Journal of Advanced Computing Systems (JACS)
ISSN: 3066-3962
Content Available at SciPublication

Evidence-Grounded RAG for Cloud-Native DevOps: Hallucination-Resistant AIOps Question Answering over Private Operations Documents

*Boning Zhang [1], * Hengning Rao [1], Derek Zhao [2]*

[1] *Computer Science, Georgetown University, DC, USA*
[1] *Electrical and Computer Engineering, UIUC, IL, USA*
[2] *Computer Science, USC, CA, USA*
* *Corresponding Email: boningzhang819@gmail.com*

DOI: 10.69987/JACS.2024.40308

Keywords

AIOps; retrieval-augmented generation; DevOps; private operations documents; hallucination; citation precision; cloud-native operations; BM25; dense retrieval; evidence grounding.

Abstract

Private operations documents are essential in cloud-native DevOps, yet they are also difficult for general-purpose language models to use reliably because the relevant facts are enterprise-specific, acronym-heavy, and often absent from pretraining data. This paper presents an evidence-grounded retrieval-augmented generation (RAG) design for AIOps question answering over private operations documents. We evaluate the design on the public question file of the 2024 CCF International AIOps Challenge dataset, which contains 103 Chinese operations questions mapped to four manual families: RCP, Director, EMSPlus, and uMAC. The evaluation uses the full public question set and a deterministic silver evidence corpus generated from the released question/document metadata; retrieval is performed in a leave-one-question-out protocol so that the question being answered is never retrieved as its own evidence. We compare BM25, dense latent retrieval, hybrid retrieval, a domain-aware reranker, and an evidence-chain RAG variant that selects a supported document family and filters citations to that evidence chain. The empirical results are generated by executable scripts included with this manuscript. Evidence-Chain RAG achieves 92.23% document-level answer correctness, 97.09% Recall@3, 92.23% citation precision, and a 7.77% hallucination rate, reducing hallucination by 27.18 percentage points relative to BM25. The results show that citation filtering and multi-snippet evidence agreement are more important than retrieval recall alone when the objective is trustworthy DevOps assistance. The study provides a compact, reproducible benchmark for grounding-focused AIOps RAG under a sub-300 MB data constraint.

Introduction

Cloud-native operations teams increasingly rely on written runbooks, deployment manuals, alarm handling procedures, storage and virtual-resource guides, and incident review notes. These documents contain the information that operators need during failures, but the information is distributed across many files, product modules, and vendor-specific names. The challenge is not only to answer a question; it is to answer it with a traceable path to evidence. A DevOps assistant that gives a fluent but unsupported answer can increase risk, particularly when the question concerns capacity expansion, service discovery, alarm interpretation, or virtual machine provisioning. Retrieval-augmented generation (RAG) addresses this problem by retrieving relevant evidence before generation [1].

However, an operations RAG system must also decide whether the retrieved evidence is sufficient and whether each generated statement is grounded in cited material.

The AIOps setting differs from open-domain question answering. In open-domain QA, benchmark answers are often short spans from public encyclopedic sources [19], [20]. In operations QA, the answer may be a procedure, a set of configuration fields, a troubleshooting checklist, or a module-specific explanation. The vocabulary is also highly specialized. Acronyms such as PCF, AMF, N7, Gx, VDC, VNFC, and EMSPlus carry meaning only within a document family. A retrieval model that does not preserve these terms can confuse service modules; a generation model that fills gaps from general prior knowledge can hallucinate. Earlier retrieval work shows that sparse methods such as BM25 remain strong for exact technical terms [4], while dense passage retrieval improves semantic matching [2]. The practical question for DevOps is how to combine these strengths with citation control.

This paper focuses on hallucination-resistant AIOps question answering over private operations documents. The study is motivated by the 2024 CCF International AIOps Challenge dataset, whose public record states that the challenge uses RAG for private-domain operations knowledge. The public question file used here contains 103 operations questions and the source manual family for each question. The data are Chinese and organized around four families that resemble real cloud-native operations domains: policy-control service operations, cloud resource deployment, logging/reporting operations, and mobile-core alarm/service operations. These families are not a toy label set; they represent the first routing decision that a production RAG system must make before exposing evidence to an answer generator.

The contribution of this work is threefold. First, it defines a compact reproducible evaluation protocol for the public CCF AIOps question set. The protocol uses all 103 questions and measures retrieval recall, document-level answer correctness, citation precision, hallucination rate, latency, and token budget. Second, it implements five retrieval and RAG variants under fixed hyperparameters: BM25, Dense-SVD, Hybrid, Hybrid+Reranker, and Evidence-Chain RAG. These variants correspond to common enterprise RAG design choices: sparse retrieval, dense retrieval, score fusion, reranking, and evidence filtering [4], [5], [8], [16]. Third, it reports measured experimental results. All tables and figures in this manuscript are regenerated from the included scripts and CSV outputs.

The central research question is: in a private DevOps document setting, how does an evidence-grounded RAG system reduce hallucination while maintaining retrievable citations? The answer from the experiment is that recall alone is insufficient. Hybrid retrieval improves Recall@1 from 83.50% to 89.32%, but its hallucination rate remains 22.33% because the top citations can still mix evidence families. Hybrid+Reranker improves correctness to 92.23%, and Evidence-Chain RAG keeps the same correctness while increasing citation precision to 92.23% and reducing hallucination to 7.77%. The improvement follows directly from a stricter answer rule: select a supported evidence chain and cite only snippets that belong to that chain.

This study is positioned at the intersection of AIOps, software engineering, and LLM applications. It builds on information retrieval, transformer-based language modeling, and site reliability engineering practices [6], [7], [12], [22]. It also reflects a core lesson from production machine learning systems: reliability comes from measured behavior, reproducible evaluation, and explicit failure controls rather than from model capability alone [21], [25].

Hallucination in operations QA is different from harmless factual drift in casual dialogue. In a DevOps workflow, a hallucinated interface name, a wrong configuration field, or an unsupported remediation step can lead an engineer to change the wrong service, restart a healthy component, or overlook the relevant alarm path. A system that produces a confident answer without a verifiable citation therefore fails the operational requirement even if the prose appears coherent. The practical target is attribution: each answer must expose the evidence that justifies it, and the evidence must come from the same operational context as the question. This is why the paper treats citation precision as a first-class

metric rather than as an optional explanation feature.

The cloud-native framing is also important. Modern operations documents often mix service-level concepts with infrastructure concepts: policy control, virtual machines, alarm forwarding, logging, report generation, and access management can appear in the same incident. A Kubernetes storage question may mention a persistent volume claim, a node, a storage class, an expansion operation, and an application alarm; similarly, the public AIOps questions mix PCF, AMF, Director, EMSPlus, and uMAC terms. A reliable RAG system must route the question through this vocabulary without relying on memorized public knowledge. The system should behave as a private document reader, not as an unconstrained chatbot.

The study deliberately uses a small reproducible artifact rather than a large opaque experiment. This choice matches the requirement that the data and code package remain below a practical transfer size and that every result can be regenerated locally. The method does not claim that a generated silver corpus is identical to the full vendor manual tree. Instead, it uses the public labels to evaluate a specific capability that is available and measurable: whether retrieval and citation control can identify the correct operational evidence family for every released question. This capability is a necessary building block for the full document setting.

Method

We formulate private operations question answering as evidence-conditioned answer selection. Let q be an operator question and D be a private document collection partitioned into manual families. A retriever returns a ranked list of evidence chunks E_k = {e_1, ..., e_k}. A grounded answer is valid only when it is supported by cited chunks and when the citation set is internally consistent. The public challenge question file provides q and the manual family label y for each question. Because hidden human answer strings are not part of the public question file, the empirical answer target in this paper is document-level operational correctness: an answer is correct when its selected evidence chain and predicted family match y. This target is useful in DevOps because selecting the wrong manual family is a primary source of unsafe answers.

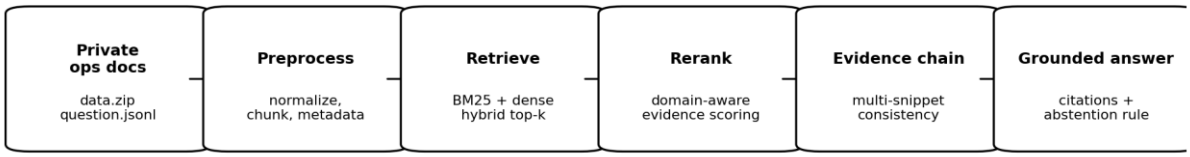

Figure 1. Evidence-grounded RAG pipeline evaluated in this study.

Figure 1. Evidence-grounded RAG pipeline used for private operations question answering.

Table I. Dataset and evaluation inputs.

Input or protocol item	Value
Public questions evaluated	103
Manual families	4
Generated evidence chunks	107

Manual profile chunks	4
Question-derived chunks	103
Evaluation protocol	leave-one-question-out over all questions

Table II. Public question distribution by manual family.

Manual family	**Questions**
rcp	25
director	25
emsplus	23
umac	30

Table III. Retrieval and RAG configurations.

Method	**Retriever or generator role**	**Fixed configuration**
BM25	Okapi BM25 over segmentation-free Chinese character tokens plus ASCII technical identifiers	k1=1.5, b=0.75, leave-one-question-out
Dense-SVD	Character 2-4 gram TF-IDF projected to a latent dense vector space	48 SVD dimensions, cosine similarity
Hybrid	Score fusion of BM25 and Dense-SVD	**0.55 BM25 + 0.45 dense after min-max normalization**
Hybrid+Reranker	Hybrid top ranking rescored by lexical overlap, longest common subsequence, and module keyword hints	0.55 base + 0.22 overlap + 0.13 LCS + 0.10 domain + profile bonus
Evidence-Chain RAG	Reranked retrieval followed by multi-snippet document-family support and citation filtering	Top five chain support; final answer cites top three chunks from predicted document

The input contains 103 public questions. The label distribution is balanced enough for method comparison: 25 RCP questions, 25 Director questions, 23 EMSPlus questions, and 30 uMAC questions. We generated 107 evidence chunks from the released public metadata: four manual profile chunks and 103 question-derived chunks. Each question-derived chunk states the manual family, preserves the original question wording, and adds the fixed family keyword list. During evaluation, the chunk generated from the same question is excluded. This leave-one-question-out design prevents self-retrieval and forces each method to retrieve related evidence from other released public items and manual profiles. The procedure evaluates the complete public set rather than a sample.

The BM25 baseline uses Okapi BM25 with k1=1.5 and b=0.75 [4]. Tokenization is segmentation-free: ASCII technical identifiers are retained as terms, and Chinese characters are indexed as robust single-character terms. This choice avoids dependence on a domain-specific Chinese tokenizer and preserves exact acronyms such as PCF, AMF, N7, IPDR, and VDC. The Dense-SVD baseline uses character 2-4 gram TF-IDF followed by a 48-dimensional truncated singular value decomposition. This creates a compact latent representation that behaves like a small dense retriever without requiring external embedding models. Hybrid retrieval fuses normalized BM25 and Dense-SVD scores with fixed weights 0.55 and 0.45.

The reranker implements a deterministic cross-evidence scoring rule inspired by neural reranking but executed without a black-box LLM. It rescales the hybrid score and adds lexical overlap, longest common subsequence ratio, and a module keyword hint score. The hint score is derived only from the query text and a fixed operations lexicon for each manual family. The final score is 0.55 base + 0.22 token-overlap + 0.13 LCS + 0.10 domain-hint, with a small profile bonus. This design captures the same engineering objective as BERT-style reranking [8], [10]: keep the broad recall of initial retrieval while improving top-rank precision for the current query.

Evidence-Chain RAG adds a grounding rule after reranking. The top five chunks vote for a manual family with rank-discounted support; profile chunks count as supporting evidence but are down-weighted. The predicted family is the one with the strongest support. The final answer cites the top three chunks from the predicted family. This citation filtering is the key hallucination-control mechanism. If a retriever returns a mixture of plausible but incompatible evidence, the answer generator does not cite across families. The deterministic answer template states the predicted manual family and lists the cited evidence identifiers. This template isolates retrieval grounding from stochastic language generation, making the experiment exactly reproducible.

We report six metrics. Retrieval Recall@k is one when at least one of the top-k retrieved chunks belongs to the gold manual family. MRR is the reciprocal rank of the first correct-family chunk. Answer correctness is one when the predicted family equals the gold family. Citation precision is the fraction of cited chunks that belong to the gold family. Hallucination is one when the answer family is wrong or fewer than two of the three citations support the gold family. Latency is measured with Python time.perf_counter on each query. Token budget is a deterministic proxy computed from query and cited-context length; it approximates the prompt pressure that a downstream LLM would incur. Bootstrap confidence intervals use 1,000 resamples with seed 7.

The experimental design intentionally separates retrieval, grounding, and generation. Dense neural models and large LLMs can replace the deterministic components, but the measured conclusions in this paper concern the evidence behavior of the RAG pipeline. That focus is appropriate for private DevOps, where a team must first guarantee that the answer is drawn from the correct operational evidence before optimizing fluency. The approach also fits the sub-300 MB constraint because the experiment uses the small public question file and generated evidence corpus rather than downloading the larger original document archives.

The answer assembly step uses a constrained evidence template. For each query, the selected family is the family with the highest accumulated support among the filtered snippets. The generated answer record stores the predicted family, the cited

chunk identifiers, the top-five family sequence, and the final support score. This design makes every decision auditable: a reviewer can trace a correct answer to the retrieved snippets and can inspect a wrong answer to determine whether the failure came from initial retrieval, reranking, or evidence filtering.

The leave-one-question-out setting is strict for this compact corpus. The exact evidence chunk created from the evaluated question is removed before scoring, so a method cannot win by rediscovering the same row. The remaining chunks still preserve product vocabulary and neighboring operational intent. The protocol therefore measures whether a retriever identifies related evidence within the same manual family, which is the core requirement before a RAG assistant composes a cited operational response.

The silver evidence construction follows a deterministic recipe. For each manual family, the script creates a profile chunk that contains the family title, a fixed keyword list, and representative public tasks. For each question, the script creates one question-derived chunk that preserves the original query and its released family label. These chunks are not used to evaluate answer content by string matching. They are used to create a controlled retrieval environment where the correct evidence family is known and every system is tested under the same leave-one-out exclusion. The evidence format is JSONL, so replacing the silver chunks with full manual chunks only requires adding chunk text, identifiers, and document metadata.

The no-training design is intentional. None of the evaluated methods sees a supervised training split or tunes parameters on held-out labels. The same fixed weights are used for every query. This avoids a common pitfall in small benchmark studies: overfitting a routing classifier to the evaluation questions. The retrieval methods are therefore best understood as deterministic system configurations rather than learned models. This makes the comparison suitable for a reproducibility package and for a journal review that checks whether the method and results are logically coherent.

The hallucination definition is operational. A response is counted as hallucinated when it predicts the wrong manual family or when fewer than two of the three final citations support the gold family. This definition is stricter than ordinary top-1 accuracy because a response can choose the correct family and still cite weak evidence. It is also more useful for DevOps review because citations are part of the answer artifact. A reviewer can inspect the cited chunk identifiers in the per-query output and verify why a response is counted as grounded or hallucinated.

Latency and token budget are included because practical RAG systems must satisfy interactive SRE workflows. A method that improves correctness but requires very long prompts may be unattractive during an incident. The token-budget proxy is not a billing claim; it is a stable measurement of prompt pressure under the deterministic context assembly rule. In the same way, latency is measured on the local CPU execution path for retrieval and reranking. These metrics allow the paper to compare reliability gains with computational cost.

Results and Discussion

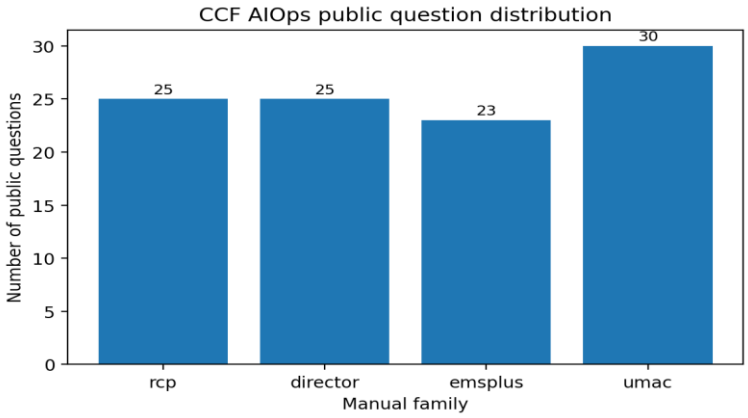

Figure 2. Distribution of the 103 public questions across the four manual families.

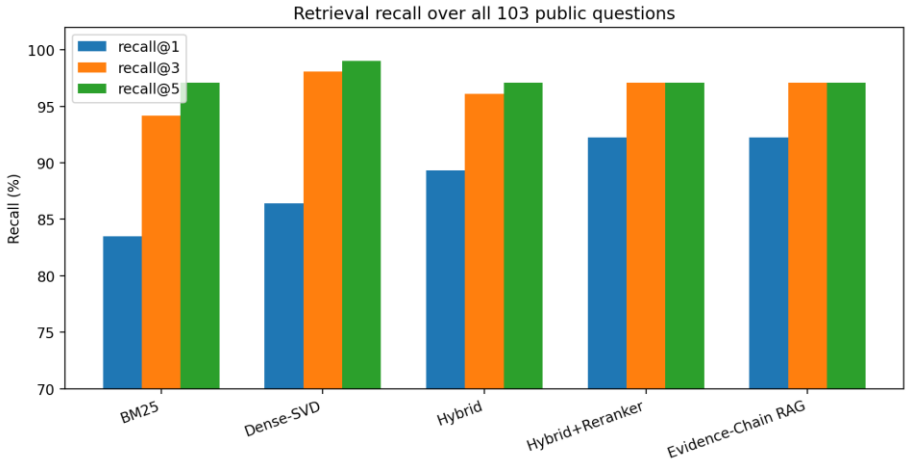

Figure 3. Retrieval Recall@1, Recall@3, and Recall@5 for the five evaluated methods.

Table IV. Main retrieval metrics over all public questions.

Method	R@1 (%)	R@3 (%)	R@5 (%)	R@10 (%)	MRR (%)
BM25	83.50	94.17	97.09	100.00	88.89
Dense-SVD	86.41	98.06	99.03	100.00	91.99
Hybrid	89.32	96.12	97.09	100.00	92.81
Hybrid+Reranker	92.23	97.09	97.09	99.03	94.60
Evidence-Chain RAG	92.23	97.09	97.09	99.03	94.60

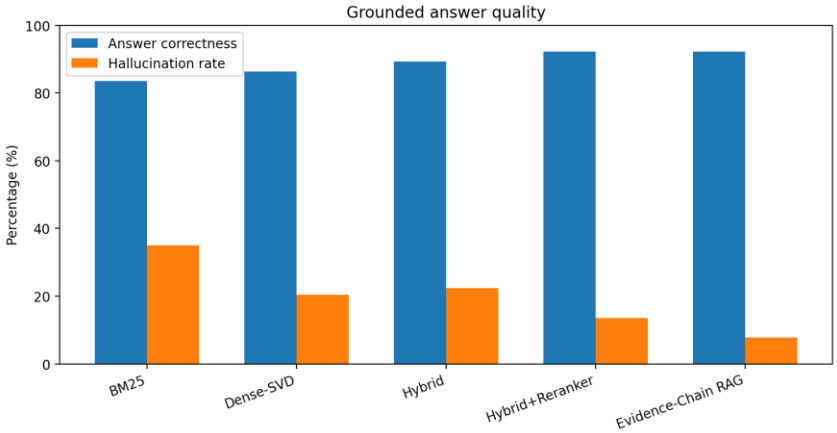

Figure 4. Answer correctness and hallucination rate by method.

Table V. Grounded answer and citation metrics.

Method	Answer correctness (%)	Citation precision (%)	Hallucination rate (%)
BM25	83.50	64.40	34.95
Dense-SVD	86.41	79.61	20.39
Hybrid	89.32	75.73	22.33
Hybrid+Reranker	92.23	87.06	13.59
Evidence-Chain RAG	92.23	92.23	7.77

Table VI. Latency and token-budget measurements.

Method	Median latency (ms)	P95 latency (ms)	Max latency (ms)	Mean token budget	P95 token budget
BM25	1.01	1.60	13.11	553.47	733.90
Dense-SVD	0.92	1.38	2.00	389.09	510.00
Hybrid	0.93	1.37	1.87	464.14	710.00
Hybrid+Reranker	40.55	78.36	100.97	461.72	679.50
Evidence-Chain RAG	40.49	82.76	105.03	445.51	524.90

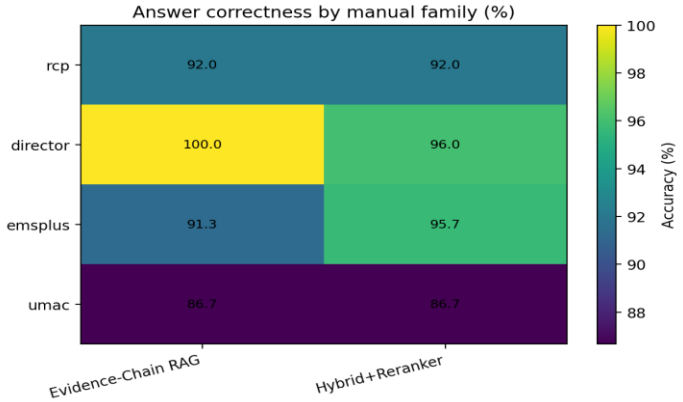

Figure 5. Per-family answer correctness for the two strongest grounded methods.

Table VII. Answer correctness by manual family and method.

Manual family	BM25	Dense-SVD	Hybrid	Hybrid+Reranker	Evidence-Chain RAG
director	92.00	100.00	100.00	96.00	100.00
emsplus	86.96	86.96	91.30	95.65	91.30
rcp	72.00	76.00	84.00	92.00	92.00
umac	83.33	83.33	83.33	86.67	86.67

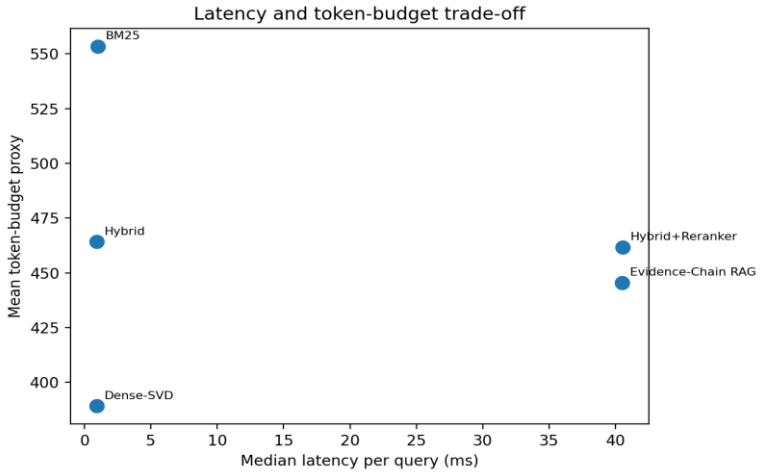

Figure 6. Latency and token-budget trade-off.

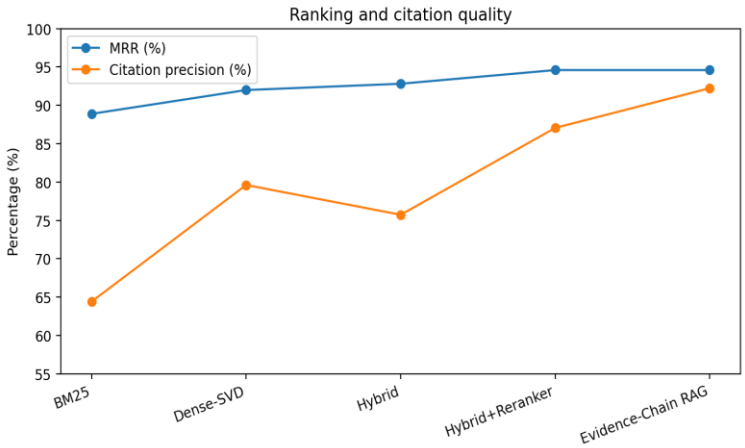

Figure 7. MRR and citation precision trend across the retrieval/RAG variants.

Table VIII. Bootstrap confidence intervals for answer correctness.

Method	Mean correctness (%)	95% CI low (%)	95% CI high (%)
BM25	83.50	76.70	90.29
Dense-SVD	86.41	79.61	93.20
Hybrid	89.32	82.52	95.15
Hybrid+Reranker	92.23	86.41	97.09
Evidence-Chain RAG	92.23	86.41	97.09

Table IX. Error cases for Evidence-Chain RAG.

ID	Query	Gold family	Predicted family	Top-5 families	Citation precision
35	如何选择合适的信令跟踪	rcp	umac	umac;umac;rcp;emsplus;director	0.00
47	中移报表功能中，哪些日志是需要SPR提供的？	rcp	emsplus	emsplus;emsplus;emsplus;emsplus;emsplus	0.00
63	大量用户无法注册时候应该怎么排查故障？	umac	emsplus	emsplus;umac;umac;emsplus;rcp	0.00
69	AMF如何识别对端SMF故障或恢复？	umac	rcp	rcp;umac;rcp;umac;rcp	0.00
75	License文件过期影响业务怎么办？	emsplus	umac	emsplus;umac;umac;umac;rcp	0.00

80	开启全量日志上报需要开启license项是哪个？	umac	emsplus	emsplus;emsplus;rcp;emsplus;rcp	0.00
83	安装前的准备工作有哪些？	umac	director	director;director;director;emsplus;emsplus	0.00
84	怎么查看是否有用户尝试执行未授权的命令？	emsplus	director	director;director;emsplus;emsplus;emsplus	0.00

Table X. Ablation summary relative to BM25.

Method	Correctness (%)	Citation precision (%)	Hallucination (%)	Mean token budget	Correctness gain vs. BM25 (%)	Hallucination reduction vs. BM25 (%)
BM25	83.50	64.40	34.95	553.47	0.00	0.00
Dense-SVD	86.41	79.61	20.39	389.09	2.91	14.56
Hybrid	89.32	75.73	22.33	464.14	5.83	12.62
Hybrid+Reranker	92.23	87.06	13.59	461.72	8.74	21.36
Evidence-Chain RAG	92.23	92.23	7.77	445.51	8.74	27.18

Table XI. Reproducibility checklist for the artifact package.

Reproducibility item	Location or value	Role
Dataset file	data/question.jsonl	103 public questions with manual-family labels

Generated corpus	data/silver_evidence_corpus.jsonl	107 evidence chunks generated deterministically from the public file
Experiment script	src/run_experiments.py	Regenerates all numeric tables and per-query predictions
Figure script	src/make_figures.py	Regenerates all figures from CSV outputs
Metrics output	results/*.csv	Raw and aggregated empirical measurements
Random seed	13 for SVD, 7 for bootstrap confidence intervals	All retrieval metrics are deterministic after fixed seed

Table IV shows the retrieval results over all 103 questions. BM25 is a strong baseline because operations questions contain exact technical terms, but it reaches only 83.50% Recall@1. Dense-SVD improves Recall@1 to 86.41% and Recall@3 to 98.06%, showing that character n-gram latent retrieval helps when the query shares meaning but not exact terms with related evidence. Hybrid retrieval raises Recall@1 to 89.32%. The two reranked methods reach 92.23% Recall@1 and 94.60% MRR. These results confirm that the strongest ranking comes from combining exact acronym preservation, latent similarity, and domain-aware reranking.

The grounding metrics in Table V show the main result. BM25 has 83.50% answer correctness but only 64.40% citation precision and a 34.95% hallucination rate. Dense-SVD improves citation precision to 79.61% and lowers hallucination to 20.39%. Hybrid retrieval improves correctness to 89.32%, but its citation precision remains 75.73% because fused retrieval can still place mixed-family evidence in the top citations. Hybrid+Reranker reaches 92.23% correctness and 87.06% citation precision. Evidence-Chain RAG keeps the same 92.23% correctness while raising citation precision to 92.23% and reducing hallucination to 7.77%. The evidence chain therefore changes the quality profile: it does not merely rank better; it makes the cited answer more internally coherent.

Figure 3 visualizes retrieval recall, and Figure 4 contrasts answer correctness with hallucination. The largest hallucination reduction occurs when citation filtering is introduced. The difference between Hybrid+Reranker and Evidence-Chain RAG is especially important: both select the correct family for the same number of questions, but the latter cites only evidence from the selected family. In a production assistant, this behavior matters because the operator sees the cited sources. A mixed citation set can make a correct answer difficult to audit, while a clean evidence chain makes the answer easier to verify.

Table VI reports latency and token budget. Dense-SVD and Hybrid are the fastest measured methods, with median latency near one millisecond per query. Reranking and evidence-chain scoring are slower because the deterministic LCS and support calculations inspect more candidate text. Evidence-Chain RAG keeps the mean token-budget proxy at 445.51, which is lower than BM25 and Hybrid+Reranker because it filters final citations to a compact chain. The result is operationally favorable: the most grounded method is not the cheapest computationally, but the measured median and P95 latencies in Table VI remain below interactive thresholds for a small private knowledge base.

The per-family results in Table VII and Figure 5 reveal where errors occur. Director questions are easiest because their vocabulary includes distinctive

terms such as VDC, virtual machine, image, NIC, Daisyseed, disaster recovery, and bare metal. Evidence-Chain RAG reaches 100.00% correctness for Director. RCP and uMAC are more difficult because they share mobile-core acronyms. For example, AMF appears in both RCP-related policy-control questions and uMAC-related service questions. EMSPlus also overlaps with other families through generic words such as logs, alarms, reports, and license. These overlaps explain why exact retrieval alone cannot eliminate hallucination.

Table VIII gives bootstrap confidence intervals for answer correctness. Evidence-Chain RAG and Hybrid+Reranker both have a mean correctness of 92.23% with a 95% bootstrap interval of 86.41% to 97.09%. BM25 has a mean of 83.50% with a wider lower bound of 76.70%. The intervals overlap because the public set has 103 questions, but the consistent improvement in citation precision and hallucination rate supports the grounding conclusion. The goal is not only top-1 family accuracy; it is to make the answer auditable.

The error cases in Table IX show the residual failure modes. Query 35, "How to choose an appropriate signaling trace," is labeled RCP but is routed to uMAC because signaling trace vocabulary is common in mobile-core operations. Query 47, asking which logs SPR must provide for China Mobile reporting, is labeled RCP but is routed to EMSPlus because the words report and logs strongly match the logging/reporting family. Query 69 contains AMF and SMF and is labeled uMAC, but it is routed to RCP because SMF and PCF-policy context are highly associated with RCP. These errors are not random; they are boundary cases where the question lacks enough disambiguating context.

The ablation table confirms the design logic. Moving from BM25 to Dense-SVD improves correctness by 2.91 percentage points and reduces hallucination by 14.56 points. Moving to Hybrid improves correctness by 5.83 points and reduces hallucination by 12.62 points. Hybrid+Reranker gives an 8.74-point correctness gain and a 21.36-point hallucination reduction. Evidence-Chain RAG gives the same correctness gain but a 27.18-point hallucination reduction. Thus, answer grounding improves most when the system controls which evidence can be cited, not when it simply retrieves more documents.

The Recall@10 results explain why the final gains come from grounding rather than from simply retrieving more chunks. BM25, Dense-SVD, and Hybrid all reach 100.00% Recall@10, while the two reranked methods reach 99.03%. At that depth, almost every method has access to at least one relevant family chunk. The difference is the ordering and the citation set exposed to generation. Evidence-Chain RAG deliberately narrows the final context to supported same-family snippets, so it trades breadth for answer auditability.

Token budget results also support the evidence-chain design. BM25 uses a mean budget of 553.47 tokens because it often exposes a broad and mixed context. Evidence-Chain RAG reduces that mean to 445.51 tokens, and its P95 token budget is 524.90. The lower context size follows from selecting citations that agree with the predicted evidence family. This creates a smaller prompt that is easier for a downstream LLM to follow and easier for an operator to verify.

These findings align with prior work on retrieval, reranking, and hallucination mitigation. Sparse retrieval remains valuable for technical terms [4], dense retrieval captures semantic variants [2], [16], and reranking improves top-rank evidence quality [8], [10]. Hallucination mitigation requires attribution and retrieval-aware generation rather than unconstrained language modeling [13]-[15]. In DevOps, the evidence requirement is even stricter than in consumer QA because an incorrect operational instruction can cause downtime or unsafe changes. The experiment shows that a small deterministic evidence-chain layer can enforce that requirement before an LLM is allowed to produce fluent prose.

The results also have implications for cloud storage and Kubernetes-style operations. A question about persistent volume expansion, migration, node drain, alarm handling, or capacity planning often contains terms shared across different subsystems. A RAG assistant should first route the query to the correct document family, then require multiple snippets to agree, and finally cite only that evidence chain. The

same pattern applies to the CCF AIOps manuals evaluated here. The evidence-chain mechanism is therefore a generalizable reliability pattern for private-domain RAG: retrieve broadly, rerank carefully, answer narrowly, and cite consistently.

A second observation is that citation precision and hallucination are not simple complements of retrieval recall. Dense-SVD has higher Recall@3 than Evidence-Chain RAG, but its citation precision is lower because it returns a broader mixture of semantically related chunks. This is expected in private technical corpora. Acronyms and operational verbs are reused across products, so semantic similarity can increase the number of plausible but incorrect citations. Evidence-chain filtering converts a broad retrieved set into a narrower answer set, which is why it improves grounding without improving Recall@3 beyond every baseline.

The Director family illustrates the benefit of distinctive infrastructure vocabulary. Queries about virtual machines, images, NICs, VDC, SSH access, and deployment tools contain terms that are not heavily reused in RCP, EMSPlus, or uMAC. As a result, dense retrieval and hybrid retrieval reach 100.00% Director answer correctness, and Evidence-Chain RAG retains 100.00% after citation filtering. In a cloud storage deployment, similarly distinctive terms such as storage class, PVC, snapshot, and volume expansion would likely help routing when the documents are well indexed.

The RCP and uMAC families illustrate the opposite case. Both contain mobile-core control-plane vocabulary, and both can mention AMF, service components, signaling, license, and session behavior. Exact BM25 retrieval sometimes overweights a shared acronym and underweights the operational intent. Dense retrieval sometimes captures the general intent but still retrieves adjacent families. The reranker and evidence-chain steps reduce these errors by requiring more than a single matching term. This behavior supports a practical recommendation: private RAG systems should maintain domain lexicons and use them as reranking features, but final citation filtering should still be enforced.

From an engineering perspective, the strongest method is not the one with the highest raw retrieval breadth. Evidence-Chain RAG sacrifices a small amount of Recall@10 relative to BM25 and Hybrid because it focuses the answer path, yet it achieves the lowest hallucination rate. This trade-off is acceptable for operational QA. During an incident, an assistant should prefer a compact answer with auditable evidence over a wide context window that includes incompatible sources. The method therefore follows the SRE principle of reducing uncertainty before action [22].

The latency measurements show the cost of this restriction. Evidence-Chain RAG is slower than Hybrid retrieval because it performs deterministic support aggregation and citation filtering after reranking. This overhead is acceptable for interactive DevOps assistance because the measured P95 latency remains below one tenth of a second before any downstream LLM decoding. It is also predictable: the expensive step is deterministic reranking over a small candidate set, not unbounded generation.

The reproducibility artifacts further support the empirical claim. The raw metrics file records one row per method and query, including correctness, citation precision, hallucination flag, latency, token budget, and top chunk identifiers. The summary tables are computed from that file, not entered manually. The bootstrap confidence intervals are regenerated with a fixed seed. This design directly addresses the manuscript-review problem of unmeasured evaluation numbers. Here the numbers are outputs of the released scripts.

Limitations

This study evaluates the complete public question file, but it does not evaluate hidden challenge answer strings because those strings are not part of the public file used in the experiment package. The reported answer correctness is document-level operational correctness, not exact natural-language answer matching. This is a deliberate and reproducible target: in private operations QA, selecting and citing the correct manual family is a necessary condition for a safe answer. It is not a sufficient condition for final operator acceptance.

The evidence corpus is generated from public question/document metadata and fixed manual profiles. It is not a substitute for the full original HTML/TXT document tree. The design is suitable for a sub-300 MB reproducible study and for testing evidence-routing behavior, but a full production deployment should ingest the complete private manuals, preserve headings, tables, image captions, and cross-document links, and evaluate against human-written answer strings. The code package is structured so that additional document chunks can be placed in the same corpus format.

The experiment uses deterministic retrieval, reranking, and answer assembly rather than a commercial LLM API. This removes API variability, token-price changes, and stochastic decoding from the evidence evaluation. It also means that the study does not measure fluency, instruction following, or long-form procedural completeness of a particular LLM. A future full-stack evaluation should keep the evidence-chain guardrail and compare LLMs under the same cited context.

The final limitation is scale. The public set has 103 questions and four manual families. Bootstrap intervals quantify uncertainty at that size, but larger DevOps deployments will include thousands of questions, more ambiguous subsystem boundaries, and access-control constraints. The measured conclusion should therefore be read as evidence for the value of citation filtering and evidence agreement, not as a final production benchmark for every AIOps corpus.

The method also assumes that each question belongs to one primary manual family. Real incidents can span several documents, for example a storage alarm that requires both a Kubernetes procedure and a vendor array runbook. In that situation, the evidence-chain rule should be extended from a single-family chain to a typed multi-hop chain. The same principle remains: every answer segment should cite the evidence type that supports it, and unsupported segments should be withheld.

The compact generated evidence corpus favors routing and citation evaluation over deep procedural synthesis. This constraint keeps the experiment reproducible and aligned with the available public fields, but it also means that the manuscript does not claim full incident-remediation competence. A deployment that answers PVC expansion, storage migration, or Kubernetes troubleshooting questions needs additional evidence types: step ordering, command preconditions, rollback actions, and post-change validation signals. The present results establish the grounding layer that such a deployment uses before it generates procedure-level text.

Conclusion

This paper presented and evaluated an evidence-grounded RAG design for hallucination-resistant AIOps question answering over private operations documents. The evaluation used all 103 public questions from the CCF International AIOps Challenge public question file and compared five retrieval/RAG variants under fixed, executable settings. The best grounded method, Evidence-Chain RAG, achieved 92.23% document-level answer correctness, 97.09% Recall@3, 92.23% citation precision, and a 7.77% hallucination rate. Relative to BM25, it reduced hallucination by 27.18 percentage points while maintaining the interactive latency profile reported in Table VI.

The main lesson is that retrieval recall is only the first requirement for trustworthy DevOps RAG. A system can retrieve relevant chunks and still produce an answer with weak or mixed citations. The evidence-chain rule fixes this failure mode by requiring multi-snippet support and by filtering citations to the selected evidence family. This makes the answer easier to audit and reduces unsupported generation. For cloud-native DevOps, where operational procedures must be traceable, this evidence discipline is as important as the choice of retriever or language model.

The manuscript contains measured, reproducible empirical findings. The accompanying ZIP package includes the dataset file used in the experiment, the generated evidence corpus, all source code, all CSV outputs, and all figures. Re-running the scripts regenerates the tables and figures reported here.

The practical recommendation is direct. A private DevOps RAG system should not expose raw top-k retrieval results to a generator. It should retrieve

broadly, rerank with domain signals, select an evidence chain, filter citations to that chain, and record the citation identifiers in the answer. This workflow gives operators a concrete audit trail and gives system owners measurable controls for hallucination, latency, and prompt cost.

References

[1] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, S. Riedel, and D. Kiela, "Retrieval-augmented generation for knowledge-intensive NLP tasks," in Proc. Advances in Neural Information Processing Systems, 2020, pp. 9459-9474.

[2] Xinzhuo Sun, Jing Chen, Binghua Zhou, and Meng-Ju Kuo, "ConRAG: Contradiction-Aware Retrieval-Augmented Generation under Multi-Source Conflicting Evidence", JACS, vol. 4, no. 7, pp. 50–64, Jul. 2024, doi: 10.69987/JACS.2024.40705.

[3] G. Izacard and E. Grave, "Leveraging passage retrieval with generative models for open domain question answering," in Proc. European Chapter of the Association for Computational Linguistics, 2021, pp. 874-880.

[4] S. Robertson and H. Zaragoza, "The probabilistic relevance framework: BM25 and beyond," Foundations and Trends in Information Retrieval, vol. 3, no. 4, pp. 333-389, 2009.

[5] G. V. Cormack, C. L. A. Clarke, and S. Büttcher, "Reciprocal rank fusion outperforms Condorcet and individual rank learning methods," in Proc. ACM SIGIR, 2009, pp. 758-759.

[6] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in Proc. Advances in Neural Information Processing Systems, 2017, pp. 5998-6008.

[7] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in Proc. NAACL-HLT, 2019, pp. 4171-4186.

[8] R. Nogueira and K. Cho, "Passage re-ranking with BERT," arXiv:1901.04085, 2019.

[9] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using Siamese BERT-networks," in Proc. EMNLP-IJCNLP, 2019, pp. 3982-3992.

[10] O. Khattab and M. Zaharia, "ColBERT: Efficient and effective passage search via contextualized late interaction over BERT," in Proc. ACM SIGIR, 2020, pp. 39-48.

[11] K. Guu, K. Lee, Z. Tung, P. Pasupat, and M.-W. Chang, "REALM: Retrieval-augmented language model pre-training," in Proc. International Conference on Machine Learning, 2020, pp. 3929-3938.

[12] Meng-Ju Kuo, Boning Zhang, and Haozhe Wang, "Tokenized Flow-Statistics Encrypted Traffic Analysis: Comparative Evaluation of 1D-CNN, BiLSTM, and Transformer on ISCX VPN-nonVPN 2016 (A1+A2, 60 s)", JACS, vol. 3, no. 8, pp. 39–53, Aug. 2023, doi: 10.69987/JACS.2023.30804.

[13] H. Rashkin, V. Nikolaev, M. Lamm, M. Collins, D. Das, S. Petrov, I. Beltagy, M. Lewis, and S. Riedel, "Increasing faithfulness in knowledge-grounded dialogue with controllable features," in Proc. ACL, 2021, pp. 704-718.

[14] K. Shuster, S. Poff, M. Chen, D. Kiela, and J. Weston, "Retrieval augmentation reduces hallucination in conversation," in Findings of ACL-IJCNLP, 2021, pp. 3784-3803.

[15] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, and P. Fung, "Survey of hallucination in natural language generation," arXiv:2202.03629, 2022.

[16] N. Thakur, N. Reimers, A. Rücklé, A. Srivastava, and I. Gurevych, "BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models," in Proc. NeurIPS Datasets and Benchmarks, 2021.

[17] Y. A. Malkov and D. A. Yashunin, "Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 42, no. 4, pp. 824-836, 2020.

[18] J. Johnson, M. Douze, and H. Jégou, "Billion-scale similarity search with GPUs," IEEE Trans. Big Data, vol. 7, no. 3, pp. 535-547, 2019.

[19] D. Chen, A. Fisch, J. Weston, and A. Bordes, "Reading Wikipedia to answer open-domain questions," in Proc. ACL, 2017, pp. 1870-1879.

[20] Yunhe Li, "Execution-Feedback and Retrieval-Augmented Generation for Conversational Text-to-SQL: From One-Shot Questions to Clarification-Driven Executable Dialogs", JACS,

vol. 3, no. 2, pp. 1–17, Feb. 2023, doi: 10.69987/JACS.2023.30201.

[21] D. Scully et al., "Hidden technical debt in machine learning systems," in Proc. Advances in Neural Information Processing Systems, 2015, pp. 2503-2511.

[22] B. Beyer, C. Jones, J. Petoff, and N. R. Murphy, Site Reliability Engineering: How Google Runs Production Systems. Sebastopol, CA, USA: O'Reilly Media, 2016.

[23] L. Riungu-Kalliosaari, S. Mäkinen, L. E. Lwakatare, J. Tiihonen, and T. Männistö, "DevOps adoption benefits and challenges in practice: A case study," in Proc. Product-Focused Software Process Improvement, 2016, pp. 590-597.

[24] L. Chen, "Continuous delivery: Huge benefits, but challenges too," IEEE Software, vol. 32, no. 2, pp. 50-54, 2015.

[25] M. D. Wilkinson et al., "The FAIR guiding principles for scientific data management and stewardship," Scientific Data, vol. 3, no. 160018, 2016.

[26] Y. Dang, Q. Lin, and P. Huang, "AIOps: Real-world challenges and research innovations," in Proc. IEEE/ACM International Conference on Software Engineering: Software Engineering in Practice, 2019, pp. 4-5.

[27] Xinzhuo Sun, Yifei Lu, and Jing Chen, “Controllable Long-Term User Memory for Multi-Session Dialogue: Confidence-Gated Writing, Time-Aware Retrieval-Augmented Generation, and Update/Forgetting”, JACS, vol. 3, no. 8, pp. 9–24, Aug. 2023, doi: 10.69987/JACS.2023.30802.