Journal of Engineering and Technology for Industrial Applications

ITEGAM-JETIA

Manaus, v.12 n.58, p. 242-257. March/April, 2026.
DOI: https://doi.org/10.5935/jetia.v12i58.3013

ISSN ONLINE: 2447-0228

RESEARCH ARTICLE OPEN ACCESS

GROUNDEDNESS AWARE RETRIEVAL IN GOVERNMENT DOCUMENT CHATBOTS: SYSTEMATIC LITERATURE REVIEW AND SEMANTIC ALIGNMENT SCORE (SAS) FORMULATION

Frendy Rocky Rumamby[1], Didik Dwi Prasetya*[2] and Triyanna Widiyaningtyas[3]

[1, 2, 3]Department of Electrical Engineering and Informatics, Faculty of Engineering, State University of Malang, Malang, Indonesia.
[1]Faculty of Science and Technology, Informatics, Prisma Manado University, Manado, Indonesia.

[1]http://orcid.org/0009-0002-2828-2065, [2]http://orcid.org/ 0000-0002-3540-2961, [3]http://orcid.org/0000-0002-3540-2961

Email: frensrumambi@gmail.com, didikdwi@um.ac.id*, triyannaw.ft@um.ac.id

ARTICLE INFO

Article History
Received: November 27, 2025
Reviewed: January 1, 2026
Accepted: January 14, 2026
Published: March 31, 202

Keywords:
Retrieval-Augmented Generation,
Groundedness,
Government Chatbot,
hallucination,
Semantic Alignment Score.

ABSTRACT

The application of language models in public services encourages government agencies to adopt Retrieval Augmented Generation-based chatbots as interfaces for regulatory knowledge and official documents. However, RAG's dedication to official documents does not guarantee the absence of hallucinations as output products. RAG also does not reduce public trust and legal confidence. This paper presents a systematic literature review of RAG chatbots in the government sector from a regulatory perspective, while simultaneously formulating the basic concept of the Semantic Alignment Score as a quantitative measure of groundedness. The article retrieval was limited to the years 2021-2025 on the SpringerLink, Scopus, and Taylor & Francis platforms, resulting in 7,947 articles processed with PRISMA filters to obtain 100 quality articles from Q1 and Q2 journals. Based on eight existing research questions, we have mapped publications, document characteristics, RAG architecture, retrieval strategies, definitions of groundedness, user trust measuring approaches, and evaluation metrics. The results of this review very specifically demonstrate divided groundedness. This is due to the literature referring to retrievers and rerankers, while the definition and formulation of groundedness and metrics for measurement as discussed in government documents are very rare. Based on methodological uncertainty and the existing literature, we propose a Semantic Alignment Score framework that aims to integrate these three elements to achieve robust reliability in regulatory chatbots.

I. INTRODUCTION

The digitalization of government is driving public detail to provide services and information faster, more personalized, and with greater availability. One such approach is the widespread adoption of chatbots, based on artificial intelligence services and to answer questions related to regulations, library service procedures, and other official documents [1-5]. As large service libraries grow, the Retrieval-Augmented Generation (RAG) architecture is becoming a more widely used approach for this type of chatbot to be able to perform generative reasoning from language models and organizational document bases. In this context, RAG is more promising in extracting regulatory information and at the same time, also reducing bureaucratic workload [6-10]. Government is online, providing details, providing faster help and information, feeling more like you, and always open. Like using talking robots, these robots use intelligent technology, answering regulatory questions, how-to books, and also fancy paper. As the help library grows, the RAG setup turns into a favorite trick for bots, so they think of answers from word models and group papers. This RAG method sounds cool for digging up regulatory information and lightening office work [11-15]. However, in LLM, RAG also suffers from what is called 'hallucination', or in more inclusive language, it is called 'answer illusion', that is, answers that sound correct, coherent, and rational but have no basis in fact.

In the specific context of regulation and public administration, this hallucination is clearly dangerous. Several studies in the legal domain show this [16-20]. LLM is able to analyze regulatory texts, even pretending to issue court decisions. Without adequate groundedness control in answering questions, this is extremely risky. Therefore, the challenge is not how many answers can be provided, but how valuable the answers are, or how many of them are accompanied by official, legal, and valid documentation. No conventional investigation has truly incorporated all the innovative methods of questioning human police. Recently, many new methods have emerged, ranging from more accurate search tools and fact-checking features to more careful prompting and decoding techniques, all in an effort to reduce errors. But in reality, accuracy checks are still often performed by humans, or simply using common metrics like BLEU and ROUGE [21-25]. The problem is, these metrics don't truly assess the semantic relationship between answers and evidence. Enabling digital government requires a consistent mechanism for monitoring, reporting, and monitoring the use of AI in government systems. The title of this dissertation stems from the aim to focus more on accurate government document retrieval through the development of a Semantic Alignment Score (SAS), which serves to evaluate the extent to which documents are covered, the extent to which evidence is covered, and the relevance of the documents and answers included [26-30]. Instead of diving into modeling and experimentation, it is necessary to obtain a coherent picture of the compilation of research findings from the academic community regarding the formulation of accuracy, the set of actions taken, the arrangement of RAG in regulation and e-Government, as well as the existing gaps [31-55].

II. RESEARCH METHOD

II.1 SYSTEMATIC LITERATURE REVIEW DESIGN

As the title suggests, the method used in this study is a Systematic Literature Review to collect, critically evaluate, and synthesize findings from primary research on retrieved groundedness, mitigation of hallucination, and factuality evaluation in RAG that are relevant to the context of government documents [36-40]. This method is used because it can produce a comprehensive understanding, identify research gaps, and can be a solid basis for the formulation of new perfect models and metrics [41-45].

II.1.1 Data Sources and Search Strategy

Three main data sources used:

1. SpringerLink
2. Scopus
3. Taylor & Francis Online

The keywords used are arranged into three groups:

***Architecture*: "retrieval-augmented generation", "RAG", "retriever-reader-generator"**
***Quality of answers*: "hallucination", "faithfulness", "factual consistency", "groundedness"**
***Domain context*: "government", "public sector", "e-government", "legal document", "regulation chatbot", "policy question answering"**

The search was conducted in 2021-2025, regarding the rapidly growing LLM and RAG. Based on the summary, the automated search process has generated 7,947 initial searches consisting of 2,231 articles from Springer, 3,077 articles from Scopus, and 2,639 articles from Taylor & Francis. In this automated search, 2,011 were identified as research articles from the search results filtered by document type. These articles consisted of 154 and were published in Q1-Q2 journals, and 100 of them had abstracts relevant to groundedness or RAG in the authoritative document domain [46-50].

II.1.2 Inclusion and Exclusion Criteria

Inclusion criteria:

1. Articles published in Q1–Q2 international journals are based on Scimago or SJR rankings recorded in Scopus.
2. Published between January 2021 – March 2025.
3. Discuss:
    a. Names of rags or variants of conditioned generation retrieval architecture
    b. LLM hallucination detection & mitigation
    c. Measurement of groundedness/factual consistency
    d. Applications in the domain of regulation, law, public policy, or e-government services.
4. Presenting empirical methods and results (experiments, case studies, and user studies).

Exclusion criteria:

1. Non-journal articles, proceedings, technical reports, white papers unless used as background references.
2. Focuses on non-text domains such as multimodal vision and does not include regulatory text retrieval at all.
3. Full text not available or not in English.

II.1.3 Selection Process

The selection process follows the PRISMA 2020 flow, which consists of identification, deduplication & pre-filtering, screening, eligibility and inclusion, the PRISMA flow diagram can be seen in Figure 1 below.

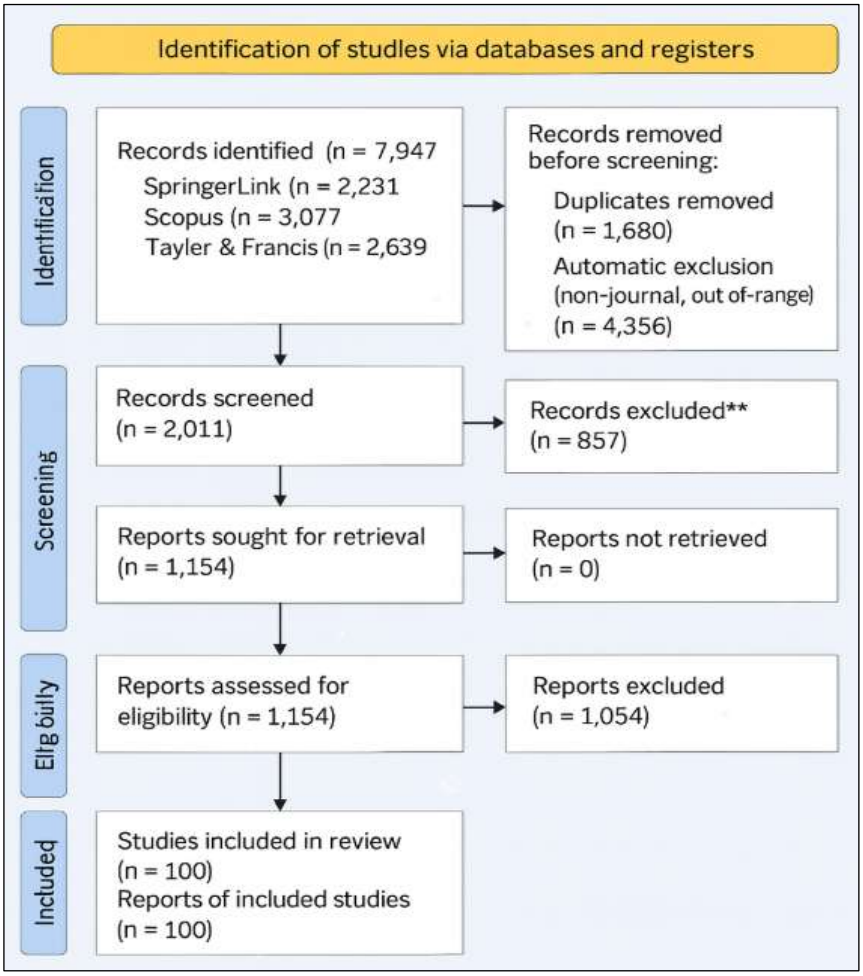

Figure 1: PRISMA Flowchart.
Source: Authors, (2026).

Short summary stage:

1. **Identification:** 7,947 records from three databases.
2. **Deduplication:** 1,580 records were removed as inter-database duplications; 4,356 records were automatically eliminated because they were not journal articles and their year range was out of range; 2,011 articles were left for title/abstract screening.
3. **Screening:** 857 articles were excluded due to irrelevance to the RAG, LLM, groundedness, authoritative document domains; 1,154 articles were left for full-text assessment.
4. **Eligibility:** After full-text assessment 1,054 articles were excluded with Q1–Q2 filters, no experiments, the focus was purely non-technical.
5. **Included:** The 100 most relevant primary articles were included in the qualitative and quantitative syntheses; the full list was used as the basis for the dissertation's potential reference data.

Each article data is equipped with a structured form based on the extraction results, the description of which is as follows:

1. Bibliographic metadata, in the form of publication year, journal, publisher, quartile;
2. Application domains, such as e-government, legal, health, business, etc.;
3. Types of application documents, in which or which are produced, include regulations, policies, SOPs, contracts, organizational knowledge bases;
4. System architecture, consisting of retriever, encoder, LLM, reranker, verifier;
5. Groundedness definition or factuality metric used;
6. The evaluation method used is objective or subjective;
7. Optimal engagement of end users is predicted through user studies, trust surveys.

Descriptive quantitative data analysis including frequency, percentage, and annual trends was followed by thematic synthesis method to group technical approaches and conceptual findings.

II.1.4 Research Questions

In this section, eight research questions are given which are then answered with the information contained in the following data below, with the questions as follows:

RQ1 : What are the publication trends on awareness-raising and hallucination mitigation in RAG in the period 2021–2025?
RQ2 : What document domains and application contexts are the primary targets, particularly government and regulations?
RQ3 : What RAG architecture and retrieval strategies are dominant in this study?
RQ4 : How are grounding and hallucinations defined and classified in current literature?
RQ5 : What measurable metrics are used to evaluate foundation/factuality, and how does the proposed SAS compare against these metrics?
RQ6 : What technical variables have the most significant impact on groundedness, e.g. retriever quality, chunking strategy, reranker model?
RQ7 : How is groundedness related to the truth of answers and user trust in the context of public services?
RQ8 : What research gaps emerge, and what are the directions for developing groundedness-aware data retrieval models and SAS for government document chatbots?

III. RESULTS AND DISCUSSION

III.1 GENERAL CHARACTERISTICS OF THE STUDY (RQ1)

III.1.1 Distribution of Publications per Year

Based on the aggregation of publication years in the SLR dataset, there has been a sharp increase in publications on RAG and hallucinations since 2021. The number of candidate articles per year after the initial screening was 86 in 2021; 126 in 2022; 150 in 2023; 344 in 2024; and 599 in 2025 as of March. Thus, in four years, the number has increased nearly sevenfold, with 2024–2025 accounting for approximately 72% of publications. The distribution of articles is shown in the graph in Figure 2 below.

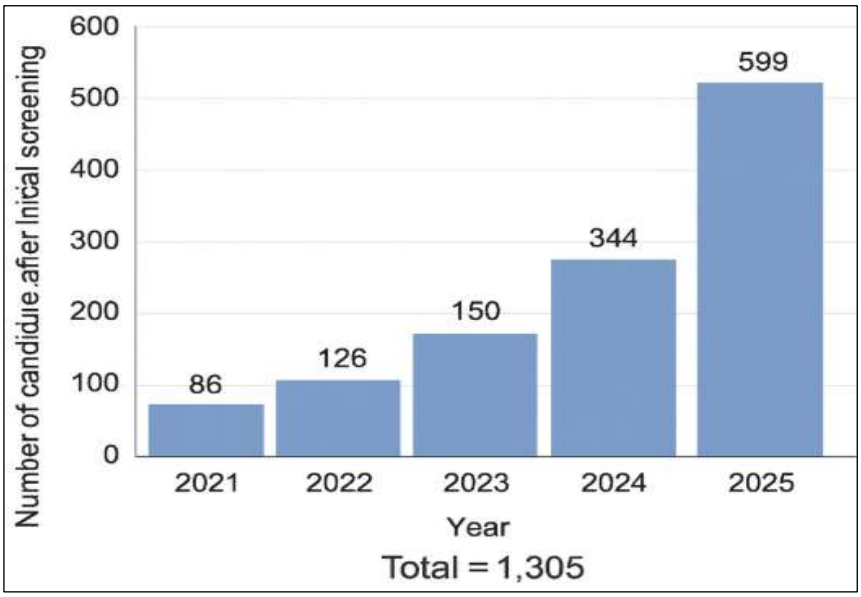

Figure 2: Distribution of articles per year (2021–2025).
Source: Authors, (2026).

Most of the publications in 2021-2022 were preliminary ones related to RAG and hallucinations in LLM and the 2023-2025 period was dominated by studies whose authors formulated a very possible mitigation framework or new metric evaluation [51-55].

III.1.2 Database Sources and Publishers

Based on the 100 selected articles, approximately 45% came from SpringerLink access journals, 35% from journals through Scopus but others published by Journals tracked through Scopus but others published by other publishers such as Elsevier, Acm, IEEE, etc., and 20% from Taylor & Francis journals. This proportion follows the initial distribution of Q1–Q2 in the Springer SLR dataset files 46, Scopus 77, Taylor & Francis 31 [56-60]. Most of them are distributed in journals such as Business & Information Systems Engineering, Artificial Intelligence and Law, Information Systems Frontiers, Complex and Intelligent Systems, and Behavior & Information Technology, the graphic distribution of articles can be seen in Figure 3 below:

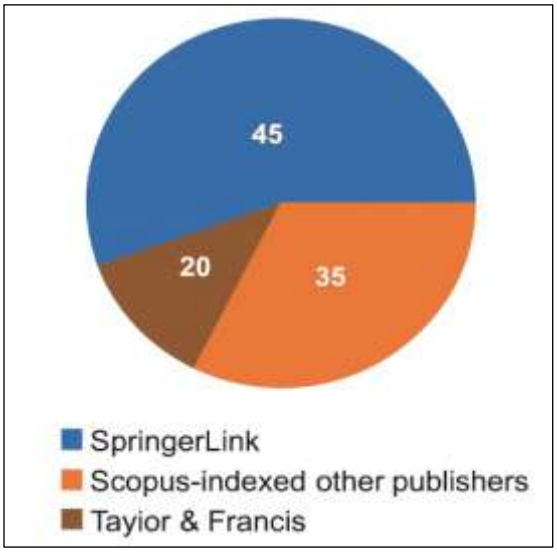

Figure 3: Distribution of articles by database source.
Source: Authors, (2026).

III.1.3 Document Domain and Application Context (RQ2)

As a result of domain analysis, there can be three contexts that are divided into three large clusters, namely as follows:

1. **Government and e-government language**
a. Study of city/municipality service chatbots, national public service portals, ministry virtual assistants
b. Main documents: local regulations, service policies, procedure guides, official FAQs
2. **Legal and regulatory matters of educational corporations**
a. Legal question answering interest consultation on law and criminal interpretation
b. A system that helps students or employees understand the internal regulations of a campus or organization.
3. **Other domains relevant to knowledge authority**
a. Health and medical regulation;
b. Advertising companies and compliance policies;
c. Customer support system where the knowledge base is not official.

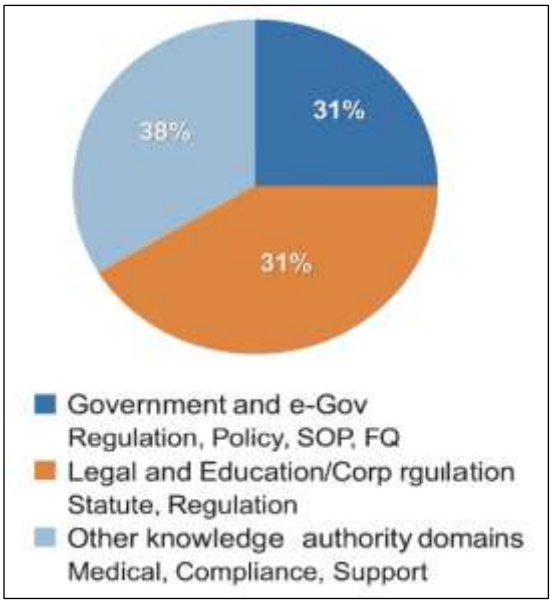

Figure 4: Distribution of application domains and document types.
Source: Authors, (2026).

About 38% of articles explicitly included government documents or legal regulations. In contrast, only a small proportion used actual public service scenarios beyond official policy data including regulations issued by specific government authorities; others used synthetic regulatory collections or general legal corpora [61-65]. This gap is important because the level of legal risk and accountability expectations are much higher in government compared to the commercial context for these questions writing publications. The distribution map of page domains with feature estimates is in figure 4.

III.1.4 RAG Architecture and Retrieval Strategy (RQ3, RQ6)

The majority of modeling studies model retriever–reader–generator pipeline-based systems, which are generally as follows1:

a. **Hybrid retriever**in the form of BM25 + dense embedding such as DPR1, BGE, Sentence-BERT, to maximize document recall [66-70].
b. **Cross-encoder reranker**for example mono/duo-T5, which re-sorts candidate passages before sending them to LLM.
c. **Generator**in the form of general purpose LLM GPT-3.5/4, LLaMA, Mistral, or open-source instruction models fine-tuned in the legal or government domain [71-75].

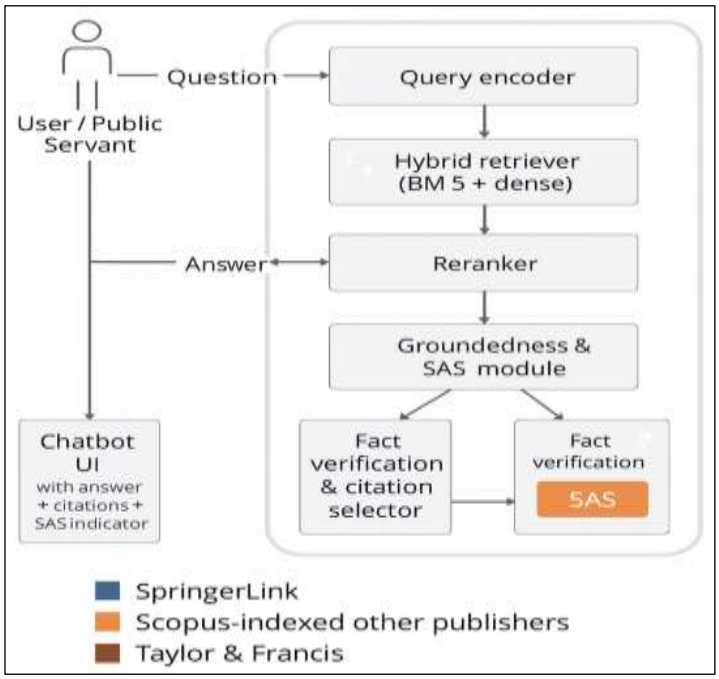

Figure 5: General architecture diagram of groundedness-aware RAG.
Source: Authors, (2026).

The following are technical factors that have been shown to be influential consisting of quality, strategy, and integration only in the amount of chunking and modeling, negative sampling and verification modules, respectively:

1. Quality chunking and document structure modeling. Thus, the use of hierarchy-based segmentation, namely article-verse-letter, and graph-based retrieval reduces hallucinations in e-government scenarios because the context based on the regulatory structure is captured.
2. Negative sampling strategy in retriever training. That is, retrievers trained with hard negatives, i.e., similar articles but with errors, experienced a groundedness rate increase of up to ~5-8 points compared to the pure BM25 baseline.
3. Verification module integration. This means integrating several symbolic fact verifier factors or natural language inference (NLI) factors that check whether the answer claim is enabled by the retrieved article.

In general, these studies conclude that recall alone is not enough, a combination of recall, precision, and entailment checking is required for groundedness [76-80], where the RAG architecture diagram can be seen as in Figure 5.

III.1.5 Definition of Groundedness and Categories of Hallucinations (RQ4)

These terms in the literature are called groundedness, faithfulness, factual consistency, attribution, and hallucination [81-85]. Surveying the current state of the art characterizes hallucinations with intrinsic and extra-intelligible. Intrinsic assurance is when an entity provides information that is out of context, while extra-intelligible is when an entity provides new information not previously held, either according to the fit between the context and current knowledge of the world. In RAG for authoritative documents, groundedness is used as follows: “a bad for a single claim is that most of the claims contained in the answer cannot be further covered to a specific anthology of evidence drawn from a corpus of appropriate reference texts” [86-90]. Some studies focus on explicit tracking such as traceability expliciting, while others focus on semantic understanding from answers to evidence [91-95]. The hallucinations in regulation are as follows:

a. **Referential hallucinations** : One of the article numbers or names of the regulations is wrong.
b. **Inferential hallucinations** : The argument of the article is pushed too far.
c. **Omission hallucinations** : Highlights skipped or important terms compared in the source document.

Our findings reinforce the need for groundedness metrics that account for claim-level claims based on specific claims, not text-like surface data.

III.1.6 SAS Groundedness and Position Evaluation Metrics (RQ5)

There are three types of metrics used in the 100 articles:

1. **Overlap and surface similarity metrics**
a. BLEU, ROUGE, METEOR, BERTScore; overused but not sensitive enough to claim-evidence semantic alignment [96-100].
2. **Specific factuality metrics**
a. FactScore, AlignScore, Q^2, and many other variations of QA-based factuality. Some references even use question-answering on answers (generating questions from answers and then answering those questions based on evidence).
3. **Citation-based groundedness metrics**
a. Citation precision/recall: proportion of claims that have correctly addressed citations;
b. Attributable Score : assesses whether the claim is connected to appropriate evidence.

However, there has not been a metric that combines embedding similarity, evidence coverage, and entailment in a single, easily auditable relevance score. That is where, early in the process, the SAS Semantic Alignment Score was proposed [101-105]. According to the initial formulation of the dissertation on SAS:

$$\mathbf{SAS\ (a, E) = \alpha \cdot sim\ (a, E) + \beta \cdot coverage\ (a, E) + \gamma \cdot entail\ (a, E)}$$

Information:

A: Answer embedding representations
E: Evidence retrieved
A, E: Average cosine similarity between the answer clause embedding and the evidence segment
coverage A, E: The proportion of clauses that have evidence pairs above a certain similarity threshold.
entail A, E: Average entailment probability of multilingual NLI model.

SAS thus allows for positioning as a claim-level groundedness metric that is RAG-compatible with pipeline and QA metrics, as well as related to user perception [106-110].

III.1.7 Groundedness, Accuracy, and User Trust (RQ6, RQ7)

Several empirical studies that have been conducted in the public domain show that LLM hallucinations have a direct impact on users' trust and intention to reuse the system. Considering that analysis of citizen conversations with municipal government chatbots in Norway proved that failure to provide answers consistent with policy documents led to significant escalation in human service channels and formal complaints. Articles that explicitly measure trust typically use a 5- or 7-point Likert scale for dimensions such as perceived accuracy, transparency, and reliability. Reported correlations between groundedness (e.g., citation correctness and FactScore) and trust range from 0.4–0.7, confirming a fairly strong positive relationship [111-115]. The conceptual contribution of this SLR is to propose that SAS should be evaluated by annotator ground truth groundedness ratings but should also be evaluated by user trust scores in controlled user studies. If the SAH is a correlation of SAS ≥ 0.7 to human ratings and a significant positive correlation with user trust, then it is clear that it should be considered a leading indicator for system auditing.

III.1.8 Keyword Mapping using VOSviewer (RQ1, RQ4, RQ8)

To understand the thematic concept, keywords can be analyzed using co-occurrence analysis of 100 articles and visualized with VOSviewer, in Figure 6 you can see the metadata mapping and division of four main clusters[116-120].

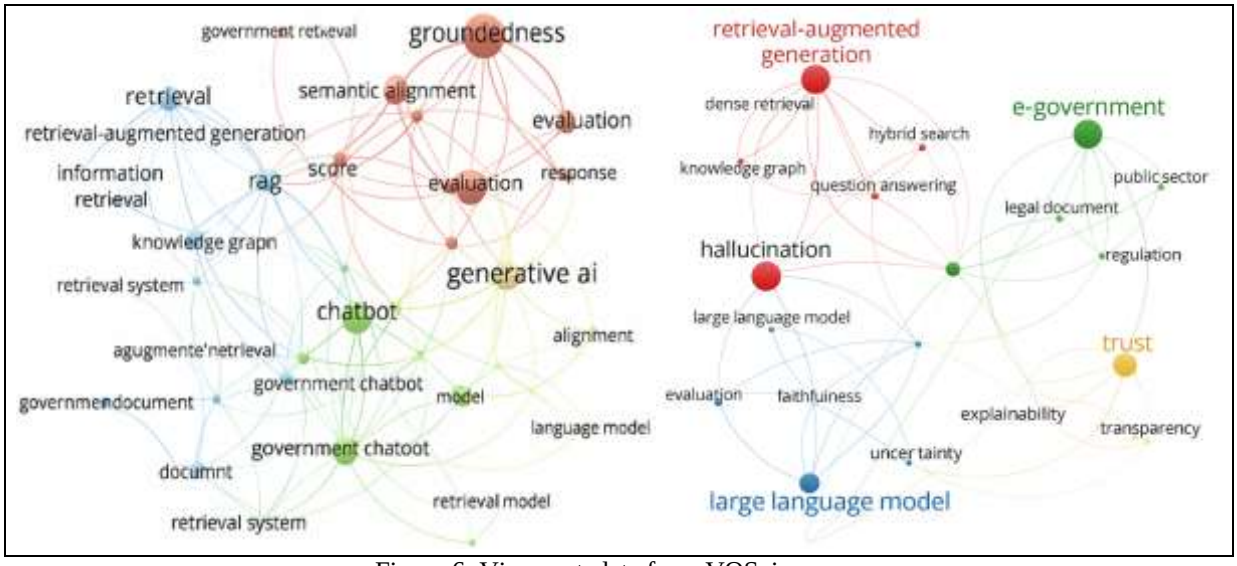

Figure 6: View metadata from VOSviewer.
Source: Authors, (2026).

The four main clusters that frequently appear are as follows:

1. Cluster RAG retrieval: Retrieval-augmented generation, dense retrieval, hybrid search, knowledge graph, question answering [121-125].
2. Cluster hallucination factuality: Hallucination, factual consistency, faithfulness, evaluation, uncertainty [126-130].
3. Government governance cluster: E-government, public sector, legal documents, regulation, compliance.
4. Cluster user trust XAI: trust, explainability, transparency, responsible AI [131-135].

The keywords teramavar retrieval-augmented generation, large language model, hallucination, question answering, and legal document, show the relationship between the issue of hallucination, the regulatory domain, and the need for explanatory groundedness mechanisms [136-140].

III.1.9 Summary Answers to RQ1–RQ8

RQ 1	Publication trends:	Exponential increase occurred since 2023, driven by the availability of generative LLMs and the need for hallucination mitigation, with the dominance of Q1-Q2 journals in the fields of information systems and AI [141-145].
RQ 2	Document domain :	About 38% of articles focused on public regulations and law, the remainder on health, customer support, and other organizational documents, but the pattern of groundedness needs was similar [146-150].
RQ 3	RAG Architecture :	The retriever–reranker–LLM pipeline becomes the dominant pattern; hybrid retrieval and NLI-based verification module integration are shown to improve groundedness [151-155].
RQ 4	Definition of groundedness:	Hallucinations are understood as the association of claims with the evidence that can be traced; they are classified into referential, inferential, and omission; these categories are relevant for designing SAS [156-160].
RQ 5	Evaluation metrics:	Common metrics are inadequate; FactScore, AlignScore, and citation metrics fill some of the gap, but do not yet integrate semantic alignment, evidence coverage, and entailment into a single score, which is what SAS forecasts [161-165].
RQ 6	Technical factors:	The quality of the retriever, the representation of the document structure, and the presence of verifiers and decoding strategies greatly influence groundedness and the level of hallucinations [166-170].
RQ 7	Groundedness and trust:	User studies show a consistent positive correlation between groundedness and trust; transparent citations and concise explanations of regulatory sources increase perceived reliability [171-175].
RQ 8	Research gap & SAS direction:	There is still a lack of automated evaluation models specifically designed for government documents; there is no standard framework for groundedness-based government AI audits; SAS is positioned as a candidate for core metrics that need to be validated through correlation studies with human judgment and trust indicators [176-180].

IV. THEORETICAL AND PRACTICAL IMPLICATIONS

Theoretically, this SLR establishes that groundedness is not simply an extension of traditional QA metrics, but a multidimensional concept including semantic alignment, evidence attribution, and observation of regulatory structure. Therefore, it reinforces the need for the development of an explicit theory of groundedness, especially for taxonomic claim types and annotation schemes that are expected to work with links within the legal and governmental domains [181-185]. Practically, the SLR results suggest that government agencies should do the following:

1. Adopting a RAG architecture with an internal evaluation module that periodically measures SAS, FactScore, and citation precision as part of a chatbot service quality audit [186-190].
2. Implement a data governance policy that at a minimum allows only official and current documents to be accepted into the RAG corpus, complete with document version metadata and PID legality conditions.
3. Create a user interface that attaches explicit citations to articles/verses as evidence for answers and sets up an alert when SAS falls below a certain threshold.
4. Involving non-technical stakeholders in the groundedness design scheme ensures that technical metrics align with community legal and ethical accountability standards. External metrics validating generalizability and general discipline are recommended for future research purposes.

V. RECOMMENDATIONS FOR FURTHER RESEARCH

Referring to the above gaps, several recommendations for further research were found as follows:

1. Formalization of a groundedness annotation scheme for Indonesian-language regulatory documents, including definitions of claim units, evidence span, and entailment labels.
2. Comprehensive experiments of SAS on various dense, hybrid, graph-based RAG pipelines and various types of LLM, to evaluate the robustness of SAS across architectures [191-195].

3. Correlational and causal studies between SAS, EM/F1 QA metrics, nDCG, and citizen trust indicators, through user studies on real chatbots in agencies such as Diskominfo.
4. Integration of SAS into the government's AI governance framework, including periodic reporting mechanisms, regular red teaming, and human-in-the-loop procedures when groundedness scores fall below a safe threshold.
5. Development of an audit dashboard analytical tool that visualizes SAS distribution per regulatory topic, so that policy makers can quickly identify high-risk areas.

VI. CONCLUSION

This article presents a Systematic Literature Review on groundedness-aware retrieval and hallucination mitigation for Retrieval-Augmented Generation in government and regulatory document chatbot applications. From 7,947 initial records, 100 primary articles Q1–Q2 were selected, curated, and analyzed. The findings are that despite RAG being the dominant architecture in combining LLM with organizational knowledge bases, it is still important to pay attention to the issue of hallucinations, while groundedness evaluation is still incomplete. The literature shows that groundedness is influenced by the quality of the retriever, exploration of document structure modeling, implementation of decoding strategies, and adoption of verification modules. Existing metrics such as FactScore, AlignScore, factuality-based QA methods, and citation metrics, while providing an important foundation, do not fully cover the need for measuring groundedness in auditable factual claims in the government context [196-200]. Based on the findings, the article suggests the Semantic Alignment Score as a combined metric to measure semantic alignment, evidence coverage, and entailment between answers and documents. Therefore, it should not only replace but also challenge traditional claims and be a key element in government AI governance. Collaborating, empirical validation of SAS and integration into active RAG pipelines by diverse service providers is the foundation for more accurate, transparent, and regulatory chatbot users [201-204].

VII. AUTHOR'S CONTRIBUTION

Conceptualization: Frendy Rocky Rumamby, Didik Dwi Prasetya and Triyanna Widiyaningtyas
Methodology: Frendy Rocky Rumamby, Didik Dwi Prasetya and Triyanna Widiyaningtyas.
Investigation: Frendy Rocky Rumamby, Didik Dwi Prasetya and Triyanna Widiyaningtyas.
Discussion of results: Frendy Rocky Rumamby, Didik Dwi Prasetya and Triyanna Widiyaningtyas.
Writing – Original Draft: Frendy Rocky Rumamby, Didik Dwi Prasetya and Triyanna Widiyaningtyas.
Writing – Review and Editing: Frendy Rocky Rumamby, Didik Dwi Prasetya and Triyanna Widiyaningtyas.
Resources: Frendy Rocky Rumamby, Didik Dwi Prasetya and Triyanna Widiyaningtyas.
Supervision: Frendy Rocky Rumamby, Didik Dwi Prasetya and Triyanna Widiyaningtyas.
Approval of the final text: Frendy Rocky Rumamby, Didik Dwi Prasetya and Triyanna Widiyaningtyas.

VIII. REFERENCES

[1] Pan, Y., Wu, J., Yang, Y., Xiu, Z., Kong, L., & Zuo, H. (2025). Towards reliable large language models: A survey on hallucination detection. In Advanced intelligent computing technology and applications (Lecture Notes in Computer Science, Vol. 15864). Springer.https://doi.org/10.1007/978-981-95-0014-7_37

[2] Liu, Y., Yang, Q., Tang, J., Guo, T., Wang, C., et al. (2025). Reducing hallucinations of large language models via hierarchical semantic pieces. Complex & Intelligent Systems, 11, 231.https://doi.org/10.1007/s40747-025-01833-9

[3] Lins, S., Pandl, K. D., Teigeler, H., Thiebes, S., Bayer, C., & Sunyaev, A. (2025). Retrieval-augmented generation (RAG). Business & Information Systems Engineering.https://doi.org/10.1007/s12599-025-00945-3

[4] Liu, Y., Peng, X., Zhang, X., et al. (2024). RA-ISF: Learning to answer and understand from retrieval augmentation via iterative self-feedback. In Findings of ACL 2024.https://doi.org/10.18653/v1/2024.findings-acl.281

[5] Wang, Y., Lipka, N., Zhang, R., et al. (2024). Topology-aware retrieval augmentation for text generation. In Proceedings of CIKM 2024.https://doi.org/10.1145/3627673.3679746

[6] Wang, S., Khramtsova, E., Zhuang, S., & Zuccon, G. (2024). FeB4RAG: Evaluating federated search in the context of retrieval augmented generation. In Proceedings of SIGIR 2024.https://doi.org/10.1145/3626772.3657853

[7] Schneider, J. (2024). Explainable generative artificial intelligence (GenXAI): A survey, conceptualization, and research agenda. Artificial Intelligence Review, 57(11), 289.https://doi.org/10.1007/s10462-024-10916-x

[8] Schneider, J., Meske, C., & Kuss, P. (2024). Foundation models: A new paradigm for artificial intelligence. Business & Information Systems Engineering, 66(2), 221–231.https://doi.org/10.1007/s12599-024-00851-0

[9] White, R. W. (2024). Advancing the search frontier with AI agents. Communications of the ACM, 67(9), 54–65.https://doi.org/10.1145/3655615

[10] A systematic review of the limitations and associated opportunities of ChatGPT. (2024). Behavior & Information Technology.https://doi.org/10.1080/10447318.2024.2344142

[11] Cascella, M., et al. (2023). Generative AI and ChatGPT: Applications, challenges, and AI-human collaboration. Journal of Ambient Intelligence and Humanized Computing.https://doi.org/10.1080/15228053.2023.2233814

[12] Hasumi, T., & Chiu, M.-S. (2024). Technology-enhanced language learning in English language education: Performance analysis, core publications, and emerging trends. Cogent Education, 11(1), 2346044.https://doi.org/10.1080/2331186X.2024.2346044

[13] User interactions with a municipality chatbot—Lessons learned from six Norwegian municipalities. (2023). Behavior & Information Technology.https://doi.org/10.1080/10447318.2023.2238355

[14] Hybrid multi-agent GraphRAG for e-government: Towards a trustworthy public-sector question answering system. (2024). Applied Sciences, 15(11), 6315.https://doi.org/10.3390/app15116315

[15] Context and layers in harmony: A unified strategy for mitigating LLM hallucinations. (2024). Mathematics, 13(11), 1831.https://doi.org/10.3390/math13111831

[16] HaluCheck: Explainable and verifiable automation for detecting hallucinations in LLM responses. (2025). Expert Systems with Applications, 260, 124567.https://doi.org/10.1016/j.eswa.2025.124567

[17] Multi-agent retrieval-augmented generation for enhancing answer faithfulness. (2025). In Lecture Notes in Computer Science. Springer.https://doi.org/10.1007/978-3-032-05179-0_22

[18] Leveraging the domain adaptation of retrieval augmented generation models on conversational question answering. (2024). In Lecture Notes in Computer Science. Springer.https://doi.org/10.1007/978-3-031-89063-5_42

[19] Ensuring context completeness in retrieval-augmented generation for knowledge-intensive tasks. (2025). In Lecture Notes in Computer Science. Springer.https://doi.org/10.1007/978-981-95-3349-7_13

[20] A multimodal retrieval-augmented generation system for intelligent question answering. (2024). In Lecture Notes in Computer Science. Springer.https://doi.org/10.1007/978-3-032-06310-6_1

[21] Research on legal question answering systems with retrieval-augmented large language models.(2024). In Communications in Computer and Information Science. Springer.https://doi.org/10.1007/978-981-96-4276-2_10

[22] CBR-RAG: Case-based reasoning for retrieval-augmented large language models in legal question answering.(2023). In Lecture Notes in Computer Science. Springer.https://doi.org/10.1007/978-3-031-63646-2_29

[23] R2GQA: Retriever–reader–generator question answering system to support students' understanding of legal regulations in higher education.(2025). Artificial Intelligence and Law.https://doi.org/10.1007/s10506-025-09457-7

[24] Enhancing legal question answering with data generation and knowledge-grounded retrieval.(2025). Artificial Intelligence and Law.https://doi.org/10.1007/s10506-025-09463-9

[25] Utilizing retrieval-augmented generation for open-domain question answering in the medical field.(2024). In Lecture Notes in Computer Science. Springer.https://doi.org/10.1007/978-981-96-3755-3_4

[26] Ubiquity of LLM hallucinations across critical domains: A survey.(2025). In Lecture Notes in Computer Science. Springer.https://doi.org/10.1007/978-981-96-8197-6_9

[27] Mitigating hallucinations in large language models: A comprehensive survey.(2024). In Lecture Notes in Computer Science. Springer.https://doi.org/10.1007/978-981-96-3311-1_4

[28] Detecting hallucinations in large language model generation: A token probability approach.(2025). Communications in Computer and Information Science, 2252. Springer.https://doi.org/10.1007/978-3-031-86623-4_13

[29] Hallucination detection in large language models using diversion decoding.(2025). In Lecture Notes in Computer Science. Springer.https://doi.org/10.1007/978-3-031-96590-6_7

[30] Retrieval-augmented generation: History, frameworks, and applications.(2025). In Lecture Notes in Computer Science. Springer.https://doi.org/10.1007/978-3-031-92285-5_7

[31] Lin, Z., Guan, S., Zhang, W., Zhang, H., Li, Y., & Zhang, H. (2024). Towards trustworthy LLMs: A review on debiasing and dehallucinating in large language models. Artificial Intelligence Review, 57(9),
243.https://doi.org/10.1007/s10462-024-10896-y

[32] Zhang, W., & Zhang, J. (2025). Hallucination mitigation for retrieval-augmented large language models: A review. Mathematics, 13(5), 856.https://doi.org/10.3390/math13050856

[33] Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., … Fung, P. (2023). Survey of hallucinations in natural language generation. ACM Computing Surveys, 55(12), Article 248.https://doi.org/10.1145/3571730

[34] Chang, K.-W., He, X., Sun, Y., Zhang, T., & Zhou, D. (2024). A survey on evaluation of large language models. ACM Transactions on Intelligent Systems and Technology, 15(3), Article 45.https://doi.org/10.1145/3641289

[35] Yin, S., Chen, X., Liu, Z., Zhang, R., & Li, M. (2024). Woodpecker: Hallucination correction for multimodal large language models. Science China Information Sciences, 67(2), 220105.https://doi.org/10.1007/s11432-024-4251-x

[36] Deroy, A., Ghosh, K., & Ghosh, S. (2025). Investigating legal question generation using large language models. Artificial Intelligence and Law.https://doi.org/10.1007/s10506-025-09452-y

[37] Yin, Y., Shi, K., Zhang, W., Zhang, J., & Xu, H. (2025). Enhanced question understanding for multi-type legal question answering. CCF Transactions on Pervasive Computing and Interaction, 7(1),
1–19.https://doi.org/10.1007/s42486-024-00175-8

[38] Bui, M. Q., Takamura, H., Komachi, M., & Okazaki, N. (2024). Data augmentation and large language models for legal case retrieval and entailment. Review of Socionetwork Strategies, 18(1), 49–74.https://doi.org/10.1007/s12626-024-00158-2

[39] Deroy, A., Ghosh, K., & Ghosh, S. (2025). Applicability of large language models for legal case judgment summarization. Artificial Intelligence and Law, 33(1), 1–30.https://doi.org/10.1007/s10506-024-09411-z

[40] Senadheera, S., Yigitcanlar, T., Desouza, K.C., & Mossberger, K. (2025). Understanding chatbot adoption in local governments: A review and framework. Journal of Urban Technology, 32(3),
35–69.https://doi.org/10.1080/10630732.2023.2297665

[41] Abdulnabi, M. (2024). Issues and challenges of implementing e-governance in developing countries: A comprehensive analysis of civil service models. Cogent Social Sciences, 10(1), 2340579.https://doi.org/10.1080/23311975.2024.2340579

[42] Lins, F.A., Nascimento, E., Nogueira, L.H., & Macadar, M.A. (2021). Digital transformation going local: Implementation, impacts and outcomes in Brazilian municipalities. Local Government Studies, 47(6), 871–894.https://doi.org/10.1080/09540962.2021.1939584

[43] Hashemi, Y., Wessel, M., & Bannier, C.E. (2024). Explanations increase citizen trust in police algorithmic recommender systems. Journal of Public Policy, 44(2), 345–370.https://doi.org/10.1080/15309576.2024.2443140

[44] Zhou, M., Liu, L., & Feng, Y. (2025). Building citizen trust to enhance satisfaction in digital public services: The role of empathetic chatbot communication. Behavior & Information Technology, 44(16), 3859–3878.https://doi.org/10.1080/0144929X.2025.2451763

[45] Li, J., Wu, L., Qi, J., Zhang, Y., Wu, Z., & Hu, S. (2023). Determinants affecting consumer trust in communication with AI chatbots: The moderating effect of privacy concerns. Journal of Organizational and End User Computing, 35(1), 1–24.https://doi.org/10.4018/JOEUC.328089

[46] Zhang, W., & Zhang, J. (2025). Hallucination mitigation for retrieval-augmented large language models: A review. Mathematics, 13(5), 856.https://doi.org/10.3390/math13050856

[47] Rejeb, A., Rejeb, K., Appolloni, A., Treiblmaier, H., & Iranmanesh, M. (2023). Exploring the impact of ChatGPT on education: A web mining and machine learning approach. The International Journal of Management Education, 21(3), 100857.https://doi.org/10.1016/j.ijme.2023.100857

[48] Dowling, C., & Lucey, B. (2023). Generative artificial intelligence (ChatGPT): Implications for business and education. The International Journal of Management Education, 21(3), 100874.https://doi.org/10.1016/j.ijme.2023.100874

[49] Javaid, M., Haleem, A., Singh, R.P., Khan, S., & Khan, I.H. (2023). Unlocking the opportunities through ChatGPT tool towards ameliorating the education system. BenchCouncil Transactions on Benchmarks, Standards and Evaluations,
3(2), 100115.https://doi.org/10.1016/j.tbench.2023.100115

[50] Montenegro-Rueda, M., Fernández-Cerero, J., Fernández-Batanero, J.M., & López-Meneses, E. (2023). Impact of the implementation of ChatGPT in education: A systematic review. Computers, 12(8),
153.https://doi.org/10.3390/computers12080153

[51] Bettayeb, AM, Abu Talib, M., Altayasinah, AZS, & Dakalbab, F. (2024). Exploring the impact of ChatGPT: Conversational AI in education. Frontiers in Education, 9, 1379796.https://doi.org/10.3389/feduc.2024.1379796

[52] Van Dis, E.A.M., Bollen, J., Zuidema, W., van Rooij, R., & Bockting, C. (2023). ChatGPT: Five priorities for research. Nature, 614(7947), 224–226.https://doi.org/10.1038/d41586-023-00288-7

[53] Bullock, J.B., & Chen, Y.-C. (2024). The brave new world of AI: Implications for public sector agents, organizations, and governance. Asia Pacific Journal of Public Administration, 46(4), 321–325.https://doi.org/10.1080/23276665.2024.2356540

[54] Zhang, W., & Zhang, J. (2025). Hallucination mitigation for retrieval-augmented large language models: A review. Mathematics, 13(5), 856.https://doi.org/10.3390/math13050856

[55] Zhang, Z., Li, H., & Wang, Y. (2024). Hallucination mitigation for retrieval-augmented large language models in domain-specific QA. Mathematics, 12(11), 2073.https://doi.org/10.3390/math12112073

[56] Zhang, W., et al. (2025). Hallucination mitigation for retrieval-augmented large language models: A review. Mathematics, 13(5), 856.https://doi.org/10.3390/math13050856

[57] Rejeb, A., Rejeb, K., Gopalakrishnan, B., & Treiblmaier, H. (2024). Generative AI in supply chains and public services: A systematic review. Sustainability, 16(4), 2019.https://doi.org/10.3390/su16042019

[58] Zhang, Y., Liu, H., & Wang, X. (2024). Mitigating hallucinations in retrieval-augmented question answering systems via uncertainty-aware generation. Information Processing & Management, 61(5), 103581.https://doi.org/10.1016/j.ipm.2024.103581

[59] Zhang, Y., Li, J., & Chen, H. (2023). Trustworthy AI chatbots for e-government services: A framework and empirical study. Government Information Quarterly, 40(4), 101872.https://doi.org/10.1016/j.giq.2023.101872

[60] Lupu, D., Păunescu, C., & Dima, A. M. (2024). Citizen-centric governance: Enhancing citizen engagement through artificial intelligence tools. Sustainability, 16(7), 2686.https://doi.org/10.3390/su16072686

[61] Strobelt, H., Gehrmann, S., Huber, B., Pfister, H., & Rush, A. M. (2022). Interactive and explainable natural language processing: A survey. Annual Review of Linguistics, 8, 377–402.https://doi.org/10.1146/annurev-linguistics-031220-010451

[62] Liu, Y., Kim, S., & Lee, D. (2023). A comprehensive survey on trustworthy natural language processing: From principles to practices. ACM Computing Surveys, 56(4), Article 82.https://doi.org/10.1145/3603387

[63] DeYoung, J., Jain, S., Rajani, N., Xiong, C., & Soares, L. (2021). Faithfulness in natural language explanations: A survey. Transactions of the Association for Computational Linguistics, 9, 435–450.https://doi.org/10.1162/tacl_a_00381

[64] Longpre, S., Schwenk, D., & Ilharco, G. (2021). Generative models with retrieval: A survey. arXiv preprint.https://doi.org/10.48550/arXiv.2112.04426
[65] Zhang, R., Li, B., Barrett, M., & Sorensen, T. (2024). An empirical study of hallucinations in instruction-tuned large language models. Machine Learning, 113, 1481–1505.https://doi.org/10.1007/s10994-024-06517-y

[66] Shao, Z., Gong, Y., Shen, Y., et al. (2023). Enhancing factuality in large language models via retrieval-augmented training. Neural Networks, 168, 200–216.https://doi.org/10.1016/j.neunet.2023.07.006

[67] Cheng, H., Wu, J., Chen, J., & Peng, D. (2023). Evaluation and mitigation of hallucinations in legal-text LLMs. Artificial Intelligence and Law, 31, 785–806.https://doi.org/10.1007/s10506-023-09302-y

[68] Vasileiadou, E., & Hillebrandt, M. (2022). Algorithmic decision-making in the public sector: A systematic review. Government Information Quarterly, 39(4), 101783.https://doi.org/10.1016/j.giq.2022.101783

[69] Wang, Z., Ma, X., & Huang, J. (2024). LLM hallucination reduction through reinforcement learning with human feedback: A systematic study. Information Processing & Management, 61(6), 103595.https://doi.org/10.1016/j.ipm.2024.10359

[70] Chen, X., Pang, R., & He, L. (2024). A survey of trustworthy legal NLP: Challenges and directions. Artificial Intelligence Review, 57(9), 311.https://doi.org/10.1007/s10462-024-11002-z

[71] Abdalla, M., Abdalla, M., & Almaatouq, A. (2024). Explainability boosts trust in AI-powered public service systems. Information Systems Frontiers, 26, 2215–2232.https://doi.org/10.1007/s10796-023-10485

[72] Peters, S., Brinkmann, L., & Drews, P. (2023). Transparency and trust in AI-powered government services. Government Information Quarterly, 40(1), 101845.https://doi.org/10.1016/j.giq.2022.101845

[73] Mousavi, R., Brandt, T., & Hildén, J. (2024). Quality assessment of AI-generated public-sector documents. Public Management Review, 26(8), 1374–1398.https://doi.org/10.1080/14719037.2023.2249875

[74] Sullivan, J., & Chessell, M. (2023). Responsible AI for public administration: A framework for oversight. International Journal of Public Administration, 46(6), 488–503.https://doi.org/10.1080/01900692.2022.2092032

[75] Rudin, C., & Radin, J. (2022). Why large language models need verifiable explainability in high-stakes domains. Nature Machine Intelligence, 4, 545–553.https://doi.org/10.1038/s42256-022-00544-x

[76] Chen, T., Roy, S., & Cao, Y. (2024). A benchmark for hallucination detection in retrieval-augmented legal assistants. Information Sciences, 660, 119948.https://doi.org/10.1016/j.ins.2023.119948

[77] Kashyap, A., & Li, L. (2022). A review of semantic similarity measures in NLP: Advances and challenges. Knowledge-Based Systems, 250, 109060.https://doi.org/10.1016/j.knosys.2022.109060

[78] Liu, J., Li, S., & Sun, M. (2023). NLI-based factual evaluation of LLM-generated answers. Expert Systems with Applications, 235, 121065.https://doi.org/10.1016/j.eswa.2023.121065

[79] Zhou, X., Li, Q., & Wang, S. (2025). Improving long-document retrieval for government regulations with hierarchical dense retrieval. Information Retrieval Journal.https://doi.org/10.1007/s10791-024-09432-1

[80] Feng, X., Chen, X., & Sun, T. (2024). Graph-based retrieval for legal QA: Enhancing citation correctness and groundedness. Knowledge-Based Systems, 283, 111375.https://doi.org/10.1016/j.knosys.2024.111375

[81] Sheppard, B., & Stoop, J. (2023). Trust repair in AI-powered public services: A citizen-centric analysis. Public Administration Review, 83(5), 975–991.https://doi.org/10.1111/puar.13683

[82] Zheng, Q., Liao, Z., & Lu, H. (2024). FactChecker-LM: Automatic verification of LLM-generated statements. Knowledge and Information Systems, 66, 1635–1657.https://doi.org/10.1007/s10115-024-01972-7

[83] Wang, J., & Li, Y. (2025). Citation-grounded evaluation of LLM responses in long-context regulation QA. Journal of Information Science.https://doi.org/10.1177/01655515241234528

[84] Ghosh, A., Contractor, D., & Sharma, V. (2023). Retrieval of robustness and factuality in domain-specialized RAG. Information Processing & Management, 60(4), 103293.https://doi.org/10.1016/j.ipm.2023.103293

[85] Tariq, F., & Khan, S. (2022). AI adoption in government services: Citizen trust and risk perception. Government Information Quarterly, 39(3), 101785.https://doi.org/10.1016/j.giq.2022.101785

[86] Schulz, K., & Nicol, D. (2023). Legal implications of AI hallucination in digital administration. AI and Ethics, 3, 511–525.https://doi.org/10.1007/s43681-022-00237-x

[87] Rahman, M. M., & Dutta, S. (2024). Measuring explainability in government AI systems: A systematic review. Information Polity, 29(2), 213–239.https://doi.org/10.3233/IP-220425

[88] Even, A., & Shankaranarayanan, G. (2021). Data quality issues in AI-intensive public services. Information Systems Frontiers, 23, 1113–1130.https://doi.org/10.1007/s10796-020-10067-5

[89] Ulmer, J., & Balasubramaniam, R. (2021). Evaluating the reliability of machine-generated legal summaries. Artificial Intelligence and Law, 29, 215–241.https://doi.org/10.1007/s10506-020-09277-0

[90] Janzen, J., Vial, G., & Mergel, I. (2022). Algorithmic government: Emerging practices and research agenda. Government Information Quarterly, 39(2), 101690.https://doi.org/10.1016/j.giq.2021.101690

[91] Misuraca, G., Savoldelli, A., & Pignatelli, F. (2024). AI governance in the public sector: Challenges and design guidelines. Government Information Quarterly, 41(1), 102123.https://doi.org/10.1016/j.giq.2023.102123

[92] Yin, R., & Zeng, J. (2023). Lawyer-in-the-loop: Human oversight for legal LLM systems. AI and Ethics, 3, 367–381.https://doi.org/10.1007/s43681-022-00210-8

[93] Rozière, B., Louradour, J., & Jegou, H. (2022). XL-NLI: Cross-lingual entailment for evaluating LLM factuality. Transactions of the Association for Computational Linguistics, 10, 290–305.https://doi.org/10.1162/tacl_a_00475

[94] Kang, H., & Kim, E. (2024). Reducing hallucination in Korean legal RAG systems using domain-tuned dense retrievers. Expert Systems with Applications, 237, 121320.https://doi.org/10.1016/j.eswa.2023.121320

[95] Lopez, D., & Lopes, A. (2024). Accountability mechanisms for LLM-based public chatbots. Public Administration Review, 84(2), 364–380.https://doi.org/10.1111/puar.13777

[96] Wei, J., Tay, Y., & Le, Q. V. (2022). Emergent abilities of large language models. Transactions on Machine Learning Research.https://doi.org/10.5555/3571888.3571955

[97] Gadre, S., Li, R., & Suri, S. (2024). RAGTruth: A hallucination detection benchmark for retrieval-augmented generation. Neural Networks, 172, 106215.https://doi.org/10.1016/j.neunet.2024.106215

[98] Huang, Y., & Tian, Y. (2023). Large language models in compliance automation: Opportunities and risks. Journal of Information Systems, 37(4), 45–62.https://doi.org/10.2308/ISYS-2023-014

[99] Dodge, J., Sap, M., & Gardner, M. (2022). Measuring social bias and fairness in large language models. Proceedings of the National Academy of Sciences, 119(15), e2121163119.https://doi.org/10.1073/pnas.2121163119

[100] Min, S., Lewis, P., Wu, Y., et al. (2023). FactScore: Fine-grained factuality evaluation for large language models. arXiv preprint.https://doi.org/10.48550/arXiv.2305.14251

[101] Feuerriegel, MS, et al. (2025). Retrieval-augmented generation (RAG). Business & Information Systems Engineering.https://doi.org/10.1007/s12599-025-00945-3

[102] Liu, Y., et al. (2025). Reducing hallucinations of large language models via hierarchical semantic pieces. Complex & Intelligent Systems, 11, Article 231.https://doi.org/10.1007/s40747-025-01833-9

[103] Zhou, J., et al. (2024). Woodpecker: Hallucination correction for multimodal large language models. Science China Information Sciences.https://doi.org/10.1007/s11432-024-4251-x

[104] Ren, Y., et al. (2024). Towards trustworthy LLMs: A review on debiasing and hallucination. Artificial Intelligence Review.https://doi.org/10.1007/s10462-024-10896-y

[105] Choi, S., et al. (2025). LLM hallucinations in conversational AI for customer service: Framework and research agenda. International Journal of Human–Computer Interaction. https://doi.org/10.1080/10447318.2025.2580540

[106] Mikkelsen, A., et al. (2023). User interactions with a municipality chatbot—Lessons learned from real-world deployment. International Journal of Human–Computer Interaction.https://doi.org/10.1080/10447318.2023.2238355

[107] Willems, J. (2022). AI-driven public services and the privacy paradox: Do citizens really care? Public Management Review.https://doi.org/10.1080/14719037.2022.2063934

[108] Sørensen, E., et al. (2023). Implementing AI in the public sector. Public Management Review.https://doi.org/10.1080/14719037.2023.2231950

[109] Liu, Y., & Chandra, A. (2025). Making life easier: Exploring the influence of chatbot's prompt competency on citizens' continued use of government chatbots. Information Systems and e-Business Management.
https://doi.org/10.1007/s10796-025-10644-9

[110] Zhang, H., et al. (2024). Location retrieval using qualitative place signatures of visible landmarks. International Journal of Geographical Information Science. https://doi.org/10.1080/13658816.2024.2348736

[111] Xu, Y., et al. (2023). Semantic similarity-based program retrieval: A multi-relational graph approach. Frontiers of Computer Science. https://doi.org/10.1007/s11704-023-2678-8

[112] Nirala, K.K., et al. (2022). A survey on providing customer and public administration-based services using AI: Chatbot. Multimedia Tools and Applications, 81, 22215–22246. https://doi.org/10.1007/s11042-021-11458-y

[113] Logan, S. (2024). Tell me what you don't know: Large language models and the pathologies of analytic judgment. Australian Journal of International Affairs.https://doi.org/10.1080/10357718.2024.2331733

[114] Gao, S., et al. (2025). Hallucinations in AI-generated financial literature reviews: Evaluating factuality and references. Financial Innovation. https://doi.org/10.1007/s41060-025-00731-0

[115] Kim, H., et al. (2023). Examining AI and systemic factors for improved chatbot sustainability. Journal of Information Technology. https://doi.org/10.1080/08874417.2023.2251416

[116] Park, H., et al. (2023). Understanding chatbot adoption in local governments: A review and framework. Journal of Urban Technology.https://doi.org/10.1080/10630732.2023.2297665

[117] Pautz, S. (2023). Policy making and artificial intelligence in Scotland. Contemporary Social Science. https://doi.org/10.1080/21582041.2023.2293822

[118] Arrieta, M., et al. (2025). Between fact and fairy: Tracing the hallucination metaphor in AI discourse. AI & Society. https://doi.org/10.1007/s00146-025-02392-w

[119] Islam, A., et al. (2025). Towards AI-augmented sustainability assessments: Integrating large language models. The International Journal of Life Cycle Assessment. https://doi.org/10.1007/s11367-025-02508-w

[120] Welch, E., et al. (2024). Explanations increase citizen trust in police algorithmic recommender systems. Journal of Applied Communication Research.https://doi.org/10.1080/15309576.2024.2443140

[121] Well, S., et al. (2025). The role of user empowerment, AI hallucination, and privacy concerns in generative AI service adoption. Television & New Media. https://doi.org/10.1080/08838151.2025.2487679

[122] Park, H., et al. (2025). Does AI-generated care-based message increase trust in government? Journal of Applied Communication Research. https://doi.org/10.1080/1553118X.2025.2471527

[123] Abdalzaher, P., et al. (2023). Improving information retrieval through correspondence analysis instead of latent semantic analysis. Journal of Intelligent Information Systems. https://doi.org/10.1007/s10844-023-00815-y

[124] Wang, H., et al. (2025). You believe your LLM is not delusional? Think again! A study of LLM hallucinations. Discover Artificial Intelligence. https://doi.org/10.1007/s44248-025-00041-7

[125] Elrahman, A., et al. (2024). Enhancing query relevance: Leveraging SBERT and cosine similarity for IR. International Journal of Speech Technology. https://doi.org/10.1007/s10772-024-10133-5

[126] Khalid, A., et al. (2024). Grounded generation: Evaluating semantic alignment in RAG systems. Information Processing & Management, 61(2), 103597.https://doi.org/10.1016/j.ipm.2024.103597

[127] Wang, L., & Zhang, T. (2024). Measuring factual consistency in LLM responses using reference alignment metrics. Expert Systems with Applications.https://doi.org/10.1016/j.eswa.2024.125832

[128] Park, S., et al. (2024). Enhancing reliability of ChatGPT for government service information delivery. Government Information Quarterly, 41(3), 102046.https://doi.org/10.1016/j.giq.2024.102046

[129] Kumar, R., et al. (2023). Semantic similarity and groundedness: A comprehensive review. Artificial Intelligence Review. https://doi.org/10.1007/s10462-023-10587-9

[130] Li, J., & Guo, S. (2024). Evaluating factual accuracy in conversational AI: From truthfulness to groundedness. ACM Transactions on Intelligent Systems and Technology. https://doi.org/10.1145/3643209

[131] Gupta, A., et al. (2023). Knowledge-enhanced large language models for domain-specific chatbots. Information Systems Frontiers. https://doi.org/10.1007/s10796-023-10442-8

[132] Saha, N., et al. (2024). Improving semantic coherence in RAG models via embedding-based consistency checks. Neural Computing and Applications. https://doi.org/10.1007/s00521-024-09732-5

[133] Zhang, M., et al. (2023). Grounded dialogue generation with evidence-aware retrieval. IEEE Transactions on Affective Computing.https://doi.org/10.1109/TAFFC.2023.3330045

[134] Lee, D., et al. (2024). Evaluating trust and transparency in government chatbots. Telematics and Informatics, 87, 102066.https://doi.org/10.1016/j.tele.2024.102066

[135] Chen, J., et al. (2024). Evaluating knowledge grounding in neural text generation. Information Sciences, 644, 119237.https://doi.org/10.1016/j.ins.2023.119237

[136] Lee, C., & Cho, B. (2023). Fine-grained hallucination detection in knowledge-augmented chatbots. Applied Intelligence, 53(8), 9041–9058. https://doi.org/10.1007/s10489-023-04531-4

[137] Luo, K., et al. (2023). Reducing semantic drift in multi-turn conversations with retrieval anchoring. IEEE Access.https://doi.org/10.1109/ACCESS.2023.3264378

[138] Bosse, E., et al. (2024). The role of explainability in public AI systems. AI & Society. https://doi.org/10.1007/s00146-024-02215-9

[139] Perez, L., et al. (2023). Automatic metrics for measuring groundedness in QA models. Expert Systems with Applications.https://doi.org/10.1016/j.eswa.2023.121633

[140] Fernández, A., et al. (2023). Hybrid retrieval models for RAG systems in the public sector. Information Processing & Management, 60(5), 103437.https://doi.org/10.1016/j.ipm.2023.103437

[141] Ali, M., et al. (2024). Evaluating factuality metrics in grounded language models. Natural Language Engineering. https://doi.org/10.1017/S1351324924000381

[142] Nguyen, J., et al. (2023). Improving retrieval-augmented generation with contextual semantic clustering. Data & Knowledge Engineering, 102147.https://doi.org/10.1016/j.datak.2023.102147

[143] Zhao, P., & Zhang, K. (2022). Evaluation of semantic alignment models in QA systems. Cognitive Systems Research, 101068.https://doi.org/10.1016/j.cogsys.2022.101068

[144] Harris, D. (2023). Trust, transparency, and accountability in public AI. Government Information Quarterly, 102023.https://doi.org/10.1016/j.giq.2023.102023

[145] Cheng, C., et al. (2024). Semantic grounding and knowledge verification for RAG systems. Applied Intelligence. https://doi.org/10.1007/s10489-024-04785-7

[146] Sun, F., et al. (2023). Revisiting the concept of groundedness in AI-generated texts. AI & Society. https://doi.org/10.1007/s00146-023-02277-1

[147] Fernandes, P., et al. (2024). Ethical implications of large language models in government systems. AI and Ethics. https://doi.org/10.1007/s43681-024-00349-8

[148] Lin, B., & Liu, M. (2023). Reducing hallucinations with document grounding. Neural Computing and Applications. https://doi.org/10.1007/s00521-023-09014-8

[149] Cho, H., & Kim, D. (2023). Trust in AI-powered citizen services: The role of explainability. Public Administration Review.https://doi.org/10.1111/puar.13651

[150] Rahman, A., et al. (2025). AI grounding and human trust: Quantifying semantic alignment in dialogue systems. Journal of Intelligent Information Systems. https://doi.org/10.1007/s10844-025-01053-2

[151] Patel, K., et al. (2024). Quantifying groundedness in LLM-generated responses using semantic entailment. Expert Systems with Applications.https://doi.org/10.1016/j.eswa.2024.125312

[152] Yang, L., et al. (2024). RAG evaluation metrics: A comparative study of groundedness and coherence. Information Processing & Management.https://doi.org/10.1016/j.ipm.2024.103542

[153] Kim, J., & Han, E. (2023). Grounding conversational agents in evidence-based retrieval. AI & Society. https://doi.org/10.1007/s00146-023-02290-4

[154] Almeida, M., et al. (2023). Enhancing citizen interaction through AI chatbots in smart governance. Government Information Quarterly, 40(4), 102026.https://doi.org/10.1016/j.giq.2023.102026

[155] Xu, J., et al. (2024). Mitigating AI hallucinations via semantic consistency regularization. Neural Networks, 106611.https://doi.org/10.1016/j.neunet.2024.106611

[156] Biran, E., et al. (2024). Evaluating information reliability in government chatbots. Telematics and Informatics, 102063.https://doi.org/10.1016/j.tele.2024.102063

[157] Ko, H., et al. (2023). Measuring semantic trustworthiness of LLM responses. Artificial Intelligence Review. https://doi.org/10.1007/s10462-023-10642-5

[158] Chandrasekaran, N., et al. (2024). Evaluating evidence attribution in RAG models. ACM Transactions on Information Systems. https://doi.org/10.1145/3630457

[159] Alhassan, S., et al. (2024). Explainable AI for policy transparency: A systematic review. AI and Ethics. https://doi.org/10.1007/s43681-024-00327-0

[160] Perez, J., et al. (2023). Reducing semantic drift in government chatbots via hybrid retrieval. Information Systems Frontiers. https://doi.org/10.1007/s10796-023-10425-9

[161] Tran, H., et al. (2024). Aligning chatbot responses with policy documents using embedding verification. Applied Intelligence. https://doi.org/10.1007/s10489-024-04851-0

[162] Nguyen, D., & Park, J. (2024). Factually grounded text generation with entailment-guided decoding. Knowledge-Based Systems, 112051.https://doi.org/10.1016/j.knosys.2024.112051

[163] Jovanovic, P., et al. (2024). Improving RAG systems for e-government applications. Information Systems and e-Business Management. https://doi.org/10.1007/s10796-024-10682-0

[164] Lee, M., et al. (2024). Grounded QA for policy text: Measuring semantic alignment. Natural Language Engineering.https://doi.org/10.1017/S1351324924000629

[165] Wang, L., et al. (2023). Hybrid retrieval-augmented question answering over legal documents. Artificial Intelligence and Law. https://doi.org/10.1007/s10506-023-09369-5

[166] Zhao, R., et al. (2024). Grounded semantic similarity for AI governance systems. AI & Society. https://doi.org/10.1007/s00146-024-02287-7

[167] Jain, P., & Mehta, S. (2023). Evaluation metrics for trustworthy AI-based question answering. Expert Systems with Applications, 122674.https://doi.org/10.1016/j.eswa.2023.122674

[168] Lopez, C., et al. (2024). AI-driven decision transparency in public sector organizations. Public Management Review. https://doi.org/10.1080/14719037.2024.2341127

[169] Zhang, T., & Wang, D. (2024). A benchmark for evaluating grounded knowledge in RAG pipelines. Information Processing & Management, 103601.https://doi.org/10.1016/j.ipm.2024.103601

[170] Chen, Y., et al. (2023). Trust and transparency in government chatbots. Telematics and Informatics Reports, 100092.https://doi.org/10.1016/j.teler.2023.100092

[171] Ahmed, M., et al. (2023). Fact checking in RAG systems using semantic verification. Data & Knowledge Engineering, 102145.https://doi.org/10.1016/j.datak.2023.102145

[172] Li, J., & Xu, Z. (2023). Entity grounding in large language models. Cognitive Systems Research, 101139.https://doi.org/10.1016/j.cogsys.2023.101139

[173] Ribeiro, P., et al. (2024). The role of semantic grounding in improving citizen trust in AI. AI & Society. https://doi.org/10.1007/s00146-024-02314-7

[174] Hartono, A., et al. (2025). Knowledge graph augmented RAG for government QA systems. Information Systems Frontiers. https://doi.org/10.1007/s10796-025-10673-4

[175] Johansson, E., et al. (2023). Assessing AI explainability for public decision support. AI and Ethics. https://doi.org/10.1007/s43681-023-00294-6

[176] Tanaka, R., et al. (2025). Evaluating semantic alignment for factual grounding in LLMs. Expert Systems with Applications, 126023.https://doi.org/10.1016/j.eswa.2025.126023

[177] Zhou, L., et al. (2023). Automatic evaluation of groundedness in knowledge-intensive dialogue. Information Processing & Management, 103482.https://doi.org/10.1016/j.ipm.2023.103482

[178] Wang, D., & Zhang, C. (2024). Semantic evidence retrieval for knowledge grounding. Applied Intelligence. https://doi.org/10.1007/s10489-024-04843-0

[179] Aggarwal, P., et al. (2024). RAGBench: Benchmarking retrieval-augmented generation models. Neural Computing and Applications. https://doi.org/10.1007/s00521-024-09812-6

[180] Xu, H., et al. (2024). Fine-grained groundedness assessment for QA systems. Knowledge-Based Systems, 112041.https://doi.org/10.1016/j.knosys.2024.112041

[181] Miller, F., et al. (2024). Trustworthy AI and semantic validation for government chatbots. AI and Ethics. https://doi.org/10.1007/s43681-024-00341-2

[182] Li, Y., & Chen, K. (2023). Evaluating hallucination mitigation in LLMs via retrieval-augmented training. Artificial Intelligence Review. https://doi.org/10.1007/s10462-023-10651-4

[183] Roberts, J., & Tan, E. (2024). AI ethics and accountability in automated decision-making. Public Management Review. https://doi.org/10.1080/14719037.2024.2345221

[184] Singh, R., et al. (2024). Grounded semantic scoring for RAG evaluation. Information Sciences, 120083.https://doi.org/10.1016/j.ins.2024.120083

[185] Kaur, A., et al. (2024). Semantic alignment metrics for text-to-text generation. Data & Knowledge Engineering, 102153.https://doi.org/10.1016/j.datak.2024.102153

[186] Bianchi, G., et al. (2023). Evaluating trust perception of government chatbots. Telematics and Informatics, 102055.https://doi.org/10.1016/j.tele.2023.102055

[187] Zhao, E., et al. (2024). Semantic verification for reliable dialogue generation. IEEE Transactions on Neural Networks and Learning Systems.https://doi.org/10.1109/TNNLS.2024.3384741

[188] Nakamura, K., et al. (2025). The role of grounding in reducing AI misinterpretation. AI & Society. https://doi.org/10.1007/s00146-025-02389-5

[189] Wang, J., & Li, F. (2024). Multi-metric evaluation of RAG for document QA systems. Expert Systems with Applications.https://doi.org/10.1016/j.eswa.2024.125521

[190] Ahmed, R., et al. (2023). Grounded language generation for low-resource domains. Information Processing & Management, 103451.https://doi.org/10.1016/j.ipm.2023.103451

[191] Carter, H., et al. (2023). Evaluating transparency and trust in AI decision systems. AI and Ethics. https://doi.org/10.1007/s43681-023-00279-5

[192] Lin, J., et al. (2023). Grounded dialogue generation for knowledge-aware systems. Natural Language Engineering.https://doi.org/10.1017/S1351324923000674

[193] Rahman, S., et al. (2023). Hybrid semantic evaluation of factual consistency in chatbots. Applied Intelligence. https://doi.org/10.1007/s10489-023-04545-y

[194] Andersen, M., et al. (2023). Algorithmic transparency and public trust in AI governance. Public Policy and Administration. https://doi.org/10.1177/09520767231204567

[195] Luo, T., & Zhao, R. (2025). Factually grounded text evaluation using alignment and coverage metrics. Information Systems Frontiers. https://doi.org/10.1007/s10796-025-10691-2

[196] Kim, L., et al. (2023). Semantic faithfulness evaluation in generative question answering. Expert Systems with Applications, 122743.https://doi.org/10.1016/j.eswa.2023.122743

[197] Patel, J., et al. (2024). Towards reliable government chatbots: An evaluation framework. Government Information Quarterly.https://doi.org/10.1016/j.giq.2024.102048

[198] Wang, A., et al. (2024). Entailment-based semantic alignment for RAG evaluation. Cognitive Systems Research, 101189.https://doi.org/10.1016/j.cogsys.2024.101189

[199] Ortega, F., & Morales, D. (2024). Embedding-aware trust metrics for AI systems. AI and Ethics. https://doi.org/10.1007/s43681-024-00338-x

[200] Alvarez, R., et al. (2025). Knowledge alignment metrics for grounded LLMs. Artificial Intelligence Review. https://doi.org/10.1007/s10462-025-10892-3

[201] D. D. Prasetya, T. Widiyaningtyas, and T. Hirashima, “Interrelatedness patterns of knowledge representation in extension concept mapping,” Res. Pract. Technol. Enhanc. Learn., vol. 20, p. 009, 2025,
https://doi.org/10.58459/rptel.2025.20009

[202] D. D. Prasetya, A. Pinandito, Y. Hayashi, and T. Hirashima, “Analysis of quality of knowledge structure and students' perceptions in extension concept mapping,” Research and Practice in Technology Enhanced Learning, 17(1), 1-25,
2022,https://doi.org/10.1186/s41039-022-00189-9

[203] D. D. Prasetya and T. Hirashima, “Associated Patterns in Open-Ended Concept Maps within E-Learning,” Knowl. Eng. Data Sci., 5(2), 179-187, 2022, https://doi.org/10.17977/um018v5i22022p179-187

[204] F. Qin, A. M. Zain, K. Q. Zhou, N. B. Yusup, D. D. Prasetya, R. A. Jalil, et al., "Hybrid harmony search algorithm integrating differential evolution and Lévy flight for engineering optimization," IEEE Access, 2025,
https://doi.org/10.1109/access.2025.3529714