International Journal of Advanced Artificial Intelligence Research (IJAAIR)
eISSN: 3087-4270
Volume03 Issue11 | Page No: - 01-10

Received: 02 Oct 2026, Revised: 15 Oct 2026, Accepted: 22 Oct2026, Published: 01 Nov 2026

DOI: https://doi.org/10.55640/ijaair-v03i11-01

Configuration-Ready, Content-Blind: Identifying the Missing Content-Health Layer in Enterprise AI Readiness Assessments for HR Service Delivery

Venkata Sai Sushmanth Yamana
Independent Researcher, Dallas, Texas, United States

ABSTRACT: As enterprises adopt agentic and generative artificial intelligence (AI) within service delivery platforms, "AI readiness" has emerged as a formal discipline, typically operationalized through automated assessment tooling that flags configuration gaps and structural data-completeness issues before deployment. This paper examines a representative instance of that tooling — ServiceNow's AI Readiness Evaluation (AIRE) application, used to assess IT Service Management (ITSM), Customer Service Management (CSM), and HR Service Delivery (HRSD) instances prior to enabling agentic AI capabilities — and argues, based on the publicly available documentation and community evidence reviewed in this paper, that its assessment scope appears to exclude a dimension that the Retrieval-Augmented Generation (RAG) research literature treats as foundational: the freshness, ownership, and accuracy of the knowledge content an AI system actually retrieves from at inference time. Drawing on ServiceNow's own Knowledge-Centered Service (KCS) methodology, which independently names "content health" as a distinct knowledge management discipline, this paper proposes a four-dimension content-layer readiness framework — ownership accountability, staleness and expiry discipline, feedback-loop capture, and retrieval-scope correctness — and outlines a concrete validation methodology using a ServiceNow Personal Developer Instance (PDI). This paper is positioned explicitly as a framework and protocol paper: the four-phase validation methodology described in Section 5 has been designed in full but has not yet been executed, and no empirical results are claimed. The intended contribution is a practical, auditable framework and a falsifiable research protocol that could help address the gap between what current AI readiness tooling checks and what the RAG literature indicates actually determines the trustworthiness of an AI-generated answer in HR service delivery.

KEYWORDS: ServiceNow; HR Service Delivery; Retrieval-Augmented Generation; AI Readiness Assessment; Knowledge Management; Content Health; Agentic AI; Enterprise AI Governance

1. Introduction

Enterprise service delivery platforms are moving quickly from ticket automation toward agentic AI: systems that interpret intent, retrieve relevant context, and act with bounded autonomy across IT, HR, and customer service domains. ServiceNow's own roadmap communications describe this shift explicitly, positioning a set of predefined "AI Specialists" across HR Service Delivery, IT operations, and case management [10]. Industry analysis similarly frames 2026 as the year enterprise platforms move from static workflow automation toward orchestrated, outcome-driven agentic processes [10].

This shift raises an obvious governance question: how does an organization know its instance is actually ready for an AI agent to act autonomously within it? To answer that question, ServiceNow shipped a formal AI Readiness Evaluation application — originally released as "Now Assist Readiness Evaluation" (NARE) and renamed AI Readiness Evaluation (AIRE) as of the Australia Patch 5 release [6] — that automatically scans

an instance and classifies findings into three categories: Data Readiness, Configuration Needed, and Now Support Recommended [4], [5]. Each finding is hyperlinked directly to the relevant configuration page, and the tool estimates remediation effort to support project planning [5].

This is a meaningful step forward for the industry: automated, actionable, effort-scored readiness checks did not widely exist a year prior to this writing. However, a closer inspection of what the "Data Readiness" category actually measures reveals a narrower scope than its label implies. ServiceNow's own published data-quality tooling for Predictive Intelligence readiness — available on its public developer GitHub repository — evaluates whether structured fields such as short_description, category, close_notes, and assignment_group are populated, sufficiently diverse, and reasonably descriptive. This is fundamentally a check for structural completeness of transactional records, not a check for whether the content of a knowledge article, once retrieved by an AI agent at inference time, remains factually accurate.

This paper argues that the distinction between structural readiness and content readiness is not a minor omission but a substantive gap with direct consequences for answer trustworthiness — precisely the distinction the RAG research literature treats as decisive for system performance. To state this distinction precisely: AIRE evaluates whether an instance is configured correctly — whether required fields are populated, whether roles and access control lists are correctly assigned, whether workflows are correctly wired — using automated checks against structured platform tables [5]. It does not evaluate whether the content of a published knowledge article is still factually accurate, currently owned, or subject to any review cadence, because none of the objects AIRE inspects (field population, role assignment, workflow configuration) encode that information. This claim is based on two pieces of evidence available in the publicly reviewed documentation rather than on confirmed, version-specific product specifications: first, ServiceNow's published Predictive Intelligence data-quality scripts, which the AIRE Data Readiness category is understood to draw upon based on publicly available developer documentation, appear in that documentation to operate on field-completeness heuristics rather than content-accuracy checks [5]; second, ServiceNow's own KCS documentation separately and independently names "content health" as a distinct discipline outside the AI-readiness product line altogether [7], and independent, third-party analysis of ServiceNow knowledge base management has separately identified ownership and feedback-handling gaps (e.g., owner fields left empty after go-live, article feedback mechanisms left unconfigured or ignored) as common, persistent causes of knowledge base staleness in production ServiceNow instances [13]. Because the exact version-to-version mapping between AIRE's internal implementation and these published scripts is not independently confirmed by this paper, the limitations described here should be read as gaps identified in the reviewed documentation and community evidence, not as confirmed omissions in any specific AIRE release. Together, these two observations — what the reviewed documentation indicates AIRE-related scripts check, and the independent, contemporaneous evidence that content problems persist unaddressed — support the claim that the omission is plausibly structural rather than incidental, pending confirmation against official, version-specific AIRE specifications.

This paper is guided by one primary research question and three supporting objectives, stated explicitly here to orient the sections that follow:

Primary research question: What dimensions of knowledge-content health, beyond the structural field completeness currently assessed by enterprise AI readiness tooling, should be evaluated before enabling agentic AI over an HR Service Delivery knowledge base, and how can each dimension be operationalized against artifacts that already exist within the platform?

Objective 1: Formally characterize the scope of a representative enterprise AI readiness tool (AIRE) and identify, with supporting evidence, what it does and does not evaluate (Sections 1–3).

Objective 2: Propose a content-layer readiness framework whose dimensions map directly to existing ServiceNow HRSD Knowledge Management objects, so that the framework is auditable without new tooling (Section 4).

Objective 3: Define a falsifiable, replicable methodology capable of empirically testing whether content-layer readiness predicts AI answer accuracy independently of structural readiness (Section 5).

The remainder of this paper is organized as follows: Section 2 reviews relevant literature on RAG system design and enterprise knowledge management, and situates this paper's contribution relative to existing frameworks; Section 3 formalizes the gap between current AI readiness tooling and content-layer requirements; Section 4 proposes a four-dimension content-health framework; Section 5 describes a proposed validation methodology; Section 6 presents an illustrative walkthrough; Section 7 discusses implications; Section 8 addresses limitations and future work; and Section 9 concludes.

2. Literature Review

2.1 Retrieval-Augmented Generation and the Role of Content Freshness

Retrieval-Augmented Generation, as formalized by Lewis et al. [1], grounds a language model's output in documents retrieved from an external corpus rather than relying solely on parametric memory. This approach has become a standard architecture for enterprise chatbots specifically because it reduces hallucination and improves the traceability of generated answers back to a source document. However, grounding a model in retrieved content is only as useful as the accuracy of that content.

Building on this foundation, a 2024 framework paper introduced FACTS (Freshness, Architectures, Cost, Testing, Security) as five pillars required to build production-grade enterprise RAG chatbots [2]. Notably, the authors validated their framework using three real enterprise chatbot deployments, one of which was an IT/HR benefits assistant — squarely the same domain addressed by ServiceNow HRSD. The explicit inclusion of Freshness as a first-class pillar, alongside security and testing, indicates that the RAG research community does not treat content currency as a secondary concern relative to system architecture or access control; it is treated as coequal.

This finding is corroborated by an independent case study published in the MDPI journal Knowledge [3], which documented a multi-tenant SaaS chatbot deployment in behavioral safety analysis. The authors explicitly named stale information as one of three primary production challenges encountered, alongside domain-specific terminology handling and role-based access control. Their solution was an automated daily extract-transform-load (ETL) pipeline designed specifically to keep the underlying knowledge base current — a deliberate, engineered response to the freshness problem rather than an incidental afterthought.

2.2 Content Quality Across RAG Application Domains

The importance of retrieval-content quality is not specific to HR or IT service management; it recurs across domains where RAG has been applied to specialized knowledge work. In clinical radiology, a proof-of-concept study found that grounding a GPT-4 model in a curated set of trauma radiology references substantially improved diagnostic and classification accuracy compared to the same model operating without curated retrieval, with the authors noting that model output quality is bounded by the quality of the underlying reference corpus [9]. Similarly, a RAG-based chatbot developed for gastrointestinal disease guidance achieved improved retrieval hit rates and answer accuracy only after fine-tuning its retrieval component against a carefully curated set of twenty-five clinical guidelines, rather than an unfiltered general corpus [8].

Security-oriented research has examined a related but distinct concern: knowledge-extraction attacks capable of recovering sensitive knowledge-base content through adversarially crafted queries [12]. While this line of work addresses confidentiality rather than accuracy, it reinforces a broader point relevant to this paper's argument: the trustworthiness of a RAG system is multidimensional, encompassing accuracy, freshness, ownership, and access-scope correctness simultaneously, rather than being reducible to any single structural metric.

2.3 Enterprise Knowledge Management Practice: Content Health as a Named Discipline

Independent of the recent AI-readiness literature, ServiceNow's own Knowledge-Centered Service (KCS) methodology, documented in its Knowledge Management Process Guide, already names this exact problem. The KCS "Evolve Loop" explicitly identifies Content Health as a discipline distinct from Process Integration (usage statistics) and Performance Assessment (search-gap identification), defined as the practice of measuring the effectiveness of each

individual knowledge article over its lifecycle [7]. This indicates that ServiceNow's own knowledge management practice already possesses the conceptual vocabulary this paper draws on — it has simply not yet been integrated into the newer, AI-specific readiness assessment product line.

Independent, third-party practitioner analysis published in the same period ServiceNow was promoting AIRE as a readiness checkpoint [4], [5] documents specific, recurring content-health failure modes in production ServiceNow knowledge bases — including owner fields left empty after go-live and article feedback mechanisms left unconfigured or ignored — that align closely with Dimensions 1 and 3 of the framework proposed in Section 4 [13]. This contemporaneity is itself informative: the readiness tooling and the practitioner-reported content problems are contemporaneous, suggesting the gap identified in this paper is not a lagging concern but a live one.

2.4 Differentiation from Existing Frameworks

Because several adjacent bodies of work address related concerns, this paper's contribution is delineated explicitly against each. Relative to KCS [7], which defines content health as a qualitative discipline practiced by knowledge coaches and authors, this paper's contribution is to operationalize a subset of that discipline as machine-checkable criteria against specific platform tables (kb_knowledge, kb_feedback_task, COE Security Configuration), so that it can be assessed automatically alongside AIRE rather than relying solely on human governance process.

Relative to RAG evaluation frameworks such as the tripartite chain-of-thought evaluator proposed for knowledge-graph-grounded chatbots, which scores query relevance, factual accuracy, and coherence of a generated answer after retrieval has already occurred [11], this paper's contribution addresses an earlier stage in the pipeline: whether the underlying corpus is fit to be retrieved from at all, prior to any specific query being issued. The two approaches are complementary rather than competing — a post-hoc answer evaluator such as RAG-Eval [11] could, in principle, be used as part of Phase 4 of the methodology proposed in Section 5 to score the retrieval-simulation outputs.

Relative to general enterprise knowledge-quality frameworks (for example, data-quality dimensions such as those codified in DAMA-DMBOK or ISO 8000, which define constructs such as accuracy, completeness, and timeliness at the level of a data record), this paper's contribution is domain-specific operationalization: translating generic timeliness and accuracy constructs into concrete, ServiceNow-HRSD-specific criteria (expiry/valid-to configuration, kb_feedback_task routing, COE-scoped retrieval permissions) rather than proposing new abstract data-quality theory. Relative to AI-governance models, which typically address organizational accountability, risk classification, and human-oversight requirements at a policy level, this paper's contribution operates one level below policy — at the level of a pre-deployment technical audit that could serve as evidence of compliance with such a policy, rather than proposing the policy itself.

3. Problem Statement: The Structural-Content Gap

Synthesizing Sections 2.1 through 2.3, this paper formalizes the following problem statement: current enterprise AI readiness tooling, exemplified by ServiceNow AIRE [4], [5], evaluates whether an instance is structurally prepared for agentic AI deployment — correct field population, correct role and access configuration, correct workflow wiring. Based on the documentation and community evidence reviewed in Section 1, this tooling does not appear to evaluate whether the knowledge content an AI agent will retrieve from at inference time is current, owned, or subject to any accuracy-review discipline; this characterization reflects the limits of publicly available evidence rather than a confirmed, version-specific product audit. The RAG literature reviewed above is consistent in treating content freshness and quality as necessary, not optional, conditions for trustworthy retrieval-grounded output [1], [2], [3], [8], [9].

The practical consequence of this gap is asymmetric risk. A structural gap — a missing required field, a misconfigured access control list — tends to fail loudly: a workflow halts, a case cannot be created, an error surfaces to an administrator. A content gap fails quietly: an AI agent retrieves a stale or inaccurate knowledge article and generates a fluent, confident, and incorrect answer, with no signal to the administrator that anything is wrong. In an HR service delivery context, where knowledge articles govern matters such as benefits eligibility, leave policy, and accommodation procedures, an incorrect but confidently delivered answer carries direct

consequences for the employee who received it and potential compliance exposure for the organization.

This yields the central research question addressed by this paper: what dimensions of knowledge content health, beyond structural field completeness, should be incorporated into enterprise AI readiness assessment for HR Service Delivery, and how can they be operationalized against artifacts that already exist within the platform?

4. Proposed Content-Layer Readiness Framework

This paper proposes four content-layer readiness dimensions, each deliberately mapped to a mechanism already available within ServiceNow HRSD's Knowledge Management application, so that the framework is designed to be auditable rather than purely conceptual.

Ownership accountability — every published knowledge article has a currently active, identifiable owner or ownership group accountable for its accuracy, verifiable against the kb_knowledge table's owner field rather than assumed from original authorship.

Staleness and expiry discipline — articles are subject to a defined review cadence, such as scheduled expiry warnings issued at fixed intervals before a valid-to date, rather than persisting indefinitely once published without re-review.

Feedback-loop capture — end-user and agent feedback signals, captured through mechanisms such as kb_feedback_task, are actively routed to the owning group and result in a measurable review action, rather than being logged and left unaddressed.

Retrieval-scope correctness — the access-control configuration governing which articles a given AI agent or virtual agent persona can retrieve is verified to match its intended audience, since an over-broad retrieval scope can surface confidential or role-restricted content through an AI-generated answer in a way a human agent's manual search would not.

Table 1 summarizes how these four dimensions relate to, and extend beyond, AIRE's existing finding categories [4], [5].

AIRE Finding Category	What It Measures	Content-Layer Dimension It Omits
Data Readiness	Structured field completeness (e.g., description, category, assignment group)	Whether article content is still factually accurate
Configuration Needed	Roles, ACLs, workflow and plugin configuration	Whether retrieval scope matches intended audience (Dimension 4)
Now Support Recommended	Instances flagged for vendor-assisted remediation	Ownership accountability and feedback-loop capture (Dimensions 1 and 3)

Table 1. Mapping of AIRE finding categories to omitted content-layer dimensions.

The middle column of Table 1 reflects publicly documented AIRE finding categories and their associated published descriptions [5]; the right-hand column reflects this paper's interpretation of what

those documented categories do not appear to cover, based on the evidence reviewed in Sections 1 and 2, rather than a confirmed audit of AIRE's internal implementation.

Each of these four dimensions maps to a concrete, queryable object within a ServiceNow instance, which distinguishes this framework from a purely aspirational model: it is intended as a checklist an implementer can execute today, using existing platform tables and configuration screens, rather than a future-state capability requiring new tooling to be built from scratch.

5. Proposed Validation Methodology

To move this framework from a conceptual proposal to a validated one, this paper proposes a controlled testbed methodology using a ServiceNow Personal Developer Instance (PDI) running the current Australia release. This paper is positioned directly as a framework and protocol contribution: the methodology below is presented as a proposed, not-yet-executed research protocol, detailed to a level intended to allow independent replication. It consists of four phases.

5.1 Phase 1: Corpus Seeding

A knowledge base is constructed with a target sample size of N = 500 published articles, stratified across five simulated HR Centers of Excellence (Payroll, Time Off, Benefits, Employee Relations, and Onboarding/Offboarding), with 100 articles per COE to permit balanced comparison. Within each COE-stratum, articles are randomly assigned via a random-number-generator seed (fixed and reported for reproducibility) to one of four ground-truth conditions: current and accurate (60 percent), stale — valid-to date deliberately set in the past (20 percent), unowned — owner or owner-group field deliberately cleared or pointed to a deactivated group (10 percent), and factually altered — a deliberately introduced factual error in the article body, logged in a private answer key (10 percent). This stratified design ensures every content-health dimension in Section 4 has a known, labeled ground truth against which detection can be scored, rather than relying on subjective post hoc judgment.

5.2 Phase 2: Structural Baseline

AIRE's existing assessment is run against the seeded instance twice, at week 0 and week 4, to establish a baseline and confirm result stability over time. The primary evaluation metric for this phase is the AIRE detection rate: the proportion of the 200 deliberately defective articles (stale, unowned, and factually altered strata combined) that AIRE flags under any finding category. Based on the gap identified in Section 3, the directional hypothesis (H1) is that the AIRE detection rate for these content-level defects will be low relative to its detection rate for structural defects seeded as a positive control — that is, a matched set of 20 articles with deliberately blanked required fields that AIRE's existing checks are expected to detect, confirming the tool is functioning correctly on the instance before its detection rate on content-level defects is interpreted — since AIRE's underlying checks are not designed to inspect article body content or ownership currency.

5.3 Phase 3: Content-Layer Audit

The four-dimension framework proposed in Section 4 is applied against the same instance via a scripted audit (a Background Script or Scripted REST endpoint) that queries kb_knowledge ownership fields, expiry/valid-to configuration, kb_feedback_task records, and COE Security Configuration policies governing the relevant knowledge bases. Detection performance is scored using standard binary-classification metrics against the Phase 1 ground truth: precision, recall, and F1-score, computed separately for each of the four dimensions and pooled overall. Inter-rater reliability for any manually adjudicated edge cases (e.g., partially stale articles) is assessed using Cohen's kappa between two independent reviewers, with a target threshold of $\kappa \geq 0.75$ before a case is added to the ground-truth set.

To ensure all four dimensions, not only ownership and expiry, are exercised by the audit, Phase 1's seeded corpus is extended with two additional defect classes: a feedback-loop-failure set of 20 articles, each seeded with one or more kb_feedback_task records marked negative and left unresolved past a defined threshold (e.g., 30 days without a review action), which the audit is scored on its ability to flag via a query for open feedback tasks older than that threshold; and a retrieval-scope-error set of 20 articles, each seeded in a knowledge base whose COE Security Configuration or associated ACL is deliberately misconfigured to grant a broader AI-agent persona access than its documented intended audience, which the audit is scored on its

ability to flag by comparing each knowledge base's configured retrieval-permitted groups against a maintained reference mapping of intended audience per COE. Detecting the factually altered articles (the fourth Phase 1 defect class) requires a mechanism beyond ownership or expiry metadata, since a factual error can appear in an otherwise current, owned, and feedback-clean article. This paper proposes detecting this class through two complementary mechanisms rather than metadata inspection alone: first, cross-referencing article content against a maintained set of authoritative source-of-truth statements for each COE, confirmed independently with the owning business team, using a scripted text-similarity or claim-matching check that flags articles whose stated values diverge from the authoritative source; second, treating the kb_feedback_task channel itself as a secondary signal, since end users or agents who encounter a factual error in practice are a plausible source of negative feedback captured under Dimension 3. The first mechanism is the primary, proactive check this protocol relies on for the Phase 1 ground truth; the second is proposed as a complementary, reactive signal rather than a replacement, since not every factual error generates user-reported feedback before an AI agent retrieves it. To avoid ambiguity in sample allocation, this paper specifies explicitly that the three 20-article sets described above — the structural positive-control set introduced in Section 5.2, the feedback-loop-failure set, and the retrieval-scope-error set — are additional to, and disjoint from, the core 500-article corpus described in Section 5.1; they are not drawn from or overlapping with the 500 articles used to establish the ownership, staleness, and factual-accuracy ground truth. The total seeded environment therefore comprises 560 articles across all defect classes. Detection-performance metrics are computed separately for distinct ground-truth sets rather than pooled into a single denominator: precision, recall, and F1-score for the ownership, staleness, and factual-alteration dimensions are computed against the 500-article core corpus and its internal 60/20/10/10 percent stratification; precision and recall for the feedback-loop-capture dimension (Dimension 3) are computed against the 20-article feedback-loop-failure set plus an equal-sized matched set of 20 clean-feedback control articles (40 articles total, not drawn from the core 500); and precision and recall for the retrieval-scope-correctness dimension (Dimension 4) are computed against the 20-article retrieval-scope-error set plus an equal-sized matched set of 20 correctly scoped control articles (40 articles total, likewise not drawn from the core 500). A single pooled F1-score across all four dimensions, where reported, is calculated as the unweighted mean of the four dimension-specific F1-scores rather than a single confusion matrix across a shared denominator, since the four dimensions are evaluated against distinct ground-truth sets with different sample sizes and defect-generation mechanisms.

5.4 Phase 4: Retrieval Simulation

A query set of 50 representative HR questions is constructed using a template-based generation method: ten question templates per COE (e.g., "What is the current policy on [topic] for [employee persona]?") instantiated with COE-specific variables, reviewed by at least one subject-matter expert for realism before use. Each query is issued against a standard RAG pipeline (a sentence-embedding retriever such as a BERT-family bi-encoder paired with a vector store, and a fixed large language model held constant across conditions to isolate corpus quality as the independent variable) under two conditions: (a) the full seeded corpus, unmodified — the baseline corpus condition, representing the corpus state that this protocol hypothesizes AIRE would report as ready based on its documented Data Readiness criteria, since the AIRE assessment itself has not yet been run as part of this not-yet-executed protocol — and (b) the same corpus with all articles flagged in Phase 3 removed or corrected — the content-audited condition. Resulting answers are scored for factual accuracy against the private answer key by two independent human raters blind to condition, using a three-point scale (correct, partially correct, incorrect), with disagreements adjudicated by a third rater. Because McNemar's test requires a binary outcome, "partially correct" responses are collapsed to "incorrect" for the primary analysis — a conservative choice that treats any factual deviation from the expected answer as a failure — and the three-point scale is retained for a separate, descriptive secondary analysis reported alongside the binary result so that the information lost in collapsing is not discarded entirely. The primary statistical test is a paired comparison (McNemar's test, given paired binary correct/incorrect outcomes per query across the two conditions) evaluating whether the content-audited condition yields significantly higher answer accuracy than the

baseline corpus condition at α = 0.05. A secondary analysis reports the raw accuracy delta with a 95 percent confidence interval.

This four-phase design was chosen specifically because each phase produces an independently falsifiable, quantitatively reported output — a detection-rate comparison (Phase 2 vs. Phase 3), a set of precision/recall/F1 scores (Phase 3), and a paired significance test with an effect-size estimate (Phase 4) — rather than relying on qualitative impression, which is intended to make the protocol directly executable and its results independently reviewable once run.

6. Illustrative Application

To make the framework concrete ahead of full empirical validation, this section presents a hypothetical, illustrative walkthrough. It is presented explicitly as a conceptual worked example rather than as empirical results, since Phases 2 through 4 described in Section 5 have not yet been executed at the time of this writing.

Consider an HR knowledge base containing 200 published articles across six Centers of Excellence. Suppose 15 percent of articles have not been reviewed in over eighteen months, 8 percent have an owner field pointing to a group that no longer exists in the instance (e.g., following a reorganization), and a further 5 percent have received negative feedback ratings that were logged but never routed to an action. Under AIRE's existing Data Readiness check, all 200 articles would likely register as structurally complete, since each has populated title, body, category, and workflow-state fields. Summing these three categories (15 percent, 8 percent, and 5 percent) gives 28 percent only under the simplifying assumption that the categories are mutually exclusive, which this paper makes explicit rather than implicit; in a real knowledge base these categories would likely overlap to some degree, making the true union somewhat smaller than the simple sum. Under the stated no-overlap assumption, the four-dimension content-layer audit proposed in Section 4 would flag an estimated 28 percent of the knowledge base for remediation before it is deemed safe as a retrieval source for an autonomous AI agent; under a more conservative assumption of partial overlap, the true figure would be somewhat lower than 28 percent but still substantially above zero, since the Data Readiness check would flag none of these 200 articles. Either way, this produces a materially different readiness picture than the structural check alone.

This illustrative gap, under the explicit no-overlap assumption stated above, is directionally consistent with, though not derived from, the freshness and ownership challenges independently reported in the practitioner-facing evidence discussed in Section 2.3 [13] and the production challenges described in the applied RAG case study reviewed in Section 2.1 [3]. It is presented here as a hypothetical worked example to demonstrate the scale of discrepancy the framework is designed to detect, not as a validated statistic or an empirical finding.

7. Discussion

The central implication of this analysis is that enterprise AI readiness, as currently operationalized by leading platform vendors, is necessary but not sufficient. Organizations that treat a passing AIRE assessment as a green light for enabling agentic AI across HR Service Delivery may be accurately assessing configuration risk while remaining fully exposed to content risk — a risk category the RAG literature treats as at least equally consequential [1], [2], [3].

This has a practical governance implication: the content-layer framework proposed here is not intended to replace AIRE or similar tooling, but to run alongside it as a complementary pre-deployment gate. Because all four proposed dimensions map to existing platform objects, the framework can, in principle, be operationalized as a companion scheduled job or reporting dashboard within the same Now Assist Center workspace that surfaces AIRE's own findings [5], reducing the operational burden of adopting it.

The framework also generalizes beyond HRSD. The same structural/content distinction applies to ITSM and CSM knowledge bases, and arguably to any vendor's AI-readiness product built primarily around field-completeness heuristics rather than content-lifecycle discipline. This suggests the gap identified here may be a category-level blind spot in the current generation of enterprise AI readiness tooling generally, not a ServiceNow-specific shortcoming.

8. Limitations and Future Work

Consistent with its framing throughout this paper, this work should be read as a framework and protocol

contribution rather than an empirical study. It has three principal limitations. First, the validation methodology specified in Section 5 has not yet been executed; Section 6's illustrative walkthrough is explicitly hypothetical and should not be read as empirical evidence. Executing the four-phase protocol on a seeded PDI, and subsequently on an anonymized production-scale dataset, is the immediate next step for this research program.

Second, the framework has been developed and scoped specifically against ServiceNow's Knowledge Management data model; while Section 7 argues for generalizability, confirming that claim would require replicating the framework against at least one additional enterprise service management platform with a comparable AI-readiness product.

Third, the illustrative percentages presented in Section 6 are constructed for explanatory purposes and should not be cited as measured findings. Future work should replace this illustration with results from the four-phase methodology, ideally across multiple seeded corpora of varying size and staleness distribution, to establish whether the magnitude of the structural-content gap is consistent across instance scale.

Future research directions also include: (a) extending the retrieval-scope correctness dimension to cover multi-agent orchestration scenarios, where several AI agents with different retrieval permissions operate over a shared knowledge base; (b) investigating whether feedback-loop capture rates correlate with measurable downstream answer-accuracy improvements over time; and (c) exploring whether a standardized content-health score, analogous to AIRE's remediation-effort scoring, could be adopted as an industry-wide metric across enterprise service management platforms.

9. Conclusion

Based on the documentation reviewed in this paper, enterprise AI readiness tooling such as AIRE appears to have made genuine progress in catching configuration-level obstacles before AI deployment. However, the RAG research literature is consistent on a point that this reviewed tooling does not appear to check: a retrieval-grounded AI system is only as trustworthy as the content it retrieves from, and structural completeness is not a reliable proxy for content accuracy. This paper has formalized that gap, proposed a four-dimension content-layer readiness framework mapped directly to existing ServiceNow HRSD Knowledge Management objects, and outlined a concrete, falsifiable methodology for validating it. Closing this gap does not require replacing existing readiness tools; it requires adding a content-layer audit alongside them, using mechanisms that already exist inside the platform's own Knowledge Management application. The absence of that layer in current tooling is not evidence that the underlying problem is unimportant — based on the documentation and community evidence reviewed in this paper, it appears that, as of this writing, this specific content-layer checklist has not yet been incorporated into mainstream AI readiness assessment tooling.

References

1. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. Advances in Neural Information Processing Systems, 33, 9459–9474.

2. Akkiraju, R., Xu, A., Bora, D., Yu, T., An, L., Seth, V., Shukla, A., Gundecha, P., Mehta, H., Jha, A., Raj, P., Balasubramanian, A., Maram, M., Muthusamy, G., Annepally, S.R., Knowles, S., Du, M., Burnett, N., Javiya, S., Marannan, A., Kumari, M., Jha, S., Dereszenski, E., Chakraborty, A., Ranjan, S., Terfai, A., Surya, A., Mercer, T., Thanigachalam, V.K., Bar, T., Krishnan, S., Kilaru, S., Jaksic, J., Algarici, N., Liberman, J., Conway, J., Nayyar, S., & Boitano, J. (2024). FACTS about building retrieval augmented generation-based chatbots. arXiv preprint arXiv:2407.07858.

3. Rivera, J., Zapata, S., Pizarro, R., & Keith, B. (2025). Enhancing chatbot performance in a SaaS platform through retrieval-augmented generation and prompt engineering: A case study in behavioral safety analysis. Knowledge, 5(4), 25. https://doi.org/10.3390/knowledge5040025

4. ServiceNow, Inc. Now Assist Readiness Evaluation (ServiceNow Store listing). Application ID

c46d62364724b250def56992e36d43c4, version 1.0.5. Retrieved from https://apprepo.service-now.com/store/app/c46d62364724b250def56992e36d43c4

5. ServiceNow, Inc. Now Assist Readiness Evaluation (application documentation). Retrieved from https://www.servicenow.com/docs/r/SRy58xeijUUqttfwLgfrHQ/PNYTQdSOMVj0UljKFgYBJg

6. ServiceNow, Inc. Now Assist Readiness Evaluation release notes, documenting the rebrand from Now Assist Readiness Evaluation to AI Readiness Evaluation. Retrieved from https://www.servicenow.com/docs/r/C17ZqvSaMwUVpMBg8Nc5sg/QOvz4bf6dYCBx51SDVf33Q

7. ServiceNow, Inc. Knowledge Management Process Guide (Rome release), Asset Number 0003644: Knowledge-Centered Service (KCS) Evolve Loop and Content Health. This paper reviewed a third-party-hosted copy of this official ServiceNow process guide at https://www.scribd.com/document/710268043/ITSM-Knowledge-Process-Guide-Rome, since an official ServiceNow-hosted copy was not independently located at the time of writing.

8. Zhou, Q., Liu, C., Duan, Y., Sun, K., Li, Y., Kan, H., Gu, Z., Shu, J., & Hu, J. (2024). GastroBot: A Chinese gastrointestinal disease chatbot based on retrieval-augmented generation. Frontiers in Medicine, 11, 1392555. https://doi.org/10.3389/fmed.2024.1392555

9. Fink, A., Nattenmüller, J., Rau, S., Rau, A., Tran, H., Bamberg, F., Reisert, M., Kotter, E., Diallo, T., & Russe, M.F. (2025). Retrieval-augmented generation improves precision and trust of a GPT-4 model for emergency radiology diagnosis and classification: A proof-of-concept study. European Radiology. PMC12226682.

10. Kellton. (2026, January 30). ServiceNow AI Agents and Agentic Workflow Automation in 2026: A Practical Guide. Kellton Tech Blog. https://www.kellton.com/kellton-tech-blog/servicenow-ai-agents-and-agentic-workflow-automation-complete-guide

11. Akindele, O.K., Mishra, B.K., & Wertheim, K.Y. (2025). A knowledge graph and a tripartite evaluation framework make retrieval-augmented generation scalable and transparent. arXiv preprint arXiv:2509.19209.

12. Qi, Z., Sahu, U., Ma, L., Han, H., Rossi, R.A., Dernoncourt, F., Halappanavar, M.M., Ahmed, N.K., Dong, Y., Zhao, Y., Zhang, Y., & Wang, Y. (2026). Benchmarking knowledge-extraction attack and defense on retrieval-augmented generation. arXiv preprint arXiv:2602.09319.

13. ScreenMeet. 9 Reasons Your ServiceNow Knowledge Base Goes Stale, and What to Do About It. https://www.screenmeet.com/blog/reasons-your-servicenow-knowledge-base-goes-stale