Indonesian Journal of Computational Language Studies
Vol. 1, No. 2, June 2026, pp. 63~76
ISSN: 3163-7884, DOI:

Beyond translated benchmarks: a culturally grounded evaluation of large language models for Indonesian language understanding

Achraf El Bouazzaoui
Embedded Electronics and Intelligent Systems, Ibn Tofail University, Morocco

Article Info

Article history:

Received Apr 21, 2026
Revised May 25, 2026
Accepted Jun 25, 2026

Keywords:

cultural commonsense;
language understanding;
large language models;
multilingual evaluation;
source-grounded benchmark.

ABSTRACT

Background: Translated benchmarks have expanded multilingual large language model evaluation, yet Indonesian remains a sociolinguistically dense setting where meaning is shaped by public institutions, regional languages, religious discourse, and culturally situated inference. **Objective:** This study aims to examine how Indonesian language understanding can be evaluated through source-grounded, culturally embedded corpus items rather than through translated English-centric tasks. **Method:** Using a qualitative-diagnostic benchmark design, this study constructs and codes an auditable corpus of public Indonesian documents, benchmark references, and contextual materials across semantic, pragmatic, cultural, regional, institutional, safety, and multimodal dimensions. **Results:** The analysis shows that semantic comprehension appears as a baseline requirement but does not sufficiently capture Indonesian understanding. Institutional register and cultural commonsense emerge as dominant dimensions, indicating that Indonesian prompts frequently require recognition of public authority, local values, collective memory, and socially constrained meaning. **Implication:** Task sensitivity is highest when documents combine policy language, cultural inference, safety cues, regional references, and pragmatic restraint, revealing why ordinary accuracy metrics may obscure fragile understanding. **Novelty:** This study contributes a culturally grounded evaluation perspective that reframes Indonesian LLM assessment as source-verifiable, context-sensitive interpretation rather than transferable language performance, and offers methodological resources for future multilingual benchmarking in low-resource and culturally plural settings within Southeast Asian digital ecologies.

Corresponding Author:

Achraf El Bouazzaoui
Embedded Electronics and Intelligent Systems, Ibn Tofail University, Morocco
Av. de L'Université, B.P. 242, Kénitra, Morocco
Email: achraf.elbouazzaoui@icfo.eu

1. INTRODUCTION

Indonesia offers a decisive test case for evaluating whether large language models can move beyond translated competence toward culturally grounded language understanding. By 2025, Indonesia's mid-year population had reached approximately 284.4 million, while the 2024 national internet-user estimate exceeded 221.5 million, making Indonesian digital communication both demographically large and platform-saturated. Yet this scale is inseparable from linguistic complexity: Badan Bahasa reported 718 regional languages in 2024, with only a small proportion classified as safe and several already endangered, critical, or extinct. In such a setting, Indonesian is not simply a national code into which global benchmarks can be translated; it is a communicative infrastructure shaped by regional languages, religious expression, bureaucratic registers,

politeness norms, public-risk discourse, and everyday multilingualism. Evaluating LLMs only through translated English-centric benchmarks therefore risks mistaking surface fluency for situated understanding, particularly when meaning depends on cultural inference rather than literal semantic equivalence.

Recent scholarship has substantially improved multilingual and Indonesian LLM evaluation, but important gaps remain. General surveys have mapped LLM evaluation across knowledge, reasoning, robustness, safety, and multilinguality, while studies beyond English have shown that model performance often declines outside high-resource settings [1], [2], [3]. Indonesian NLP has also developed significant resources: IndoNLG evaluates Indonesian generation tasks; IndoMMLU tests language and cultural knowledge through Indonesian educational questions; IndoCulture foregrounds geographically influenced commonsense reasoning; Komodo and LoraxBench expand attention to regional languages; and IndoSafety introduces culturally grounded safety evaluation across formal, colloquial, and local-language varieties [4], [5], [6], [7], [8], [9]. Regionally, BHASA and SEA-HELM further consolidate Southeast Asian linguistic and cultural evaluation [10], [11]. However, less is known about how Indonesian public, institutional, cultural, and regional discourse can be combined into a diagnostic corpus that evaluates pragmatic interpretation, register sensitivity, and culturally situated reasoning without relying primarily on translated benchmark items.

This study addresses that gap by developing a culturally grounded evaluation of LLMs for Indonesian language understanding. Rather than asking only whether models can answer Indonesian prompts, this study asks what kinds of Indonesian meaning they can and cannot interpret when prompts are drawn from publicly verifiable cultural, institutional, regional, and communicative contexts. Specifically, this study examines three interrelated questions: first, how selected LLMs respond to Indonesian-language items constructed from legitimate public sources; second, which dimensions of understanding—semantic, pragmatic, cultural, regional, institutional, and safety-related—generate the most visible interpretive difficulty; and third, what error patterns emerge when model outputs are compared with source-grounded expectationsThis study does not claim to represent all Indonesian cultures or all forms of Indonesian language use. Instead, it builds an auditable corpus and coding framework through which model performance can be examined as a situated interaction between linguistic form, cultural knowledge, institutional context, and evaluative design.

This study advances the argument that culturally grounded evaluation must treat Indonesian language understanding as a layered competence rather than as monolingual task completion. The working proposition is that LLMs may perform adequately on standard Indonesian prompts while still failing when interpretation requires local commonsense, indirect pragmatic inference, regional-cultural recognition, or sensitivity to formal public registers. This proposition extends the concerns raised by IndoMMLU and IndoCulture, which show that Indonesian evaluation cannot be reduced to generic multilingual transfer, and aligns with SEA-HELM's broader claim that Southeast Asian evaluation requires linguistic, cultural, and safety dimensions to be assessed together [7], [8], [11]. The implication is methodological as much as empirical: benchmark validity depends not only on language coverage but also on how test items encode social knowledge, communicative norms, and source-verifiable context. This study therefore positions translated benchmarks as useful but insufficient, and proposes culturally grounded Indonesian evaluation as a necessary step toward more accountable multilingual LLM assessment.

2. LITERATURE REVIEW

2.1. Evaluation in language model research

Evaluation has become a central problem in large language model research because model fluency can conceal fragile reasoning, uneven multilingual competence, and culturally narrow assumptions. Early evaluation frameworks tended to treat performance as measurable through task accuracy, benchmark scores, or human preference, whereas more recent work understands evaluation as a multidimensional inquiry into reasoning, robustness, safety, factuality, and language coverage [1], [3]. This distinction matters for Indonesian because language understanding cannot be reduced to correct lexical mapping. Studies beyond English show that LLMs may perform impressively in high-resource languages while becoming less reliable in multilingual or low-resource contexts [2], [12].

Multilingual LLM evaluation is usually organized through several measurable dimensions: linguistic accuracy, task generalization, reasoning difficulty, domain transfer, instruction following, and safety. Benchmarks such as M3Exam, MMLU-ProX, BenchMAX, Global MMLU, and Global PIQA extend evaluation across languages, disciplines, and reasoning types, but they also reveal a persistent tension between comparability and cultural specificity [13], [12], [14], [15], [16]. Translated or parallel benchmarks enable cross-model comparison, yet they may preserve the epistemic structure of the source language. A more defensible evaluation therefore requires indicators that distinguish translation adequacy from native interpretive competence, especially when prompts involve context, pragmatic inference, or socially embedded knowledge.

2.2. Cultural understanding

Cultural understanding has emerged as a corrective to benchmark universalism, but its definition remains contested. Some studies define culture as everyday knowledge distributed across countries and languages; others treat it as situated commonsense, social norm interpretation, moral reasoning, superstition, affective expectation, or worldview alignment [17], [18], [19], [20]. This diversity is productive, but it also creates methodological risk: culture can be flattened into trivia, national stereotypes, or static identity labels. Work on Korean, Turkish, Arabic, Indian, Cantonese, and Persian benchmarks shows that culturally grounded evaluation must move beyond factual recall toward interpretive and contextual reasoning [21], [22], [23], [24], [25].

The main indicators of cultural evaluation now include cultural commonsense, pragmatic appropriateness, moral and safety judgement, locally meaningful affect, multimodal cultural recognition, and sensitivity to social relations. BLEnD, CultureScope, CURE, WorldView-Bench, and culturally specific benchmarks demonstrate that culture is not a single variable but a layered evaluative domain involving knowledge, inference, stance, and context [26], [17], [27], [28]. However, the field still struggles to balance thick cultural description with scalable benchmarking. This tension is crucial for Indonesian evaluation because many meanings depend on indirectness, hierarchy, regional language influence, religious expression, bureaucratic register, and locally familiar public discourse rather than explicit cultural labels.

2.3. Indonesian language understanding

Indonesian language understanding occupies a distinctive position within multilingual NLP because Indonesian is both a national language and a contact zone among hundreds of regional languages. Existing work has produced important resources for Indonesian generation, educational knowledge, multilingual transfer, regional-language exploration, sentiment, emotion, safety, and cultural commonsense [29], [4], [5], [6], [7], [8], [9]. Southeast Asian initiatives such as BHASA, MERaLiON-TextLLM, and SEA-HELM further situate Indonesian within a regional ecology of multilingual evaluation [30], [10], [11]. Yet Indonesian evaluation remains methodologically open: many tasks still privilege answer correctness over culturally situated interpretation.

The relevant indicators for Indonesian evaluation should therefore include semantic comprehension, pragmatic inference, regional and multilingual awareness, institutional-register sensitivity, safety judgement, and source-grounded cultural reasoning. IndoMMLU demonstrates the value of Indonesian educational and cultural knowledge, IndoCulture foregrounds geographically influenced commonsense, IndoSafety adds locally grounded safety dimensions, while LoraxBench and Komodo draw attention to the multilingual reality beneath national-language evaluation [4], [5], [7], [8], [9]. This study builds from these contributions but takes a sharper theoretical position: Indonesian LLM evaluation should not begin with translated benchmarks and then add culture as an afterthought. It should begin with Indonesian as a culturally embedded communicative system and evaluate models through auditable, source-grounded, context-sensitive tasks.

3. METHOD

The unit of analysis in this study is the source-grounded Indonesian document item, defined as a publicly verifiable text, video page, policy communication, cultural article, benchmark record, or contextual document from which Indonesian-language understanding prompts can be constructed. The corpus consists of 18 deduplicated items: 13 primary sources, four secondary benchmark sources, and one contextual source. Each item is assigned a unique document code, metadata record, text record, and coding record. This structure enables this study to examine Indonesian language understanding not as isolated sentence processing, but as interpretation across cultural discourse, institutional register, regional meaning, public-risk communication, and multilingual benchmark comparison.

This study uses a qualitative-diagnostic benchmark design. Rather than treating Indonesian as a translated target language, this design constructs evaluation items from native Indonesian public discourse and compares model responses against source-grounded expectations. The design follows recent calls for multilingual and culturally sensitive LLM evaluation, where benchmark validity depends on linguistic coverage, task design, safety, and cultural specificity [1], [2], [14], [11]. This study is not designed to infer causal effects of model architecture or training data. Its purpose is to identify observable interpretive patterns under controlled prompt conditions.

Table 1. Dataset composition

Code	Data Type	Source	Location	Period	Number	Characteristics
IDNLLM-PRIM-0001	Official article	Indonesia.go.id	Surakarta, Central Java	2022–2026	1	Grebeg Sudiro, tolerance, gotong royong, local economy
IDNLLM-PRIM-0002	Official article	Indonesia.go.id	Banyuwangi, East Java	2022–2026	1	Seblang Bakungan, Osing identity, ritual heritage
IDNLLM-PRIM-0003	Press release	Kemendikdasmen	National; regional languages	2022–2026	1	Revitalisation, corpus, documentation, AI
IDNLLM-PRIM-0004	Official news	Badan Bahasa	National; regional languages	2022–2026	1	Language policy, identity, preservation
IDNLLM-PRIM-0005	Policy communication	Bank Indonesia	National	2022–2026	1	Rupiah, national identity, public education
IDNLLM-PRIM-0006	Information page	OJK	National	2022–2026	1	Financial literacy, legality, consumer protection
IDNLLM-PRIM-0007	Video page	OJK-TV	Padang, West Sumatra	2022–2026	1	Scam warning, youth, visual public literacy
IDNLLM-PRIM-0008	Official article	BNPB/DIBI	Simeulue, Aceh	2022–2026	1	Smong, disaster mitigation, oral tradition
IDNLLM-PRIM-0009	Information page	Rumah Pendidikan	National	2022–2026	1	Education, literacy, conceptual language
IDNLLM-PRIM-0010	Official article	Indonesia.go.id	National	2022–2026	1	AI ethics, Pancasila, national character
IDNLLM-PRIM-0011	Hoax listing	Komdigi	National	2022–2026	1	Misinformation, verification, public warning
IDNLLM-PRIM-0012	Opinion column	Kemenag	National	2022–2026	1	Religious moderation, tolerance, local tradition
IDNLLM-PRIM-0013	Official news	Kemenag	National	2022–2026	1	Religious moderation, community programme
IDNLLM-CONT-0014	Event report	Sekretariat Kabinet	Jakarta; national	2022–2026	1	Pancasila, national identity, diversity
IDNLLM-SEC-0015	Benchmark paper	IndoMMLU	Indonesia	2022–2026	1	Indonesian language and knowledge evaluation
IDNLLM-SEC-0016	Benchmark paper	IndoCulture	Eleven Indonesian provinces	2022–2026	1	Geographically influenced cultural commonsense
IDNLLM-SEC-0017	Benchmark paper	IndoSafety	Indonesia and local languages	2022–2026	1	Culturally grounded safety evaluation
IDNLLM-SEC-0018	Benchmark paper	SEA-HELM	Southeast Asia	2022–2026	1	Regional multilingual and cultural evaluation

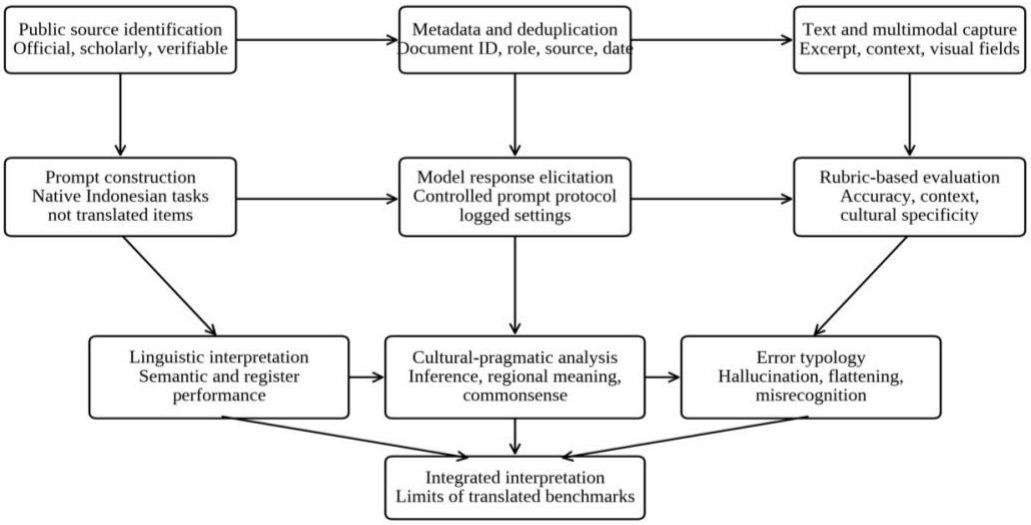

Figure 1. Operational matrix of source-grounded Indonesian LLM evaluation

The information sources were selected from official public institutions, national public-information portals, regulatory bodies, cultural and disaster-mitigation sources, religious-public communication, and peer-reviewed or publicly accessible benchmark records. Primary data were used to construct Indonesian prompts; secondary data were used to situate this study within Indonesian and Southeast Asian LLM evaluation; contextual data were used to clarify national ideological and cultural reference points. The corpus also includes one video page, for which visual and audiovisual fields were prepared: image description, visual actor, visual symbol, colour, layout, logo position, frame, timestamp, transcript, and screenshot reference. These fields support multimodal extension without collapsing video evidence into text alone.

Data collection followed six procedures. First, relevant institutions, documents, media channels, and benchmark records were identified through targeted keywords on Indonesian language, public discourse, cultural knowledge, safety, regional languages, and LLM evaluation. Second, each source was screened for public accessibility, institutional legitimacy, date, region, and relevance. Third, metadata were captured in corpus_metadata.csv, while short excerpts, source-grounded summaries, quotation notes, and multimodal fields were stored in corpus_text.csv. Fourth, deduplication used canonical URLs, document titles, and deduplication keys. Fifth, each item received a stable document code. Sixth, coding fields were prepared in corpus_coding.csv for later model-output evaluation.

Data analysis proceeds through three linked stages. The first stage is source-verifiability analysis, which checks source type, access status, data role, region, and duplication. The second stage is culturally grounded Indonesian understanding analysis, which codes each item for semantic understanding, pragmatic inference, cultural commonsense, regional or multilingual awareness, institutional register, safety sensitivity, and multimodal relevance. The third stage is model-response analysis, applied only after model outputs are collected, using rubric scores for accuracy, context sensitivity, cultural specificity, and hallucination risk. Qualitative error interpretation then identifies whether a response shows literalisation, cultural flattening, regional misrecognition, institutional-register failure, unsafe inference, or source-unverified fabrication.

4. RESULTS

4.1. Source-grounded distribution of Indonesian evaluation dimensions

Indonesian language understanding emerges here as a layered evaluative object rather than a single linguistic skill. The corpus is organized around source-grounded items that require models to interpret official, cultural, regional, regulatory, religious, educational, and benchmark-related discourse. This composition follows recent critiques of multilingual evaluation, which show that benchmark performance cannot be reduced to language coverage or translated task accuracy alone [1], [2], [14]. The strongest methodological point is that every item retains a verifiable source basis, while several items simultaneously carry cultural, institutional, and regional meanings. Consequently, this corpus is positioned to evaluate not only whether a model produces plausible Indonesian, but whether its response remains anchored in Indonesian communicative contexts.

Table 2. Coded distribution of Indonesian evaluation dimensions across the corpus

Analytical Domain	Subdomain	Operational Indicator	Frequency	Percentage	Data-Role Variation	Representative Evidence	Document Code(s)
Semantic comprehension	Source-grounded meaning	The item requires recognition of explicit meaning, topic, and referential content in Indonesian.	18	100.0%	Primary = 13; Secondary = 4; Context = 1	"Literasi mencakup kemampuan berpikir kritis."	IDNLLM-PRIM-0009; all items
Institutional register	Public, policy, regulatory, or official discourse	The item contains formal institutional language requiring recognition of communicative purpose, authority, and public-facing register.	14	77.8%	Primary = 13; Secondary = 0; Context = 1	"Rupiah sebagai alat pembayaran yang SAH."	IDNLLM-PRIM-0005; IDNLLM-CONT-0014
Cultural commonsense	Ritual, identity, religion, heritage, or public values	The item requires locally situated cultural interpretation beyond literal lexical meaning.	13	72.2%	Primary = 8; Secondary = 4; Context = 1	"Smong adalah tradisi lisan."	IDNLLM-PRIM-0008; IDNLLM-SEC-0016
Regional/ multilingual awareness	Local language, place-based meaning, regional identity	The item involves regional language, local identity, geographically situated culture, or multilingual awareness.	7	38.9%	Primary = 6; Secondary = 1; Context = 0	"Bahasa daerah merupakan bagian dari identitas bangsa."	IDNLLM-PRIM-0004; IDNLLM-SEC-0016
Safety/ sensitivity	Public risk, misinformation, consumer harm, AI ethics	The item requires sensitivity to risk, misinformation, legality, public harm, or ethically constrained interpretation.	5	27.8%	Primary = 4; Secondary = 1; Context = 0	"Cek 2L (Legal dan Logis)."	IDNLLM-PRIM-0007; IDNLLM-SEC-0017
Pragmatic inference	Indirect meaning, implied public action, tolerance, dialogic positioning	The item requires interpretation of implied meaning, social intention, or public communicative stance.	4	22.2%	Primary = 4; Secondary = 0; Context = 0	"Toleransi membuka ruang dialog."	IDNLLM-PRIM-0012; IDNLLM-0013
Multimodal visual relevance	Visual-public communication	The item contains or implies visual/video elements requiring description of actor, symbol, layout, colour, logo, frame, or transcript.	4	22.2%	Primary = 4; Secondary = 0; Context = 0	"AI membawa banyak manfaat, tetapi juga memicu tantangan baru."	IDNLLM-PRIM-0010; IDNLLM-PRIM-0011

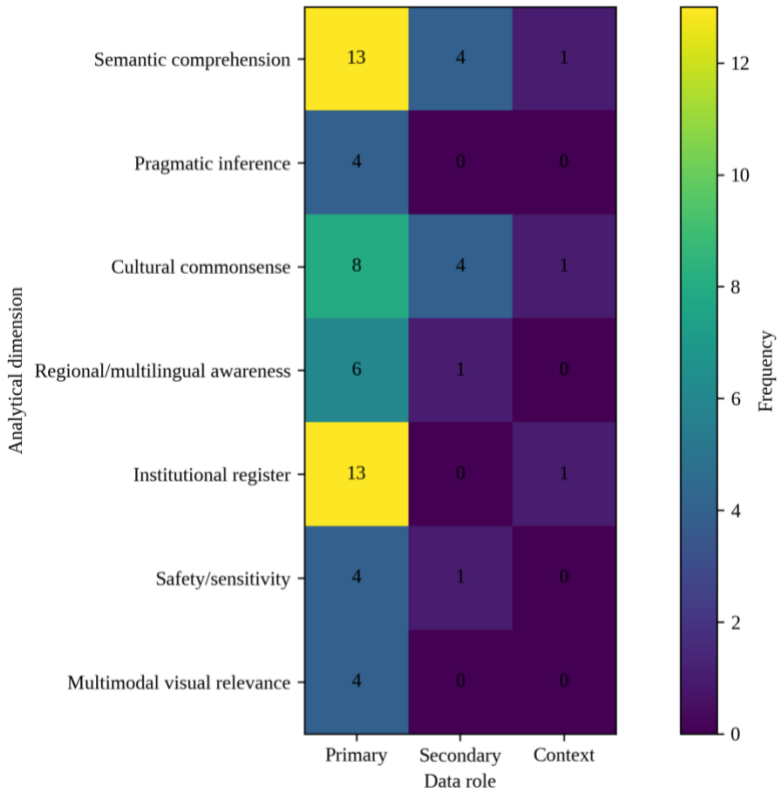

Figure 2. Distribution of coded Indonesian evaluation dimensions by data role

The dominant pattern is the density of semantic, institutional, and cultural dimensions. Semantic comprehension appears across all 18 items, institutional register appears in 14 items, and cultural commonsense appears in 13 items. Regional and multilingual awareness is present in seven items, while safety or sensitivity is present in five. Pragmatic inference and multimodal visual relevance are less frequent, each appearing in four items, but their lower frequency does not indicate lower analytical value. Instead, these categories mark higher-risk interpretive zones where literal translation is least sufficient. The distribution therefore shows that this corpus is not simply a collection of Indonesian texts; it is a structured evaluative field in which public authority, cultural knowledge, regional identity, and risk-sensitive communication intersect.

Theoretically, these patterns support the argument that culturally grounded Indonesian LLM evaluation should begin from communicative embeddedness rather than from translated benchmark transfer. IndoMMLU and IndoCulture have already shown that Indonesian evaluation requires educational, cultural, and geographically influenced knowledge, while IndoSafety and SEA-HELM extend this concern toward safety and Southeast Asian multilinguality [5], [7], [8], [11]. This corpus adds a complementary position: Indonesian understanding should be assessed through the interaction of semantic comprehension, institutional register, cultural commonsense, regional awareness, and pragmatic inference. The implication is that benchmark validity depends on how well evaluation items preserve the social conditions under which Indonesian meanings become intelligible.

4.2. Evaluation task types and their source functions

The corpus does not distribute evaluation tasks randomly; it arranges them around the communicative functions through which Indonesian meanings become publicly intelligible. Literature mapping accounts for four items, while institutional-pragmatic and semantic-policy comprehension account for three items each. Cultural commonsense and sensitive-pragmatic reasoning each appear in two items, and the remaining categories appear once. This pattern reflects this study's methodological commitment to source-grounded evaluation: secondary benchmark documents situate the inquiry, primary institutional and cultural documents provide task material, and contextual material anchors national interpretive frames. This organization aligns with recent multilingual evaluation work that questions whether translated benchmarks can adequately measure locally embedded reasoning [1], [14], [11].

Table 3. Distribution of evaluation task types across corpus documents

Evaluation Task Type	Source Function	Frequency	Percentage	Data-Role Distribution	Dimension Load	Representative Evidence	Document Code(s)
Literature mapping	Locates this study within Indonesian and Southeast Asian LLM benchmark development	4	22.2%	Primary = 0; Secondary = 4; Context = 0	2–3	"multi-task language understanding benchmark for Indonesian culture and languages"	IDNLLM-SEC-0015; IDNLLM-SEC-0016; IDNLLM-SEC-0017; IDNLLM-SEC-0018
Institutional-pragmatic comprehension	Tests interpretation of public-facing institutional discourse, regulation, warning, and civic instruction	3	16.7%	Primary = 3; Secondary = 0; Context = 0	3–6	"Rupiah sebagai alat pembayaran yang SAH."	IDNLLM-PRIM-0005; IDNLLM-PRIM-0006; IDNLLM-PRIM-0007
Semantic-policy comprehension	Tests understanding of policy, literacy, language planning, and educational concepts	3	16.7%	Primary = 3; Secondary = 0; Context = 0	2–4	"Korpus yang berisi kumpulan data bahasa."	IDNLLM-PRIM-0003; IDNLLM-PRIM-0004; IDNLLM-PRIM-0009
Cultural commonsense reasoning	Tests interpretation of ritual, collective work, tolerance, heritage, and local identity	2	11.1%	Primary = 2; Secondary = 0; Context = 0	4–5	"Gotong royong menjadi bahasa bersama yang menjembatani keberagaman."	IDNLLM-PRIM-0001; IDNLLM-PRIM-0002
Sensitive-pragmatic reasoning	Tests interpretation of tolerance, religious moderation, and socially careful meaning	2	11.1%	Primary = 2; Secondary = 0; Context = 0	4–4	"Toleransi membuka ruang dialog."	IDNLLM-PRIM-0012; IDNLLM-PRIM-0013
Claim verification and label interpretation	Tests interpretation of misinformation labels, public correction, and verification cues	1	5.6%	Primary = 1; Secondary = 0; Context = 0	4–4	"[HOAKS] Rupiah Sengaja Dilemahkan Bank Indonesia."	IDNLLM-PRIM-0011
Contextual comprehension	Supplies ideological and national-context anchoring for diversity discourse	1	5.6%	Primary = 0; Secondary = 0; Context = 1	3–3	"Pancasila menyatukan dengan segala perbedaan."	IDNLLM-CONT-0014
Local-term contextual reasoning	Tests whether a local term can be interpreted through source-grounded cultural context	1	5.6%	Primary = 1; Secondary = 0; Context = 0	4–4	"Smong adalah tradisi lisan."	IDNLLM-PRIM-0008
Policy-discourse reasoning	Tests interpretation of AI ethics, national character, and public policy language	1	5.6%	Primary = 1; Secondary = 0; Context = 0	5–5	"AI membawa banyak manfaat, tetapi juga memicu tantangan baru."	IDNLLM-PRIM-0010

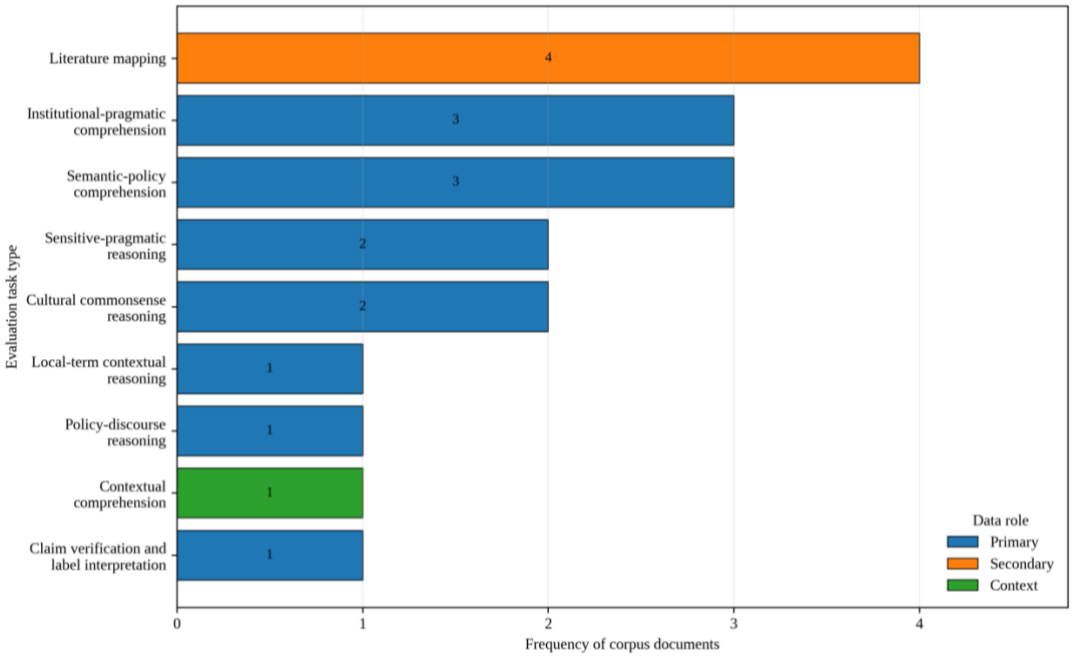

Figure 3. Task-type distribution by data role

The most visible pattern is the dominance of public and institutional language as a testing environment for Indonesian understanding. Institutional-pragmatic comprehension and semantic-policy comprehension together account for six documents, indicating that formal Indonesian is central to the corpus rather than peripheral. Cultural and religious-pragmatic tasks add four further documents, producing a corpus in which bureaucratic clarity, local value, tolerance, ritual, and public protection co-exist. The secondary literature items remain analytically separate, because their function is to define benchmark positioning rather than to serve as native prompt evidence. This distribution prevents the corpus from collapsing into either folklore-based cultural testing or purely technical benchmark comparison.

Theoretically, task-type distribution strengthens the argument that culturally grounded Indonesian evaluation should be built around communicative function, not only topic classification. IndoMMLU, IndoCulture, IndoSafety, and SEA-HELM demonstrate that Indonesian and Southeast Asian evaluation must include educational knowledge, cultural commonsense, safety, and regional multilinguality [5], [7], [8], [11]. The present distribution extends that insight by showing how those domains can be operationalized as evaluative tasks: institutional-pragmatic comprehension, semantic-policy comprehension, cultural commonsense reasoning, sensitive-pragmatic reasoning, and verification-based interpretation. This means that the relevant question is not whether a model "knows Indonesian," but whether it can interpret Indonesian as public, cultural, regional, and ethically constrained discourse.

4.3. Evaluation sensitivity and dimensional load of the corpus

Evaluation sensitivity is unevenly distributed across the corpus because not all Indonesian documents demand the same degree of interpretive layering. Literature mapping has the highest frequency but a lower mean dimension load, while policy-discourse reasoning, cultural commonsense reasoning, institutional-pragmatic comprehension, and sensitive-pragmatic reasoning carry higher evaluative density. This distinction matters because benchmark design is not strengthened merely by adding more documents; it is strengthened by identifying which documents combine multiple interpretive demands. Recent work on multilingual and culturally grounded evaluation similarly suggests that task validity depends on how prompts encode knowledge, context, safety, and social meaning rather than on linguistic translation alone [1], [14], [17].

The clearest pattern is that primary public sources carry the highest sensitivity load. The most dimensionally dense document is IDNLLM-PRIM-0007, which combines institutional register, financial warning, regional or youth-facing communication, pragmatic inference, safety, and multimodal relevance. Cultural and religious-public items also show high density because they require interpretation of ritual, tolerance, gotong royong, moderation, local wisdom, and public values. Secondary benchmark sources remain important, but their main function is comparative and conceptual rather than directly diagnostic. The distribution therefore separates two kinds of evidence: benchmark literature that frames the problem and native Indonesian public discourse that operationalizes the problem as testable language-understanding material.

The theoretical implication is that culturally grounded Indonesian evaluation should prioritize dimensional density, not only topic diversity. IndoMMLU and IndoCulture demonstrate that Indonesian knowledge and cultural commonsense require explicit benchmarking, while IndoSafety and SEA-HELM show that safety and regional multilinguality are integral to Southeast Asian model evaluation [5], [7], [8], [11]. The present sensitivity profile extends that argument by showing that Indonesian prompts become methodologically stronger when they combine semantic content with public register, cultural inference, regional grounding, and risk-sensitive interpretation. This supports this study's core position: translated benchmarks may test language transfer, but source-grounded Indonesian tasks test situated understanding.

Table 4. Evaluation sensitivity by task type and coded dimension load

Evaluation Task Type	Frequency	Percentage	Mean Dimension Load	Load Range	Dominant Data Role	Highest-Load Document(s)	Key Interpretive Demand
Literature mapping	4	22.2%	2.5	2–3	Secondary	IDNLLM-SEC-0016; IDNLLM-SEC-0017	Benchmark comparison, Indonesian positioning, cultural and safety framing
Institutional-pragmatic comprehension	3	16.7%	4.0	3–6	Primary	IDNLLM-PRIM-0007	Public authority, warning, legality, implied action, youth-facing literacy
Semantic-policy comprehension	3	16.7%	3.3	2–4	Primary	IDNLLM-PRIM-0003; IDNLLM-PRIM-0004	Policy meaning, language planning, literacy, preservation
Cultural commonsense reasoning	2	11.1%	4.5	4–5	Primary	IDNLLM-PRIM-0001	Ritual meaning, tolerance, gotong royong, local economy, regional identity
Sensitive-pragmatic reasoning	2	11.1%	4.0	4–4	Primary	IDNLLM-PRIM-0012; IDNLLM-PRIM-0013	Religious moderation, tolerance, dialogic stance, careful public language
Policy-discourse reasoning	1	5.6%	5.0	5–5	Primary	IDNLLM-PRIM-0010	AI ethics, national character, Pancasila, public policy
Claim verification and label interpretation	1	5.6%	4.0	4–4	Primary	IDNLLM-PRIM-0011	Hoax label, evidentiality, verification, institutional trust
Local-term contextual reasoning	1	5.6%	4.0	4–4	Primary	IDNLLM-PRIM-0008	Local wisdom, disaster mitigation, oral tradition, cultural memory
Contextual comprehension	1	5.6%	3.0	3–3	Context	IDNLLM-CONT-0014	National identity, diversity, ideological reference

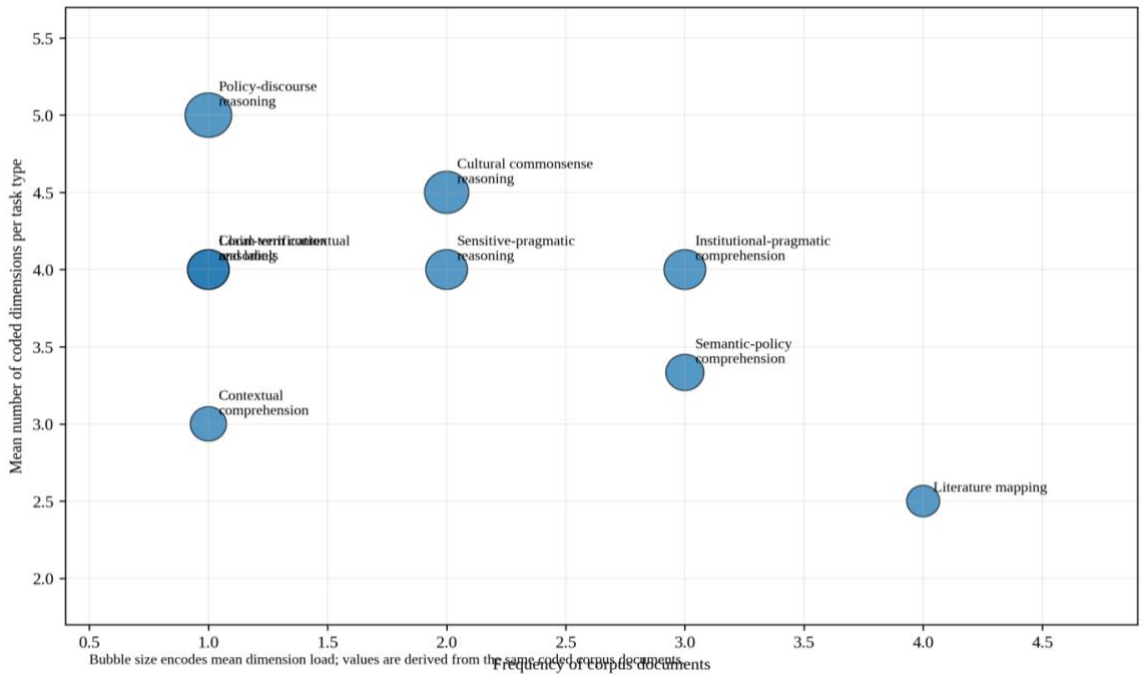

Figure 2. Evaluation sensitivity by task frequency and mean dimension load

5. DISCUSSION

The distribution of coded dimensions shows that Indonesian language understanding cannot be evaluated as a purely linguistic achievement. Its most important implication is that semantic comprehension functions as an entry condition, not as sufficient evidence of understanding. When institutional register, cultural commonsense, regional awareness, safety sensitivity, and pragmatic inference appear alongside semantic meaning, model evaluation must move from surface correctness toward contextual accountability. This finding supports the critique advanced in multilingual evaluation studies that translated benchmarks often preserve source-language assumptions while giving an appearance of cross-lingual comparability [1], [2], [14]. The functional value of this finding lies in its diagnostic clarity: an LLM may correctly identify topic or reference but still misread public authority, local tradition, or indirect social meaning. This means that Indonesian evaluation requires layered scoring, because failures may occur after semantic access has already been achieved.

The underlying structure behind this pattern is Indonesia's sociolinguistic organization itself. Indonesian operates simultaneously as a national language, an institutional language, a medium of digital communication, and a contact zone shaped by regional languages and local cultural frames. This explains why the corpus does not separate semantics from culture, register, or regional meaning. The pattern is consistent with IndoCulture, which foregrounds geographically influenced cultural commonsense, and with LoraxBench and Komodo, which show that Indonesian evaluation must account for the multilingual ecology beneath national-language use [4], [8], [9]. Yet this study also complicates benchmark traditions that treat multilingual evaluation primarily as language transfer. The relevant structure is not only low-resource scarcity or translation quality, but the embedding of Indonesian meanings in public discourse, regional memory, institutional authority, and pragmatic social relations.

The distribution of evaluation task types further implies that benchmark design should be organized around communicative function rather than topic coverage alone. Institutional-pragmatic, semantic-policy, cultural commonsense, sensitive-pragmatic, verification-based, and local-term tasks each place different interpretive demands on a model. This matters because task labels such as "question answering" or "classification" may obscure the social work performed by Indonesian prompts. A public warning, a religious-moderation statement, a cultural ritual description, and a benchmark paper all require different forms of interpretive discipline. This finding extends IndoMMLU and SEA-HELM, which emphasize Indonesian and Southeast Asian evaluation across knowledge, language, and cultural dimensions [7], [11]. Its implication is methodological: culturally grounded evaluation should not merely add cultural topics to existing formats, but should design tasks according to how Indonesian texts function in public, cultural, regulatory, and regional settings.

The structure underlying these task functions is an asymmetry between benchmarking as comparison and benchmarking as interpretation. Secondary benchmark documents make this study legible within wider

LLM evaluation debates, but primary public sources make Indonesian understanding testable as situated discourse. This difference explains why literature-mapping items are analytically important yet less suitable as direct prompt material than institutional or cultural documents. Global MMLU, BenchMAX, and MMLU-ProX seek comparability across languages and domains, while culturally specific benchmarks such as HAE-RAE Bench, KULTURE Bench, HKCanto-Eval, and Cetvel show that locally grounded tasks can reveal limitations hidden by general multilingual scores [21], [22], [12], [14], [25], [15]. This study therefore occupies a middle position: it preserves auditability and comparison while resisting the assumption that translation is the natural starting point for Indonesian evaluation.

The sensitivity profile of task types carries a stronger implication for model assessment: not all documents contribute equally to evaluative power. Items with high dimensional load are especially valuable because they require models to coordinate semantic content, institutional role, cultural inference, pragmatic restraint, and risk awareness. This does not make lower-load items irrelevant; rather, it clarifies their function as baseline, contextual, or comparative evidence. The finding aligns with cultural evaluation research arguing that culture should be treated as thick context, not as isolated factual recall [17], [18], [27], [28]. It also resonates with IndoSafety, where safety is culturally and linguistically grounded rather than universally reducible to abstract harm categories [5]. The implication is that Indonesian LLM evaluation should weight interpretive density carefully, because complex documents expose failures that ordinary accuracy metrics may smooth over.

The reason dimensionally dense tasks matter is that LLM failure often emerges at the boundary between fluency and situated judgement. A model may produce grammatically fluent Indonesian while flattening regional culture, overgeneralizing religious moderation, misreading institutional warning labels, or inventing unsupported context. This pattern does not prove a universal limitation of all LLMs, but it identifies where evaluation should look for fragile understanding. It also qualifies more optimistic accounts of multilingual benchmark expansion: broader language coverage is necessary, yet coverage alone does not guarantee cultural or pragmatic adequacy [1], [3], [10]. The contribution of this study is therefore theoretical as well as methodological. It reframes Indonesian LLM evaluation as source-grounded interpretation under culturally specific constraints, showing why translated benchmarks are useful for comparison but insufficient for assessing whether models understand Indonesian as lived, institutional, regional, and public language.

6. CONCLUSION

This study shows that Indonesian language understanding in large language models cannot be adequately evaluated through translated benchmarks alone. Its most important lesson is that semantic fluency is only the lowest layer of competence; culturally grounded evaluation must also examine institutional register, pragmatic inference, regional awareness, safety sensitivity, and source-grounded cultural commonsense. By constructing an auditable corpus from public Indonesian discourse, this study shifts benchmark design from language transfer to situated interpretation. Its scholarly contribution lies in renewing the methodological question of multilingual LLM evaluation: not simply whether models can answer in Indonesian, but whether they can interpret Indonesian as a lived, institutional, regional, and culturally mediated communicative system.

This study is limited by its diagnostic corpus design, which prioritizes depth, verifiability, and interpretive density over national representativeness. The corpus cannot claim to cover all Indonesian regions, languages, institutions, genres, or cultural traditions, and this study does not make causal claims about model architecture, training data, or system-level performance. Future research should expand the corpus across more regional languages, multimodal platforms, institutional domains, and locally validated cultural scenarios. Subsequent work should also test multiple LLMs under controlled prompting conditions, involve expert and community-based annotators, and compare source-grounded Indonesian evaluation with translated benchmark performance to clarify where multilingual fluency diverges from culturally situated understanding.

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

Achraf El Bouazzaoui: 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 Indonesian documents and benchmark materials.

ETHICAL APPROVAL

This research did not involve human or animal subjects; it analyzed publicly accessible Indonesian documents and benchmark 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] Y.-C. Chang, X. Wang, J. Wang, Y. Wu, K. Zhu, H. Chen, et al., "A survey on evaluation of large language models," *ACM Transactions on Intelligent Systems and Technology*, vol. 15, pp. 1–45, 2023, doi: 10.1145/3641289.

[2] V. D. Lai, N. T. Ngo, A. P. B. Veyseh, H. Man, F. Dernoncourt, T. Bui, and T. H. Nguyen, "ChatGPT beyond English: Towards a comprehensive evaluation of large language models in multilingual learning," *arXiv*, 2023, doi: 10.48550/arxiv.2304.05613.

[3] H. Naveed, A. Khan, S. Qiu, M. Saqib, S. Anwar, M. Usman, N. Barnes, and A. S. Mian, "A comprehensive overview of large language models," *ACM Transactions on Intelligent Systems and Technology*, vol. 16, pp. 1–72, 2023, doi: 10.1145/3744746.

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

[5] M. F. Azmi, M. D. A. Kautsar, A. Wicaksono, and F. Koto, "IndoSafety: Culturally grounded safety for LLMs in Indonesian languages," *arXiv*, 2025, doi: 10.48550/arxiv.2506.02573.

[6] S. Cahyawijaya, G. I. Winata, B. Wilie, K. Vincentio, X. Li, A. Kuncoro, et al., "IndoNLG: Benchmark and resources for evaluating Indonesian natural language generation," *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021, doi: 10.18653/v1/2021.emnlp-main.699.

[7] F. Koto, N. Aisyah, H. Li, and T. Baldwin, "Large language models only pass primary school exams in Indonesia: A comprehensive test on IndoMMLU," *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, 12359–12374*, 2023, doi: 10.48550/arxiv.2310.04928.

[8] 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.

[9] L. Owen, V. Tripathi, A. Kumar, and B. Ahmed, "Komodo: A linguistic expedition into Indonesia's regional languages," *arXiv*, 2024, doi: 10.48550/arxiv.2403.09362.

[10] 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.

[11] 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.

[12] X. Huang, W. Zhu, H. Hu, C. He, L. Li, S. Huang, and F. Yuan, "BenchMAX: A comprehensive multilingual evaluation suite for large language models," *arXiv*, 2025, doi: 10.48550/arxiv.2502.07346.

[13] T. Chang, C. Arnett, A. Eldesokey, A. Sadallah, A. Kashar, A. Daud, et al., "Global PIQA: Evaluating physical commonsense reasoning across 100+ languages and cultures," *arXiv*, 2025, doi: 10.48550/arxiv.2510.24081.

[14] S. Singh, A. Romanou, C. Fourrier, D. I. Adelani, J. G. Ngui, D. Vila-Suero, et al., "Global MMLU: Understanding and addressing cultural and linguistic biases in multilingual evaluation," *arXiv*, 2024, doi: 10.48550/arxiv.2412.03304.

[15] W. Xuan, R. Yang, H. Qi, Q. Zeng, Y. Xiao, Y. Xing, et al., "MMLU-ProX: A multilingual benchmark for advanced large language model evaluation," *arXiv*, 2025, doi: 10.48550/arxiv.2503.10497.

[16] W. Zhang, S. M. Aljunied, C. Gao, Y. K. Chia, and L. Bing, "M3Exam: A multilingual, multimodal, multilevel benchmark for examining large language models," *arXiv*, 2023, doi: 10.48550/arxiv.2306.05179.

[17] J. Myung, N. Lee, Y. Zhou, J. Jin, R. A. Putri, D. Antypas, et al., "BLEnD: A benchmark for LLMs on everyday knowledge in diverse cultures and languages," *arXiv*, 2024, doi: 10.48550/arxiv.2406.09948.

[18] S. Shen, L. Logeswaran, M. Lee, H. Lee, S. Poria, and R. Mihalcea, "Understanding the capabilities and limitations of large language models for cultural commonsense," *arXiv*, 2024, doi: 10.48550/arxiv.2405.04655.

[19] A. Mushtaq, I. Taj, M. Naeem, I. Ghaznavi, and J. Qadir, "WorldView-Bench: A benchmark for evaluating global cultural perspectives in large language models," *arXiv*, 2025, doi: 10.48550/arxiv.2505.09595.

[20] S. Z. Ridoy, A. T. Wasi, and K. A. Tonmoy, "BengaliMoralBench: A benchmark for auditing moral reasoning in large language models within Bengali language and culture," *arXiv*, 2025, doi: 10.48550/arxiv.2511.03180.
[21] T. C. Cheng, C. S. Cheng, C.-M. Lau, E. Lam, C. Y. Wong, H. O. Yu, and C. H. Chong, "HKCanto-Eval: A benchmark for evaluating Cantonese language understanding and cultural comprehension in LLMs," *arXiv*, 2025, doi: 10.48550/arxiv.2503.12440.
[22] Y. A. Er, I. Kesen, G. Şahin, and A. Erdem, "Cetvel: A unified benchmark for evaluating language understanding, generation and cultural capacity of LLMs for Turkish," *arXiv*, 2025, doi: 10.48550/arxiv.2508.16431.
[23] A. Maji, R. Kumar, A. Ghosh, A. Anushka, and S. Saha, "SANSKRITI: A comprehensive benchmark for evaluating language models' knowledge of Indian culture," *arXiv*, 2025, doi: 10.48550/arxiv.2506.15355.
[24] O. Nacar, S. Sibaee, S. Ahmed, S. B. Atitallah, A. Ammar, Y. AlHabashi, et al., "Towards inclusive Arabic LLMs: A culturally aligned benchmark in Arabic large language model evaluation," *Proceedings, 387–401*, 2025.
[25] X. Wang, J. Yeo, J.-H. Lim, and H. Kim, "KULTURE Bench: A benchmark for assessing language model in Korean cultural context," *arXiv*, 2024, doi: 10.48550/arxiv.2412.07251.
[26] H.-S. Lee, C.-C. Chang, C.-Y. Chen, and Y.-H. Hsu, "Evaluating cultural knowledge processing in large language models: A cognitive benchmarking framework integrating retrieval-augmented generation," *arXiv*, 2025, doi: 10.48550/arxiv.2511.01649.
[27] T. Vo and O. Koyejo, "CURE: Cultural understanding and reasoning evaluation: A framework for "thick" culture alignment evaluation in LLMs," *arXiv*, 2025, doi: 10.48550/arxiv.2511.12014.
[28] J. Zhang, S. Jiang, S. Guo, S. Chen, Y. Xiao, H. Feng, et al., "CultureScope: A dimensional lens for probing cultural understanding in LLMs," *arXiv*, 2025, doi: 10.48550/arxiv.2509.16188.
[29] H. Ahmadian, T. F. Abidin, H. Riza, and K. Muchtar, "Hybrid models for emotion classification and sentiment analysis in Indonesian language," *Applied Computational Intelligence and Soft Computing*, vol. 2024, Art. no. 2826773, 2024, doi: 10.1155/2024/2826773.
[30] X. Huang, T. K. Vangani, M. D. Pham, X. Zou, B. Wang, Z. Liu, and A. Aw, "MERaLiON-TextLLM: Cross-lingual understanding of large language models in Chinese, Indonesian, Malay, and Singlish," *arXiv*, 2024, doi: 10.48550/arxiv.2501.08335.

BIOGRAPHY OF AUTHOR

Achraf El Bouazzaoui a project engineer at ICFO (The Institute of Photonic Sciences) within the Optoelectronics research group. He holds a PhD in Embedded Electronics and Intelligent Systems from Ibn Tofail University in Morocco. His research interests include Embedded Electronics & Systems; FPGA Implementation & Hardware Acceleration; Intelligent Systems & Neural Networks; Optoelectronics. He can be contacted at email: achraf.elbouazzaoui@icfo.eu