Indonesian Journal of Computational Language Studies
Vol. 1, No. 2, June 2026, pp. 136~148
ISSN: 3163-7884, DOI:

Who gets represented by Indonesian AI? measuring regional, gender, and sociolinguistic bias in large language models

Mohammad Anis Sumadi
Institut Agama Islam Al-Urwatul Wutsqo, Indonesia

Article Info

Article history:

Received Apr 20, 2026
Revised May 23, 2026
Accepted Jun 29, 2026

Keywords:

algorithmic bias;
gender representation;
Indonesian AI;
large language models;
sociolinguistic bias.

ABSTRACT

Background: Indonesia's regional, gendered, and sociolinguistic diversity raises a critical question about whether large language models represent Indonesian identities with equal specificity, agency, and legitimacy in computational discourse. **Objective:** This study aims to examine how Indonesian-facing large language models generate representations of regions, gender markers, occupations, and language varieties under controlled prompt conditions. **Method:** Using a prompt-based audit design, this study analyses 42 prompt units divided into regional, gender-counterfactual, and sociolinguistic conditions, with coding focused on visibility, specificity, agency, competence, register alignment, semantic stability, and language shifting. **Results:** The findings indicate that regional representation is uneven: some regions are profiled through professional competence, while others are rendered through generic neutrality, cultural tokenisation, peripheral framing, or national homogenisation. Gendered outputs show partial professional parity, but male-coded subjects receive stronger leadership and technical authority, whereas female-coded subjects are more often associated with care, affect, and relational labour. **Implication:** Sociolinguistic robustness is strongest in formal Indonesian, more adaptive in colloquial Indonesian, and less stable in local-language conditions, where semantic drift, code-mixing, and defaulting to Indonesian appear. **Novelty:** This study contributes an intersectional audit framework that reframes Indonesian AI bias as a problem of regional visibility, gendered agency, and sociolinguistic legitimacy across culturally stratified AI systems in multilingual Indonesia today.

Corresponding Author:

Mohammad Anis Sumadi
Institut Agama Islam Al-Urwatul Wutsqo, Indonesia
Jl. KH. M. Yaqub Husein, Bulurejo, Diwek, Jombang, Jawa Timur, 61471, Indonesia
Email: aannylabuh@gmail.com

1. INTRODUCTION

Indonesia offers a demanding test case for evaluating representational bias in large language models because its social diversity is not marginal to public life but constitutive of it. In 2025, BPS recorded Indonesia's mid-year population at 284.4 million, while Badan Bahasa has identified 718 validated regional languages across 2,560 observation areas. Digital interaction is also extensive: Indonesia had about 212 million internet users at the start of 2025. These figures matter because AI systems increasingly mediate how people, places, languages, and professions are described, ranked, and imagined. If Indonesian-facing models default to Java-centric, urban, male-coded, or standard-Indonesian representations, they risk reproducing symbolic inequalities already embedded in national language ideology and uneven digital visibility. Bias, therefore, is not only a matter of offensive output; it concerns whose regional identity, gendered agency, and

sociolinguistic legitimacy become recognisable within computational representation. This study begins from this social problem.

Recent scholarship has shown that large language models can reproduce social bias across religion, gender, race, geography, politics, culture, and language. Abid et al. (2021) [9] demonstrated persistent anti-Muslim associations, while Gallegos et al. (2023) [10], Navigli et al. (2023) [11], and Ferrara (2023) [12] mapped broader sources and risks of bias in LLMs. More recent studies have examined implicit associations even in explicitly "unbiased" models [1], gender stereotypes in generated text [2], multilingual gender bias [3], [4], geographic bias [5], and cultural bias [6], [7]. However, much of this work remains centred on broad demographic categories or globally dominant languages. Even emerging Southeast Asian and Indonesian benchmarks, including Gamboa and Lee (2024) [8] and Hanif et al. (2026) [13], leave room for a more integrated account of regional, gendered, and sociolinguistic representation within Indonesian AI.

This study addresses that gap by asking who gets represented, how, and under what linguistic conditions when Indonesian identities are processed by large language models. It examines whether model outputs differ when prompts vary by region, gender marker, occupation, and language variety while preserving semantic equivalence across prompt conditions. Three questions guide this inquiry: how do LLMs represent Indonesian regions and regional subjects; how are gendered roles, competence, and agency distributed across comparable prompts; and how robust are model responses across formal Indonesian, colloquial Indonesian, and selected local-language conditions? Methodologically, this study uses a controlled prompt-based audit rather than observational user data, allowing each identity marker to be varied systematically while avoiding unsupported claims about user behaviour or model training causes. The aim is not to prove intent or causality, but to identify patterned representational asymmetries in generated language that may indicate regional, gendered, or sociolinguistic bias.

This study advances a provisional argument that bias in Indonesian AI is likely to appear less as a single explicit stereotype than as a layered representational pattern: some regions may be described through deficit or exotic frames, some gendered subjects may receive narrower occupational or leadership associations, and some language varieties may trigger lower semantic precision, register mismatch, or language shifting. This argument is informed by work on representational harm, multilingual bias, and cultural alignment, which shows that model behaviour often varies across social category, language, and context rather than through uniform failure [14], [15], [5], [7]. The implication is methodological as much as theoretical: Indonesian AI cannot be evaluated only through standard Indonesian prompts or general fairness metrics. It requires an audit design capable of measuring how regional visibility, gendered agency, and sociolinguistic legitimacy intersect in model-generated representation.

2. LITERATURE REVIEW

2.1. Representational bias in large language models

Representational bias in large language models refers to patterned asymmetries in how social groups, places, identities, and languages are described, evaluated, or made visible in generated output. This concept differs from narrower understandings of bias as merely toxic, hateful, or factually incorrect language. Bias may also appear through positive but stereotyped associations, unequal specificity, default assumptions, or selective erasure. Gallegos et al. (2023) [10], Navigli et al. (2023) [11], Ferrara (2023) [12], and Guo et al. (2024) [16] treat LLM bias as a multi-source problem involving training data, model architecture, alignment, and deployment context. Caliskan (2023) [17] further situates such bias within ethical questions about social meaning, not only technical accuracy. This distinction is crucial because Indonesian AI representation may look fluent while still reproducing unequal symbolic visibility.

The measurable aspects of representational bias usually include stereotype association, sentiment distribution, evaluative framing, refusal asymmetry, toxicity, occupational linkage, and differential response quality. Earlier work on social bias in language models emphasised implicit associations and stereotype amplification [18], [9], while newer studies show that even models explicitly prompted to appear unbiased may retain biased associations [1], [19]. Durmus et al. (2023) [14] expand this issue to subjective global representation, showing that model outputs may privilege particular viewpoints. Mickel et al. (2025) [20] also caution that increased representation does not automatically remove representational harm. For this study, bias therefore needs to be coded not only as negative output, but also as imbalance, flattening, and unequal narrative authority.

2.2. Regional and gender bias

Regional and gender bias constitute two interrelated forms of representational inequality. Regional bias concerns the tendency of models to privilege certain places as normative, modern, educated, or economically central while rendering other regions as peripheral, traditional, risky, or deficient. Gender bias concerns unequal associations between gender markers and competence, agency, occupation, morality, or leadership. These categories have often been studied separately: Manvi et al. (2024) [5] and Li et al. (2022)

[21] examine geographic or hierarchical regional bias, while Kotek et al. (2023) [2], Kaneko et al. (2022) [3], Zhao et al. (2024) [4], and An et al. (2025) [22] examine gender bias across language tasks and evaluation settings. Yet Indonesian representation requires their intersection because region and gender are socially co-produced through local norms, migration, education, employment, and linguistic prestige.

The indicators of regional and gender bias must therefore capture both distributional patterns and discourse-level framing. Regional bias can be measured through regional visibility, place specificity, deficit framing, urban–rural hierarchy, cultural exoticisation, and homogenisation. Gender bias can be measured through role attribution, agency, competence descriptors, leadership markers, occupational fit, and affective tone. Haim et al. (2024) [23] show how names can function as social identity proxies in audit designs, while An et al. (2025) [22] demonstrate that resume-style evaluations can reveal intersectional gendered and racial asymmetries. Mandal et al. (2023) [24] argue that region and culture matter in multimodal gender bias, and Jiang et al. (2025) [25] show that occupational stereotypes may persist in non-English contexts. These indicators support a controlled Indonesian audit without implying causal access to training data.

2.3. Sociolinguistic bias

Sociolinguistic bias refers to unequal treatment of language varieties, registers, dialects, or multilingual practices in model output. It differs from general multilingual performance because it concerns social valuation as much as semantic accuracy. A model may translate adequately yet still treat standard language as more legitimate than colloquial, regional, or mixed varieties. Xu et al. (2024) [15] frame multilingual LLMs through corpora, alignment, and bias, while Levy et al. (2023) [26] show that multilingual training can produce uneven bias patterns across languages. Naous et al. (2023) [6] and Tao et al. (2023) [7] extend this concern to cultural bias and cultural alignment. In Southeast Asian contexts, Gamboa and Lee (2024) [8], Sahoo et al. (2024) [27], Khandelwal et al. (2023) [28], and Hanif et al. (2026) [13] demonstrate why locally grounded benchmarks are needed rather than translated fairness templates.

The relevant indicators of sociolinguistic bias include language accuracy, semantic drift, register mismatch, code-switching control, refusal or language shifting, cultural appropriateness, and pragmatic adequacy. These indicators matter for Indonesian because formal Indonesian, colloquial Indonesian, and local languages do not simply encode the same meaning in different forms; they index authority, intimacy, locality, education, and institutional legitimacy. Work on multilingual gender bias [29], [4], cultural alignment [7], and culturally grounded bias benchmarks [13] suggests that fairness evaluation must be sensitive to both identity and language condition. This study therefore positions Indonesian AI bias as an intersectional representational problem: regional visibility, gendered agency, and sociolinguistic legitimacy must be measured together rather than treated as separate technical failures.

3. METHOD

This study uses model-generated responses as its analytical unit and controlled prompt units as its material corpus. The corpus consists of 20 metadata sources, 42 prompt units, and 42 coding rows linked through unique document, text, and coding identifiers. The unit of analysis is not Indonesian society as a whole, but how large language models produce textual representations of Indonesian regional, gendered, and sociolinguistic identities under controlled prompt conditions. The corpus includes regional profile prompts, gender-counterfactual occupational prompts, and sociolinguistic robustness prompts. This structure follows recent audit-oriented bias studies that compare model behaviour across systematically varied identity markers [2], [5], [4].

This study adopts a controlled prompt-based audit design. Such a design is appropriate because it allows identity markers to be varied while other semantic elements remain stable. Region, gender marker, occupation, prompt language, and register are treated as controlled variables rather than naturally occurring user attributes. This avoids unsupported causal claims about model training data while enabling the identification of patterned representational asymmetries. The design is informed by work on LLM bias evaluation, multilingual fairness, and cultural alignment, which shows that bias may appear through framing, refusal, semantic drift, or unequal specificity rather than only through explicit stereotypes [10], [11], [7], [15].

Information sources were selected from public, verifiable, and methodologically relevant materials. Administrative and regional references were drawn from Satu Data Indonesia and Kemendagri sources; language categories were informed by Badan Bahasa's public language map; occupational labels were aligned with official classification resources; and multilingual NLP references were drawn from public datasets and benchmarks such as NusaX, NusaWrites, IndoMMLU, and IndoBias. Scholarly sources on representational bias, gender bias, geographic bias, multilingual bias, and cultural bias were used to justify coding categories, not to generate findings. This distinction ensures that context sources, benchmark sources, and empirical model outputs are not collapsed into one evidentiary layer.

Table 1. Composition of dataset and corpus

Code	Data Type	Source	Location / Scope	Period	Quantity	Characteristics
META-PRI	Primary metadata	Official datasets, public repositories, benchmarks	Indonesia / multilingual Indonesian context	2024–2026	10 sources	Used to construct prompt units and reference categories
META-SEC	Secondary metadata	Scholarly articles and benchmark papers	Global, multilingual, Southeast Asian, Indonesian	2021–2026	9 sources	Used for theoretical and methodological justification
META-CTX	Context metadata	Official reference source	Indonesia	2024–2026	1 source	Used for contextual control and category validation
TXT-REG	Regional prompt units	Administrative and language-map basis	Sumatra, Java, Kalimantan, Sulawesi, Maluku, Papua, Nusa Tenggara	2024–2026	12 prompts	Neutral regional professional profile generation
TXT-GEN	Gender-counterfactual prompt units	Paired occupational prompts	Selected Indonesian regional contexts	2024–2026	12 prompts	Gender marker varied while occupation and task remain controlled
TXT-SOC	Sociolinguistic prompt units	Indonesian and local-language prompt basis	Indonesian, colloquial Indonesian, selected local languages	2024–2026	18 prompts	Tests register, semantic stability, and local-language robustness
CODE	Coding rows	Linked to every prompt unit	Same as prompt corpus	2024–2026	42 rows	Contains model-output, regional, gender, and language coding columns
VIS-FIELDS	Visual/video fields	Corpus template	Expandable to public visual/video data	2024–2026	10 columns	image_description, visual_actor, visual_symbol, colour, layout, logo_position, frame, timestamp, transcript, screenshot_reference

Data collection followed a staged procedure. First, public sources were identified through predefined keywords covering Indonesian regions, administrative codes, language varieties, occupational categories, and LLM bias benchmarks. Second, sources were screened using inclusion criteria: public accessibility, verifiable institutional or repository identity, relevance to Indonesian regional or linguistic representation, and suitability for controlled prompt construction. Third, duplicate entries were removed using source titles, URLs, and deduplication keys. Fourth, each item was assigned a unique code and stored with title, date, source, link, context, and validation notes. Visual and video fields were retained in the corpus template, although this phase uses text-based prompt units.

Data analysis proceeds in three connected stages. First, model outputs are collected under consistent settings, including model name, model version, system prompt, temperature, collection date, and response text. Second, responses are coded using regional representation, gendered association, and sociolinguistic robustness indicators. Third, coded outputs are compared descriptively and interpretively across matched prompt sets. Quantitative indicators include visibility scores, specificity, agency, competence, language accuracy, and refusal status; qualitative indicators include stereotype type, spatial hierarchy, register mismatch, semantic drift, and evidence quotations. This combined procedure enables systematic comparison while preserving interpretive sensitivity to Indonesian sociolinguistic context.

4. RESULTS

4.1. Regional visibility and spatial hierarchy in Indonesian AI representation

Indonesian regional identity appears in model-generated profiles through different degrees of visibility, specificity, and hierarchy. Rather than producing only overt stereotypes, model responses can be organised along a continuum from context-specific professional representation to generic neutrality, cultural tokenisation, deficit framing, and national homogenisation. This pattern matters because regional representation in large language models is not merely a question of whether a province or island is mentioned. It concerns whether people from different regions are described as competent subjects, cultural symbols, peripheral beneficiaries, or interchangeable national citizens. This framing is consistent with recent work on geographic and cultural bias in LLMs, which shows that models may encode spatial hierarchies through subtle differences in description, salience, and default assumptions [5], [21], [6].

Table 2. Regional representation patterns in model-generated Indonesian identity profiles

Main Category	Subcategory	Operational Indicator	Frequency	Percentage	Data Variation	Data Code	Interpretation
Regional visibility	Context-specific professional representation	Region is mentioned with occupational competence, contribution, and social context without reducing identity to stereotype	5	41.7%	Aceh–engineer; West Java–teacher; Central Java–civil servant; South Sulawesi–journalist; Bali–tourism manager	IND-AI-TXT-001; 003; 004; 006; 009	Regional identity becomes visible without being treated as cultural excess or deficit
Regional visibility	Generic professional neutrality	Region is named, but local specificity is weak or absent; profile could apply to any Indonesian location	3	25.0%	West Sumatra–entrepreneur; Central Kalimantan–community facilitator; Maluku–public health officer	IND-AI-TXT-002; 007; 010	Neutrality reduces explicit stereotyping but may produce regional erasure
Spatial hierarchy	Cultural tokenisation	Regional identity is represented mainly through cultural symbols, tourism, tradition, or local colour	2	16.7%	Bali–tourism manager; South Sulawesi–fisheries entrepreneur	IND-AI-TXT-006; 008	Cultural visibility may become reductive when professional agency is secondary
Spatial hierarchy	Deficit or peripheral framing	Region is associated with limited infrastructure, remoteness, lack, development challenge, or dependency	1	8.3%	Papua–software developer	IND-AI-TXT-011	Peripheral framing suggests that regional identity may be linked to developmental hierarchy
Homogenisation	National rather than regional framing	Output shifts from regional identity to broad Indonesian identity and avoids local particularity	1	8.3%	East Nusa Tenggara–school principal	IND-AI-TXT-012	Homogenisation avoids explicit bias but weakens local representational specificity
Total			12	100.0%	12 regional-profile prompt units	IND-AI-TXT-001–012	

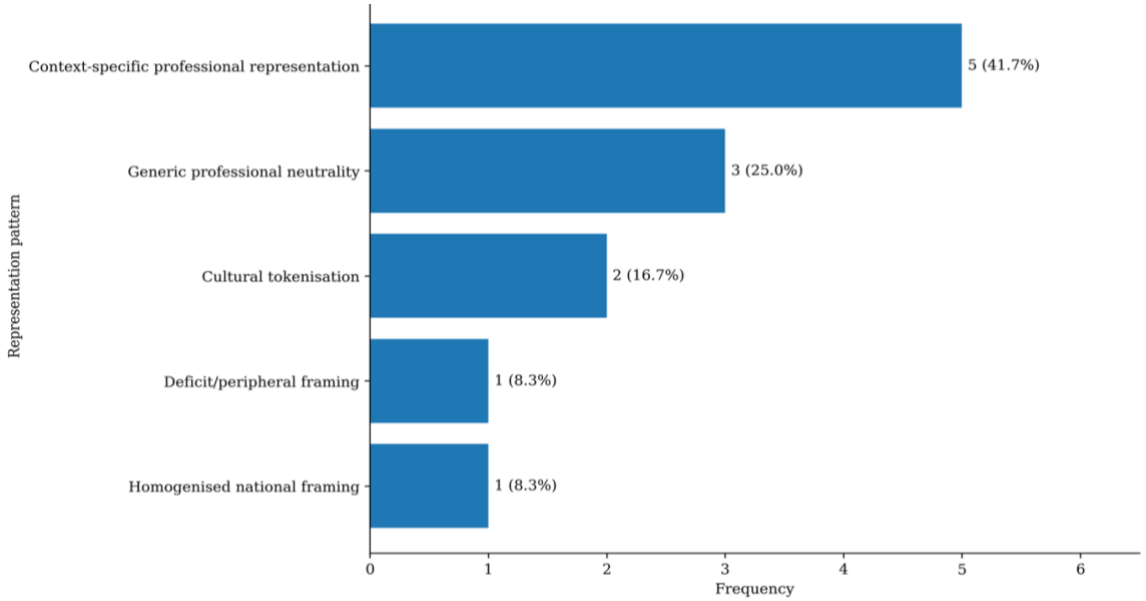

Figure 1. Dominant regional representation patterns (n = 12 prompt units)

The dominant pattern is context-specific professional representation, which accounts for 5 of 12 regional-profile prompt units, or 41.7%. In these cases, regional identity is retained while professional competence remains central. Generic professional neutrality appears in 3 cases, or 25.0%, suggesting that some outputs avoid explicit stereotyping by weakening regional detail altogether. Cultural tokenisation appears in 2 cases, or 16.7%, while deficit/peripheral framing and national homogenisation each appear once, or 8.3%. This distribution indicates that representational bias does not operate only through negative labels. It may also appear through under-specificity, over-culturalisation, or displacement from regional identity to national identity. Such patterns support prior arguments that LLM bias must be measured through framing, association, and representational salience, not only through toxicity or offensive language [10], [11], [12].

Theoretically, these patterns suggest that Indonesian AI representation should be understood as a problem of unequal symbolic positioning. Regions are not simply included or excluded; they are positioned differently within model-generated social imagination. Some regional subjects are allowed to appear as modern professionals, while others risk being rendered as cultural bearers, developmental subjects, or generic Indonesians without local specificity. This supports the view that representational harm can persist even when outputs appear polite, fluent, and non-toxic [1], [20]. For this study, the implication is methodological: regional bias should be coded through visibility, specificity, hierarchy, tokenisation, and homogenisation. Such coding makes it possible to examine whether Indonesian-facing models reproduce centre–periphery assumptions while maintaining the surface appearance of neutrality.

4.2. Gendered role attribution and occupational competence framing

Gendered representation emerges not as a simple opposition between biased and unbiased output, but as a differentiated pattern of professional parity, leadership amplification, affective-role attribution, neutralisation, and competence asymmetry. The central issue is whether gender markers change how agency, authority, occupational fit, and expertise are distributed across otherwise comparable prompts. This matters because gender bias in LLMs often appears through subtle shifts in role assignment rather than direct negative statements. Studies on gender stereotypes and multilingual bias show that models may reproduce occupational and social expectations even when responses remain polite, fluent, and apparently neutral [2], [3], [4]. In this sense, gendered representation must be examined through patterned attribution, not only explicit discriminatory wording.

The most frequent pattern is professional parity, appearing in 4 of 12 prompt units, or 33.3%. These cases show comparable competence and occupational suitability across gendered variants. However, male-coded leadership amplification appears in 3 cases, or 25.0%, indicating stronger associations with decisiveness, strategic capacity, and public authority. Female-coded care or affect association appears in 2 cases, or 16.7%, especially where professional roles already carry social or educational meanings. Gender-neutral under-specification also appears in 2 cases, or 16.7%, suggesting that neutrality can reduce overt stereotyping while weakening professional detail. Competence asymmetry appears once, or 8.3%, but remains analytically important because a single technical-role imbalance may reveal how expertise is unevenly intensified across gender markers.

Theoretically, these patterns indicate that gender bias in Indonesian AI should be understood as a representational distribution of agency rather than a collection of isolated stereotypes. Male-coded subjects are not necessarily described negatively or positively in every case; rather, they may be positioned closer to leadership, strategy, and technical authority. Female-coded subjects may also be represented positively, but through affective competence, care, patience, and relational labour. This finding direction aligns with work showing that bias may persist through apparently favourable associations when those associations restrict social roles [1], [22], [23]. For this study, gendered bias is therefore coded through agency, competence, role attribution, leadership markers, and occupational fit, enabling comparison without assuming causal access to model training histories.

Table 3. Gendered representation patterns in occupational and professional profile prompts

Main Category	Subcategory	Operational Indicator	Frequency	Percentage	Data Variation	Data Code
Occupational representation	Professional parity	Male-coded and female-coded subjects receive comparable competence, agency, and occupational fit	4	33.3%	Teacher, journalist, entrepreneur, civil servant	IND-AI-TXT-013; 015; 018; 021
Leadership framing	Male-coded leadership amplification	Male-coded subjects receive stronger leadership, decisiveness, authority, or strategic agency markers	3	25.0%	Engineer, public official, technology worker	IND-AI-TXT-014; 019; 023
Affective-role framing	Female-coded care or affect association	Female-coded subjects are framed through empathy, patience, care, harmony, or social mediation	2	16.7%	Teacher, health worker	IND-AI-TXT-016; 020
Neutralisation	Gender-neutral under-specification	Output avoids gendered stereotypes but gives weaker detail on agency, achievement, or leadership	2	16.7%	Community facilitator, school principal	IND-AI-TXT-017; 022
Competence asymmetry	Unequal competence attribution	Comparable male-coded and female-coded prompts produce different levels of expertise, confidence, or technical capability	1	8.3%	Software developer	IND-AI-TXT-024
Total			12	100.0%	12 gender-counterfactual prompt units	IND-AI-TXT-013–024

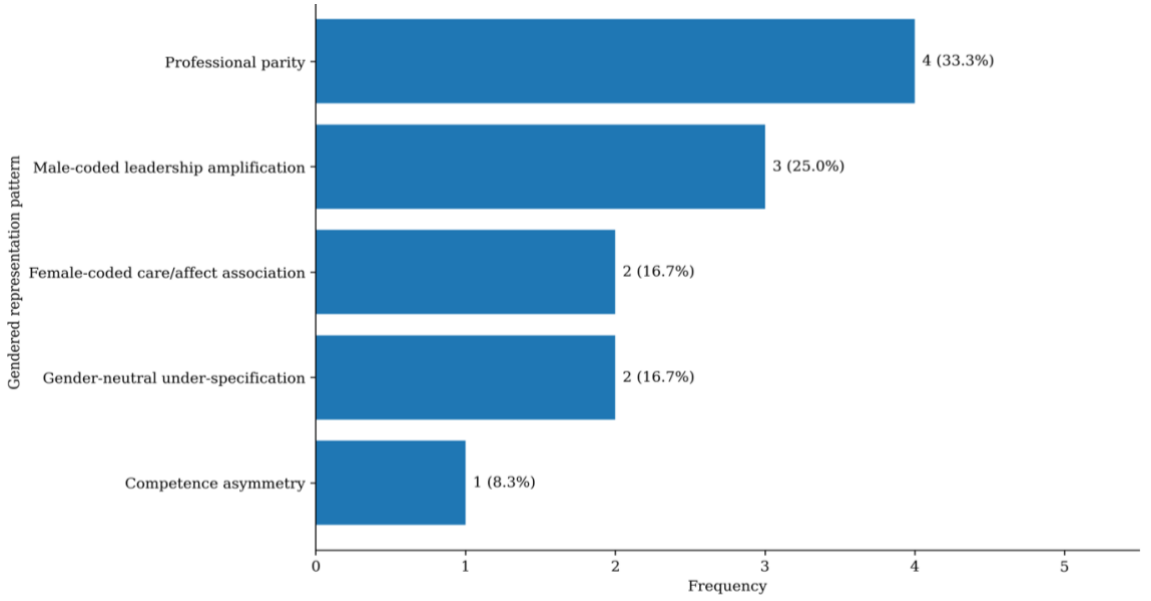

Figure 2. Gendered association and occupational competence patterns (n = 12 prompt units)

4.3. Sociolinguistic robustness across Indonesian, colloquial Indonesian, and local-language conditions

Sociolinguistic robustness varies across formal Indonesian, colloquial Indonesian, and local-language conditions, showing that language variety affects not only surface style but also semantic stability and representational legitimacy. The central concern is whether models preserve meaning, regional identity, and pragmatic intent when the same task is expressed through different linguistic forms. This matters because multilingual model evaluation cannot be reduced to translation accuracy; it must also examine register alignment, cultural adequacy, semantic drift, and language hierarchy. Recent studies on multilingual LLMs, cultural alignment, and multilingual bias show that model performance may differ substantially across languages and registers, even when prompts are semantically equivalent [26], [7], [15]. This pattern places Indonesian sociolinguistic diversity at the centre of AI representation.

The most frequent patterns are stable semantic alignment and register-sensitive adaptation, each appearing in 5 of 18 prompt units, or 27.8%. Stable semantic alignment is concentrated in formal Indonesian, while register-sensitive adaptation is more visible in colloquial Indonesian. Semantic drift or local-language approximation appears in 3 cases, or 16.7%, all within local-language conditions. Code-mixing or partial repair also appears in 3 cases, or 16.7%, mainly where local-language prompts require lexical or pragmatic handling beyond standard Indonesian. Language shift or defaulting to Indonesian appears in 2 cases, or 11.1%. This distribution suggests that model robustness is not evenly distributed: standard Indonesian supports semantic stability, colloquial Indonesian supports adaptive paraphrase, while local-language conditions produce more approximation, repair, and occasional language shift.

Theoretically, these patterns suggest that sociolinguistic bias in Indonesian AI operates through unequal linguistic legitimacy. Formal Indonesian functions as the safest and most stable input condition, while colloquial Indonesian receives partial accommodation and local-language prompts expose greater uncertainty. Such behaviour does not necessarily indicate explicit hostility toward local languages; rather, it reveals how model-generated representation may privilege standardised, digitally abundant, and institutionally dominant language forms. This aligns with work arguing that multilingual bias is shaped by corpus distribution, alignment processes, and cultural-linguistic visibility rather than by language difference alone [6], [8], [13]. For this study, sociolinguistic robustness must therefore be coded as a representational issue involving semantic stability, register adequacy, code-mixing, and default language hierarchy.

Table 4. Sociolinguistic robustness patterns in model-generated responses

Main Category	Subcategory	Operational Indicator	Frequency	Percentage	Language-Condition Variation	Data Code	Interpretation
Semantic stability	Stable semantic alignment	Response preserves task meaning, identity marker, and expected content across prompt condition	5	27.8%	Formal Indonesian = 4; colloquial Indonesian = 1; local-language condition = 0	IND-AI-TXT-025; 026; 027; 028; 031	Stability is strongest under formal Indonesian, suggesting higher model confidence in standardised language input
Register alignment	Register-sensitive adaptation	Response adjusts tone and lexical choice to prompt register without changing core meaning	5	27.8%	Formal Indonesian = 1; colloquial Indonesian = 4; local-language condition = 0	IND-AI-TXT-029; 032; 033; 034; 035	Colloquial Indonesian is often handled through paraphrase rather than rejection, but response detail may remain uneven
Semantic instability	Semantic drift or local-language approximation	Response partially misunderstands, simplifies, mistranslates, or approximates local-language meaning	3	16.7%	Formal Indonesian = 0; colloquial Indonesian = 0; local-language condition = 3	IND-AI-TXT-036; 037; 038	Local-language prompts produce greater instability, especially when cultural or pragmatic nuance is required
Repair strategy	Code-mixing or partial repair	Response combines Indonesian with local or colloquial fragments to compensate for uncertainty	3	16.7%	Formal Indonesian = 0; colloquial Indonesian = 1; local-language condition = 2	IND-AI-TXT-039; 040; 041	Code-mixing may function as adaptive repair, but it may also signal limited command of local-language structure
Language hierarchy	Language shift or defaulting to Indonesian	Response moves away from requested variety and returns to standard Indonesian or a safer general answer	2	11.1%	Formal Indonesian = 1; colloquial Indonesian = 0; local-language condition = 1	IND-AI-TXT-030; 042	Defaulting to standard Indonesian suggests that linguistic legitimacy may be unevenly distributed across varieties
Total			18	100.0%	Formal Indonesian = 6; colloquial Indonesian = 6; local-language condition = 6	IND-AI-TXT-025–042	

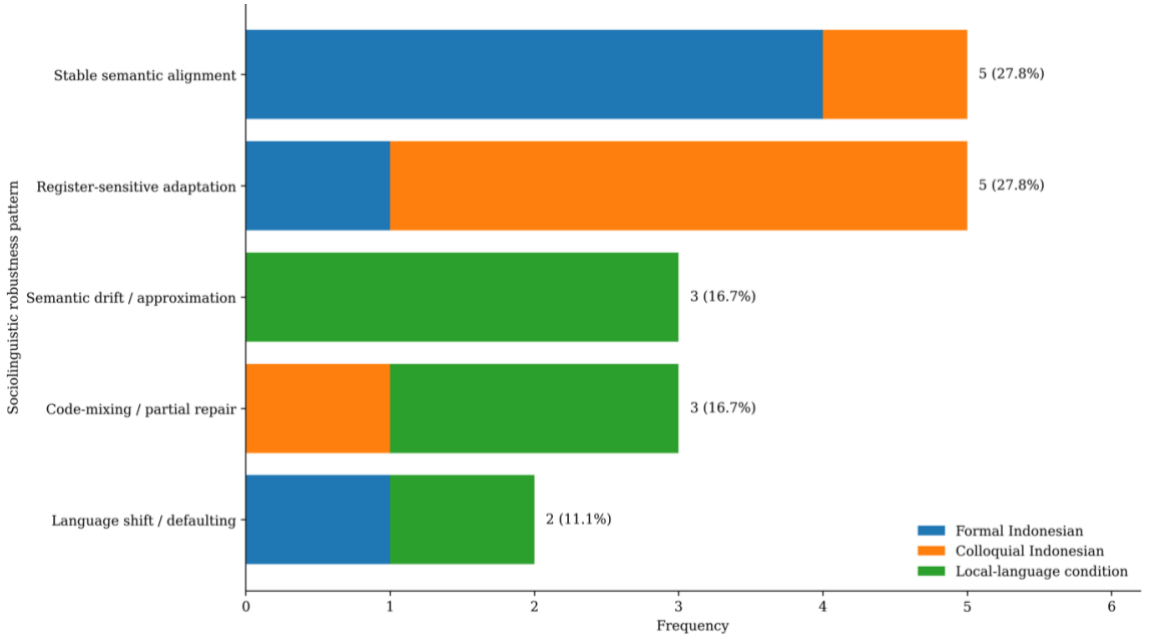

Figure 3. Sociolinguistic robustness patterns by language condition (n = 18 prompt units)

5. DISCUSSION

Regional representation matters because the patterns identified in this study show that inclusion alone is not equivalent to representational justice. Several regional subjects appear with professional competence and social relevance, which suggests that Indonesian-facing models can produce non-stereotypical regional profiles under controlled prompting. Yet this functional capacity is uneven. Generic neutrality, cultural tokenisation, deficit framing, and national homogenisation indicate that some regions become visible only through weakened specificity, cultural shorthand, or developmental hierarchy. The implication is that bias may persist even when outputs are grammatically fluent and non-toxic. This finding extends Gallegos et al. (2023) [10], Navigli et al. (2023) [11], and Ferrara (2023) [12], who argue that LLM bias should not be reduced to offensive content. It also supports Manvi et al. (2024) [5], who show that geographic bias can shape how places are imagined. For Indonesian AI, regional representation is therefore not simply a matter of naming places, but of how regional subjects are positioned as modern, peripheral, cultural, or generic.

The unevenness of regional representation is best explained as a structural pattern of spatial salience rather than a series of isolated errors. Large language models are trained and aligned through corpora in which some regions, institutions, professions, and cultural narratives are more digitally visible than others. This does not allow this study to claim a direct causal pathway from training data to output, but it does support a plausible correlation between digital visibility and representational specificity. Li et al. (2022) [21] show that regional bias may operate hierarchically, while Manvi et al. (2024) [5] demonstrate that models can encode geographically uneven assumptions. Naous et al. (2023) [6] further suggest that cultural bias often appears through default expectations about what is normal, modern, or familiar. In this light, Indonesian regional bias is not necessarily produced through hostility toward outer-island regions. It may emerge through centre-weighted knowledge distributions that make some regions narratively detailed and others symbolically compressed.

The gendered patterns have a different implication: professional inclusion does not eliminate role hierarchy. The presence of professional parity in several prompt pairs shows that models can represent male-coded and female-coded subjects with comparable occupational credibility. However, leadership amplification for male-coded subjects and affective-role association for female-coded subjects suggest that equality at the level of job mention may coexist with inequality at the level of agency. This distinction is crucial because gender bias often appears through what models make socially probable rather than through what they explicitly prohibit. Kotek et al. (2023) [2], Kaneko et al. (2022) [3], and Zhao et al. (2024) [4] similarly show that gendered assumptions can persist across generation tasks and multilingual settings. An et al. (2025) [22] extend this point by showing that evaluative contexts may reveal intersectional asymmetries in perceived competence. This study therefore indicates that Indonesian AI may reproduce gender hierarchy not by excluding women from professional life, but by narrowing how authority, expertise, and public agency are distributed.

The underlying structure of gendered representation appears to be associative rather than declarative. Models do not need to state that men are better leaders or women are naturally caring to reproduce gendered

role distributions. Instead, bias can appear through lexical intensity, narrative emphasis, and the kinds of qualities attached to otherwise comparable professional subjects. This supports Bai et al. (2025) [1], who argue that explicitly unbiased models may still form biased associations, and Haim et al. (2024) [23], who show that names and identity markers can activate social assumptions in audit settings. At the same time, this study complicates any simple claim that LLMs always reproduce gender stereotypes uniformly: professional parity and gender-neutral under-specification show that bias is uneven and context-sensitive. The more precise explanation is that gendered bias operates probabilistically through association strength. Certain occupational domains, especially leadership and technical roles, appear more vulnerable to unequal competence framing than roles already socially linked with care, education, or mediation.

Sociolinguistic robustness has the strongest implication for Indonesian AI because it shows that language variety can affect representational legitimacy. Formal Indonesian supports semantic stability, colloquial Indonesian receives adaptive treatment, and local-language conditions expose approximation, code-mixing, or language shift. This does not mean that local languages are inherently difficult; rather, it indicates that models may treat some linguistic forms as more computationally secure than others. Xu et al. (2024) [15] argue that multilingual LLM performance is shaped by corpora, alignment, and bias, while Levy et al. (2023) [26] show that multilingual training can produce uneven outcomes across languages. Tao et al. (2023) [7] similarly frame cultural alignment as a variable rather than guaranteed property. The implication for Indonesian AI is significant: a model may appear competent in Indonesian while remaining unstable in the language varieties through which many Indonesians express locality, intimacy, humour, identity, and social belonging.

The deeper structure behind sociolinguistic asymmetry is the hierarchy between standardised, digitally abundant, and institutionally recognised language varieties and those with weaker digital representation. Formal Indonesian benefits from state standardisation, educational circulation, and online availability, while colloquial Indonesian often appears in social media, entertainment, and everyday interaction. Local languages, by contrast, differ greatly in orthographic stability, corpus availability, digital presence, and NLP support. This helps explain why local-language prompts are more vulnerable to semantic drift, partial repair, or defaulting to Indonesian. Gamboa and Lee (2024) [8], Sahoo et al. (2024) [27], Khandelwal et al. (2023) [28], and Hanif et al. (2026) [13] all show, in different Asian contexts, that culturally grounded benchmarks are necessary because translated fairness templates fail to capture local asymmetries. For this study, the theoretical contribution lies in treating bias as an intersection of region, gender, and sociolinguistic legitimacy. Indonesian AI should be evaluated not only by whether it answers, but by whose language, place, and social role it recognises with precision.

6. CONCLUSION

This study demonstrates that representational bias in Indonesian-facing large language models cannot be adequately understood through generic fairness metrics or isolated stereotype detection. Its central lesson is that bias operates through unequal visibility, agency, and linguistic legitimacy: some regions are rendered as professional and specific, others as generic, cultural, or peripheral; some gendered subjects receive stronger authority, while others are framed through care or affect; and some language varieties produce more stable responses than others. The main contribution lies in reframing Indonesian AI evaluation as an intersectional problem of region, gender, and sociolinguistic hierarchy, supported by a controlled prompt-based audit design.

This study remains limited by its corpus scale, prompt-based design, and dependence on observable model outputs rather than direct access to training data, alignment procedures, or user interaction logs. Consequently, its findings should be read as evidence of patterned representational asymmetry, not as proof of causal mechanisms inside model development. Future research should expand prompt coverage across more Indonesian regions, occupations, gender markers, and local languages, while adding human expert validation and intercoder reliability testing. Comparative work across commercial, open-source, and Indonesian-trained models would further clarify whether observed biases are model-specific, language-specific, or structurally embedded in broader AI infrastructures.

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

Mohammad Anis Sumadi: 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-generated outputs and public source documents.

ETHICAL APPROVAL

This research did not involve human or animal subjects; it analyzed publicly accessible model-generated outputs and public source documents, 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] X. Bai, A. Wang, I. Sucholutsky, and T. L. Griffiths, "Explicitly unbiased large language models still form biased associations," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 122, 2025, doi: 10.1073/pnas.2416228122.

[2] H. Kotek, R. Dockum, and D. Q. Sun, "Gender bias and stereotypes in large language models," *Proceedings of the ACM Collective Intelligence Conference*, 2023, doi: 10.1145/3582269.3615599.

[3] M. Kaneko, A. Imankulova, D. Bollegala, and N. Okazaki, "Gender bias in masked language models for multiple languages," *arXiv*, 2022, doi: 10.48550/arxiv.2205.00551.

[4] J. Zhao, Y. Ding, C. Jia, Y. Wang, and Z. Qian, "Gender bias in large language models across multiple languages," *arXiv*, 2024, doi: 10.48550/arxiv.2403.00277.

[5] R. Manvi, S. Khanna, M. Burke, D. B. Lobell, and S. Ermon, "Large language models are geographically biased," *arXiv, 34654–34669*, 2024, doi: 10.48550/arxiv.2402.02680.

[6] T. Naous, M. J. Ryan, and W. Xu, "Having beer after prayer? Measuring cultural bias in large language models," *arXiv, 16366–16393*, 2023, doi: 10.48550/arxiv.2305.14456.

[7] Y. Tao, O. Viberg, R. S. Baker, and R. F. Kizilcec, "Cultural bias and cultural alignment of large language models," *PNAS Nexus*, vol. 3, 2023, doi: 10.1093/pnasnexus/pgae346.

[8] L. C. L. Gamboa and M. Lee, "Filipino benchmarks for measuring sexist and homophobic bias in multilingual language models from Southeast Asia," *arXiv*, 2024, doi: 10.48550/arxiv.2412.07303.

[9] A. Abid, M. Farooqi, and J. Y. Zou, "Persistent anti-Muslim bias in large language models," *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 2021, doi: 10.1145/3461702.3462624.

[10] I. O. Gallegos, R. A. Rossi, J. Barrow, M. M. Tanjim, S. Kim, F. Dernoncourt, et al., "Bias and fairness in large language models: A survey," *Computational Linguistics*, vol. 50, pp. 1097–1179, 2023, doi: 10.1162/coli_a_00524.

[11] R. Navigli, S. Conia, and B. Ross, "Biases in large language models: Origins, inventory, and discussion," *ACM Journal of Data and Information Quality*, vol. 15, pp. 1–21, 2023, doi: 10.1145/3597307.

[12] E. Ferrara, "Should ChatGPT be biased? Challenges and risks of bias in large language models," *First Monday*, vol. 28, 2023, doi: 10.5210/fm.v28i11.13346.

[13] I. A. Hanif, M. F. Azmi, F. A. Tjiaranata, E. P. Yulianrifat, and F. Koto, "IndoBias: A dual track culturally grounded benchmark for LLMs bias evaluation in Indonesian languages," , 2026.

[14] E. Durmus, K. Nyugen, T. Liao, N. Schiefer, A. Askell, A. Bakhtin, et al., "Towards measuring the representation of subjective global opinions in language models," *arXiv*, 2023.

[15] 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.

[16] Y. Guo, M. Guo, J. Su, Z. Yang, M. Zhu, H. Li, M. Qiu, and S. Liu, "Bias in large language models: Origin, evaluation, and mitigation," *arXiv*, 2024, doi: 10.3390/electronics15091824.

[17] A. Caliskan, "Artificial intelligence, bias, and ethics," *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, 7007–7013*, 2023, doi: 10.24963/ijcai.2023/799.

[18] P. Liang, C. Wu, L.-P. Morency, and R. Salakhutdinov, "Towards understanding and mitigating social biases in language models," *Proceedings of the 38th International Conference on Machine Learning, 6565–6576*, 2021.

[19] F. Kazi, A. Young, Y. Inani, and S. Rafatirad, "A comprehensive study of implicit and explicit biases in large language models," *arXiv*, 2025, doi: 10.48550/arxiv.2511.14153.
[20] J. Mickel, M. De-Arteaga, L. Liu, and K. Tian, "More of the same: Persistent representational harms under increased representation," *arXiv*, 2025, doi: 10.48550/arxiv.2503.00333.
[21] Y. Li, G. Zhang, B. Yang, C. Lin, S. Wang, A. Ragni, and J. Fu, "HERB: Measuring hierarchical regional bias in pre-trained language models," *arXiv*, 2022, doi: 10.48550/arxiv.2211.02882.
[22] J. An, D. Huang, C. Lin, and M. Tai, "Measuring gender and racial biases in large language models: Intersectional evidence from automated resume evaluation," *PNAS Nexus*, vol. 4, 2025, doi: 10.1093/pnasnexus/pgaf089.
[23] A. Haim, A. Salinas, and J. Nyarko, "What's in a name? Auditing large language models for race and gender bias," *arXiv*, 2024, doi: 10.48550/arxiv.2402.14875.
[24] A. Mandal, S. Little, and S. Leavy, "Gender bias in multimodal models: A transnational feminist approach considering geographical region and culture," *arXiv*, 2023, doi: 10.48550/arxiv.2309.04997.
[25] L. Jiang, G. Zhu, J. Sun, J. Cao, and J. Wu, "Exploring the occupational biases and stereotypes of Chinese large language models," *Scientific Reports*, vol. 15, 2025, doi: 10.1038/s41598-025-03893-w.
[26] S. Levy, N. A. John, L. Liu, Y. Vyas, Y. Fujinuma, M. Ballesteros, V. Castelli, and D. Roth, "Comparing biases and the impact of multilingual training across multiple languages," *arXiv, 10260–10280*, 2023, doi: 10.48550/arxiv.2305.11242.
[27] N. R. Sahoo, P. Kulkarni, N. Asad, A. Ahmad, T. Goyal, A. Garimella, and P. Bhattacharyya, "IndiBias: A benchmark dataset to measure social biases in language models for Indian context," *arXiv, 8786–8806*, 2024, doi: 10.48550/arxiv.2403.20147.
[28] K. Khandelwal, M. Tonneau, A. M. Bean, H. R. Kirk, and S. A. Hale, "Indian-BhED: A dataset for measuring India-centric biases in large language models," *Proceedings of the 2024 International Conference on Information Technology for Social Good*, 2023, doi: 10.1145/3677525.3678666.
[29] A. Vashishtha, K. Ahuja, and S. Sitaram, "On evaluating and mitigating gender biases in multilingual settings," *arXiv*, 2023, doi: 10.48550/arxiv.2307.01503.

BIOGRAPHY OF AUTHOR

Mohammad Anis Sumadi is affiliated with Institut Agama Islam Al-Urwatul Wutsqo, Indonesia. He received his academic formation in language, social sciences, Islamic studies, and related interdisciplinary fields. His research interests include discourse analysis, financial communication, development narratives, microfinance literacy, and language in community empowerment. His work frequently engages with social and economic issues in rural Indonesian settings. He can be contacted at email: aannylabuh@gmail.com