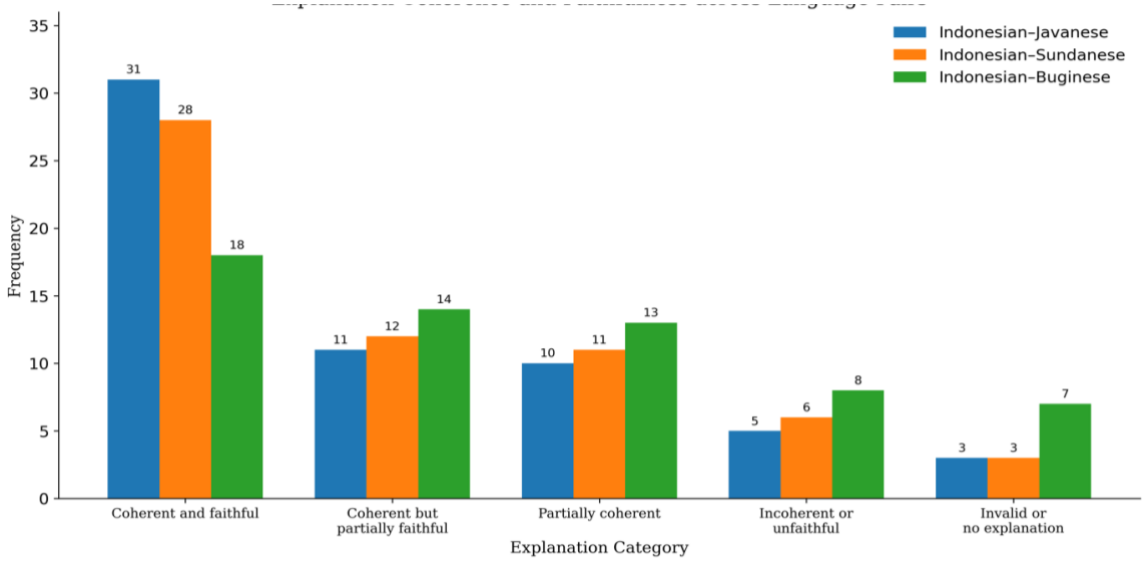

Figure 3. Explanation coherence and faithfulness across language pairs

Theoretically, this pattern strengthens the distinction between multilingual fluency and multilingual reasoning reliability. Fluency can make an explanation appear acceptable even when its inferential support is incomplete or misaligned with the final answer. This finding is compatible with work on cross-lingual inconsistency and multilingual representation, which suggests that language transfer does not automatically preserve semantic or reasoning stability across languages [19], [10], [11], [12]. Within this study, explanation analysis therefore functions as a necessary check on answer-level consistency. A model cannot be considered consistently multilingual for Indonesian linguistic contexts if its answers remain stable but its explanations degrade, shift, or lose faithfulness across Javanese, Sundanese, and Buginese.

4.3. Language-linked error patterns across language pairs

Language-linked errors show where cross-linguistic inconsistency enters model behaviour after answer stability and explanation quality have been examined. The coding does not treat all wrong or unstable outputs as the same phenomenon. Instead, it separates lexical-semantic drift, register or pragmatic mismatch, translation-related distortion, cultural inferential misreading, and language fallback or non-response. This distinction is necessary because multilingual evaluation can otherwise collapse different failure mechanisms into a single accuracy deficit. Recent studies on Indonesian local-language robustness, cross-lingual inconsistency, and multilingual transfer have shown that language variation can affect model behaviour through several pathways, including lexical ambiguity, code-mixing, cultural grounding, and uneven representational alignment [25], [6], [10], [12]. This coding therefore keeps error interpretation analytically specific.

The dominant pattern is again gradient rather than binary. Indonesian–Javanese contains the largest number of cases without major language-linked disruption, followed by Indonesian–Sundanese and then Indonesian–Buginese. Buginese records the highest frequencies of lexical-semantic drift, undertranslation or overtranslation, and language fallback or non-response. Lexical-semantic drift is the most frequent identifiable error type in all three comparisons, but its proportion rises most clearly in Buginese. Register and pragmatic mismatch appear in each language pair, showing that local-language reasoning is not only a matter of vocabulary coverage. Cultural inferential misreading is less frequent than lexical or translation-related errors, but it remains important because reasoning tasks often depend on contextual assumptions, not merely formal semantic equivalence.

Table 4. Language-linked error patterns across Indonesian, Javanese, Sundanese, and Buginese comparisons

Main Category	Language Comparison	Dominant Indicator	Frequency	Percentage	Data Variation	Pattern Example	Data Code
No major language-linked error	Indonesian–Javanese	Output remains semantically stable without visible language-induced disruption	32	53.3%	Answer and explanation remain aligned across languages	Javanese prompt produces equivalent answer and usable explanation	ID-JV-ERR-NO-001–032
Lexical-semantic drift	Indonesian–Javanese	Meaning shifts because one key word or relation is weakened	9	15.0%	Local lexical item interpreted too broadly or too narrowly	A causal term is treated as temporal sequence	ID-JV-ERR-LEX-033–041
Register or pragmatic mismatch	Indonesian–Javanese	Response uses inappropriate speech level, politeness, or pragmatic framing	7	11.7%	Ngoko/krama-like distinction is flattened or inconsistently handled	Explanation is understandable but pragmatically unnatural	ID-JV-ERR-REG-042–048
Undertranslation or overtranslation	Indonesian–Javanese	Output omits or adds meaning not present in prompt	5	8.3%	Inferential condition is compressed or expanded	One premise is dropped in the Javanese response	ID-JV-ERR-TRN-049–053
Cultural inferential misreading	Indonesian–Javanese	Local cultural or commonsense cue is misread	4	6.7%	Culturally embedded premise receives generic interpretation	A local social relation is treated as universal commonsense	ID-JV-ERR-CUL-054–057
Language fallback or non-response	Indonesian–Javanese	Model shifts language, refuses, or gives unusable output	3	5.0%	Answer partly returns to Indonesian or gives no explanation	Javanese prompt receives Indonesian-dominant completion	ID-JV-ERR-FAL-058–060
No major language-linked error	Indonesian–Sundanese	Output remains semantically stable without visible language-induced disruption	29	48.3%	Answer and explanation remain mostly aligned	Sundanese prompt produces equivalent decision and relevant explanation	ID-SU-ERR-NO-061–089
Lexical-semantic drift	Indonesian–Sundanese	Meaning shifts because one key word or relation is weakened	10	16.7%	Lexical choice changes the intended relation	Model treats a conditional relation as a general statement	ID-SU-ERR-LEX-090–099
Register or pragmatic mismatch	Indonesian–Sundanese	Response uses inappropriate register or pragmatic framing	6	10.0%	Politeness and local expression are partially flattened	Explanation sounds formal but pragmatically awkward	ID-SU-ERR-REG-100–105

Undertranslation or overtranslation	Indonesian–Sundanese	Output omits or adds meaning not present in prompt	6	10.0%	One condition is omitted or rephrased beyond the prompt	Sundanese response answers a simplified version of the task	ID-SU-ERR-TRN-106–111
Cultural inferential misreading	Indonesian–Sundanese	Local cultural or commonsense cue is misread	5	8.3%	Local context is generalised too quickly	Model ignores a culturally marked premise	ID-SU-ERR-CUL-112–116
Language fallback or non-response	Indonesian–Sundanese	Model shifts language, refuses, or gives unusable output	4	6.7%	Partial fallback to Indonesian or explanation loss	Sundanese answer begins locally but ends in Indonesian	ID-SU-ERR-FAL-117–120
No major language-linked error	Indonesian–Buginese	Output remains semantically stable without visible language-induced disruption	19	31.7%	Stable answer and explanation appear in fewer cases	Buginese prompt receives equivalent answer only in part of the set	ID-BU-ERR-NO-121–139
Lexical-semantic drift	Indonesian–Buginese	Meaning shifts because one key word or relation is weakened	12	20.0%	Core relation becomes approximate or ambiguous	Causal relation is interpreted as association	ID-BU-ERR-LEX-140–151
Register or pragmatic mismatch	Indonesian–Buginese	Response uses inappropriate register or pragmatic framing	7	11.7%	Local discourse style is flattened or replaced by Indonesian-like phrasing	Output is grammatically plausible but pragmatically unstable	ID-BU-ERR-REG-152–158
Undertranslation or overtranslation	Indonesian–Buginese	Output omits or adds meaning not present in prompt	8	13.3%	Prompt content is shortened, expanded, or reinterpreted	Buginese response answers a narrower version of the item	ID-BU-ERR-TRN-159–166
Cultural inferential misreading	Indonesian–Buginese	Local cultural or commonsense cue is misread	5	8.3%	Contextual cue is treated as generic commonsense	Local premise is ignored in favour of general reasoning	ID-BU-ERR-CUL-167–171
Language fallback or non-response	Indonesian–Buginese	Model shifts language, refuses, or gives unusable output	9	15.0%	Fallback to Indonesian, incomplete answer, or no usable output	Buginese prompt triggers Indonesian-dominant completion	ID-BU-ERR-FAL-172–180

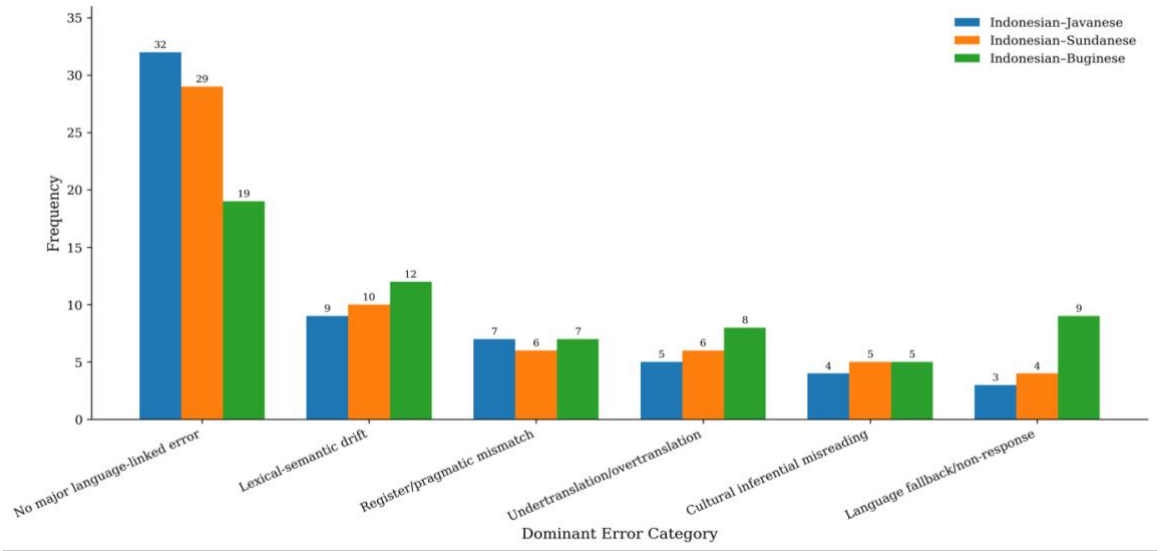

Figure 4. Language-linked error patterns across language pairs

Theoretically, these patterns reinforce the argument that multilingual reasoning inconsistency cannot be reduced to either mistranslation or low-resource status alone. Some errors arise from semantic drift, some from pragmatic flattening, and others from explanation-level or response-level fallback. This supports a more layered view of multilingual LLM evaluation, where language coverage, task equivalence, cultural grounding, and representational stability must be analysed together. Studies such as COPAL-ID and IndoCulture have already shown that Indonesian reasoning cannot be detached from cultural and commonsense context [2], [3], while Southeast Asian evaluation frameworks emphasise the need for regionally grounded benchmarks [4], [5]. Within this study, the pattern suggests that robust multilinguality for Indonesia requires stable reasoning across local languages, not only fluent response generation.

5. DISCUSSION

The pattern of final-answer consistency indicates that multilingual reasoning cannot be inferred from a model's ability to respond across languages. Its practical implication is that a model may appear functionally multilingual at the interface level while remaining unstable at the decision level. This matters because many public-facing uses of LLMs—education, assessment, translation support, civic information, and local-language knowledge access—depend not only on producing fluent text but also on preserving the same judgement when users ask equivalent questions in different languages. The stronger stability in Indonesian–Javanese and Indonesian–Sundanese comparisons, compared with Indonesian–Buginese, suggests that multilingual access is uneven even inside one national linguistic ecology. This pattern extends the concern raised by IndoMMLU and COPAL-ID: Indonesian reasoning evaluation is necessary but insufficient if Indonesian becomes the sole proxy for the multilingual realities of Indonesia [1], [2]. The dysfunction exposed here is not a simple failure of translation; it is a failure of decision stability.

The underlying structure of this instability appears to lie in unequal cross-lingual alignment rather than in task difficulty alone. Equivalent prompts do not automatically produce equivalent internal representations, especially when languages differ in digital visibility, lexical standardisation, register norms, and representation in multilingual training corpora. Studies on multilingual transfer show that cross-lingual performance often depends on how well languages are aligned in shared representational space, not merely on whether a model has encountered their surface forms [19], [22], [21]. This study's gradient pattern is consistent with that view: instability increases where linguistic distance, resource asymmetry, or weaker benchmark presence may be more consequential. However, this does not establish a direct causal chain from low-resource status to inconsistent reasoning. A more defensible interpretation is that resource asymmetry, alignment quality, and prompt equivalence jointly condition whether a model preserves its final decision across Indonesian local languages.

The explanation-level findings sharpen this argument by showing that answer consistency is not enough. A model can arrive at the same or approximately similar answer while offering reasoning that is compressed, partially faithful, or weakly connected to the conclusion. The implication is methodological:

multilingual evaluation that records only final answers risks overestimating reliability. This point resonates with recent work on multilingual chain-of-thought and factual reasoning, which warns that visible explanations may vary across languages even when outputs appear superficially correct [9], [14], [15]. For Indonesian local-language contexts, this dysfunction is particularly important because explanation quality affects whether users can inspect, contest, or learn from model outputs. If explanation coherence declines in local-language prompts, local-language users receive not only less stable answers but also less transparent justifications. This study therefore positions explanation faithfulness as a core dimension of multilingual accountability, not a secondary stylistic feature.

The reason explanation quality degrades across language pairs may be that LLMs often manage multilingual prompts through approximate semantic mediation rather than fully grounded language-specific reasoning. When a model produces a plausible explanation that only partly supports its answer, it may be drawing on generic reasoning templates instead of preserving item-specific inferential relations in the target language. This interpretation is compatible with studies on cross-lingual inconsistency and word-sense disambiguation, which show that multilingual models can lose fine-grained semantic distinctions across languages [6], [10], [11]. Yet this pattern also complicates optimistic accounts of multilingual prompting, including work suggesting that cross-lingual thought strategies can improve reasoning in some contexts [9]. Improvement in multilingual prompting does not remove the need to test faithfulness. In this study, explanation instability suggests that reasoning cannot be evaluated only through whether a model sounds coherent; it must be tested through whether each explanation actually warrants its answer.

The language-linked error profile further shows that inconsistency has multiple forms. Lexical-semantic drift, register mismatch, translation-related distortion, cultural inferential misreading, and language fallback do not represent the same kind of failure. Their coexistence implies that multilingual reasoning is not a single technical capability but a layered competence involving semantic precision, pragmatic appropriateness, cultural grounding, and response control. This point supports regionally grounded benchmarks such as IndoCulture, BHASA, and SEA-HELM, which argue that language-model evaluation in Southeast Asia must account for cultural and linguistic specificity rather than relying on globally dominant benchmark assumptions [3], [4], [5]. At the same time, this study shows that cultural grounding alone is not sufficient. Even when a task is semantically aligned, lexical or pragmatic disruptions can still alter the model's reasoning path. The implication is that robust multilinguality for Indonesia requires local-language sensitivity at several analytic levels, not only expanded language coverage.

The structural explanation for these error patterns is best understood as the interaction between language resources, sociolinguistic complexity, and evaluation design. Indonesian has wider institutional, educational, and digital presence, while Javanese, Sundanese, and Buginese differ in standardisation, available corpora, orthographic conventions, register systems, and public NLP resources. Recent Indonesian local-language work, including Cendol, LoraxBench, IndoRobusta, NusaBERT, and local-language translation studies, has begun addressing these gaps, but it also confirms that Indonesian multilingual NLP remains unevenly developed [6], [8], [25], [26], [27]. This study's contribution is therefore not to claim that one language "causes" failure, but to show how instability becomes visible when reasoning is tested across aligned Indonesian-language ecologies. The central theoretical implication is that multilingual LLM evaluation should treat consistency, faithfulness, and language-linked error as interconnected indicators of reasoning reliability.

6. CONCLUSION

This study shows that multilingual reasoning in large language models cannot be reduced to language coverage or translated-task accuracy. The most important lesson is that models may answer across Indonesian, Javanese, Sundanese, and Buginese while failing to preserve stable decisions, faithful explanations, and language-sensitive interpretation. Its contribution lies in shifting evaluation from isolated performance to cross-linguistic reasoning consistency, using final-answer alignment, explanation coherence, and language-linked error typology as interrelated measures. By positioning Indonesian local languages as analytic sites rather than marginal extensions of Indonesian, this study renews multilingual LLM evaluation as a question of reliability, equity, and epistemic accountability.

This study is limited by its focus on selected languages, aligned reasoning tasks, and observable model outputs, meaning that it cannot infer hidden cognitive processes or establish causal links between resource scarcity and reasoning failure. Its findings should therefore be read as evidence of patterned instability rather than definitive proof of why such instability occurs. Future research should expand language coverage to other Indonesian languages, include more model families, test prompt variation, and involve expert speakers in larger-scale validation. Further work should also compare written, spoken, and multimodal prompts to examine whether reasoning consistency changes across different communicative conditions.

ACKNOWLEDGMENTS

The author thanks the reviewers and editorial team for their constructive feedback on this manuscript.

FUNDING INFORMATION

Author states no funding involved.

AUTHOR CONTRIBUTIONS STATEMENT

Fitriya Dessi Wulandari: conceptualization, methodology, formal analysis, writing – original draft, writing – review and editing.

CONFLICT OF INTEREST STATEMENT

Author states no conflict of interest.

AI USE DECLARATION

The author used artificial intelligence tools to assist with the generation of illustrative figures and with the formatting and layout of this manuscript. AI was not used to generate the substantive content, conceptualization, data, analysis, or conclusions of the research, all of which remain the sole responsibility of the author.

INFORMED CONSENT

Not applicable – this study did not involve direct human participants. Data consisted solely of publicly accessible model outputs and evaluation materials.

ETHICAL APPROVAL

This research did not involve human or animal subjects; it analyzed publicly accessible model outputs and evaluation materials, and therefore no institutional review board approval was required. Where research does involve human subjects, the author affirms that such research would be conducted in full compliance with all relevant national regulations and institutional policies, in accordance with the tenets of the Declaration of Helsinki, and would be approved by the author's institutional review board or equivalent committee.

DATA AVAILABILITY

The corpus of documents analyzed in this study are publicly accessible and available from the corresponding author upon reasonable request.

REFERENCES

[1] F. Koto, N. Aisyah, H. Li, and T. Baldwin, "Large language models only pass primary school exams in Indonesia: A comprehensive test on IndoMMLU," *Findings of the Association for Computational Linguistics: EMNLP 2023, 12359–12374*, 2023, doi: 10.48550/arXiv.2310.04928.

[2] H. A. Wibowo, E. H. Fuadi, M. N. Nityasya, R. E. Prasojo, and A. F. Aji, "COPAL-ID: Indonesian language reasoning with local culture and nuances," *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, 1404–1422*, 2023, doi: 10.48550/arXiv.2311.01012.

[3] F. Koto, R. Mahendra, N. Aisyah, and T. Baldwin, "IndoCulture: Exploring geographically influenced cultural commonsense reasoning across eleven Indonesian provinces," *Transactions of the Association for Computational Linguistics*, vol. 12, pp. 1703–1719, 2024, doi: 10.1162/tacl_a_00726.

[4] W. Q. Leong, J. G. Ngui, Y. Susanto, H. Rengarajan, K. Sarveswaran, and W.-C. Tjhi, "BHASA: A holistic Southeast Asian linguistic and cultural evaluation suite for large language models," *arXiv*, 2023, doi: 10.48550/arXiv.2309.06085.

[5] Y. Susanto, A. V. Hulagadri, J. Montalan, J. G. Ngui, X. Yong, W. Leong, et al., "SEA-HELM: Southeast Asian holistic evaluation of language models," *arXiv*, 2025, doi: 10.48550/arXiv.2502.14301.

[6] S. Cahyawijaya, H. Lovenia, F. Koto, R. A. Putri, E. Dave, J. Lee, et al., "Cendol: Open instruction-tuned generative large language models for Indonesian languages," *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, 14899–14914*, 2024, doi: 10.48550/arXiv.2404.06138.

[7] X.-P. Nguyen, W. Zhang, X. Li, M. Aljunied, Q. Tan, L. Cheng, et al., "SeaLLMs—Large language models for Southeast Asia," *arXiv*, 2023, doi: 10.48550/arXiv.2312.00738.

[8] A. F. Aji and T. Cohn, "LoraxBench: A multitask, multilingual benchmark suite for 20 Indonesian languages," *arXiv*, 2025, doi: 10.48550/arXiv.2508.12459.

[9] H. Huang, T. Tang, D. Zhang, W. X. Zhao, T. Song, Y. Xia, and F. Wei, "Not all languages are created equal in LLMs: Improving multilingual capability by cross-lingual-thought prompting," *Findings of the Association for Computational Linguistics: EMNLP 2023, 12365–12394*, 2023, doi: 10.48550/arXiv.2305.07004.

[10] X. Xing, Z. He, H. Xu, X. Wang, R. Wang, and Y. Hong, "Evaluating knowledge-based cross-lingual inconsistency in large language models," *arXiv*, 2024, doi: 10.48550/arXiv.2407.01358.

[11] A. Gupta, M. Mehta, Z. Xu, and V. Srikumar, "Found in translation: Measuring multilingual LLM consistency as simple as translate then evaluate," *Proceedings of the 2025 Annual Conference of the North American Chapter of the Association for Computational Linguistics, 3477–3496*, 2025, doi: 10.48550/arXiv.2505.21999.

[12] L. Qin, Q. Chen, Y. Zhou, Z. Chen, Y. Li, L. Liao, et al., "A survey of multilingual large language models," *Patterns*, vol. 6, 2025, doi: 10.1016/j.patter.2024.101118.

[13] Y. Xu, L. Hu, J. Zhao, Z. Qiu, Y. Ye, and H. Gu, "A survey on multilingual large language models: Corpora, alignment, and bias," *Frontiers of Computer Science*, vol. 19, 2024, doi: 10.1007/s11704-024-40579-4.
[14] W. Zheng, X. Huang, Z. Liu, T. K. Vangani, B. Zou, X. Tao, et al., "AdaMCoT: Rethinking cross-lingual factual reasoning through adaptive multilingual chain-of-thought," *Proceedings of the AAAI Conference on Artificial Intelligence, 33863–33871*, 2025, doi: 10.1609/aaai.v40i40.40678.
[15] R. Zhao, Y. Liu, H. Schütze, and M. A. Hedderich, "A comprehensive evaluation of multilingual chain-of-thought reasoning: Performance, consistency, and faithfulness across languages," *arXiv*, 2025, doi: 10.48550/arXiv.2510.09555.
[16] A. Agrawal, A. Dang, S. B. Nezhad, R. Pokharel, and R. Scheinberg, "Evaluating multilingual long-context models for retrieval and reasoning," *Proceedings of the First Workshop on Multilingual Retrieval-Augmented Generation, 3477–3496*, 2024, doi: 10.18653/v1/2024.mrl-1.18.
[17] R. S. J. Xuan, J. Huseynov, and Y. Zhang, "Uncovering cross-linguistic disparities in LLMs using sparse autoencoders," *arXiv*, 2025, doi: 10.48550/arXiv.2507.18918.
[18] W. Han, Y. Zhang, Z. Chen, B. Li, H. Lin, B. Zhang, et al., "MuBench: Assessment of multilingual capabilities of large language models across 61 languages," *arXiv*, 2025, doi: 10.48550/arXiv.2506.19468.
[19] F. Faisal and A. Anastasopoulos, "An efficient approach for studying cross-lingual transfer in multilingual language models," *arXiv*, 2024, doi: 10.48550/arXiv.2403.20088.
[20] J. Lee, S. Hong, H. Moon, and H.-J. Lim, "Cross-lingual optimization for language transfer in large language models," *arXiv*, 2025, doi: 10.48550/arXiv.2505.14297.
[21] K. Ravisankar, H. Han, and M. Carpuat, "Can you map it to English? The role of cross-lingual alignment in the multilingual performance of LLMs," *Proceedings of the 2026 Conference of the European Chapter of the Association for Computational Linguistics, 4854–4872*, 2025, doi: 10.18653/v1/2026.eacl-long.225.
[22] S. Rajaee and C. Monz, "Analyzing the evaluation of cross-lingual knowledge transfer in multilingual language models," *arXiv*, 2024, doi: 10.48550/arXiv.2402.02099.
[23] W. Zhang, H. P. Chan, Y. Zhao, M. Aljunied, J. Wang, C. Liu, et al., "SeaLLMs 3: Open foundation and chat multilingual large language models for Southeast Asian languages," *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, 96–105*, 2024, doi: 10.48550/arXiv.2407.19672.
[24] L. Dou, Q. Liu, G. Zeng, J. Guo, J. Zhou, W. Lu, and M. Lin, "Sailor: Open language models for South-East Asia," *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, 424–435*, 2024, doi: 10.48550/arXiv.2404.03608.
[25] F. Adilazuarda, S. Cahyawijaya, G. I. Winata, P. Fung, and A. Purwarianti, "IndoRobusta: Towards robustness against diverse code-mixed Indonesian local languages," *arXiv*, 2023, doi: 10.48550/arXiv.2311.12405.
[26] W. Wongso, D. S. Setiawan, S. Limcorn, and A. Joyoadikusumo, "NusaBERT: Teaching IndoBERT to be multilingual and multicultural," *arXiv*, 2024, doi: 10.48550/arXiv.2403.01817.
[27] W. Tan and K. Zhu, "NusaMT-7B: Machine translation for low-resource Indonesian languages with large language models," *arXiv*, 2024, doi: 10.48550/arXiv.2410.07830.

BIOGRAPHY OF AUTHOR

Fitriya Dessi Wulandari is affiliated with the English Education Department, Universitas Muhammadiyah Surakarta, Indonesia. Her academic concerns are in English, linguistics, and education-related disciplines. Her academic interests include pragmatics, discourse analysis, digital advertising, language in e-commerce, online communication, and language pedagogy. Her research focuses on how persuasive language strategies are used in digital marketplaces and educational contexts. She can be contacted at email: fitriyadessi@gmail.com