Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 24/08/2026 - **Accepted:** 17/10/2026
Vol: 01.05
DOI: 10.66104/zkmx8c79
Pages: 1-42

Institutional Document Chatbots for User Support A Critical Narrative Review of Retrieval-Augmented Generation Accuracy, Grounding, and Hallucination

Luan Dias Bomfim

Faculdade Presidente Antônio Carlos, Brazil

Corresponding author: Luand2545@gmail.com

Orcid ID: https://orcid.org/0009-0009-8091-5289

Abstract

Institutional chatbots can retrieve regulations, service instructions, and frequently asked questions before generating responses. However, access to documents does not establish that an answer is correct, complete, current, or authorized. This critical narrative review asks whether document-grounded chatbots improve answer accuracy and reduce unsupported information compared with language models without institutional retrieval. Targeted scholarly web searches and primary-source verification were completed on October 15, 2026, with a targeted public-record verification on October 16, 2026. The original 20-source corpus was expanded to 23 sources: 20 research publications, one conference research abstract, one enterprise benchmark preprint, and one official risk-management publication. Evidence was separated into applied information-service studies, benchmark and component studies, evaluation methods, adversarial research, and governance guidance. Published applications support conditional improvements, while benchmarks demonstrate persistent failures in grounding, citation, rejection of unanswerable questions, and use of time-sensitive or sparse private evidence. No common effect size was calculated because comparators, questions, models, and outcome definitions differ. Accuracy, source faithfulness, completeness, citation validity, temporal applicability, and authorized disclosure and verified

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 24/08/2026 - **Accepted:** 17/10/2026
Vol: 01.05
DOI: 10.66104/zkmx8c79
Pages: 1-42

task resolution must be evaluated separately. The review proposes an unvalidated institutional evaluation protocol using matched questions, a fixed model, independent document-based reference answers, blinded assessment, and explicit abstention denominators. The literature supports carefully governed RAG as a means of improving bounded institutional assistance; it does not establish universal superiority or elimination of hallucinations. Local comparative testing, Portuguese-language validation, document lifecycle management, and measurement of total user and staff effort remain necessary before deployment claims can be justified.

Keywords: retrieval-augmented generation; institutional chatbots; information retrieval; grounding; hallucination; user support

1. Introduction

An institutional information service must answer questions whose validity depends on local rules rather than general plausibility. A response about enrollment, service eligibility, document submission, or a deadline may sound reasonable while being unsupported by the institution's current instructions. Retrieval-augmented generation (RAG) offers a technical route for supplying relevant external evidence during generation. Its foundational formulation combined parametric language-model knowledge with retrieved memory and improved knowledge-intensive tasks under the evaluated conditions (Lewis et al., 2020).

The practical question is therefore more demanding than whether a chatbot can locate a document. A useful institutional assistant must retrieve the relevant rule, interpret its conditions, identify its applicability, communicate the answer, and recognize when the available evidence is insufficient. These operations can fail independently. RGB explicitly distinguishes robustness to noise, rejection of unsupported requests, integration of information, and handling of counterfactual material, illustrating why retrieval access alone cannot be treated as successful assistance (Chen et al., 2024).

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 24/08/2026 - **Accepted:** 17/10/2026
Vol: 01.05
DOI: 10.66104/zkmx8c79
Pages: 1-42

Applied studies provide grounds for investigating this approach. A university-specific chatbot study reports better accuracy and reliability than general-purpose model comparators, while a climate-information chatbot evaluates access to a specialist institutional archive using automated measures and subject-matter assessment. Their findings are relevant to document-based assistance, but different settings and evaluation designs prevent a universal estimate of benefit (Biswas et al., 2025; Olawore et al., 2025).

The research question is: Under what conditions do chatbots grounded in institutional documents provide more accurate answers and less unsupported information than comparable systems without institutional retrieval? Subsidiary questions concern failure mechanisms, evaluation validity, appropriate abstention, governance, and transfer to Portuguese-language services. The article contributes a selective evidence synthesis and a prospective evaluation protocol. It reports no newly implemented chatbot, experiment, participant sample, or observed institutional improvement.

2. Literature Review

2.1 Institutional retrieval and generative assistance

For this review, an institutional document chatbot is a conversational interface that retrieves material from a defined organizational collection before generating an answer. The collection may contain policies, manuals, approved web pages, forms, and service instructions. Retrieval can identify passages; generation can combine and explain them. This operational definition excludes a generic chatbot merely prompted to act as an institution and distinguishes a generative assistant from a search interface returning passages without synthesis (Lewis et al., 2020; Rau et al., 2024).

The proposed analytical boundary is the authorized collection at a specified time, rather than all knowledge available on the internet. An answer may correctly

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 24/08/2026 - **Accepted:** 17/10/2026
Vol: 01.05
DOI: 10.66104/zkmx8c79
Pages: 1-42

describe a widespread practice while being wrong for the institution. Conversely, a response can faithfully repeat a retrieved passage that has been superseded. This review therefore treats institutional correctness as agreement with an independently adjudicated, applicable documentary reference, while source faithfulness concerns support in the actual retrieved context. These definitions are proposed for the present synthesis.

2.2 Accuracy faithfulness and completeness

RAGTruth shows that generated answers can include claims unsupported by or contradictory to retrieved content even when retrieval is available. Its nearly 18,000 manually annotated responses support studying hallucination at both response and word levels. The corpus is evidence that grounding failures persist, rather than a direct estimate of the error rate of any institution's chatbot (Niu et al., 2024).

Completeness is a separate concern. A response can contain only supported statements while omitting the exception that determines eligibility. Sub-question coverage research evaluates whether responses address the important facets of a question rather than merely producing relevant prose. This distinction motivates assessing required conditions, exclusions, deadlines, and next steps in institutional answers, instead of interpreting a faithful but partial response as fully correct (Xie et al., 2025).

This review operationalizes unsupported content as an asserted factual claim not entailed by the authorized evidence provided to the generator. Contradicted claims are coded separately. A supported claim can still be institutionally incorrect because its source is obsolete or inapplicable; an unsupported claim can happen to be true. Neither category should be silently substituted for the other. Table 1 defines the assessment dimensions used in the proposed protocol.

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 24/08/2026 - **Accepted:** 17/10/2026
Vol: 01.05
DOI: 10.66104/zkmx8c79
Pages: 1-42

2.3 Citation and institutional authority

A citation makes a claim inspectable only if the identified source exists, the location is accessible to the intended user, and the cited passage supports the claim. ALCE evaluates generated text with citations and demonstrates that citation support remains incomplete in tested systems. Its results justify checking entailment and coverage instead of counting links as proof of grounding (Gao et al., 2023).

Institutional authority requires an additional rule: the cited document must have the appropriate status, effective date, audience, and scope. This is a proposed requirement for local evaluation, not a property automatically supplied by a vector database. A correct explanation of an old policy must fail the current-policy task if it changes the user's action. Source identity, source applicability, and generated claim support should consequently remain separate fields in the reference standard.

3. Review Methods

3.1 Design and actual discovery route

A critical narrative and thematic review was conducted through targeted scholarly web discovery and verification on primary publication pages, completed on October 15, 2026, using the America/Sao_Paulo calendar date. Searches combined RAG, institutional or enterprise documents, university chatbots, customer support, accuracy, hallucination, attribution, unanswerability, and evaluation. ACL Anthology, NeurIPS proceedings, AAAI proceedings, Springer, an institutional repository, arXiv, and NIST were used to retrieve or verify selected publications. These venues were not all independently searched through their internal search interfaces.

The actual route used web search engines and site-restricted queries, followed by publication-page inspection, metadata extraction, and targeted full-text examination. It did not include direct structured searches in Scopus, Web of Science, ACM

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 24/08/2026 - **Accepted:** 17/10/2026
Vol: 01.05
DOI: 10.66104/zkmx8c79
Pages: 1-42

Digital Library, or IEEE Xplore. No complete deduplicated identification and exclusion ledger was constructed; initial hits and screening counts are therefore not reported. This manuscript does not claim PRISMA compliance or systematic exhaustiveness. Appendix A records representative queries actually used and the review's selection rules.

A limited supplementary verification was performed on October 16, 2026, through public publication records and site-restricted web searches targeting ACM and IEEE material. Direct access to the Scopus search page returned an access error, and the IEEE search-results endpoint was inaccessible. No authenticated Scopus, Web of Science, ACM, or IEEE structured search or database export was completed. The supplementary route is therefore reported as public-record verification, not as a structured-database search. It added CRAG, RAGChecker, and a directly relevant university-policy chatbot research abstract. Appendix A documents the queries, access limitations, and selection decisions.

3.2 Eligibility and corpus composition

Eligible sources addressed at least one of the following: a document-grounded user information service; an original RAG benchmark or component experiment; evaluation of grounding, attribution, completeness, or rejection; adversarial manipulation through retrieved content; or official generative-AI governance. Primary research and publisher records were preferred. Wikipedia pages, general blogs, commercial promotional claims without an original research report, duplicate versions, and results lacking a sufficiently clear relation to the review question were not retained as evidence.

The original corpus contained 20 sources: 18 peer-reviewed publications, the WixQA preprint, and one official NIST publication. The supplementary verification added two peer-reviewed NeurIPS 2024 papers and one SIGCSE 2026 Student Research Competition abstract, producing a 23-source inventory. Twenty-two of 23

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 24/08/2026 - **Accepted:** 17/10/2026
Vol: 01.05
DOI: 10.66104/zkmx8c79
Pages: 1-42

sources (95.7%) date from 2023–2026. The primary-role categories are three applied information-service records (including the conference abstract), eleven benchmark/component studies, seven evaluation-method studies, one adversarial study, and one governance publication. Each source is counted once. Only two applied studies were present in the original inventory; the added short abstract supplies limited additional application evidence rather than a full independent validation. Table 2 identifies every retained source and its transfer boundary.

Olawore et al. (2025) was assessed through the publisher and institutional records; detailed experimental figures inaccessible in those records were not reconstructed. A duplicate repository record contained inconsistent date and DOI information. The Springer chapter record was used for the 2025 publication year and complete DOI. WixQA is cited as a preprint because the verified arXiv record did not establish a peer-reviewed replacement. The two 2026 hallucination-detection papers were verified as published ACL/Findings contributions, not treated as preprints (Cohen et al., 2025; Huang et al., 2026; Jia et al., 2026; Olawore et al., 2025).

3.3 Extraction and appraisal

Extraction recorded setting, source collection, evaluated task, comparator where present, outcome definition, principal result, publication status, and transfer limitation. Full texts were examined for selected quantitative claims, particularly the UKCP evaluation and GaRAGe results. Other inclusions were restricted to contributions verifiable in original publication records and abstracts. The review does not claim a complete full-text risk-of-bias appraisal of every retained work.

Appraisal considered comparator fairness, independence of reference answers, use of real versus synthetic questions, human versus automated judging, model and corpus specificity, treatment of unanswerable cases, and external applicability. These are narrative appraisal criteria, not a validated grading instrument.

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 24/08/2026 - **Accepted:** 17/10/2026
Vol: 01.05
DOI: 10.66104/zkmx8c79
Pages: 1-42

Selection and synthesis did not undergo independent duplicate screening, extraction, or inter-rater assessment. These methodological limitations are retained separately from the generative-AI disclosure.

Results were synthesized by mechanism and evidentiary role. Benchmark percentages, automated scores, human ratings, and operational outcomes were not pooled because they estimate different quantities. Official guidance informs governance requirements rather than supplying an empirical effect size. All prospective procedures in Sections 5.1 and 5.2 are author-proposed and have not been implemented or validated in this article.

4. Results

4.1 Applied document information services

Olawore et al. (2025) describes a university chatbot integrating local information with an LLM, orchestration, vector storage, and a conversational interface. The publication record reports improvements over general model comparators. This is directly relevant to institutional information provision, but the accessible record does not support calculating a standardized gain, independently auditing the complete evaluation, or claiming portability to other institutions. The study is retained as bounded application evidence.

Biswas et al. (2025) evaluates a UKCP18 archive chatbot using 250 synthetic question-context-answer triplets, 50 authentic helpdesk question-answer pairs, and a panel of ten subject-matter experts. In its automated evaluation, one pipeline achieved mean coverage of 0.38 compared with 0.26 for the generic GPT-3.5 baseline. These are coverage scores, not percentages of wholly correct answers. The study also identifies incomplete responses and conversational limitations. It supports improvement for its specialist information task without establishing universal service superiority.

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 24/08/2026 - **Accepted:** 17/10/2026
Vol: 01.05
DOI: 10.66104/zkmx8c79
Pages: 1-42

The supplementary verification identified a university-policy chatbot evaluated with institutional documents and interaction logs (Hayrapetyan, 2026). The conference abstract reports automated relevance and groundedness assessment, but does not establish independently adjudicated documentary correctness, a matched non-RAG effect, or verified administrative task completion. It is classified as an applied conference research abstract and interpreted more cautiously than a complete evaluation report.

These applications address the central question more directly than generic benchmarks, yet their endpoints and settings differ. The synthesis therefore supports a conditional claim: curated external evidence can improve a bounded information service. It does not establish that every RAG implementation is better than a strong search interface, the same model with carefully supplied context, or trained human support. The proposed local comparison must include those relevant alternatives.

4.2 Enterprise questions and private grounding

WixQA contributes a customer-support benchmark grounded in a snapshot of the Wix Help Center. It separates 200 expert-written questions, 200 expert-validated simulated pairs, and 6,222 synthetic pairs. This distinction matters because a question generated directly from one article can be easier to ground than an authentic user's incomplete description. The preprint is enterprise-relevant and vendor-authored; it is not an independent randomized evaluation of support outcomes (Cohen et al., 2025).

GaRAGe contains 2,366 questions and more than 35,000 grounding-passage annotations covering private documents and web sources. Under its evaluated configurations, reported maxima include a 60% Relevance-Aware Factuality Score, 31% true-positive rate for deflection, and 58.9% attribution F1. The measures are distinct and must not be read as one accuracy rate. Difficulties with sparse private evidence

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 24/08/2026 - **Accepted:** 17/10/2026
Vol: 01.05
DOI: 10.66104/zkmx8c79
Pages: 1-42

and time-sensitive questions are particularly relevant to institutional collections (Soro-doc et al., 2025).

For this synthesis, those findings motivate separating documentary availability from effective use. A policy may exist in the corpus yet be missed by retrieval, retrieved without its exception, or included but ignored by generation. The appropriate diagnosis should preserve these stages. Otherwise, a team may change the model when the real defect is a missing document, or modify chunking when the reference policy itself is unresolved.

CRAG broadens the evidence on temporal applicability by evaluating factual questions with different levels of dynamism and complexity (Yang et al., 2024). Its results support testing changing facts and difficult evidence combinations; benchmark improvements and residual failures do not establish institutional service outcomes. This evidence motivates current-versus-historical policy items in Section 5.1.5 and update tests in Section 5.2.3.

4.3 Retrieval and context do not guarantee integration

BERGEN shows why reproducible RAG comparison requires explicit datasets, retrieval collections, retrievers, rerankers, generators, and evaluation measures. Improvements attributed to RAG may instead reflect several simultaneous changes. Its component-oriented approach supports controlled local comparisons and ablations, rather than choosing a pipeline from one overall score (Rau et al., 2024).

Long-context experiments show that the location of relevant information can affect model performance, with weaknesses when evidence occurs in the middle of long inputs. That result is tied to the evaluated models and tasks; it should not become an immutable law about every later model. It nevertheless provides a testable reason to inspect evidence ordering and distracting material instead of assuming that a larger context window solves documentary reasoning (Liu et al., 2024).

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 24/08/2026 - **Accepted:** 17/10/2026
Vol: 01.05
DOI: 10.66104/zkmx8c79
Pages: 1-42

RGB evaluates six models across four RAG abilities and identifies difficulties beyond tolerance of noise, particularly rejection, integration, and false information. For institutional evaluation, the implication is to include passages that are topically similar but apply to another program, period, or user group. Semantic similarity is not equivalent to applicability, and a high retrieval score cannot independently certify the answer (Chen et al., 2024).

4.4 Abstention language and conversation

UAEval4RAG introduces six categories of unanswerable queries and evaluates performance across four datasets. No single tested configuration optimally balanced answerable and unanswerable requests across all knowledge bases. A chatbot that refuses everything may avoid invented details while failing its service purpose; a chatbot that always answers can appear helpful while making unsupported claims (Peng et al., 2025).

NoMIRACL evaluates relevant and non-relevant contexts across 18 languages and finds substantial difficulty balancing recognition of useful evidence against hallucination when evidence is absent. Its multilingual scope supports testing language-specific behavior. It does not establish Portuguese institutional performance without local questions, terminology, and reference answers (Thakur et al., 2024).

CORAL evaluates multi-turn retrieval, generation, and citation labeling, including topic shifts. Although its conversation construction differs from an institutional helpdesk, it supports treating dialogue as more than repeated single-turn answering. A locally correct first answer does not validate follow-up answers when the user changes institution, program, date, or eligibility condition (Cheng et al., 2025).

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 24/08/2026 - **Accepted:** 17/10/2026
Vol: 01.05
DOI: 10.66104/zkmx8c79
Pages: 1-42

4.5 Automated evaluation and recent verification methods

RAGAs separates retrieval and generation assessment using reference-free measures, while ARES combines learned judges with human annotations and prediction-powered inference. Both provide tools for scalable evaluation. Their methodological value does not make an automated score an institutional ground truth; a local assessor still needs to verify rules, dates, exceptions, and consequential errors (Es et al., 2024; Saad-Falcon et al., 2024).

Recent work extends evidence-based verification. Jia et al. (2026) evaluates entity alignment against retrieved evidence and counterfactual perturbations. Huang et al. (2026) grounds hallucination-detection reasoning in quoted source segments. These are advances in detecting particular error classes, not demonstrations that institutional assistance has become hallucination-free. Their relevance is methodological: a convincing explanation from a verifier should itself be checked against inspectable evidence.

RAGChecker provides separate diagnostic metrics for retrieval and generation, with meta-evaluation against human judgments (Ru et al., 2024). It strengthens the rationale for component-specific diagnosis in Sections 4.3 and 5.1.5. Its agreement with human ratings in the evaluated tasks does not replace independent local assessment of policy validity, authorization, or task resolution.

4.6 Security and governance

Retrieved material can introduce malicious instructions as well as useful facts. Zhan et al. (2025) bypasses eight evaluated defenses with adaptive indirect prompt-injection attacks. This is adversarial capability evidence under specific experimental conditions, not the incidence of breaches in institutional chatbots. It supports keeping document content separate from system authority and testing adaptive manipulation rather than relying on benign accuracy alone.

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 24/08/2026 - **Accepted:** 17/10/2026
Vol: 01.05
DOI: 10.66104/zkmx8c79
Pages: 1-42

The NIST generative-AI profile supplies voluntary lifecycle risk-management guidance. It is retained to inform documentation, accountability, evaluation, and monitoring, not as proof of RAG effectiveness or certification of the protocol proposed here. Local release decisions should therefore consider information quality and security separately, especially when documents contain restricted institutional or personal information (Autio et al., 2024).

5. Discussion

5.1 Proposed Institutional Evaluation Protocol

5.1.1 Purpose and hypotheses

The following protocol is an original, unvalidated synthesis for a future institution-specific study. Its primary question is whether adding retrieval improves documentary correctness for the same model on the same eligible questions. Its second question is whether the intervention reduces unsupported substantive claims while retaining useful coverage. Neither hypothesis has been tested in this manuscript.

The primary estimand should be the paired difference in the proportion of answerable questions receiving a fully correct, complete, current response without a critical error. A co-primary grounding endpoint can be the proportion of generated substantive claims unsupported by the authorized evidence. Because claim count can change with response length, an answer-level indicator of any unsupported material must also be reported. These proposed definitions prevent a verbose system from hiding errors behind a large denominator.

The protocol should distinguish incremental retrieval benefit from total product benefit. A fixed-model comparison estimates the effect of the specified retrieval package. A comparison between different commercial products estimates a bundled system difference. Neither should be described as the other. Configuration changes,

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 24/08/2026 - **Accepted:** 17/10/2026
Vol: 01.05
DOI: 10.66104/zkmx8c79
Pages: 1-42

prompts, permissions, retrieval depth, context budget, and dates should be recorded before testing.

5.1.2 Comparators and ablations

The minimum experimental comparison uses the same model without institutional retrieval and with the proposed RAG pipeline. Both conditions should receive equivalent task instructions, including permission to state uncertainty. The baseline must not be deliberately encouraged to guess. The model identity, decoding settings, available tools, output-length limits, and question wording should otherwise be matched wherever the intervention permits.

Two additional comparators strengthen practical interpretation: institutional search or an approved FAQ interface, and RAG with independently supplied reference passages. The latter is a diagnostic oracle-context condition, not a deployable competitor. It helps distinguish retrieval bottlenecks from failures in generation under adequate evidence. A human support comparator can be added when the objective concerns service replacement or staff assistance, with equivalent access to current documents.

Ablations may remove reranking, query rewriting, citations, or the rejection mechanism one component at a time. These experiments should answer a predefined question rather than generate a large search over the held-out test set. Development questions should be used to tune the system; the final test set should remain untouched until the pipeline is frozen. Table 3 summarizes the proposed comparison conditions.

5.1.3 Corpus construction and documentary reference standard

The institution should inventory source owners, approval status, effective and expiry dates, audience, version, supersession relations, and access permissions.

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 24/08/2026 - **Accepted:** 17/10/2026
Vol: 01.05
DOI: 10.66104/zkmx8c79
Pages: 1-42

Current and historical documents must be distinguishable. Historical questions can legitimately require a past rule; current questions should use the applicable current version. When documents conflict and no authority resolves the conflict, the reference answer should require clarification or escalation rather than invent a universal precedence rule.

A frozen corpus snapshot should support the main comparison. A separate update set should test policy changes after the freeze. Document parsing should be checked for tables, footnotes, forms, headings, scanned text, and exceptions detached from the main clause. Each reference answer should identify the minimum supporting passages and the essential conditions needed for a successful user action. Generic world knowledge should not silently fill gaps in institution-specific rules.

Reference answers should be prepared independently of the candidate chatbot by at least two qualified document owners or trained assessors in the proposed future study. Their competence should cover the policy domain and the distinction between documentary support and interpretation. They should label answerability, required answer elements, applicable sources, acceptable clarification, and severity of possible errors before seeing system identities.

Disagreements should be recorded before adjudication. Agreement statistics should match the scale: an appropriate kappa for categorical labels, weighted agreement for ordered judgments, or a justified intraclass correlation for continuous ratings. A consensus reference is useful but must not be presented as independent agreement. Low agreement should trigger examination of unclear criteria or conflicting source material, rather than removal of difficult cases merely to improve reported performance. Before testing, prespecify a minimum acceptable inter-rater reliability of 0.75 for the selected chance-corrected or reliability coefficient where applicable, and report its confidence interval and scale. This is a proposed minimum for future

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 24/08/2026 - **Accepted:** 17/10/2026
Vol: 01.05
DOI: 10.66104/zkmx8c79
Pages: 1-42

validation, not an observed result of this review or a universal validation threshold. Values below 0.75 should trigger clarification of the rubric, assessor training, and independent reassessment before final adjudication; consensus must not conceal the original disagreement.

5.1.4 Question sampling and leakage control

The proposed question set should combine de-identified authentic support requests with expert-written edge cases. Synthetic questions may enlarge coverage but should remain separately labeled. Near-duplicate questions, paraphrases, and related policy families must not be split casually between development and testing. Otherwise, familiarity with a policy template can produce optimistic performance unrelated to generalization.

Sampling should include ordinary questions, multi-document conditions, exceptions, obsolete policies, missing information, false premises, ambiguous user identity, restricted documents, and follow-up conversations. Difficulty and potential consequence should be recorded in advance. A question about opening hours and one about an appeal deadline should not receive an identical harm interpretation merely because both produce one factual answer.

For a Brazilian implementation, original Portuguese requests should be included, preserving natural misspellings, acronyms, indirect questions, and local administrative vocabulary. Translating an English benchmark is insufficient to establish performance on local requests. Accessibility testing should separately examine whether users can understand the answer and follow its cited source, including on mobile interfaces and assistive technologies.

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 24/08/2026 - **Accepted:** 17/10/2026
Vol: 01.05
DOI: 10.66104/zkmx8c79
Pages: 1-42

5.1.5 Metrics denominators and abstention

Table 4 proposes a multidimensional outcome set. Documentary accuracy should be assessed at the question level, while faithfulness should additionally be assessed at the atomic-claim level. A response can pass one and fail the other. Citation precision concerns support from the cited passage; citation coverage concerns whether claims needing support have supporting citations. Neither measure alone determines overall institutional correctness.

Answerable and unanswerable questions should have separate denominators. Appropriate refusal is the number of unanswerable questions correctly rejected or escalated divided by all unanswerable questions. False refusal is the number of answerable questions unnecessarily rejected divided by all answerable questions. Coverage is the proportion of eligible questions receiving a substantive answer, reported alongside correctness among answered questions. Clarification should be a distinct output, with task completion assessed after a prespecified number of turns.

Documentary accuracy is defined prospectively as Accuracy_doc = N(fully correct, complete, current responses without a critical error) / N(answerable questions). Unnecessary refusals remain in the answerable denominator and are not counted as correct completed responses.

Unsupported content is reported at two levels: UCR_claim = N(unsupported substantive claims) / N(substantive generated claims); UCR_answer = N(eligible generated answers containing at least one unsupported substantive claim) / N(eligible generated answers). Claim segmentation, answer eligibility, and handling of partial responses must be frozen before testing. Contradicted claims are classified separately using a prespecified coding rule.

Refusal metrics are AR = N(correct refusals or escalations) / N(unanswerable questions), and FR = N(unnecessary refusals) / N(answerable questions). Report AR

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 24/08/2026 - **Accepted:** 17/10/2026
Vol: 01.05
DOI: 10.66104/zkmx8c79
Pages: 1-42

and FR together. Clarification is a separate output and is successful only if the dialogue reaches independently verified correct completion within a prespecified turn limit; record appropriate escalation, unresolved cases, and inappropriate abandonment separately.

For evidence-dependent questions, CompleteRetrieval@k = N(questions with every required evidence unit in the top-k results) / N(questions requiring documentary evidence). Report the multi-document subset separately. Necessary evidence units are identified independently before testing; a relevant passage alone does not establish complete recovery. All percentages must include counts and denominators. Zero-denominator ratios are not estimable, never perfect scores.

A critical-error rate should report responses containing at least one preclassified high-consequence mistake, together with the number of eligible responses and uncertainty intervals. Examples include an incorrect mandatory deadline, invented fee, omitted eligibility exclusion, or unauthorized disclosure. These are proposed local categories rather than a validated universal severity scale. High usefulness or speed must not offset a confidentiality violation in a composite score.

Severity anchors must be specified before outputs are scored. Critical errors are likely to cause legal, financial, eligibility, privacy, or irreversible administrative harm. Major errors cause material rework or delay. Minor errors are incomplete or suboptimal details unlikely to alter entitlement or the final action. Institutions may adapt these anchors prospectively, document examples and disagreement rules, and retain confidentiality violations as separate release failures regardless of aggregate usefulness.

Retrieval diagnostics should identify whether sufficient applicable evidence existed in the corpus, whether it appeared in the retrieved set, and whether generation used it correctly. Recall at k and reciprocal rank are informative only relative to defined

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 24/08/2026 - **Accepted:** 17/10/2026
Vol: 01.05
DOI: 10.66104/zkmx8c79
Pages: 1-42

relevance judgments; multi-document questions also require checking that the necessary combination of sources was retrieved. Error coding should preserve corpus gaps, parsing loss, retrieval failure, interpretation failure, citation failure, outdated-policy use, and permission failure separately.

5.1.6 Statistical analysis and reproducibility

Sample size should be justified using the smallest relevant paired improvement and the tolerated uncertainty around critical errors. This review proposes no universal sample size or passing percentage. A pilot may estimate discordant pairs and variance, but pilot questions used for tuning should not reappear as an untouched confirmatory test. Primary endpoints, exclusions, multiplicity handling, and stopping rules should be specified before final evaluation.

For one prespecified binary outcome per question under each condition, let b count questions correct only with RAG and c count questions correct only without RAG. The paired absolute difference is Delta = (b − c) / N(answerable paired questions). Prespecify a two-sided McNemar test for discordant pairs, using its exact binomial version when discordant counts are small, together with a 95% confidence interval for the paired absolute difference. Do not substitute independent-proportion inference. If questions share policy families, use a policy-family cluster bootstrap for the difference and its interval, or a prespecified marginal model with cluster-robust inference; ordinary McNemar inference alone does not address this dependence. Repeated outputs, users, and conversations require corresponding clustered or hierarchical analysis. Claims within one answer are not independent observations. These are prospective analysis choices, not analyses performed in this review.

Repeated generations should quantify instability under a fixed configuration. The report should distinguish between-question variation from repeated-output variation. If costs require subsampling for human review, the sampling design and any

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 24/08/2026 - **Accepted:** 17/10/2026
Vol: 01.05
DOI: 10.66104/zkmx8c79
Pages: 1-42

weighting must be explicit. Automated judges may assist triage, but their false acceptance and false rejection should be checked against independent human labels on held-out examples. Critical and disputed cases should remain inspectable.

A nonsignificant difference is not evidence of equivalence or noninferiority. Any such claim requires a defensible prespecified margin and adequate precision. Likewise, observing no critical errors in a finite test does not establish zero risk. The report should disclose model and embedding identifiers, prompts, corpus snapshot, document parser, chunking, retrieval and reranking settings, access policy, judge configuration, software versions, and evaluation date.

5.2 Governance and Service Deployment

5.2.1 A proposed supervised workflow

The proposed workflow links document approval, authorized retrieval, evidence assessment, provisional generation, claim verification, user-facing release, escalation, and monitoring. It is not a validated framework, certification, or formally endorsed institutional standard. Table 5 specifies responsibilities and observable records rather than a numerical risk score.

For bounded public information, source-linked answers may be released after the institution's predefined controls. Questions involving unresolved exceptions, competing rules, individualized eligibility, or consequential administrative interpretation should reach the responsible office. The chatbot should identify what is missing and the appropriate contact route. It should not turn a general explanation into a binding institutional decision without explicit authority.

An illustrative question is whether a student may submit a certificate after a deadline. If one document gives a general date and another allows an exception only for a specified program, a useful answer must establish program and period or request

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 24/08/2026 - **Accepted:** 17/10/2026
Vol: 01.05
DOI: 10.66104/zkmx8c79
Pages: 1-42

clarification. Repeating the date without the exception may be faithful to one passage but incomplete for the task. This scenario is hypothetical and contributes no observed performance result.

5.2.2 Permissions and adversarial testing

Public policies and restricted records should have distinct access paths. Entitlements must be checked before protected passages become visible to a generator, external reranker, or other processing service. An authorized employee's ability to read a document does not imply that the assistant may disclose it to another user. This protocol proposes separate tests for retrieval permission and output release.

Adversarial tests should insert malicious instructions into synthetic or approved test documents, including attempts to override policy, fabricate contact details, reveal protected content, or redirect users to an unauthorized destination. Tests should record attacker control, attempt budget, success definition, and the stage at which a violation occurred. Broad attack-success rates should not be interpreted without those denominators. Logs, cached responses, and conversation history should also be included in the threat model.

Provider retention, use of inputs for training, processing location, administrative access, deletion, and incident handling should be documented before restricted material is supplied. These are proposed governance checks informed by lifecycle risk management, not a legal compliance determination. Applicable Brazilian requirements and institution-specific responsibilities require competent assessment; this review does not claim that a generic technical design establishes legal compliance (Autio et al., 2024).

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 24/08/2026 - **Accepted:** 17/10/2026
Vol: 01.05
DOI: 10.66104/zkmx8c79
Pages: 1-42

5.2.3 Deployment outcomes and maintenance

An offline advantage should be followed by silent evaluation, controlled assisted use, and monitored rollout. Silent operation allows staff to assess generated responses without exposing users to unapproved advice. A later comparative service study can measure task resolution, time to verified resolution, unnecessary escalation, repeated contacts, correction work, and staff intervention. Satisfaction should be reported separately from correctness and successful action.

Document maintenance is part of the intervention. A changed deadline should propagate through the source collection, index, cache, and generated-answer memory within an institution-defined objective. Evaluation should measure time to effective update and time to effective revocation. Periodic regression questions should include previous failures, new exceptions, and current high-consequence policies, without converting the regression set into the only evidence of generalization.

Measure policy-update latency and authorization-revocation latency separately wherever practical. For each event, record the authoritative change time and the time at which the source collection, index, caches, and conversational/generated-answer memory consistently enforce the change. A new deadline concerns informational freshness; withdrawn access concerns security. Retest previously cached outputs and restricted passages after revocation, and report both end-to-end and component delays. CRAG motivates temporal stress testing but does not supply local update guarantees (Yang et al., 2024).

Net service effort should include the user's verification time, the employee's correction time, corpus maintenance, monitoring, and incident response. A reduction in first-response latency can coexist with a longer path to a reliable solution. Resource reporting should distinguish model calls, retrieval calls, reranking, storage, and human

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 24/08/2026 - **Accepted:** 17/10/2026
Vol: 01.05
DOI: 10.66104/zkmx8c79
Pages: 1-42

labor. The proposed deployment criterion is better verified service under acceptable risk, rather than more conversations or fewer referrals alone.

5.3 Interpretation of the Findings

The evidence supports a qualified yes: document retrieval can improve answers when relevant institutional information would otherwise be unavailable to the generator. Applied studies support feasibility and bounded gains, while enterprise and grounding benchmarks identify persistent weaknesses. The defensible conclusion is conditional on source quality, retrieval, interpretation, question type, and evaluation design; it is not a universal percentage improvement (Biswas et al., 2025; Cohen et al., 2025; Olawore et al., 2025; Sorodoc et al., 2025).

Most evidence supporting this evaluation architecture is transferable evidence from RAG benchmarks and evaluation research. Direct evidence from sustained institutional user-support deployments remains sparse: the original corpus contained only two applied studies, and the supplementary university-policy record is a short conference abstract rather than a complete field evaluation. The expanded inventory therefore improves breadth without establishing long-term institutional effectiveness or validating the proposed protocol.

The central contribution proposed here is to treat assistance as a chain of independently testable obligations. The institution must know whether the question was answerable, whether the correct evidence was available and retrieved, whether the answer followed it, and whether the user was entitled to receive it. Collapsing that chain into one score makes remediation harder and can hide consequential failures.

For information systems research, the most informative comparison may therefore be modest: one institution, a bounded document collection, matched conditions, independent references, and a clearly defined service outcome. Such a study can establish local incremental value without claiming universal generalization. Conversely,

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 24/08/2026 - **Accepted:** 17/10/2026
Vol: 01.05
DOI: 10.66104/zkmx8c79
Pages: 1-42

an elaborate architecture without independent reference answers cannot demonstrate that users receive more accurate institutional information.

5.4 Avoiding Misleading Success Claims

Three proposed safeguards follow from the synthesis. First, gains should survive a strong baseline rather than a poorly instructed model. Second, reduced unsupported content should not result only from refusing more questions. Third, successful retrieval should not be confused with complete task resolution. These safeguards connect evaluation design to the actual question the institution needs answered.

Future studies should also avoid interpreting model-based faithfulness as an independent audit of policy truth. Automated evaluation methods offer scalable signals, and recent entity- and evidence-based detectors extend verification techniques. Their outputs remain measurements with failure modes, requiring local calibration and independent review for consequential uses (Es et al., 2024; Huang et al., 2026; Jia et al., 2026; Saad-Falcon et al., 2024).

5.5 Limitations and Research Priorities

This review is selective, English-dominant, and concentrated in accessible computational-linguistics publications. Its original corpus includes only two applied information-service studies; the supplementary applied conference abstract adds limited detail, with uneven full-text access, and does not establish institution-wide operational effects. Publication-page summaries support bounded inclusion but limit methodological appraisal. The enterprise benchmark preprint and vendor-linked material require particular caution. No direct structured-database verification could be completed. No meta-analysis, independent duplicate screening or extraction, inter-rater assessment, or empirical validation of the proposed protocol was performed.

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 24/08/2026 - **Accepted:** 17/10/2026
Vol: 01.05
DOI: 10.66104/zkmx8c79
Pages: 1-42

The corpus underrepresents long-term institutional deployments, Portuguese-language support, accessibility outcomes, document images and complex forms, and downstream administrative harm. Model and corpus changes also limit temporal transfer. Future work should prioritize multicenter institutional comparisons, authentic questions, document updates, independent human assessment, and transparent reporting of rejected and unresolved requests. Preregistered paired studies and later field evaluations would provide stronger evidence than further satisfaction-only demonstrations.

6. Conclusion

Institutional documents can make a chatbot better informed, but they do not automatically make it correct, complete, current, or safe. The reviewed literature supports conditional improvements in bounded document-based assistance and demonstrates residual failures in evidence use, citation, abstention, and private or time-sensitive grounding. The question is best answered through a local, matched comparison with independent documentary references and explicit error denominators. The evaluation protocol proposed here separates factual accuracy, faithfulness, completeness, temporal applicability, authorization, and verified service outcomes. It remains an unvalidated research proposal. Adoption claims should depend on demonstrated local benefit and maintained safeguards rather than the presence of RAG alone.

Data Availability

No original participant or institutional-service dataset was collected or analyzed. The evidence inventory and representative search queries are provided in Table 2 and Appendix A. Public source locations are supplied in the reference list. The proposed protocol contains no empirical results.

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 24/08/2026 - **Accepted:** 17/10/2026
Vol: 01.05
DOI: 10.66104/zkmx8c79
Pages: 1-42

Declaration on Generative AI Use

Generative AI was used as an assistive tool, primarily to help identify potentially relevant references and translate text into English. Luan Dias Bomfim conducted the review and remained responsible for source selection, critical appraisal, evidence synthesis, interpretation, writing, and final editorial decisions. AI outputs were not treated as independent scientific evidence or as a substitute for author judgment. The author reviewed the final manuscript and assumes responsibility for its content, references, interpretation, and submitted version. AI assistance did not constitute independent duplicate screening, an additional reviewer, or specialist certification.

Funding

This research received no external funding.

Conflict of Interest

The author declares no conflicts of interest.

Author Contributions

Luan Dias Bomfim was solely responsible for conceptualization, methodology, literature searching and selection, appraisal and synthesis of the evidence, writing the original draft, review and editing, and preparation of the final manuscript. Generative AI was used only as an assistive tool, as disclosed in the dedicated declaration.

Ethical Approval

Not applicable. This literature review reports no new participant recruitment, identifiable personal data, or institutional intervention.

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 24/08/2026 - **Accepted:** 17/10/2026
Vol: 01.05
DOI: 10.66104/zkmx8c79
Pages: 1-42

Table 1 Assessment dimensions and interpretive boundaries

Dimension	Operational question	What a positive result does not establish
Documentary accuracy	Does the response agree with the independently adjudicated applicable rule?	That every generated statement is supported by its displayed citation.
Faithfulness	Is each substantive claim entailed by the authorized supplied evidence?	That the source is current or appropriate to this user.
Completeness	Are all required answer elements and exceptions covered?	That the user understands the response or can complete the task.
Citation validity	Does the cited location exist and support the associated claim?	That uncited claims or omitted conditions are acceptable.
Temporal applicability	Is the rule valid for the requested date and policy version?	That an otherwise applicable answer may be disclosed to this user.
Appropriate abstention	Does the system clarify, refuse, or escalate when evidence is insufficient?	That high refusal is useful on answerable questions.
Verified task outcome	Does independent assessment confirm completion of the intended task?	That the answer complied with all disclosure permissions.
Authorization/disclosure compliance	Were retrieval and disclosure authorized for this user and processing context?	That the user successfully completed the intended task.

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 24/08/2026 - **Accepted:** 17/10/2026
Vol: 01.05
DOI: 10.66104/zkmx8c79
Pages: 1-42

Note. Author-proposed operational definitions informed by the distinct evaluation problems examined in Gao et al. (2023), Niu et al. (2024), Peng et al. (2025), and Xie et al. (2025). This is not a validated scale.

Table 2 Complete retained evidence inventory

Source and primary role	Contribution used in this review	Principal transfer boundary
Lewis et al. (2020); benchmark/component	Foundational retrieval-plus-generation comparison.	Open-domain tasks do not validate institutional policy assistance.
Gao et al. (2023); evaluation method	Claim citation and support assessment.	Citation metrics differ from policy applicability.
Chen et al. (2024); benchmark/component	Four distinct RAG abilities in RGB.	Evaluated models and English/Chinese tasks.
Es et al. (2024); evaluation method	Reference-free component assessment.	Automated signals require local validation.
Liu et al. (2024); benchmark/component	Sensitivity to evidence position in context.	Model- and task-specific experiments.
Niu et al. (2024); benchmark/component	Human annotation of residual hallucinations.	Corpus error rates are not deployment prevalence.
Rau et al. (2024); benchmark/component	Reproducible pipeline comparison.	Configuration benchmarks do not measure service resolution.

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 24/08/2026 - **Accepted:** 17/10/2026
Vol: 01.05
DOI: 10.66104/zkmx8c79
Pages: 1-42

Source and primary role	Contribution used in this review	Principal transfer boundary
Saad-Falcon et al. (2024); evaluation method	Automated judging with human calibration.	Judge performance must transfer to the local domain.
Thakur et al. (2024); benchmark/component	Multilingual relevance and rejection evaluation.	Local Portuguese policies remain untested here.
Biswas et al. (2025); applied service	Specialist archive comparison with expert assessment.	Climate domain and small expert panel.
Cheng et al. (2025); benchmark/component	Multi-turn retrieval, response, and citation tasks.	Constructed conversations differ from institutional requests.
Cohen et al. (2025); benchmark/component	Enterprise support datasets with distinct origins.	Vendor-authored preprint; synthetic and authentic subsets differ.
Olawore et al. (2025); applied service	University-specific RAG application.	Detailed evaluation not fully available in accessed records.
Peng et al. (2025); benchmark/component	Unanswerability and refusal trade-offs.	Knowledge-base-specific performance.
Sorodoc et al. (2025); benchmark/component	Private/web grounding and deflection assessment.	Benchmark measures are not one general accuracy rate.

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 24/08/2026 - **Accepted:** 17/10/2026
Vol: 01.05
DOI: 10.66104/zkmx8c79
Pages: 1-42

Source and primary role	Contribution used in this review	Principal transfer boundary
Xie et al. (2025); evaluation method	Coverage of essential sub-questions.	Preference alignment is distinct from policy correctness.
Zhan et al. (2025); adversarial study	Adaptive indirect prompt-injection testing.	Attack success is not institutional incident frequency.
Huang et al. (2026); evaluation method	Evidence-grounded hallucination detection.	Detector improvement is not chatbot certification.
Jia et al. (2026); evaluation method	Entity-evidence alignment verification.	Entity-level checks do not establish complete service validity.
Autio et al. (2024); governance guidance	Generative-AI lifecycle risk management.	Guidance supplies no RAG effect estimate.
Yang et al. (2024); benchmark/component	CRAG: temporal dynamism and question complexity.	Factual benchmark results do not establish policy validity or service outcomes.
Ru et al. (2024); evaluation method	RAGChecker: separate retrieval/generation diagnostics and human meta-evaluation.	Local policy and authorization judgments still need independent assessment.
Hayrapetyan (2026); applied service abstract	University-policy chatbot assessed using interaction logs and automated measures.	Short conference abstract; no verified task-completion or matched causal effect.

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 24/08/2026 - **Accepted:** 17/10/2026
Vol: 01.05
DOI: 10.66104/zkmx8c79
Pages: 1-42

Note. The original 20-source inventory is retained above; three supplementary records expand it to 23: 3 applied records (including 1 conference abstract), 11 benchmark/component studies, 7 evaluation methods, 1 adversarial study, and 1 governance publication. The original corpus had only 2 applied studies. Cohen et al. is a preprint; Hayrapetyan is a Student Research Competition abstract; Autio et al. is official guidance; the remaining 20 are peer-reviewed research publications. Roles describe use in this review; sources are not double-counted.

Table 3 Proposed comparison conditions

Condition	Purpose	Control needed
Same model without retrieval	Estimate performance without institution-specific retrieval.	Same questions, output budget, uncertainty permission, and decoding settings.
Same model with RAG	Estimate incremental value of the specified retrieval package.	Freeze corpus, retriever, prompt, permissions, and model configuration.
Search or approved FAQ	Compare with an existing information-access alternative.	Equivalent user tasks and independent completion criteria.
Reference passages supplied	Diagnose generation under sufficient evidence.	Expert-selected passages; label as oracle-context diagnostic.
Human support, if relevant	Assess service or assisted-workflow implications.	Comparable evidence access and inclusion of verification time.

Note. These are proposed experimental conditions, not systems implemented in this manuscript. Product-level comparisons and fixed-model retrieval comparisons estimate different effects.

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 24/08/2026 - **Accepted:** 17/10/2026
Vol: 01.05
DOI: 10.66104/zkmx8c79
Pages: 1-42

Table 4 Proposed outcomes and explicit denominators

Outcome	Definition or reporting unit	Safeguard
Fully correct response	Correct, complete, current answers / all answerable questions.	Report abstentions and critical failures separately.
Unsupported-claim rate	Unsupported substantive claims / all substantive generated claims.	Also report answers with any unsupported claim.
Contradiction rate	Claims contradicting applicable evidence / substantive claims.	Separate from merely absent support.
Citation precision	Supporting valid claim-citation links / all assessed claim-citation links.	Inspect the actual cited passage.
Citation coverage	Supported, cited claims / all claims requiring citation.	Do not reward irrelevant link volume.
Appropriate refusal	Correct refusals or escalations / all unanswerable questions.	Independently label answerability first.
False refusal	Unnecessary refusals / all answerable questions.	Include in utility reporting.
Answer coverage	Questions receiving substantive answers / all eligible questions.	Pair with accuracy among answered questions.
Critical-error rate	Responses with at least one critical error / eligible responses.	Prespecify severity and report uncertainty.

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 24/08/2026 - **Accepted:** 17/10/2026
Vol: 01.05
DOI: 10.66104/zkmx8c79
Pages: 1-42

Outcome	Definition or reporting unit	Safeguard
Evidence retrieval	Applicable evidence recovered at k / judged necessary evidence units.	Test multi-source sufficiency, not similarity alone.
Verified resolution time	User and staff time until independently confirmed completion.	Include correction, escalation, and repeated contact.
Policy-update latency	Time until changed policy is correctly enforced across components.	Measure separately from authorization revocation; include source, index, cache, and memory.
Authorization-revocation latency	Time until revoked rights prevent retrieval and disclosure.	Test cached responses and history; report component and end-to-end delays.
Verified task outcome	Independently confirmed correct completions / eligible service tasks.	Separate confidentiality compliance; prespecify unresolved and escalated cases.
Authorization/disclosure compliance	Compliant outputs / outputs assessed for authorization.	Report every confidentiality violation separately from usefulness.
CompleteRetrieval@k	Questions with all required evidence in top-k / evidence-dependent questions.	Report multi-document subset; predefine required evidence units.
Clarification success	Verified completions after clarification / eligible clarification dialogues.	Prespecify turn limit; separate escalation, unresolved outcome, and abandonment.

Note. All measures are prospective proposals. Undefined ratios should be reported as not estimable with their denominators, not assigned perfect scores. Refusal, clarification, partial response, and completed answer are distinct outputs. No universal pass threshold or composite score is proposed.

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 24/08/2026 - **Accepted:** 17/10/2026
Vol: 01.05
DOI: 10.66104/zkmx8c79
Pages: 1-42

Table 5 Proposed institutional responsibilities and records

Stage	Responsible function	Evidence to retain
Source approval	Document owner	Version, validity dates, audience, authority, and supersession.
Authorized ingestion and retrieval	Information systems and security	Parsing checks, corpus snapshot, access rules, and retrieval traces.
Answer preparation	Configured application	Model settings, retrieved passage identifiers, provisional claims, and citations.
Consequential interpretation	Competent service professional	Decision basis, corrections, unresolved issues, and escalation reason.
Release and user support	Service owner	Displayed answer, source access, task outcome, and feedback route.
Monitoring and change control	Joint service and technical owners	Error trends, update tests, incidents, and revalidation after changes.

Note. Author-developed, unvalidated workflow. Model-generated explanations do not become independent institutional evidence. Retention should minimize personal and confidential content while preserving sufficient accountability.

Appendix A Search transparency and selection rules

The following representative queries were used during targeted web discovery on October 15, 2026. They are web queries, not exports from structured bibliographic databases:

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 24/08/2026 - **Accepted:** 17/10/2026
Vol: 01.05
DOI: 10.66104/zkmx8c79
Pages: 1-42

1. site.aclanthology.org retrieval augmented generation hallucination evaluation RAGTruth 2024

2. site.aclanthology.org RAGBench RGB retrieval augmented generation benchmark 2024 2025

3. site.aclanthology.org retrieval augmented generation institutional university chatbot evaluation 2025

4. site.arxiv.org retrieval augmented generation enterprise customer support benchmark 2025 2026

5. retrieval augmented generation institutional chatbot evaluation university 2025 2026 site.aclanthology.org

6. site.aclanthology.org RAGAS ARES ALCE CRAG evaluation retrieval augmented generation

7. site.aclanthology.org 2026 retrieval augmented generation hallucination evaluation institutional

8. "Development and Evaluation of a University Chatbot" "10.1007"

9. "WixQA" "2026" publication

10. site.aclanthology.org "Lost in the Middle"

Queries led to exact-title checks and original publication records. Selection favored direct relevance to institutional document assistance or a clearly specified transferable evaluation problem. Source versions were deduplicated at publication level in the final 20-item inventory, without claiming a full historical screening ledger. No quantitative limits on initial hits, language-filtered exports, or database-wide coverage were recorded. Search-result omission does not establish that a publication is irrelevant. A future systematic extension should preregister its databases, full strings, screening process, and appraisal procedure.

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 24/08/2026 - **Accepted:** 17/10/2026
Vol: 01.05
DOI: 10.66104/zkmx8c79
Pages: 1-42

Supplementary verification on October 16, 2026 used: site.proceedings.neurips.cc 2024 CRAG Comprehensive RAG Benchmark; site.proceedings.neurips.cc 2024 RAGChecker; site.dl.acm.org "institutional" "retrieval-augmented" chatbot evaluation; site.ieeexplore.ieee.org "university chatbot" "RAG"; and "WixQA" "2026" publication. Exact-title records were checked at NeurIPS, ACL Anthology, arXiv, and the official SIGCSE 2026 program. These are web queries, not internal database strings.

Access record: the Scopus basic-search page returned HTTP 403; the IEEE search-results URL was inaccessible. No direct Web of Science or ACM structured search was completed. Site-restricted discovery of ACM/IEEE records does not constitute database searching. A direct structured verification remains a limitation and should be completed through institutional access if available; no hit, exclusion, or export counts are invented.

Selection decisions: CRAG was included for temporal dynamism and complex evidence use; RAGChecker for retrieval/generation diagnosis; Hayrapetyan (2026) for direct institutional-policy relevance, with its abstract-only limits retained. RAGBench was considered but not added because the supplementary expansion prioritized the two requested peer-reviewed benchmarks and a directly institutional record. No new publication-status claim about RAGBench is made. Newly discovered peripheral architecture papers were not used to inflate the corpus.

Metadata checks reconfirmed the original publication records for the 2025 ACL papers, the Springer university-chatbot chapter, the published ACL 2026 records and DOIs for Huang et al. and Jia et al., and the arXiv 2025 record for WixQA. No peer-reviewed replacement was established for WixQA. Publication-record access does not equal a full-text appraisal. Independent subject-specialist and professional language review remain external checks before production.

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 24/08/2026 - **Accepted:** 17/10/2026
Vol: 01.05
DOI: 10.66104/zkmx8c79
Pages: 1-42

Appendix B Illustrative question specification

For each proposed test question, record: question identifier; authentic, expert-written, or synthetic origin; language; policy family; requested date; user role; answerability; authorized corpus snapshot; required passages; essential answer elements; acceptable clarification or escalation; severity categories; independent labels before consensus; and final adjudication rationale. Store candidate outputs with condition identity concealed during assessment.

An illustrative item asks: "Can I submit this certificate after the published deadline?" The reference specification would identify the applicable period, certificate type, program, rule, exceptions, and responsible office. If these facts cannot be established from the question and authorized collection, requesting clarification can be the correct action. No deadline, entitlement, or approval is invented in this example. This item is an original illustration, not an empirical observation or a question validated by an institution.

References

Autio, C., Schwartz, R., Dunietz, J., Jain, S., Stanley, M., Tabassi, E., Hall, P., & Roberts, K. (2024). Artificial intelligence risk management framework: Generative artificial intelligence profile (NIST AI 600-1). National Institute of Standards and Technology. https://doi.org/10.6028/NIST.AI.600-1

Biswas, A., Chahout, H., Pigram, T., Dong, H., Williams, H. T., Fung, F., & Xie, H. (2025). Evaluating retrieval augmented generation to communicate UK climate change information. In Proceedings of the 2nd Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2025) (pp. 126–141). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.climatenlp-1.9

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 24/08/2026 - **Accepted:** 17/10/2026
Vol: 01.05
DOI: 10.66104/zkmx8c79
Pages: 1-42

Chen, J., Lin, H., Han, X., & Sun, L. (2024). Benchmarking large language models in retrieval-augmented generation. Proceedings of the AAAI Conference on Artificial Intelligence, 38(16), 17754–17762. https://doi.org/10.1609/aaai.v38i16.29728

Cheng, Y., Mao, K., Zhao, Z., Dong, G., Qian, H., Wu, Y., Sakai, T., Wen, J.-R., & Dou, Z. (2025). CORAL: Benchmarking multi-turn conversational retrieval-augmented generation. In Findings of the Association for Computational Linguistics: NAACL 2025 (pp. 1308–1330). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.findings-naacl.72

Cohen, D., Burg, L., Pykhnivskyi, S., Gur, H., Kovynov, S., Atzmon, O., & Barkan, G. (2025). WixQA: A multi-dataset benchmark for enterprise retrieval-augmented generation [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2505.08643

Es, S., James, J., Espinosa Anke, L., & Schockaert, S. (2024). RAGAs: Automated evaluation of retrieval augmented generation. In Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations (pp. 150–158). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.eacl-demo.16

Gao, T., Yen, H., Yu, J., & Chen, D. (2023). Enabling large language models to generate text with citations. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (pp. 6465–6488). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.emnlp-main.398

Hayrapetyan, L. (2026). Label-free evaluation of retrieval-augmented generation in a university policy chatbot [Student Research Competition abstract]. SIGCSE TS 2026. https://doi.org/10.1145/3770761.3777161

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 24/08/2026 - **Accepted:** 17/10/2026
Vol: 01.05
DOI: 10.66104/zkmx8c79
Pages: 1-42

Huang, Z., Wen, D., Zhu, Y., Lian, X., Liang, Y., Hao, K., Li, N., Zhang, L., Zhang, Q., Wen, J.-R., Dou, Z., & Wu, F. (2026). RLSeek: Evidence-grounded reasoning for RAG hallucination detection. In Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 32329–32347). Association for Computational Linguistics. https://doi.org/10.18653/v1/2026.acl-long.1492

Jia, R., Fang, Z., Wu, M., Lu, J., & Zhang, Y. (2026). Evidence-aligned entity verification for hallucination detection in retrieval-augmented generation. In Findings of the Association for Computational Linguistics: ACL 2026 (pp. 29538–29554). Association for Computational Linguistics. https://doi.org/10.18653/v1/2026.findings-acl.1477

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. In Advances in Neural Information Processing Systems (Vol. 33, pp. 9459–9474). Curran Associates. https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html

Liu, N. F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., & Liang, P. (2024). Lost in the middle: How language models use long contexts. Transactions of the Association for Computational Linguistics, 12, 157–173. https://doi.org/10.1162/tacl_a_00638

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 24/08/2026 - **Accepted:** 17/10/2026
Vol: 01.05
DOI: 10.66104/zkmx8c79
Pages: 1-42

Niu, C., Wu, Y., Zhu, J., Xu, S., Shum, K., Zhong, R., Song, J., & Zhang, T. (2024). RAGTruth: A hallucination corpus for developing trustworthy retrieval-augmented language models. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 10862–10878). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.acl-long.585

Olawore, K., McTear, M., & Bi, Y. (2025). Development and evaluation of a university chatbot using deep learning: A RAG-based approach. In A. Følstad, S. Papadopoulos, T. Araujo, E. L.-C. Law, E. Luger, S. Hobert, & P. B. Brandtzaeg (Eds.), Chatbots and human-centered AI (pp. 96–111). Springer. https://doi.org/10.1007/978-3-031-88045-2_7

Peng, X., Choubey, P. K., Xiong, C., & Wu, C.-S. (2025). Unanswerability evaluation for retrieval augmented generation. In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 8452–8472). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.acl-long.415

Rau, D., Déjean, H., Chirkova, N., Formal, T., Wang, S., Clinchant, S., & Nikoulina, V. (2024). BERGEN: A benchmarking library for retrieval-augmented generation. In Findings of the Association for Computational Linguistics: EMNLP 2024 (pp. 7640–7663). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.findings-emnlp.449

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 24/08/2026 - **Accepted:** 17/10/2026
Vol: 01.05
DOI: 10.66104/zkmx8c79
Pages: 1-42

Ru, D., Qiu, L., Hu, X., Zhang, T., Shi, P., Chang, S., Cheng, J., Wang, C., Sun, S., Li, H., Zhang, Z., Wang, B., Jiang, J., He, T., Wang, Z., Liu, P., Zhang, Y., & Zhang, Z. (2024). RAGChecker: A fine-grained framework for diagnosing retrieval-augmented generation. Advances in Neural Information Processing Systems, 37. https://doi.org/10.52202/079017-0692

Saad-Falcon, J., Khattab, O., Potts, C., & Zaharia, M. (2024). ARES: An automated evaluation framework for retrieval-augmented generation systems. In Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers) (pp. 338–354). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.naacl-long.20

Sorodoc, I.-T., Ribeiro, L. F. R., Blloshmi, R., Davis, C., & de Gispert, A. (2025). GaRAGe: A benchmark with grounding annotations for RAG evaluation. In Findings of the Association for Computational Linguistics: ACL 2025 (pp. 17030–17049). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.findings-acl.875

Thakur, N., Bonifacio, L., Zhang, C., Ogundepo, O., Kamalloo, E., Alfonso-Hermelo, D., Li, X., Liu, Q., Chen, B., Rezagholizadeh, M., & Lin, J. (2024). "Knowing when you don't know": A multilingual relevance assessment dataset for robust retrieval-augmented generation. In Findings of the Association for Computational Linguistics: EMNLP 2024 (pp. 12508–12526). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.findings-emnlp.730

Xie, K., Laban, P., Choubey, P. K., Xiong, C., & Wu, C.-S. (2025). Do RAG systems cover what matters? Evaluating and optimizing responses with sub-question coverage. In Proceedings of the 2025 Conference of the Nations of the Americas

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 24/08/2026 - **Accepted:** 17/10/2026
Vol: 01.05
DOI: 10.66104/zkmx8c79
Pages: 1-42

Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers) (pp. 5836–5849). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.naacl-long.301

Yang, X., Sun, K., Xin, H., Sun, Y., Bhalla, N., Chen, X., Choudhary, S., Gui, R. D., Jiang, Z. W., Jiang, Z., Kong, L., Moran, B., Wang, J., Xu, Y. E., Yan, A., Yang, C., Yuan, E., Zha, H., Tang, N., ... Dong, X. L. (2024). CRAG—Comprehensive RAG benchmark. Advances in Neural Information Processing Systems, 37. https://doi.org/10.52202/079017-0335

Zhan, Q., Fang, R., Panchal, H. S., & Kang, D. (2025). Adaptive attacks break defenses against indirect prompt injection attacks on LLM agents. In Findings of the Association for Computational Linguistics: NAACL 2025 (pp. 7116–7132). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.findings-naacl.395

CC BY