SvedbergOpen
DISSEMINATION OF KNOWLEDGE
International Journal of Artificial Intelligence and Machine Learning
Publisher's Home Page: https://www.svedbergopen.com/
Research Paper Open Access

The Verified Query Repository: A Human-in-the-Loop Architectural Pattern for Eliminating Hallucination in Enterprise NLP Systems

Arjun Danda Sureshbabu

Independent Researcher, USA

ABSTRACT
Enterprise deployments of large language model-powered natural language processing systems produce a class of failure that is systematically more damaging than errors that surface visibly. The response is fluent. It is confident. And it is factually wrong, consumed by decision-makers who have no mechanism to detect the error before it propagates through organizational processes. Existing mitigation strategies, including prompt engineering, retrieval-augmented generation, and fine-tuning, reduce the probability of this failure without providing structural guarantees for the specific queries where reliability is non-negotiable. This paper introduces the Verified Query Repository (VQR), a human-in-the-loop architectural pattern that addresses hallucination in enterprise NLP systems by replacing generation with retrieval for high-stakes queries whose responses have been previously validated by human domain experts. The VQR stores curated, human-approved query-response pairs with provenance metadata, similarity-matched against incoming queries at runtime. Queries exceeding a defined similarity threshold retrieve the validated response directly, bypassing large language model inference for that query class. The paper presents the VQR entry schema, query routing architecture, human validation workflow, and staleness management framework, alongside a governance model covering ownership structures, curation quality standards, and coverage expansion strategies. VQR reliability depends on curation quality and staleness management rather than model capability. Organizations deploying VQR without the governance infrastructure to sustain it will find the repository degrades at the pace that business change imposes. VQR is positioned as a complement to retrieval-augmented generation, providing structural reliability guarantees for high-stakes query classes that probabilistic mitigation approaches cannot deliver.

Keywords: Enterprise NLP, Hallucination Reduction, Human-in-the-Loop, Large Language Model, Query Validation, Retrieval Augmented Generation, Verified Query Repository.

I. INTRODUCTION

Before an organizational decision can be corrupted by a hallucinated NLP response, someone has to act on it. When an enterprise system returns a fluent, confident, and incorrect figure for a finance query, the damage does not begin at generation -- it begins at consumption. A decision-maker receiving a plausible response has no reliable mechanism for detecting the error without independently verifying the underlying data, a step that defeats the purpose of the natural language interface. The figure enters a report, an executive presentation, or a compliance filing before the discrepancy surfaces, if it surfaces at all (Ji et al., 2023). For systems processing billions of dollars in revenue, a hallucinated response that appears correct is not a low-frequency edge case to be tolerated. It is a systematic exposure whose consequences scale with the organizational weight placed on the output.

Probabilistic mitigation strategies address the average, not the specific. Prompt engineering structures input context to reduce the frequency of inaccurate generation. Retrieval-augmented generation grounds responses in retrieved documents to reduce parametric knowledge gaps. Fine-tuning adjusts model weights toward more

accurate behavior across the training distribution (Ouyang et al., 2022; Lewis et al., 2020). Each reduces the average hallucination rate. None provides a structural guarantee for a specific query: that a validated correct response, once produced, will be returned again when the same query recurs. For enterprise use cases where a single hallucinated response to a repeated high-stakes query can trigger an audit finding or a material financial restatement, average performance is the wrong metric entirely.

Where probabilistic mitigation ends and guaranteed correctness must begin, a distinct class of enterprise queries comes into focus: specific, frequently repeated, high-stakes, and verifiable. A Verified Query Repository stores human-validated query-response pairs and returns them directly when incoming queries match above a defined similarity threshold, bypassing LLM inference for the matched query class. The VQR does not improve the model's accuracy on novel queries. Rather, it removes the model from the response pipeline for query types where human-validated accuracy is both required and achievable (Asai et al., 2024). The structural move is deliberate: generation is not constrained -- it is bypassed.

Organized around three connected claims, what follows builds the case for VQR as a distinct reliability mechanism. VQR is structurally distinct from general retrieval-augmented generation because it retrieves validated responses rather than retrieved context, bypassing generation rather than enriching it. VQR requires governance infrastructure -- curation workflows, staleness management, ownership structures -- to sustain its reliability guarantees as business facts and rules evolve. Combining VQR with a generative NLP system produces a hybrid architecture providing stronger reliability guarantees for high-stakes queries while preserving generative capability for the novel query types that represent the majority of enterprise NLP traffic.

II. RELATED WORK

2.1 Hallucination in LLM-Based Enterprise Systems

Token prediction, not factual reasoning, drives the outputs of transformer-based language models. The generation mechanism selects the statistically most likely next token given context -- a process that yields fluent, coherent, and factually wrong text whenever the model's parametric knowledge is incomplete, outdated, or inconsistent with domain-specific definitions the query assumes. Factual fabrication and source misattribution represent the most consequential subtypes in enterprise NLP contexts: fabrication presents incorrect figures as facts, while misattribution creates the appearance of verifiability by tracing an incorrect claim to a source that does not support it (Dang et al., 2025). Neither failure carries a surface signal of uncertainty. Both reach the decision-maker looking identical to a correct response.

Across the enterprise deployment landscape, organizations apply identical mitigation strategies to all query types, treating a finance reporting query with the same reliability approach as an exploratory analysis request. This undifferentiated treatment accepts that high-stakes queries receive the same probabilistic reliability guarantee as low-stakes ones -- a mismatch between organizational risk distribution and system design that VQR addresses by routing query classes to reliability mechanisms calibrated to their stakes level (Gao et al., 2024). Detection improvements have advanced independently: metamorphic testing identifies hallucination by checking whether semantically equivalent query variants produce consistent responses (Yang et al., 2025a), and self-evaluation frameworks prompt models to assess their own response confidence before delivery (Sun et al., 2024). These mechanisms detect hallucinations after generation. They cannot guarantee that a previously validated correct response will be returned for a repeated query.

The organizational impact of hallucination scales with query stakes, not query volume. A hallucinated response to an exploratory analysis query carries low organizational consequence. A hallucinated figure in a finance reporting workflow carries consequence proportional to the decisions made on it (Zhao et al., 2023). When benchmark conditions favor the most capable commercially available models, production failures still accumulate at rates that cannot satisfy enterprise finance and compliance requirements: GPT-4 achieves only 54.89% execution accuracy on the BIRD benchmark against a human baseline of 92.96%, and the gap widens in real-world enterprise deployments where business-specific metric definitions and implicit constraints are not encoded in the evaluation schema (Gao et al., 2024). The performance ceiling for probabilistic approaches, the data suggest, sits well below what finance and compliance workflows actually require.

2.2 Retrieval-Augmented Generation and Human-in-the-Loop Systems

Grounding LLM generation in documents retrieved from an external knowledge base reduces the extent to which the model relies on potentially incomplete parametric knowledge during response production (Lewis et al., 2020). Self-RAG advances this by training the model to selectively retrieve and to critique its own generated

responses against retrieved passages, producing a generate-then-verify cycle that further constrains hallucination (Asai et al., 2024). Optimising the hyperparameters of the vector database underlying the RAG system improves retrieved context relevance: average context similarity scores climb from approximately 0.494 before optimisation to 0.616 after, narrowing the gap between retrieved documents and ideal responses (Koyun & Gurkan, 2024). Yet all three approaches share a structural feature that limits their reliability ceiling. Every query still routes through LLM generation. Context enrichment improves what the model generates; it does not eliminate generation as a failure point (Wang et al., 2024).

When human judgment has been applied in NLP quality assurance, the dominant application has been at the training stage. Reinforcement learning from human feedback adjusts model weights toward preferred outputs (Ouyang et al., 2022). Constitutional AI applies AI-generated critique to reduce harmful outputs (Bai et al., 2022). Active learning routes uncertain training examples to human annotators (Das et al., 2023). Human judgment enters the model during training; it does not persist as a retrievable governance artifact during inference. The pattern across all three approaches is consistent: human expertise is consumed to produce a more reliable model, not to create a durable, auditable record of validated responses. VQR applies human judgment at a different architectural point -- not to train a model that is statistically less likely to hallucinate, but to curate a set of responses for which hallucination is structurally impossible because no generation occurs. The gap no existing architecture closes is this: guaranteeing that a previously validated correct response to a high-stakes query will be returned again when the query recurs, with a provenance chain demonstrating human validation. General RAG, Self-RAG, and human-in-the-loop training approaches each address part of the enterprise hallucination problem. None of them reaches the part VQR addresses -- the specific, repeated, high-stakes query class where generatively produced responses cannot provide the audit trail regulated environments require (Pan et al., 2024). No named, principled architectural pattern combining retrieval-based response preservation with governed human validation has been established in the enterprise NLP deployment literature. VQR names that pattern and specifies its operational requirements.

III. THE VQR ARCHITECTURE

3.1 VQR Entry Schema

Seven fields in each VQR entry serve distinct reliability functions that together make responses traceable, scoped, and time-bounded. The canonical query stores a normalized query representation -- stripped of surface variation, session-specific phrasing, and coreference -- against which incoming queries are similarity-matched at runtime. Provenance metadata accompanies the validated response: validator identity, validation timestamp, and source citations supporting factual claims. A confidence score records the validator's accuracy and completeness assessment of the entry at approval time. The domain scope field restricts each entry's applicability to the business domain for which it was validated, preventing cross-domain misapplication when metric definitions differ by context. The staleness threshold specifies the revalidation period, set by the domain steward based on how rapidly the underlying facts are expected to change. A version field tracks revision history, enabling rollback when revalidated responses supersede prior entries (Balaguer et al., 2024).

Table 1. VQR Entry Schema: Fields, Data Types, Reliability Purpose, and Staleness Triggers

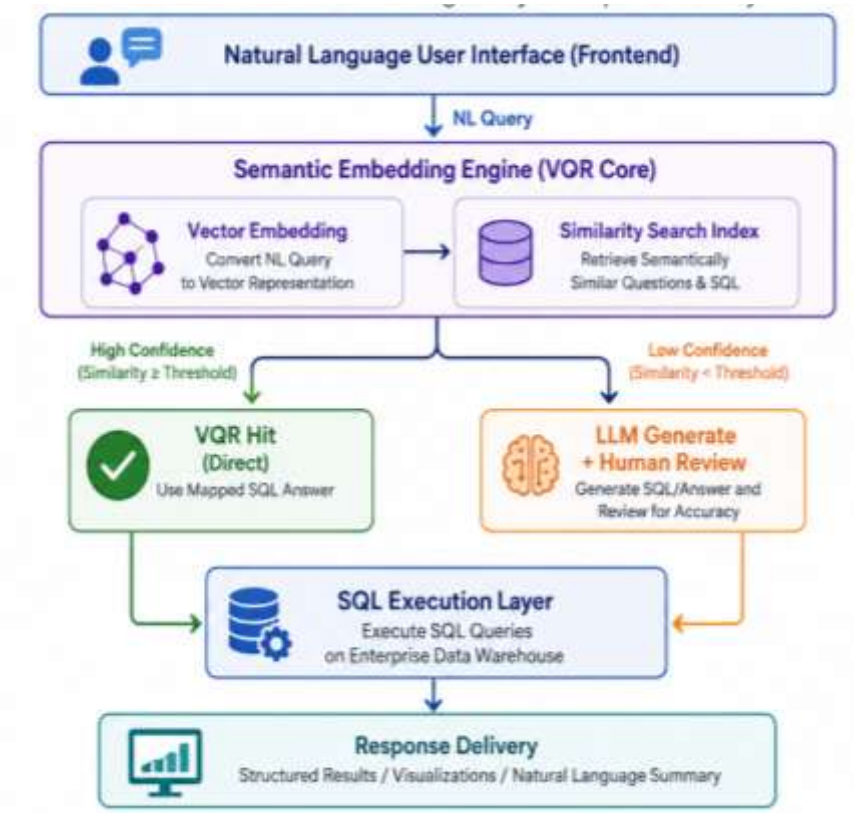
Field	Data Type	Reliability Purpose	Staleness Trigger
Canonical Query	Text (normalized)	Surface-variation-stripped query used for similarity matching at runtime	Revalidate on metric rename or scope change
Validated Response	Text + citations	Human-approved answer with inline source references	Revalidate when cited source is updated
Validator Identity	User ID + role	Named domain expert who approved the entry	Update on role change; revalidate on departure
Validation Timestamp	ISO 8601 datetime	Date and time of most recent validation	Triggers time-based expiry calculation
Source Citations	Structured refs	Verifiable sources supporting every factual claim	Revalidate when source document is revised
Confidence Score	0.0-1.0 float	Validator's accuracy and completeness assessment	Recalibrate after re-validation
Domain Scope	Enum (controlled)	Restricts entry applicability to validated business domain	Revalidate on domain boundary change
Staleness Threshold	Duration (days)	Revalidation period set by domain steward	Adjusted by steward as business cadence changes
Version	Semantic version	Revision history enabling rollback to prior entries	Incremented on each revalidation

Taken together, these nine fields convert a query-response pair from an ephemeral generation event into a governed, auditable repository entry with a defined lifecycle. The schema does not improve the quality of the response at entry time -- the human validator does that. What the schema provides is the infrastructure needed to track quality over time, attribute accountability, and detect the moment a once-accurate entry no longer reflects current business reality. Staleness, not initial validation, is where most VQR deployments encounter their first sustained operational challenge.

3.2 Query Matching and Similarity Thresholds

Encoded as dense vector representations using a semantic embedding model, incoming queries are compared against canonical query embeddings using cosine similarity. Three similarity bands govern routing: high-confidence matches retrieve the validated response and return it directly; intermediate-confidence matches retrieve it with an uncertainty flag, prompting user confirmation before downstream use; low-confidence queries route to LLM generation (Asai et al., 2024). The three-band structure is not a binary gate. It produces a continuous reliability spectrum, allowing the system to surface partial match information to users rather than forcing a hard VQR-or-generation decision at every query.

Figure 1. VQR Query Routing Architecture: Decision Path from Incoming Query to Response Delivery



Threshold calibration carries asymmetric consequences in high-stakes domains. A threshold set too high generates false negatives -- valid matches not retrieved, routing previously validated queries to generation. A threshold set too low generates false positives -- an incorrect match retrieved, returning a validated response to a query it does not address. For finance applications, the cost asymmetry favors conservative thresholds: a false negative preserves the possibility of correct generation; a false positive guarantees a mismatch between the query's actual intent and the response it receives (Yang et al., 2025b). The same logic does not hold symmetrically across all domains. Domains where the cost of a missed match exceeds the cost of a mismatch may warrant lower thresholds, calibrated against observed error rates rather than set once at deployment.

3.3 The Human Validation Workflow

Rather than automated ingestion, query-response pairs enter the VQR through a governed validation workflow that structures expert judgment into an auditable decision record. Responses flagged as high-stakes candidates -- by domain classification, query type, or generation confidence score -- are routed to human validators before repository entry. Four criteria govern each validation decision: factual accuracy requires cross-referencing the response against the cited source as of the validation date; metric correctness requires confirming the metric definition against the current metric registry; business rule compliance requires confirming implicit constraints against current business policy documentation; source traceability requires that every factual claim carries a citation resolving to a verifiable source (Das et al., 2023). A response that clears all four criteria earns repository entry; one that fails any criterion is returned for revision.

Validator-domain matching is not optional. A finance analyst validates finance domain entries; an HR data steward validates HR domain entries. This matching ensures metric correctness and business rule compliance assessments are made by someone with the organizational knowledge the evaluation requires -- not by a generalist reviewer applying general accuracy criteria. The deeper issue is accountability: when a validated response is acted on in a compliance filing, the audit trail names the validator, the validation date, and the criteria applied. Generalist validation severs the link between organizational expertise and evidentiary accountability that makes the audit trail defensible.

3.4 Staleness Detection and Repository Maintenance

Silently and continuously, business facts, metric definitions, and organizational rules evolve after a VQR entry is validated. Two mechanisms operate in parallel to detect this drift before it affects response quality. Time-

based expiry flags entries that have exceeded their staleness threshold, routing them to domain stewards for revalidation scheduling. Event-triggered invalidation responds to business changes -- a new metric definition, a policy update, a product launch affecting entity definitions -- by immediately flagging affected entries for priority revalidation rather than waiting for time-based expiry (Balaguer et al., 2024). The combination matters: time-based expiry catches gradual drift; event triggering catches discrete discontinuities that invalidate entries without warning.

Table 2. VQR Staleness Management: Triggers, Detection Mechanisms, and Repository Actions

Staleness Trigger	Detection Mechanism	Repository Action	Signal Source
Time-based expiry	Staleness threshold exceeded	Flag for steward revalidation scheduling	Staleness threshold field + system clock
Metric definition change	Event-triggered invalidation	Immediate priority flag; queries return staleness warning	Metric registry change event
Policy update	Event-triggered invalidation	Domain steward review; affected entries suspended	Business policy change notification
Source document revision	Citation monitoring	Re-verify factual claims against revised source	Source change detected by citation monitor
Entity definition change	Event-triggered invalidation	Priority revalidation; related entries cross-checked	Product/entity registry update event
Validator departure	Governance audit	Entries assigned to departing validator queued for reassignment	HR system offboarding event

Flagged entries are not immediately removed. They remain in the repository marked as pending revalidation, and queries matching them retrieve the response with a staleness warning rather than a full confidence signal. This soft-deprecation approach maintains partial reliability benefit during the revalidation window without presenting an outdated response as currently validated. The organizational cost of this approach is the steward's revalidation workload -- a workload that accumulates if governance processes are not resourced to process flagged entries at the pace business change generates them.

IV. INTEGRATION WITH ENTERPRISE NLP SYSTEMS

4.1 The Hybrid Architecture

Operating as complementary rather than competing components, VQR and a generative LLM divide the query space by stakes level. High-stakes matched queries retrieve validated responses through VQR; novel queries, low-confidence matches, and standard-risk domain queries route to LLM generation. The hybrid architecture preserves generative capability for the exploratory and novel queries that represent the majority of enterprise NLP traffic, while structurally eliminating generation for the specific query class where hallucination causes the greatest organizational harm (Lewis et al., 2020). The sequencing is the critical design decision: for queries in high-stakes domains, VQR lookup is attempted before LLM generation is engaged. Generation serves as fallback, not default -- a reversal of the typical priority ordering that makes the reliability guarantee possible (Asai et al., 2024).

4.2 Domain Scoping and Query Classification

Domain classification -- a governance decision operationalising the organization's risk tolerance for hallucination across different query types -- determines which queries attempt VQR retrieval and which route directly to generation. Finance reporting, compliance queries, HR metrics, and executive dashboards are configured as VQR-first domains. Exploratory analysis, unstructured text summarization, and general information retrieval route directly to generation. Classifying a domain as VQR-first commits the organization to maintaining validated coverage for that domain's most frequent high-stakes queries -- an ongoing governance obligation that makes domain classification a policy decision rather than a routing configuration (Gao et al., 2024). Organizations that treat it as the latter typically discover the governance obligation only after VQR coverage has degraded to the point where high-stakes queries are falling through to generation by default.

Table 3. Domain Classification and Routing Policy: VQR-First vs. LLM-Direct Domains

Query Domain	Routing Policy	Risk Rationale	Governance Requirements
Finance Reporting	VQR-first	Single hallucinated figure can trigger audit finding or material restatement	Finance analyst validators; quarterly revalidation cadence minimum
Compliance Queries	VQR-first	Regulatory exposure from incorrect compliance guidance cannot be remediated after the fact	Compliance officer validation; mandatory revalidation on regulatory update
HR Metrics	VQR-first	Headcount and compensation figures feed executive reporting with direct fiduciary implications	HR data steward validation; revalidation on policy cycle
Executive Dashboards	VQR-first	Outputs drive strategic decisions with long-lead consequence for organizational resource allocation	Senior domain expert validation; monthly staleness review
Exploratory Analysis	LLM-direct	Query diversity makes VQR coverage infeasible; errors surface through analyst review	Standard LLM quality monitoring; no VQR obligation
Text Summarization	LLM-direct	Structural accuracy requirements absent; fluency and coverage are primary quality dimensions	Standard LLM quality monitoring
General Information	LLM-direct	Query volume and diversity preclude VQR coverage; low organizational consequence per query	Standard LLM quality monitoring

4.3 Confidence Scoring and Escalation Paths

Across three confidence bands, the system produces a continuous reliability spectrum rather than a binary VQR-or-generation split. High-confidence matches return the validated response directly with provenance metadata available on demand. Intermediate-confidence matches return the validated response with an uncertainty flag, surfacing the match confidence and canonical query matched against, enabling users to confirm the match before placing organizational weight on the response. Low-confidence queries route to LLM generation, with the generated response optionally flagged for human review and potential VQR entry (Wang et al., 2024). Users receive calibrated reliability signals for every response -- validated retrieval, uncertain retrieval, or generative output -- enabling informed judgment about how much weight to place on each response type (Feng et al., 2024). What this misses, if left unattended, is the intermediate band: responses returned with uncertainty flags that users habitually dismiss will erode the reliability benefit the flag was designed to preserve.

Table 4. Confidence Band Definitions: Thresholds, Response Behaviors, and User Signals

Confidence Band	Similarity Threshold	Response Behavior	User Signal
High Confidence	>= 0.85	Validated response returned directly	Provenance metadata available on demand; no uncertainty flag surfaced to user
Intermediate Confidence	0.65-0.84	Validated response returned with uncertainty flag	Match confidence score and canonical query surfaced; user confirms before organizational use
Low Confidence	< 0.65	Routes to LLM generation	Generated response optionally flagged for human review and potential VQR seeding

4.4 Instrumentation and Feedback Loops

Sustaining VQR quality over time requires instrumented pipelines capturing four signal types. Routing logs record which queries matched VQR at which confidence level and which routed to generation. User correction signals capture queries where users rejected or modified the returned response, indicating potential match failures or stale entries. VQR coverage metrics track the proportion of high-stakes queries in each domain receiving validated responses. Revalidation cycle completion rates measure the proportion of flagged entries revalidated within governance-defined windows (Balaguer et al., 2024). These signals feed governance dashboards enabling stewards to identify coverage gaps, detect emerging query types requiring VQR entry, and track staleness management performance against defined targets. Without them, stewards are managing a repository they cannot see -- and coverage degradation is invisible until a high-stakes query returns an outdated response.

V. GOVERNANCE FRAMEWORK FOR VQR OPERATIONS

5.1 Ownership and Accountability Structures

Three functional roles sustain VQR operations, each with distinct scope and accountability. Domain validators -- subject matter experts matched to specific domain scopes -- apply the four-criterion validation standard and approve entries within their domain authority. Repository stewards manage entry lifecycle, monitor staleness flags, track coverage metrics, and coordinate revalidation scheduling across domains. Platform engineers maintain matching infrastructure, manage threshold calibration, and ensure lookup latency does not degrade the NLP system's user experience (Das et al., 2023). The critical accountability gap in VQR deployments is typically the stewardship role: organizations assigning validation to domain experts but neglecting entry lifecycle stewardship accumulate stale entries without a governance mechanism to detect or remediate the degradation (Balaguer et al., 2024).

Stewardship failure has a recognizable signature in production deployments. Entry staleness flags accumulate without triggering revalidation because no one owns the scheduling function. Coverage metrics stop being reviewed because no one is accountable for coverage targets. High-stakes queries begin routing to LLM generation not because the VQR was retired but because the validated entries supporting them expired without replacement. The pattern resolves to a single organizational fact: technical deployment does not create the human accountability that sustained repository quality requires. Governance structures that are documented but unresourced perform identically, in practice, to governance structures that were never established.

5.2 Curation Quality Standards

Implemented as explicit validation checklists rather than applied through expert judgment alone, curation quality standards produce the consistency that makes VQR outputs defensible across validators and audit cycles. The four validation criteria -- factual accuracy, metric correctness, business rule compliance, and source

traceability -- are operationalised into structured checklists that validators complete for each candidate entry (Das et al., 2023). Checklist completion generates the audit trail that makes VQR outputs defensible in regulated environments: a finance analyst acting on a VQR response in a compliance filing can demonstrate that the response was validated by a named domain expert against documented criteria on a specific date, with verifiable source citations. Generatively produced responses cannot provide this provenance chain -- and in regulated environments, the inability to provide it is itself a compliance exposure.

5.3 Coverage Expansion and Cold-Start Management

Cold-start is the first operational test of VQR governance. A newly deployed VQR provides no retrieval benefit until curation workflows have processed sufficient high-stakes queries, and passive accumulation -- waiting for queries to surface organically and then validating them -- leaves the highest-risk query classes unprotected longest. Query log analysis conducted before deployment identifies the organization's most frequently posed high-stakes queries, prioritizing them for initial validation to bootstrap coverage where hallucination risk is highest and query volume justifies validation investment (Gao et al., 2024). Coverage targets -- defined as the proportion of high-stakes queries in each domain expected to be served by validated responses -- provide governance-visible progress metrics, with targets calibrated by domain risk level creating a prioritization framework aligning validation resources with organizational hallucination risk distribution.

5.4 Audit Trail and Compliance Applications

For regulated industries where demonstrating the provenance of a reported figure is a compliance requirement, provenance metadata accompanying every VQR response converts the repository from a reliability tool into a regulatory asset. Validator identity, validation timestamp, domain scope, source citations, and version history constitute an auditable chain of accountability for enterprise NLP outputs (Yang et al., 2025b). A financial institution demonstrating that every figure in its NLP-generated management reporting came from a human-validated VQR entry, with traceable sources, occupies a materially different compliance position than one relying on generative responses whose provenance cannot be traced beyond the model's training distribution. The organizational cost of maintaining that position is the governance infrastructure described in this paper -- a cost that compares favorably with the cost of a single material misstatement traceable to an unvalidated NLP output.

VI. DISCUSSION

6.1 VQR as Complement to RAG

RAG and VQR address different parts of the enterprise hallucination problem, and the two operate most effectively in combination. RAG enriches the context available to the generative model for novel queries, reducing hallucination rates on the broad, diverse query space where VQR coverage cannot feasibly be maintained. VQR provides deterministic validated responses for the specific, repeated, high-stakes query space where generative reliability is insufficient. In a mature deployment, both are active: RAG as the reliability floor for generative responses across the full query distribution; VQR as the guarantee for the high-stakes subset where the floor is not sufficient (Asai et al., 2024; Lewis et al., 2020). The limitation of treating them as substitutes is that neither alone addresses what the other does: RAG cannot guarantee a specific validated response for a repeated query; VQR cannot provide coverage for novel queries outside its entry set.

6.2 Organizational Readiness Requirements

VQR deployment succeeds where the governance infrastructure already exists to sustain it. Domain experts available for validation workflows, query log infrastructure identifying high-stakes candidates, governance processes for staleness management, and platform engineering capacity for matching infrastructure -- these conditions, not the technical deployment itself, determine whether a VQR delivers sustained reliability (Balaguer et al., 2024). Organizations that deploy VQR as a technical feature without establishing the organizational processes keeping it accurate will find the repository degrades faster than it grows. The technical deployment is the easier half of the implementation problem. Governance deployment is where most enterprise deployments encounter their most persistent and least-documented challenges.

6.3 Limitations

VQR effectiveness scales with the volume of repeated high-stakes queries in the organization's query distribution. Organizations whose high-stakes query patterns are highly diverse will see lower VQR coverage and proportionally lower reliability benefit. During the cold-start period -- before VQR coverage reaches operational thresholds in high-stakes domains -- the system provides no retrieval benefit for those queries, and

the governance burden of initial seeding falls entirely on human validators without the benefit of accumulated coverage to reduce their workload. The cost-benefit trade-off of human validation workflows, specifically the validator time required to sustain VQR quality against the reliability benefit delivered, was not empirically evaluated in this paper and will vary substantially across organizations depending on query volume, domain complexity, and existing governance maturity. That gap in the empirical record is a limitation that constrains adoption guidance.

6.4 Future Directions

Reducing the cold-start burden by bootstrapping the repository from content that has already undergone human validation in other contexts -- audit-certified financial reports, compliance-approved documentation, validated analytics outputs -- could accelerate initial coverage without requiring net-new validation investment (Mallen et al., 2023). Adaptive similarity threshold calibration, adjusting thresholds by domain and query type based on observed false positive and false negative rates, would improve retrieval precision without manual threshold management overhead. Federated VQR architectures enabling query-response sharing across organizational units or multi-tenant enterprise deployments, under appropriate data governance controls, would extend the coverage benefits of validation investment beyond single-organization deployment contexts (Izacard et al., 2023). Each direction reduces a specific friction point identified in the operational framework described in this paper; none of them eliminates the governance obligation that remains the condition of sustained VQR reliability.

CONCLUSION

Hallucination in enterprise NLP concentrates where it does the most damage: specific, verifiable, high-stakes queries posed with an expectation of accurate answers and acted on without independent verification. Probabilistic mitigation strategies reduce average hallucination rates across the full query distribution -- a genuine improvement that still leaves the narrowest, most consequential slice of the query space without a reliability guarantee. The organizational implication of this gap is not that better models are needed. It is that the architecture needs a different component for a different problem: one that does not generate for the query class that cannot tolerate the probability of generation failure.

What the governance argument and the technical argument share is a single insight: reliability in this context is not a model property. It is an organizational one. A VQR deployed without accountable curation, proactive staleness management, and coverage expansion disciplines will degrade as business facts evolve -- producing the same class of failure it was designed to prevent, only with a validated-entry label on the incorrect response. Organizations that build the governance infrastructure as a governed process with accountable owners, documented standards, and measurable coverage targets will sustain the reliability guarantees VQR is designed to provide. Those that treat governance as a configuration detail will not find that the architecture compensates for the absence. The audit trail, the validation workflow, the staleness detection mechanisms -- each is an organizational capability before it is a technical feature, and each is what makes VQR a structural reliability guarantee rather than another probabilistic bet.

REFERENCES

1. Asai, A., Wu, Z., Wang, Y., Sil, A., & Hajishirzi, H. (2024). Self-RAG: Learning to retrieve, generate, and critique through self-reflection. International Conference on Learning Representations (ICLR). https://arxiv.org/abs/2310.11511
2. Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., Drain, D., Fort, S., Ganguli, D., Henighan, T., Joseph, N., Kadavath, S., Kernion, J., Conerly, T., El-Showk, S., Elhage, N., Hatfield-Dodds, Z., Hernandez, D., Hume, T., ... Kaplan, J. (2022). Constitutional AI: Harmlessness from AI feedback. arXiv preprint. https://arxiv.org/abs/2212.08073
3. Balaguer, A., et al., (2024). RAG vs fine-tuning: Pipelines, tradeoffs, and a case study on agriculture. arXiv preprint. https://arxiv.org/abs/2401.08406
4. T Schimanski et al., (2024). Towards faithful and robust evaluation of LLMs with hallucination-aware benchmarks. arXiv preprint. https://arxiv.org/abs/2402.08277
5. Z. Zhang (2022). A Survey of Active Learning for Natural Language Processing. Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. https://aclanthology.org/2022.emnlp-main.414.pdf

6. Yuhao Zhang, Zhongliang Yang, Linna Zhou (2026). Robust Uncertainty Quantification for Factual Generation of Large Language Models. arXiv. https://arxiv.org/html/2601.00348v1
7. Gao, Y., et al., (2024). Retrieval-augmented generation for large language models: A survey. arXiv preprint. https://arxiv.org/abs/2312.10997
8. Izacard, G., et al., (2023). Atlas: Few-shot learning with retrieval augmented language models. arXiv. https://arxiv.org/abs/2208.03299
9. Ji, Z., et al., (2023). Survey of hallucination in natural language generation. ACM Computing Surveys, 55(12), 1-38. https://doi.org/10.1145/3571730
10. Yu Wang, et al., (2024). RAG+: Enhancing Retrieval-Augmented Generation with Application-Aware Reasoning. Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing. https://aclanthology.org/2025.emnlp-main.1630.pdf
11. Lewis, P., et al., (2021). Retrieval-augmented generation for knowledge-intensive NLP tasks. arXiv. https://arxiv.org/pdf/2005.11401
12. Mallen, A., et al., (2023). When Not to Trust Language Models: Investigating Effectiveness of Parametric and Non-Parametric Memories. arXiv. https://arxiv.org/pdf/2212.10511
13. Ouyang, L., et al., (2022). Training language models to follow instructions with human feedback. arXiv. https://arxiv.org/pdf/2203.02155
14. Pan, L., et al., (2023). Fact-checking complex claims with program-guided reasoning. arXiv preprint. https://arxiv.org/pdf/2305.12744
15. Sun, Z., et al., (2023). Recitation-augmented language models. arXiv. https://arxiv.org/pdf/2210.01296
16. Wang, Y., et al., (2024) Self-knowledge guided retrieval augmentation for large language models. arXiv. https://arxiv.org/pdf/2310.05002
17. Yang, A., et al., (2025a). Qwen2.5 technical report. arXiv preprint. https://arxiv.org/pdf/2412.15115
18. Yang, Z., et al., (2025b). HotpotQA: A dataset for diverse, explainable multi-hop question answering. arXiv. https://arxiv.org/abs/1809.09600
19. Zhao, W. X., et al., (2023). A survey of large language models. arXiv. https://arxiv.org/pdf/2303.18223