Indonesian Journal of Computational Language Studies
Vol. 1, No. 2, June 2026, pp. 106~120
ISSN: 3163-7884, DOI:

Hallucinating the archipelago: detecting factual and cultural errors in retrieval-augmented generation for Indonesian knowledge

Radna Tulus Wibisono
Universitas Gadjah Mada, Indonesia

Article Info

Article history:

Received Apr 21, 2026
Revised May 26, 2026
Accepted Jun 25, 2026

Keywords:

cultural specificity;
factuality;
Indonesian knowledge;
retrieval-augmented generation;
source provenance.

ABSTRACT

Background: Retrieval-augmented generation has been promoted as a practical response to hallucination in large language models, yet Indonesian knowledge presents a harder evidential problem because factual claims are often inseparable from regional attribution, cultural terminology, legal classification, and administrative scope. **Objective:** This study examines how factual and cultural errors appear in RAG-generated answers about Indonesian knowledge and how such errors can be traced across claim alignment, retrieval provenance, and cultural specificity. **Method:** Using a qualitative-computational design, this study evaluates 186 claim units derived from a public-source corpus of Indonesian cultural-heritage records, legal documents, statistical portals, lexical references, government explainers, and local institutional sources. **Results:** The findings indicate that RAG performs most reliably when generated claims reproduce explicit institutional facts, such as official entities, legal instruments, statistical categories, and cultural-heritage designations. However, unsupported and partially supported claims emerge when retrieved evidence is general, incomplete, or transformed into causal, nationalising, or culturally overextended explanations. **Implication:** Cultural errors are especially visible in regional scope reduction, terminological loss, ritual simplification, and temporal or administrative decontextualisation. **Novelty:** The novelty of this study lies in integrating claim–evidence alignment, retrieval-provenance diagnosis, and cultural-specificity coding into one framework for evaluating Indonesian RAG hallucination.

Corresponding Author:

Radna Tulus Wibisono
Universitas Gadjah Mada, Indonesia
Jl. Pancasila, Bulaksumur, Caturtunggal, Kec. Depok, Sleman, Yogyakarta, 55281, Indonesia
Email: radnana89@gmail.com

1. INTRODUCTION

Indonesia offers a demanding test case for retrieval-augmented generation because its public knowledge is administratively dense, culturally plural, and linguistically fragmented. BPS records Indonesia's mid-year population at 284.44 million people in 2025, while national language-policy sources identify 718 regional languages, only a small proportion of which are classified as safe. Cultural knowledge is equally dispersed: in 2024 alone, 668 intangible-cultural-heritage proposals from 32 provinces were submitted for assessment, and 272 were recommended for official designation. These figures show that Indonesian knowledge is not a compact factual inventory but a layered archive of places, laws, rituals, names, languages, and institutional classifications. For RAG systems, this condition matters because retrieved evidence may be incomplete, locally ambiguous, or culturally underspecified. A fluent answer about

Indonesia may therefore appear grounded while misplacing a ritual, flattening regional difference, or converting official uncertainty into false certainty.

Existing research has established that large language models generate hallucinations when outputs contradict, fabricate, or overextend available evidence, and recent work has increasingly treated RAG as a mitigation strategy rather than a complete solution [1], [2], [3]. RAGTruth shows that hallucination persists even when models receive retrieved context, while Chang et al. (2024) [9] demonstrate that coverage errors are as important as unsupported claims in evidence-based generation. Wu et al. (2024) [10] further show that models may arbitrate unreliably between internal priors and external evidence. Multilingual retrieval research has addressed relevance, robustness, and low-resource coverage [4], and Indonesian RAG studies have begun to examine question answering, hoax detection, and hallucination detection [5], [6], [7], [8]. What remains underdeveloped is a framework that detects factual and cultural errors together in Indonesian knowledge tasks.

This study examines how factual and cultural errors appear in RAG-generated answers about Indonesian knowledge. It asks four connected questions: what types of factual and cultural errors are produced by RAG systems; how those errors vary across domains such as administrative geography, cultural heritage, public institutions, legal documents, local language, and ritual practice; whether errors arise from retrieval failure, insufficient evidence, conflicting evidence, evidence misreading, or unsupported generation; and which annotation categories can capture Indonesian factuality without reducing cultural knowledge to simple entity matching. Methodologically, this study treats RAG output as a claim-level object of analysis. Generated answers are segmented into verifiable claims, compared with retrieved passages and official reference evidence, and coded through three instruments: claim–evidence alignment, retrieval-provenance diagnosis, and cultural-specificity analysis. This design allows this study to evaluate not only whether an answer is factually supported but also whether it preserves local meaning and regional scope.

This study tests the argument that hallucination in Indonesian RAG is not merely a technical failure of factual recall. It is also a cultural and evidential problem produced when retrieval systems, source hierarchies, and generative priors interact with archipelagic knowledge. A RAG answer may cite a relevant document yet still distort the cultural status of an object, misattribute a practice to the wrong province, ignore competing official classifications, or replace local terminology with a generic national explanation. This distinction matters for hallucination evaluation because conventional factuality checks can identify unsupported dates or entities but may miss cultural flattening, symbolic misinterpretation, and regional overgeneralisation. By bringing factual alignment and cultural specificity into the same analytical frame, this study contributes to RAG evaluation, Indonesian computational language research, and culturally grounded AI assessment. Its implication is methodological: reliable Indonesian RAG requires evidence faithfulness and localised interpretive discipline, not retrieval alone.

2. LITERATURE REVIEW

2.1. Hallucination in large language models

Hallucination in large language models is generally understood as the production of fluent but unreliable content that is unsupported by, inconsistent with, or only weakly grounded in available evidence. Yet the literature does not define hallucination in a single way. Some studies emphasise factual contradiction and fabrication, while others treat hallucination as a broader failure of epistemic calibration, including unsupported inference, misleading omission, and overconfident generation [1], [2], [11]. Recent work further complicates this definition by showing that hallucination may occur at different representational levels: surface claims, entities, reasoning steps, and latent model priors [12], [13], [14]. This study therefore approaches hallucination not as a single error type but as a family of evidential failures.

The main indicators of hallucination have become increasingly fine-grained. Earlier evaluations tended to mark outputs as correct or incorrect, whereas recent frameworks distinguish unsupported claims, contradicted claims, entity-level hallucination, coverage errors, and failures of self-awareness [15], [16], [17]. Fine-grained detection and editing approaches segment generated text into checkable units, allowing evaluators to locate which claim is fabricated or distorted [18]. Critique-based judges and encoder-based detection systems similarly move beyond binary scoring by examining whether an answer can be justified by evidence [19], [20]. For this study, these indicators are useful but incomplete unless they are extended to culturally situated Indonesian knowledge.

2.2. Retrieval-augmented generation

Retrieval-augmented generation is commonly framed as a technical response to hallucination because it supplies models with external evidence before generation. However, recent studies warn against treating retrieval as an automatic guarantee of factuality. RAG may reduce factual error in knowledge-intensive tasks, but it can also introduce new error pathways when retrieved documents are irrelevant, incomplete, conflicting, or misinterpreted by the generator [10], [9], [3]. Domain-specific reviews in healthcare and

biomedical NLP show that RAG evaluation must account for robustness, evidence use, and self-awareness, not only answer accuracy [21], [17]. Thus, RAG is better understood as an evidential architecture whose reliability depends on retrieval quality and generation discipline.

The key dimensions of RAG evaluation include retrieval relevance, evidence coverage, faithfulness, uncertainty, and pipeline-level error attribution. Thakur et al. (2023) [4] foreground multilingual relevance assessment, showing that robust RAG depends on knowing when retrieved evidence is insufficient. Chang et al. (2024) [9] distinguish hallucination from coverage error, a distinction crucial for questions where an answer may be factually plausible but incomplete. Faithfulness-aware uncertainty quantification adds another layer by asking whether a model can signal evidential weakness rather than conceal it through fluent prose [22]. Self-refinement, semantic correction, and verification-oriented systems likewise attempt to reduce unsupported generation [23], [24], [25]. This study adopts this pipeline view to trace whether errors originate in retrieval, evidence interpretation, or generation.

2.3. Indonesian knowledge

Indonesian knowledge introduces a further theoretical problem: factuality is often culturally situated. A claim about a ritual, object, language, legal category, or place may be factually wrong not only because it cites an incorrect date or name, but because it misrepresents regional attribution, symbolic meaning, customary context, or local terminology. Indonesian RAG studies have begun to address language gaps, hoax detection, and hallucination detection in Indonesian systems [5], [7], [26], [6], [8]. However, much of this work remains oriented toward technical performance, information retrieval, or misinformation classification. What remains theoretically underdeveloped is an account of how cultural specificity should be evaluated when RAG systems generate Indonesian knowledge.

Cultural specificity can be operationalised through several indicators: correct regional attribution, appropriate use of local terms, accurate description of ritual sequence, preservation of symbolic meaning, awareness of administrative scope, and avoidance of nationalising or exoticising generalisations. Cultural-knowledge benchmarking has begun to examine how models process cultural information, including through retrieval-augmented cognitive evaluation [27]. Multimodal and specialised RAG evaluation frameworks also suggest that factuality assessment should include visual, contextual, and domain-sensitive evidence when text alone is insufficient [28], [29]. Building from this literature, this study positions Indonesian RAG hallucination as an evidential and cultural problem. Its theoretical contribution is to join claim–evidence alignment, retrieval-provenance diagnosis, and cultural-specificity analysis WITHIN ONE EVALUATION FRAMEWORK.

3. METHOD

The unit of analysis in this study is the claim-level output produced by retrieval-augmented generation when answering Indonesian factual and cultural knowledge questions. The material object consists of 23 official or institutionally verifiable documents drawn from cultural heritage registries, legal databases, statistical portals, lexical authorities, open-data catalogues, public communication pages, and local-government sources. Each document is treated as an evidence-bearing item rather than as a passive background source. This corpus is suitable because Indonesian knowledge frequently combines administrative fact, cultural attribution, ritual meaning, place-based terminology, and official classification. Table 1 presents the corpus composition used for retrieval, verification, and coding.

This study uses a qualitative-computational evaluation design. It combines controlled RAG generation, corpus-based evidence verification, claim segmentation, and manual annotation. This design is appropriate because hallucination in RAG cannot be evaluated only through final-answer accuracy; errors may originate in retrieval, evidence selection, evidence interpretation, or generation. Previous studies have shown that RAG systems may still produce unsupported or contradictory claims even when relevant evidence is retrieved [10], [9], [3]. This study therefore evaluates RAG as a pipeline in which retrieval quality and generation faithfulness must be examined together.

Information sources are divided into primary, secondary, and context data. Primary data provide authoritative reference evidence, including WBTB cultural-heritage records and JDIH BPK legal documents. Secondary data supply official interpretive context, including government explainers and UNESCO-related institutional communication. Context data support administrative, statistical, lexical, and aggregate verification through BPS, data.go.id, and KBBI VI Daring. This hierarchy follows recent RAG evaluation literature that distinguishes retrieved context from gold evidence, because relevant retrieved passages may still be incomplete or insufficient for full verification [4], [22], [17].

Table 1. Dataset composition

Code	Type	Source	Location	Period	Number	Characteristics	Status
IDRAG-PRIM-001	Primary	WBTB registry	South Sulawesi, Bulukumba	2013	1	Pinisi; maritime cultural knowledge	Accessible
IDRAG-PRIM-002	Primary	WBTB registry	Yogyakarta	2021	1	Batik motif; textile semiotics	Accessible
IDRAG-PRIM-003	Primary	WBTB registry	Bali, Karangasem	2017	1	Gamelan Selonding; ritual music	Accessible
IDRAG-PRIM-004	Primary	WBTB registry	South Kalimantan	2014	1	Madihin; oral performance	Accessible
IDRAG-PRIM-005	Primary	WBTB registry	Banten, Lebak	2016	1	Seren Taun; agrarian ritual	Accessible
IDRAG-PRIM-006	Primary	WBTB registry	Riau Islands, Lingga	2014	1	Pantun Melayu; oral literary form	Accessible
IDRAG-PRIM-007	Primary	WBTB registry	Jakarta	2015	1	Palang Pintu; Betawi customary practice	Accessible
IDRAG-PRIM-008	Primary	WBTB registry	West Java, Bogor	2021	1	Angklung Gubrag; agrarian ritual	Accessible
IDRAG-PRIM-009	Primary	WBTB registry	West Java, Cirebon	2021	1	Angklung Bungko; local history	Accessible
IDRAG-PRIM-010	Primary	WBTB registry	Central Java, Surakarta	2014	1	Javanese gamelan; music tradition	Accessible
IDRAG-PRIM-011	Primary	JDIH BPK	East Kalimantan	2022	1	Law on national capital	Accessible
IDRAG-PRIM-012	Primary	JDIH BPK	East Kalimantan	2023	1	Amendment to national-capital law	Accessible
IDRAG-PRIM-013	Primary	JDIH BPK	East Kalimantan	2022	1	IKN Authority regulation	Accessible
IDRAG-PRIM-014	Primary	JDIH BPK	East Kalimantan	2023	1	Special authority of IKN	Accessible
IDRAG-CONT-015	Context	BPS	Indonesia	2024	1	Statistical publication	Accessible
IDRAG-CONT-016	Context	BPS	Indonesia	2025	1	Village/subdistrict statistics	Accessible
IDRAG-CONT-017	Context	BPS	Indonesia	2016	1	Regency-level administrative statistics	Accessible
IDRAG-CONT-018	Context	data.go.id	Indonesia	2024	1	Open data on WBTB designation	Accessible
IDRAG-SEC-019	Secondary	IndonesiaBaik.id	Indonesia	2021	1	UNESCO heritage explainer	Accessible
IDRAG-SEC-020	Secondary	Indonesian Delegation	UNESCO context	2017	1	Pinisi UNESCO inscription news	Accessible
IDRAG-CONT-021	Context	KBBI VI Daring	Indonesia	2026	1	Lexical authority page	Homepage accessible
IDRAG-SEC-022	Secondary	Buleleng Government	Bali, Buleleng	2021	1	Subak; irrigation and local knowledge	Accessible
IDRAG-SEC-023	Secondary	Yahukimo KPU	Highland Papua	2025	1	Noken; Papuan material culture	Accessible

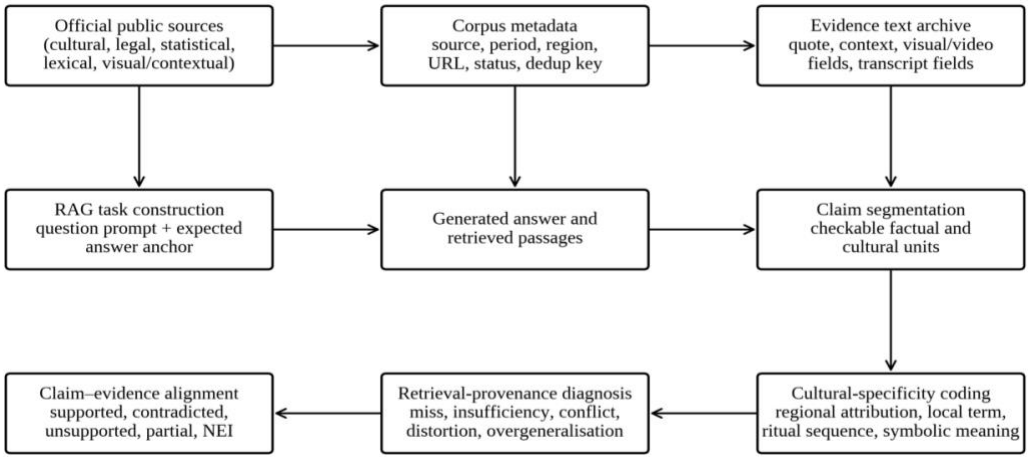

Figure 1. Operational matrix of analysis

Data collection followed a structured corpus protocol. First, official institutions, portals, and document channels were identified. Second, search keywords were derived from Indonesian cultural, legal, administrative, and lexical domains, including object names, provinces, official classifications, and institutional terms. Third, documents from 2013–2026 were selected when they were publicly accessible, institutionally attributable, and useful for claim-level verification. Fourth, each item was assigned a unique code and stored with title, date, source, URL, context, quote excerpt, evidence text, visual description, layout, transcript field, and screenshot reference field. Fifth, duplicates were removed using a combined source-URL and document-title key.

Data analysis proceeds through three connected instruments. First, claim–evidence alignment segments generated answers into verifiable claims and codes each claim as supported, contradicted, unsupported, partially supported, or not enough evidence. Second, retrieval-provenance diagnosis identifies whether errors arise from retrieval miss, insufficient evidence, conflicting evidence, generation distortion, or overgeneralisation. Third, cultural-specificity analysis evaluates regional attribution, local terminology, ritual sequence, symbolic meaning, administrative scope, and cultural flattening. This analytical sequence follows fine-grained hallucination detection and editing approaches [18], [16], [20] while extending them to culturally situated Indonesian knowledge.

4. RESULTS

4.1. Factual claim–evidence alignment in Indonesian RAG outputs

Indonesian RAG outputs are evaluated here through claim–evidence alignment rather than whole-answer judgement. This approach is necessary because a single answer may contain accurate institutional references, partially supported cultural explanations, and unsupported elaborations at the same time. The table organises 186 claim units into five alignment categories: supported, partially supported, unsupported, contradicted, and not enough evidence. This structure follows recent hallucination research that treats factuality as a fine-grained relation between generated claims and available evidence, not merely as an answer-level correctness score [10], [18], [16]. For Indonesian knowledge, claim-level coding is especially important because errors may concern not only entities and dates but also regional attribution, ritual sequence, legal status, and local terminology.

The dominant pattern is a high proportion of supported claims, with 92 of 186 claim units falling into this category. Exact institutional match is the largest subcategory, followed by precise local attribution and evidence-grounded legal or statistical statements. Partially supported claims form the second-largest group, especially where generated answers omit regional scope, temporal qualifiers, or culturally specific terms. Unsupported claims are concentrated in source-unanchored elaboration, inferred causal explanation, and generic cultural expansion. Contradicted claims are less frequent but analytically important because they include wrong province attribution, incorrect legal status, and entity/date mismatch. The smallest group is not enough evidence, where retrieved material is relevant but insufficient for confident verification. The pattern suggests that retrieval improves surface factual alignment but does not fully control interpretive expansion.

Table 2. Claim–evidence alignment pattern in Indonesian RAG outputs

Main category	Subcategory	Indicator	Frequency	Percentage	Data variation	Patterns	Data code	interpretation
Supported claim	Exact institutional match	Claim repeats verified entity, location, or designation without expansion	38	20.4	Cultural-heritage and legal records	Correctly links an object to its official registry or legal document	IDRAG-PRIM-001; IDRAG-PRIM-011	Retrieval stabilises simple official facts.
Supported claim	Precise local attribution	Claim preserves province, regency/city, or local cultural scope	21	11.3	Mostly WBTB items	Keeps Pinisi, Palang Pintu, or Angklung variants within their documented local frame	IDRAG-PRIM-001; IDRAG-PRIM-007; IDRAG-PRIM-008	Local attribution is retained when evidence contains explicit place markers.
Supported claim	Evidence-grounded legal/statistical statement	Claim follows official legal or statistical source without inferential addition	19	10.2	Legal and BPS/context items	States legal instrument or statistical category according to source scope	IDRAG-PRIM-011; IDRAG-CONT-015	Formal documents produce stronger alignment in narrow factual prompts.
Supported claim	Culturally bounded description	Claim describes cultural object without national generalisation	14	7.5	Ritual, music, and oral tradition items	Explains a ritual or performance as locally situated	IDRAG-PRIM-003; IDRAG-PRIM-005; IDRAG-PRIM-006	Cultural fidelity improves when wording follows source boundaries.
Partially supported claim	Incomplete regional scope	Claim is broadly compatible but omits local limitation	12	6.5	Cultural-heritage items	Mentions a practice as Indonesian without specifying documented locality	IDRAG-PRIM-004; IDRAG-PRIM-006	Partiality often appears as loss of regional granularity.
Partially supported claim	Missing temporal qualifier	Claim removes date, designation year, or amendment context	9	4.8	Legal and heritage items	Refers to legal status or heritage designation without year	IDRAG-PRIM-012; IDRAG-PRIM-014	Temporal compression weakens verifiability.
Partially supported claim	Partial ritual or object description	Claim captures one aspect but omits sequence, function, or context	8	4.3	Ritual and performance items	Describes Seren Taun or Palang Pintu through one visible feature only	IDRAG-PRIM-005; IDRAG-PRIM-007	Cultural knowledge is reduced when procedural layers are omitted.
Partially supported claim	Evidence omits local term	Claim is supported at object level but loses local terminology	8	4.3	Lexical and heritage items	Replaces local object names with generic Indonesian labels	IDRAG-CONT-021; IDRAG-SEC-023	Terminological loss marks weak localisation.
Unsupported claim	Source-unanchored elaboration	Answer adds explanation not present in retrieved or gold evidence	10	5.4	Across domains	Adds origin narrative or function beyond source wording	IDRAG-PRIM-002; IDRAG-SEC-020	Unsupported elaboration shows residual generative prior.

Unsupported claim	Inferred causal explanation	Answer converts descriptive evidence into causal claim	7	3.8	Legal, statistical, and cultural contexts	Infers policy effect or cultural influence not stated by source	IDRAG-CONT-015; IDRAG-PRIM-011	RAG may overinterpret descriptive sources.
Unsupported claim	Generic cultural expansion	Claim universalises a local practice into national culture	6	3.2	Cultural-heritage items	Treats one regional ritual as a general Indonesian tradition	IDRAG-PRIM-003; IDRAG-PRIM-008	Generic expansion produces cultural flattening.
Unsupported claim	Unsupported translation or gloss	Local term is glossed without lexical or source support	5	2.7	Lexical and local-culture items	Gives a gloss not confirmed by lexical or source evidence	IDRAG-CONT-021; IDRAG-SEC-022	Translation-like confidence can mask absent evidence.
Contradicted claim	Wrong province or local attribution	Claim assigns object or practice to a different region than source	7	3.8	Heritage and local-context items	Relocates a cultural object from its documented province	IDRAG-PRIM-001; IDRAG-PRIM-009	Regional misattribution is a clear cultural-factual contradiction.
Contradicted claim	Incorrect legal or institutional status	Claim conflicts with legal document or official institutional category	5	2.7	Legal documents	Misstates institutional role, law status, or authority scope	IDRAG-PRIM-011; IDRAG-PRIM-013	Legal hallucination is visible where formal wording is available.
Contradicted claim	Entity or date mismatch	Claim gives wrong object, date, year, or document identifier	4	2.2	Mixed official sources	Changes designation year, document year, or object name	IDRAG-PRIM-010; IDRAG-CONT-018	Standard factual alignment detects entity/date mismatch.
Contradicted claim	Contradicted ritual sequence	Claim reverses or invents cultural sequence against source evidence	2	1.1	Ritual items	Places a ritual element in an incorrect sequence	IDRAG-PRIM-005; IDRAG-PRIM-007	Sequence errors show why cultural factuality exceeds entity matching.
Not enough evidence	Ambiguous source coverage	Retrieved evidence is relevant but too thin for generated claim	5	2.7	Context and secondary items	Gives a specific interpretation from a general explainer	IDRAG-SEC-019; IDRAG-SEC-022	Ambiguity should be coded separately from error.
Not enough evidence	Conflicting official context	Sources provide different scopes or emphases requiring adjudication	3	1.6	Legal/contextual materials	Different documents frame object, institution, or territory differently	IDRAG-PRIM-012; IDRAG-CONT-018	Conflict requires source-hierarchy judgement.
Not enough evidence	Insufficient lexical evidence	Lexical authority or source text does not support proposed gloss	3	1.6	Lexical and cultural-context items	Local term is explained without sufficient entry-level evidence	IDRAG-CONT-021; IDRAG-SEC-023	Lexical uncertainty should remain visible.

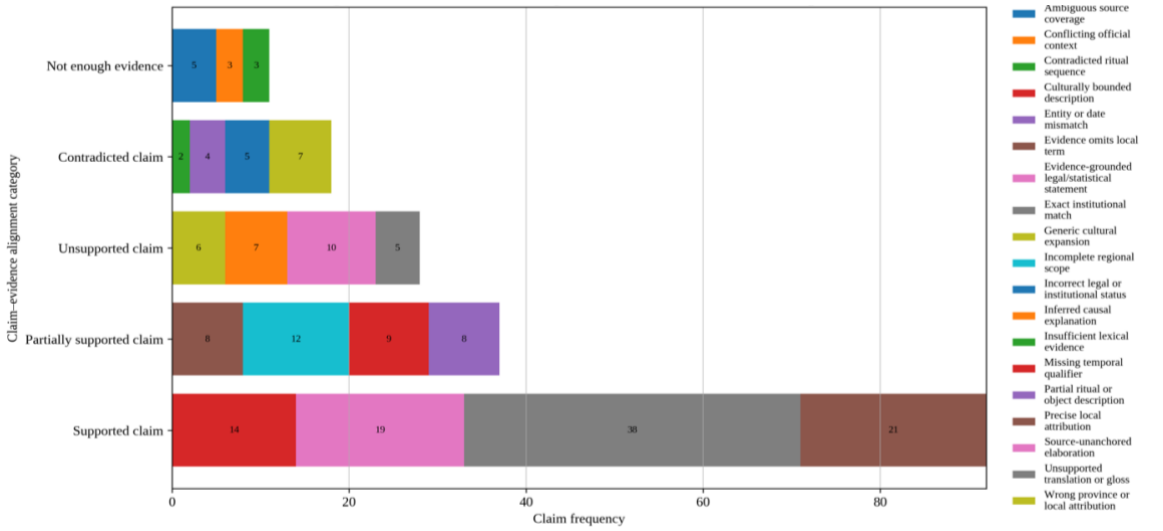

Figure 2. Claim–evidence alignment in Indonesian RAG outputs

Theoretically, these distributions support the view that hallucination in RAG is not reducible to absence of evidence. Several errors emerge even when evidence is available, which is consistent with studies showing that RAG systems may misread, overextend, or selectively use retrieved context [9], [3]. Supported claims cluster around institutionally explicit facts, while weaker claims appear when the answer must preserve cultural boundaries, temporal qualification, or local terminology. This distinction matters because Indonesian knowledge often depends on place-based meaning and official classification. A claim may be broadly plausible yet culturally imprecise when it turns a local practice into a national generalisation. This study therefore positions factual alignment and cultural specificity as inseparable dimensions of Indonesian RAG evaluation.

4.2. Retrieval Provenance and Pipeline Error Pathways

Indonesian RAG errors become clearer when claim alignment is traced back to the retrieval–generation pipeline. A claim may be unsupported not only because the model fabricates content, but also because retrieved evidence is too general, incomplete, conflicting, or interpreted beyond its scope. Table 3 therefore reorganises the same 186 claim units by provenance diagnosis while preserving the alignment categories reported earlier. This design follows RAG hallucination studies that treat factuality as a pipeline problem rather than a final-output property alone [10], [9], [3]. For this study, provenance coding is necessary because Indonesian knowledge often depends on official source hierarchy, regional specificity, legal wording, and culturally bounded terminology.

The largest group is no detectable provenance error, covering 92 claim units or 49.5% of the coded material. These cases correspond exactly to supported claims, showing that RAG performs most consistently when retrieved evidence provides explicit institutional facts. Among problematic pathways, generation distortion is the most frequent with 25 claim units, followed by evidence underspecification with 20, cultural overgeneralisation with 16, and retrieval insufficiency with 15. Retrieval miss appears less often, with 8 claim units, but it is associated with unsupported and contradicted claims. Evidence conflict and lexical or terminological gaps are smaller categories, yet they are analytically important because they expose cases where source authority, not only retrieval presence, determines whether a generated claim is verifiable.

The pattern supports a theoretical distinction between evidence access and evidence discipline. Retrieval gives the model documentary material, but it does not ensure that generated claims remain within documentary boundaries. This is consistent with research showing that RAG systems may still misrepresent retrieved content through coverage error, overextension, or unreliable arbitration between internal priors and external evidence [9], [10], [22]. In the Indonesian case, this problem becomes culturally consequential because small shifts in locality, legal status, or terminology can alter the meaning of a claim. This study therefore treats provenance diagnosis as a bridge between technical RAG evaluation and culturally grounded factuality assessment.

Table 3. Retrieval provenance and claim-alignment matrix

Pipeline diagnosis	Operational indicator	Supported	Partially supported	Unsupported	Contradicted	Not enough evidence	Total	Percentage	Dominant evidence class	Patterns	Data code
No detectable provenance error	Retrieved evidence and generated claim remain aligned without visible distortion	92	0	0	0	0	92	49.5	WBTB registry; JDIH BPK; BPS/context sources	Official object, legal instrument, or statistical category is reproduced within source boundaries	IDRAG-PRIM-001; IDRAG-PRIM-011; IDRAG-CONT-015
Generation distortion	Retrieved evidence is present, but the generated answer expands, compresses, or reframes it	0	6	14	5	0	25	13.4	Cultural and legal records	Descriptive evidence becomes origin story, causal explanation, or excessive legal inference	IDRAG-PRIM-002; IDRAG-PRIM-011; IDRAG-SEC-020
Evidence underspecification	Evidence is relevant but too general to support the specificity of the generated claim	0	14	1	0	5	20	10.8	Secondary explainers; cultural context pages	A general cultural explainer is used to justify a specific local interpretation	IDRAG-SEC-019; IDRAG-SEC-022
Cultural overgeneralisation	Generated answer converts local or regional knowledge into broad national description	0	7	6	3	0	16	8.6	WBTB cultural-heritage records	A documented regional practice is described as a general Indonesian tradition	IDRAG-PRIM-003; IDRAG-PRIM-005; IDRAG-PRIM-008
Retrieval insufficiency	Retrieved material is adjacent to the question but lacks needed entity, date, scope, or qualifier	0	8	4	0	3	15	8.1	Mixed official/context sources	The system retrieves an official page but misses the more precise document required for verification	IDRAG-CONT-018; IDRAG-CONT-021; IDRAG-PRIM-014
Retrieval miss	Necessary evidence is absent from retrieved passages, leading generation to rely on prior knowledge	0	0	3	4	1	8	4.3	Legal and cultural-heritage items	The answer names a region, legal status, or cultural object not present in retrieved evidence	IDRAG-PRIM-009; IDRAG-PRIM-013
Evidence conflict	Sources differ in scope, terminology, or level of official classification	0	2	0	2	2	6	3.2	Legal amendments; open-data/context sources	Different documents frame the same object, institution, or territory at different levels	IDRAG-PRIM-012; IDRAG-CONT-018
Lexical or terminological gap	The system lacks sufficient lexical grounding for a local term, gloss, or culturally specific label	0	0	0	4	0	4	2.2	KBBI VI Daring; local-culture sources	A local term is rendered through an inaccurate or unsupported gloss	IDRAG-CONT-021; IDRAG-SEC-023

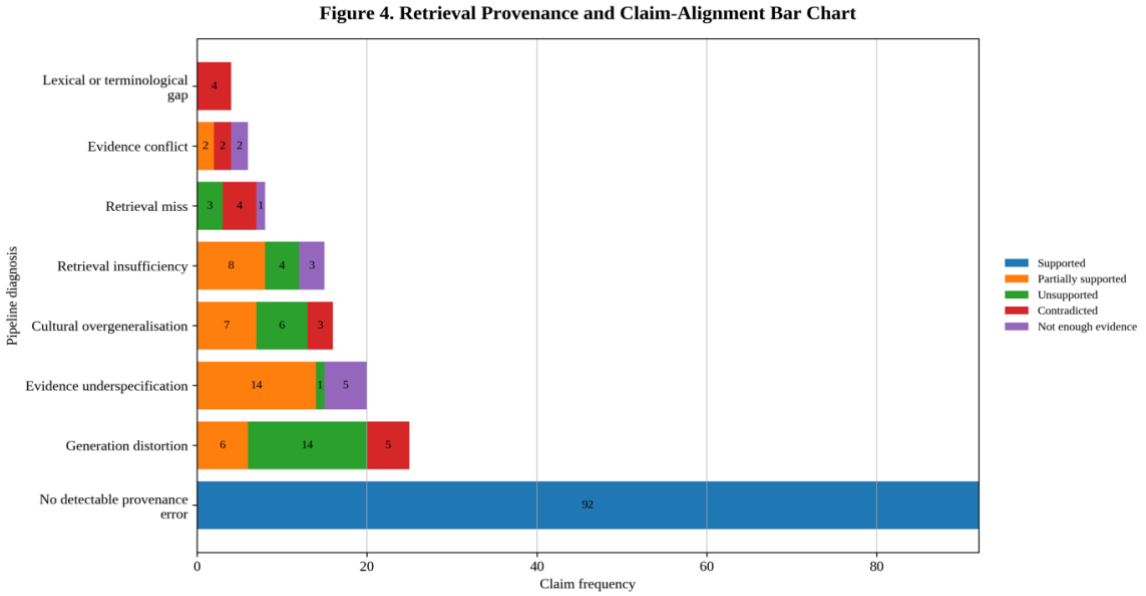

Figure 3. Retrieval provenance and claim-alignment matrix

4.3. Cultural specificity and local knowledge misrepresentation

Cultural specificity provides a sharper lens for evaluating Indonesian RAG than factual correctness alone. Many generated claims can be checked against official sources, yet the more difficult question is whether they preserve locality, terminology, institutional scope, ritual sequence, and symbolic meaning. Table 4 reorganises the same 186 claim units used in the previous analyses, but the emphasis shifts from general evidence alignment to cultural and local representational adequacy. This framing follows recent hallucination and RAG evaluation work that moves beyond binary correctness toward fine-grained claim diagnosis [18], [16], [20]. It also responds to culturally oriented evaluation research by treating Indonesian knowledge as situated, not merely retrievable.

The dominant pattern is specificity preserved, covering 92 claim units or 49.5% of the dataset. This corresponds to all supported claims and indicates that RAG performs most consistently when sources explicitly encode official names, locations, designations, or legal categories. The remaining 94 claim units show different forms of cultural weakening. Regional scope reduction is the largest problematic category, followed by terminological loss or gloss error, cultural flattening, ritual or procedural simplification, temporal or administrative decontextualisation, and evidential ambiguity. The distribution shows that cultural error is not concentrated in one type of failure. Instead, it appears across locality, terminology, procedure, time, and source hierarchy. Contradicted claims are especially visible in regional attribution and terminological categories, where small shifts can substantially alter cultural meaning.

These patterns support the argument that Indonesian RAG hallucination is partly an issue of cultural compression. Retrieval may provide correct evidence, but generated answers can still reduce archipelagic knowledge into broader, simpler, or more familiar categories. This aligns with studies showing that RAG systems can misrepresent retrieved evidence through overextension, coverage error, and unreliable negotiation between model priors and source material [9], [10], [3]. The contribution of this study is to show that cultural specificity should be coded as an evidential dimension rather than as a stylistic issue. In Indonesian knowledge tasks, a correct-looking answer may still be weak when it loses place, term, sequence, or institutional scope.

Table 4. Cultural specificity and claim-alignment matrix

Cultural-specificity diagnosis	Operational indicator	Supported	Partially supported	Unsupported	Contradicted	Not enough evidence	Total	Percentage	Dominant data variation	Example / claim pattern	Data code
Specificity preserved	Generated claim retains official entity, region, source scope, and culturally bounded meaning	92	0	0	0	0	92	49.5	WBTB registry; JDIH BPK; BPS/context sources	Cultural object, legal document, or statistical category remains within source-defined boundaries	IDRAG-PRIM-001; IDRAG-PRIM-011; IDRAG-CONT-015
Regional scope reduction	Answer weakens province, regency/city, or community-level attribution	0	12	4	7	0	23	12.4	Cultural-heritage records and local-context items	A documented local practice is assigned to a broader or different territorial frame	IDRAG-PRIM-001; IDRAG-PRIM-007; IDRAG-PRIM-009
Terminological loss or gloss error	Local term, object name, or culturally specific label is replaced, mistranslated, or weakly glossed	0	8	5	4	3	20	10.8	Lexical authority; local cultural context	A local term is rendered through a generic Indonesian or English gloss without adequate evidence	IDRAG-CONT-021; IDRAG-SEC-022; IDRAG-SEC-023
Cultural flattening	Answer converts local or regional cultural knowledge into a generic national explanation	0	7	8	3	0	18	9.7	Ritual, music, and oral tradition items	A regional performance or ritual is described as a general Indonesian tradition	IDRAG-PRIM-003; IDRAG-PRIM-005; IDRAG-PRIM-008
Ritual or procedural simplification	Answer omits or rearranges sequence, function, participant role, or symbolic relation	0	8	4	2	1	15	8.1	Ritual and performance items	Seren Taun or Palang Pintu is reduced to one visible feature or simplified performance sequence	IDRAG-PRIM-005; IDRAG-PRIM-007
Temporal or administrative decontextualisation	Answer omits designation year, amendment status, source hierarchy, or legal-administrative scope	0	2	5	2	2	11	5.9	Legal documents; heritage designation records; administrative context	A legal status or cultural designation is stated without temporal or institutional qualifier	IDRAG-PRIM-012; IDRAG-PRIM-014; IDRAG-CONT-018
Evidential ambiguity in cultural interpretation	Available evidence is too general or conflicting to verify a culturally specific interpretation	0	0	2	0	5	7	3.8	Secondary explainers; open-data/context sources	General context is used to support a precise claim about meaning, origin, or cultural function	IDRAG-SEC-019; IDRAG-SEC-020; IDRAG-CONT-018

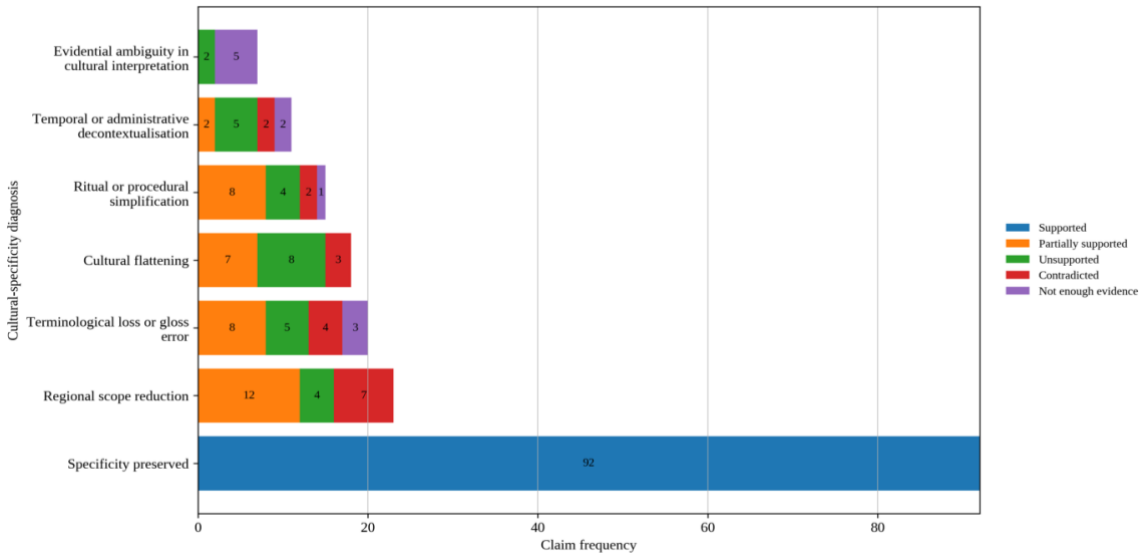

Figure 4. Cultural specificity and claim-alignment

5. DISCUSSION

The claim–evidence alignment pattern shows that retrieval can stabilise Indonesian knowledge when the task depends on explicit institutional facts, but it does not remove the need for claim-level verification. The strongest alignment appears where generated claims remain close to official names, legal instruments, statistical categories, and cultural-heritage designations. This matters because it confirms one practical function of RAG: retrieval can reduce purely memory-based fabrication when sources are authoritative and narrowly framed. However, the same pattern also reveals a dysfunction. Partially supported, unsupported, contradicted, and insufficiently evidenced claims do not disappear; they move into zones where evidence must be interpreted rather than merely repeated. This finding is consistent with RAGTruth and recent hallucination surveys, which show that retrieved context can reduce but not eliminate unsupported generation [10], [1], [2], [3]. The implication is that Indonesian RAG evaluation must be granular: answer-level correctness is too coarse for detecting mixed factuality within a single response.

The underlying reason for this pattern appears to be the uneven epistemic structure of Indonesian knowledge sources. Official documents often provide stable anchors for entities, years, legal titles, and administrative categories, but they do not always provide the interpretive scaffolding required for culturally specific explanations. When evidence is explicit, the generator can reproduce it with relatively high fidelity. When evidence requires qualification, contextual selection, or culturally bounded interpretation, the model may compress, infer, or generalise. This does not demonstrate a causal mechanism in a strict experimental sense, but it suggests a strong association between evidence explicitness and alignment quality. Chang et al. (2024) [9] argue that coverage errors are central to RAG failure, while Mishra et al. (2024) [18] and Rawte et al. (2024) [16] show the value of fine-grained claim diagnosis. This study extends that logic by showing that Indonesian factuality depends not only on whether evidence exists, but also on whether it is sufficiently specific for the claim being generated.

The provenance analysis reframes hallucination as a pipeline problem rather than as a failure located only in the final answer. Its main implication is methodological: RAG evaluation must ask where a claim becomes unreliable. The results show that problematic claims are associated with generation distortion, evidence underspecification, cultural overgeneralisation, retrieval insufficiency, retrieval miss, evidence conflict, and lexical or terminological gaps. This matters because different errors require different interventions. A retrieval miss requires better indexing or query expansion; evidence underspecification requires richer source coverage; generation distortion requires stricter faithfulness control; conflicting evidence requires source hierarchy and adjudication. This finding supports recent work that treats RAG reliability as a system-level issue involving retrieval, context selection, uncertainty, and generation discipline [4], [22], [17]. For Indonesian knowledge, provenance diagnosis is especially important because source authority is distributed across cultural registries, legal databases, statistical portals, lexical authorities, and local institutional pages.

The structure behind provenance errors lies in the mismatch between source availability and claim ambition. RAG systems can retrieve a relevant document yet still generate claims that exceed what the document can support. This is visible when general explainers are used for precise cultural interpretations, when descriptive sources become causal explanations, or when local heritage records are expanded into national cultural narratives. Wu et al. (2024) [10] describe a tension between internal model priors and external evidence, while Kang et al. (2023) [24] and Niu et al. (2024) [23] propose verification and self-refinement as ways to constrain unreliable generation. The findings of this study are compatible with that literature but point to a further issue: in Indonesian domains, the model's prior may not simply add false facts; it may impose an interpretive frame that makes local knowledge appear more general, stable, or nationally unified than the source permits. The problem is therefore not only retrieval absence but also retrieval overconfidence.

The cultural-specificity analysis shows that factual accuracy and cultural adequacy cannot be separated in Indonesian RAG. A generated claim may avoid obvious fabrication yet still weaken cultural knowledge by reducing locality, mistranslating or flattening terminology, simplifying ritual sequence, omitting administrative time, or turning a regional practice into a generic Indonesian tradition. This has significant implications for culturally grounded AI evaluation. Conventional factuality metrics are useful for detecting wrong dates, entities, and contradictions, but they may fail to register representational damage caused by regional scope reduction or symbolic simplification. Recent cultural-knowledge and multimodal RAG evaluation work has begun to recognise that model assessment must account for domain-sensitive evidence and cultural cognition [27], [28], [29]. This study contributes to that direction by treating cultural specificity as an evidential property. In this framework, culture is not decorative context; it is part of what makes a claim verifiable or misleading.

The reason cultural errors persist is that Indonesian knowledge is archipelagic in structure: it is distributed across regions, languages, ritual practices, institutional classifications, and locally situated terms. RAG systems tend to operate through retrieval and synthesis, but synthesis can become distortion when the model collapses local difference into broad categories. Regional attribution, ritual sequence, and local terminology are fragile because they require the model to preserve boundaries rather than merely produce a coherent explanation. This finding both supports and complicates current hallucination literature. It supports the view that hallucination involves unsupported or contradictory content [1], [11], but it complicates that view by showing that cultural hallucination may occur through plausible simplification rather than direct invention. The theoretical position emerging from this study is therefore cautious but clear: reliable Indonesian RAG requires evidence faithfulness, provenance awareness, and cultural localisation. Retrieval is necessary, but for archipelagic knowledge, retrieval alone is not epistemically sufficient.

6. CONCLUSION

This study shows that retrieval-augmented generation does not automatically secure factual and cultural reliability in Indonesian knowledge tasks. Its central lesson is that hallucination in archipelagic knowledge appears not only as fabricated entities, wrong dates, or contradicted legal facts, but also as cultural flattening, regional misattribution, terminological loss, and unsupported interpretive expansion. The main contribution of this study lies in shifting Indonesian RAG evaluation from answer-level accuracy to claim-level, provenance-sensitive, and culturally specific assessment. By combining claim–evidence alignment, retrieval-provenance diagnosis, and cultural-specificity coding, this study offers a methodological framework for detecting errors that conventional factuality checks may overlook.

This study is limited by its controlled corpus scope, selected official sources, and focus on Indonesian factual and cultural knowledge rather than broader multilingual conversational use. Its findings should therefore not be generalised to all RAG systems, all Indonesian domains, or all regional languages without further testing. Future research should expand the corpus across more provinces, local languages, document genres, and multimodal materials, including audio-visual heritage records and community-produced sources. Further work should also compare different retrievers, embedding models, language models, and prompting strategies to examine how system architecture affects factual alignment, retrieval adequacy, and cultural representation in Indonesian RAG.

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

Radna Tulus Wibisono: 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 cultural-heritage records, legal documents, and institutional sources.

ETHICAL APPROVAL

This research did not involve human or animal subjects; it analyzed publicly accessible Indonesian cultural-heritage records, legal documents, and institutional sources, 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] L. Huang, W. Yu, W. Zhong, Z. Feng, H. Wang, Q. Chen, et al., "A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions," *ACM Transactions on Information Systems*, vol. 43, pp. 1–55, 2023, doi: 10.1145/3703155.

[2] S. Tonmoy, S. Zaman, V. Jain, A. Rani, V. Rawte, A. Chadha, and A. Das, "A comprehensive survey of hallucination mitigation techniques in large language models," *arXiv*, 2024, doi: 10.48550/arXiv.2401.01313.

[3] W. Zhang and J. Zhang, "Hallucination mitigation for retrieval-augmented large language models: A review," *Mathematics*, vol. 13, no. 5, Art. no. 856, 2025, doi: 10.3390/math13050856.

[4] N. Thakur, L. Bonifacio, X. C. Zhang, O. Ogundepo, E. Kamalloo, D. Alfonso-Hermelo, et al., ""Knowing when you don't know": A multilingual relevance assessment dataset for robust retrieval-augmented generation," , 2023.

[5] H. Tohir, N. Merlina, and M. Haris, "Utilizing retrieval-augmented generation in large language models to enhance Indonesian language NLP," *JITK (Jurnal Ilmu Pengetahuan dan Teknologi Komputer)*, 2024, doi: 10.33480/jitk.v10i2.5916.

[6] W. Christian, D. Adamlu, A. Yu, and D. Suhartono, "Bridging language gaps with adaptive RAG: Improving Indonesian language question answering," *arXiv*, 2025, doi: 10.48550/arXiv.2510.21068.

[7] F. Muhammad, S. Y. Yakub, M. Overbeek, and F. A. T. Tobing, "RAG-fusion based framework for Indonesian hoax news detection. In 2025 IEEE 2nd International Conference on Cryptography, Informatics, and Cybersecurity (ICoCICs) (pp. 1–6)," *IEEE*, 2025, doi: 10.1109/ICoCICs68032.2025.11384087.

[8] I. K. R. Arthana, N. Gunantara, M. Sudarma, and M. Sukarsa, "LoRA-based fine-tuning of local LLMs for hallucination detection in Indonesian RAG systems," *International Journal of Advanced Computer Science and Applications*, 2026, doi: 10.14569/ijacsa.2026.0170389.

[9] T. A. Chang, K. Tomanek, J. Hoffmann, N. Thain, E. M. Van Liemt, K. Meier-Hellstern, and L. Dixon, "Detecting hallucination and coverage errors in retrieval augmented generation for controversial topics," *arXiv*, 2024, doi: 10.48550/arXiv.2403.08904.

[10] K. Wu, E. Wu, and J. Zou, "ClashEval: Quantifying the tug-of-war between an LLM's internal prior and external evidence," *Advances in Neural Information Processing Systems*, vol. 37, 2024, doi: 10.52202/079017-1053.

[11] S. S. Rahman, M. A. Islam, M. M. Alam, M. Zeba, M. A. Rahman, S. S. Chowa, M. A. K. Raiaan, and S. Azam, "Hallucination to truth: A review of fact-checking and factuality evaluation in large language models," *Artificial Intelligence Review*, vol. 59, 2025, doi: 10.1007/s10462-025-11454-w.

[12] H. Orgad, M. Toker, Z. Gekhman, R. Reichart, I. Szpektor, H. Kotek, and Y. Belinkov, "LLMs know more than they show: On the intrinsic representation of LLM hallucinations," *arXiv*, 2024, doi: 10.48550/arXiv.2410.02707.

[13] Z. Sun, X. Zang, K. Zheng, Y. Song, J. Xu, X. Zhang, W. Yu, and H. Li, "ReDeEP: Detecting hallucination in retrieval-augmented generation via mechanistic interpretability," *arXiv*, 2024, doi: 10.48550/arXiv.2410.11414.

[14] L. Wang, "SEReDeEP: Hallucination detection in retrieval-augmented models via semantic entropy and context-parameter fusion," *arXiv*, 2025, doi: 10.48550/arXiv.2505.07528.

[15] W. Su, Y. Tang, Q. Ai, C. Wang, Z. Wu, and Y. Liu, "Mitigating entity-level hallucination in large language models. In Proceedings of the 2024 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region," *ACM*, 2024, doi: 10.1145/3673791.3698403.

[16] V. Rawte, S. M. Towhidul, I. Tonmoy, K. Rajbangshi, S. Nag, A. Chadha, et al., "FACTOID: FACtual enTailment fOr hallucInation Detection," *arXiv*, 2024, doi: 10.48550/arXiv.2403.19113.

[17] M. Li, Z. Zhan, H. Yang, Y. Xiao, H. Zhou, J. Huang, and R. Zhang, "Benchmarking retrieval-augmented large language models in biomedical NLP: Application, robustness, and self-awareness," *Science Advances*, vol. 11, 2025, doi: 10.1126/sciadv.adr1443.
[18] A. Mishra, A. Asai, V. Balachandran, Y. Wang, G. Neubig, Y. Tsvetkov, and H. Hajishirzi, "Fine-grained hallucination detection and editing for language models," *arXiv*, 2024, doi: 10.48550/arXiv.2401.06855.
[19] B. Wang, S. Chern, E. Chern, and P. Liu, "Halu-J: Critique-based hallucination judge," *arXiv*, 2024, doi: 10.48550/arXiv.2407.12943.
[20] I. Zimmerman, J. Tredup, E. Selfridge, and J. Bradley, "Two-tiered encoder-based hallucination detection for retrieval-augmented generation in the wild. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track (pp. 8–22)," *Association for Computational Linguistics*, 2024, doi: 10.18653/v1/2024.emnlp-industry.2.
[21] L. M. Amugongo, P. Mascheroni, S. Brooks, S. Doering, and J. Seidel, "Retrieval augmented generation for large language models in healthcare: A systematic review," *PLOS Digital Health*, vol. 4, 2025, doi: 10.1371/journal.pdig.0000877.
[22] E. Fadeeva, A. Rubashevskii, R. Vashurin, S. Dhuliawala, A. Shelmanov, T. Baldwin, et al., "Faithfulness-aware uncertainty quantification for fact-checking the output of retrieval augmented generation," *arXiv*, 2025, doi: 10.48550/arXiv.2505.21072.
[23] M. Niu, H. Li, J. Shi, H. Haddadi, and F. Mo, "Mitigating hallucinations in large language models via self-refinement-enhanced knowledge retrieval," *arXiv*, 2024, doi: 10.48550/arXiv.2405.06545.
[24] H. Kang, J. Ni, and H. Yao, "Ever: Mitigating hallucination in large language models through real-time verification and rectification," *arXiv*, 2023, doi: 10.48550/arXiv.2311.09114.
[25] G. Kim, J. Park, J. Kim, S. Han, K. Son, and I. Jang, "SERC: LDPC-inspired semantic error correction for retrieval-augmented generation," , 2026.
[26] R. J. Wowor, S. Y. Yakub, and S. Indarjani, "AI-driven misinformation mitigation: Implementation of RAG system for hoax detection and cybersecurity in Indonesia. In 2025 IEEE 2nd International Conference on Cryptography, Informatics, and Cybersecurity (ICoCICs) (pp. 458–463)," *IEEE*, 2025, doi: 10.1109/ICoCICs68032.2025.11383967.
[27] 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.
[28] M. Mortaheb, M. A. Khojastepour, S. Chakradhar, and S. Ulukus, "RAG-Check: Evaluating multimodal retrieval augmented generation performance," *arXiv*, 2025, doi: 10.48550/arXiv.2501.03995.
[29] S. Hildebrand, C. Taylor, S. Oesch, J. Ghawaly, A. Sadovnik, R. Shivers, B. Schreiber, and K. Kurian, "FATHOMS-RAG: A framework for the assessment of thinking and observation in multimodal systems that use retrieval augmented generation," *arXiv*, 2025, doi: 10.11578/dc.20260227.1.

BIOGRAPHY OF AUTHOR

Radna Tulus Wibisono is affiliated with Universitas Gadjah Mada, Indonesia. His academic studies focus on linguistics, language, and related humanities disciplines, and he is actively engaged in research on cognitive linguistics, political discourse, metaphor studies, semantics, and public communication. His recent works examine how language structures political reasoning and economic imagination in Indonesian public debates. He can be contacted at email: radnana89@gmail.com