Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

TRIDOMRAG-BENCH: A Tri-Domain Benchmark for Evaluating Retrieval-Augmented Generation in Healthcare, Legal, and Finance

Abhinav Mishra

Abhinav06579@gmail.com

NIMS University

Vineet Mehan

NIMS University

Nilesh Bhosle

NIMS University

Research Article

Keywords: Retrieval-Augmented Generation, Hallucination Evaluation, Domain-Specific Benchmark, Faithfulness, Healthcare NLP, Legal AI, Financial NLP

Posted Date: June 2nd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9859942/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

TRIDOMRAG-BENCH: A Tri-Domain Benchmark for Evaluating Retrieval-Augmented Generation in Healthcare, Legal, and Finance

Abhinav Mishra
NIMS University Rajasthan, Jaipur, India
abhinav06579@gmail.com

Vineet Mehan*, Nilesh Bhosle*
NIMS University Rajasthan, Jaipur, India
mehanvineet@gmail.com, bhoslenp@gmail.com

***Abstract*—Retrieval-augmented generation (RAG) now powers most knowledge-intensive NLP pipelines, yet the frameworks used to evaluate these systems have not kept pace with the technology. A diversity of domains is treated as one-size-fits-all in current benchmarks, but it is not. For instance, the RAGAS metrics are generic and use natural language inference (NLI) scoring completely. It does not takes care of the numerical accuracy, which is a key consideration in financial establishments. In addition, there is no existing benchmark that assesses the same RAG system in healthcare, legal, and financial sectors through a single unified lens.**

In this paper, we introduce TriDomRAG, the first benchmark to assess RAG systems in these three important areas under a common protocol. We propose a composite score, the Domain Hallucination Score (DHS) which combines faithfulness, attribution, and terminology which assess these three characteristics and technical terminology via domain calibrated weights. We used 1200 question-answer pairs for our benchmark.

We draw our datasets from experts, such as PubMedQA (healthcare), Open Australian Legal QA (legal), and a merged FinanceBench and 10-K corpus (finance). With the use of GPT-4o-mini, we tested six different RAG configurations and observed that dense retrieval with chunk-level retrieval and rewriting yields the best results. Dense chunk retrieval achieves a DHS of 0.585, outperforming BM25 by 1.6 percentage points. Our results emphasize that real data is not something that can be compromised; e.g., substituting synthetic legal data with legal corpora from Australia improved scores for attribution from almost zero to 0.678. Lastly, we discuss the well-known shortcomings of conventional RAGAS metrics in the financial domain and present a numerical proximity extension of the DHS. All code and data have been released as open source at https://github.com/abhi6579/tridomrag-bench.

***Index Terms*—Retrieval-Augmented Generation, Hallucination Evaluation, Domain-Specific Benchmark, Faithfulness, Healthcare NLP, Legal AI, Financial NLP**

I. INTRODUCTION

Retrieval-Augmented Generation [1] has shifted the landscape of knowledge-intensive NLP. By grounding model outputs in retrieved evidence, RAG reduces reliance on parametric memorization. This grounding matters most in high-stakes professional settings, where a clinical tool that hallucinates a drug interaction, a legal assistant that misattributes case law, or a financial engine that fabricates earnings figures does not produce a "wrong answer" but it causes direct harm.

Despite the clear stakes, evaluation frameworks have struggled to keep pace. Current leaders include RAGAS [2], RAGChecker [3], and TruLens [4] rest on domain-agnostic assumptions that break down in specialized settings:

The Numerical Blindspot. RAGAS breaks answers into sentences and checks each one. This does not work for "$19.88 billion". There are no sentences to break. RAGAS returns no score. FinanceBench documented this problem [9].

The Jargon Trap. A doctor reads "MI" as a heart attack. A trader reads it as market impact. Generic embeddings blur these differences. They inflate retrieval scores while missing real errors.

The Evaluation Gap. Healthcare has MedRAG, Law has LegalBench, and Finance has FinanceBench. No single benchmark tests all three together. No one knows how a RAG system handles all three at once.

Motivating Example: RAGAS Failure on Finance
Query: *"What was Apple's net income in Q3 FY2023?"*
Context: *"Apple reported net income of $19.88 billion..."*
Gold answer: `$19.88 billion`
RAGAS faithfulness: *Unscored* (0 NLI sentence pairs extracted)
DHS score: 0.94

Fig. 1: RAGAS cannot score numerical answers. DHS addresses this via the numerical proximity scorer F_N (Eq. (2)) for Type-N answers.

II. CONTRIBUTIONS

- **C1 : TRIDOMRAG-BENCH Benchmark.** 1,200 QA pairs from PubMedQA, Open Australian Legal QA, and FinanceBench+financial-qa-10K. This is the first tri-domain RAG evaluation suite.
- **C2 : Domain Hallucination Score (DHS).** A metric that combines NLI faithfulness, numerical proximity for Type-N answers, BM25 attribution, and domain-terminology precision. Each domain has its own weights.

- **C3 : Multi-Generator Baseline Results.** The first cross-domain evaluation of six RAG configurations with two generators (GPT-4o-mini and Llama-3.3-70B) at a 1,200-sample scale. All results are public.
- **C4 : Failure Taxonomy.** Four failure types: wrong chunk, numerical hallucination, terminology mismatch, and attribution gap. We analyzed them across all configurations and domains.
- **C5 : Open Release:** All benchmark data, evaluation code, and baseline results are under the MIT license at https://github.com/abhi6579/tridomrag-bench.

III. RELATED WORK

A. Retrieval-Augmented Generation

Lewis et al. [1] introduced the foundational Retrieval-Augmented Generation framework. They showed that retrieving evidence makes QA more factual. Izacard and Grave [16] encoded each retrieved passage separately before fusion. RETRO [17] scaled retrieval to trillion-token databases. REPLUG [15] treats the language model as a black box. These advances are important but the evaluation has been stayed domain-agnostic. That is the gap we addressed.

B. RAG Evaluation Frameworks

RAGAS [2] uses an LLM as a judge. It measures faithfulness, answer relevancy, and context metrics. But its NLI faithfulness needs sentence answers. It cannot score numerical responses. The FinanceBench paper documents this limitation [9].

RAGChecker [3] gives claim-level diagnostics across 10 generic domains. It does not calibrate scores per domain. It does not handle numerical answers or terminology.

TruLens [4] monitors production systems. It checks context relevance and groundedness. No domain-specific components. ROUGE-L [14] and BERTScore [13] are lightweight baselines. They lack faithfulness decomposition and domain awareness.

ARES [5] trains domain-specific classifiers using synthetic data. But it needs per-domain training corpora. Multi-domain extension is expensive.

FaithJudge [6] benchmarks LLM-as-judge for faithfulness. It is rigorous but domain-agnostic. It does not handle terminology or numerical answers.

C. Domain-Specific Benchmarks

Healthcare. MedRAG [7] assembles five biomedical corpora with four retrieval methods. Retrieval quality directly affects diagnostic accuracy. PubMedQA [12] gives 1,000 expert-labelled QA pairs. We use it for our healthcare split.

Legal. LegalBench [8] spans 162 legal reasoning tasks. But it evaluates classification accuracy, not RAG faithfulness. Louis and Spanakis [23] show that legal term disambiguation is critical. This motivates our F4 (attribution gap) category. LexGLUE [22] shows that domain pre-training helps legal NLP.

Finance. FinanceBench [9] gives 150 QA pairs from SEC filings. LLMs achieve only moderate accuracy. Numerical reasoning is the limit. FinGPT [20] and BloombergGPT [21] show that domain pre-training helps financial NLU.

D. Hallucination and Faithfulness

Ji et al. [18] distinguish two types of hallucinations. Intrinsic hallucinations contradict the source. Extrinsic ones make unsupported claims. We identify a third type: terminological hallucination.

The answer is semantically faithful but uses the wrong domain vocabulary. "Profit" instead of "EBITDA." Generic metrics accept this. Professionals do not.

FActScore [19] decomposes text into atomic facts. Mündler et al. [24] find that larger LLMs are more prone to self-contradiction.

Gap Summary. No existing benchmark covers healthcare, law, and finance together. None handles numerical and entity answer types. None checks domain-terminology precision. None of them provides the structured cross-domain failure taxonomy. TRIDOMRAG-BENCHis built to fix all four.

IV. PROBLEM FORMULATION

Definition 1 (Domain-Specific RAG Evaluation)**.** *Let* $\mathsf{D} = \{\mathsf{D}_{\mathsf{med}}, \mathsf{D}_{\mathsf{leg}}, \mathsf{D}_{\mathsf{fin}}\}$ *be a set of domain corpora. A RAG system* R *has a retriever* $\mathsf{f}_r : \mathsf{Q} \rightarrow \mathsf{C}$ *and a generator* $\mathsf{f}_g : (\mathsf{Q}, \mathsf{C}) \rightarrow \mathsf{A}$. *We want an evaluation function* $\mathsf{E} : (\mathsf{R}, \mathsf{D}_i, \mathsf{q}, \mathsf{a}*) \rightarrow [0, 1]$ *that is (1) sensitive to the domain, (2) aware of answer type, and (3) comparable across* $\mathsf{D}_i \in \mathsf{D}$.

Answer Type Taxonomy: We define three types.

Type-T (Textual): Free-form sentence answers. Needs semantic entailment.

Type-N (Numerical): Numbers, percentages, financial figures. Needs range-tolerant scoring.

Type-E (Entity): Drug names, statute titles, and company identifiers. They need entity-normalized matching.

Why does this matter? Correctness is not uniform. In finance, a missing decimal point is catastrophic. In medicine, "Advil" and "ibuprofen" are the same. In law, citing the exact section is mandatory. We judge systems by the standards of their profession.

V. THE TRIDOMRAG-BENCH BENCHMARK

A. Design Principles

We built the benchmark around four principles.

First, every question must come from real professional materials. No synthetic data. Second, we keep the natural answer-type distribution of each source. Finance retains its 57.8% numerical questions because that is what makes finance different. Third, we picked sources that range from simple fact retrieval to complex reasoning. Fourth, every QA pair has a source document. This makes attribution scoring possible for every sample.

B. Data Sources

Healthcare – PubMedQA [12]. We use 400 QA pairs from the labeled split. Each has a retrievable evidence paragraph from a PubMed abstract. The dataset covers cardiovascular disease, oncology, pharmacotherapy, and endocrinology. It provides yes/no/maybe answers with long-form explanations.

Legal – Open Australian Legal QA [10]. We use 400 QA pairs from Hugging Face. Each is grounded in real Australian legislation and case law. This choice was experimentally decisive.

An early run used synthetic legal data. Attribution scores were near zero. Switching to the real Australian corpus raised attribution to 0.678. That is a 67.8-point improvement. Authentic legal corpora are not optional.

Finance – FinanceBench + Financial QA-10-K [9, 11]. We use all 150 QA pairs from FinanceBench (annual reports, 10-K, 10-Q filings). We add 250 pairs from financial-qa-10K (SEC EDGAR 10-K extracts). The combined 400 samples cover a broad range of financial reporting.

C. Benchmark Statistics

We assign answer types automatically. Answers containing only numbers (plus currency or percentage symbols) become Type-N. Answers that match named-entity patterns become Type-E. Everything else is Type-T. Table I shows the distribution.

TABLE I: Answer-type distribution per domain.

Domain	Total	Type-T	Type-N	Type-E
Healthcare	400	318 (79.5%)	42 (10.5%)	40 (10.0%)
Legal	400	276 (69.0%)	28 (7.0%)	96 (24.0%)
Finance	400	112 (28.0%)	231 (57.8%)	57 (14.3%)
Total	**1,200**	**706**	**301**	**193**

Finance is dominated by Type-N answers. That is why we added numerical proximity to DHS. Legal has the most Type-E answers (statute and case names).

D. Comparison with Existing Benchmarks

Table II compares TRIDOMRAG-BENCHto other resources. Ours is the only one with tri-domain coverage, numerical QA support, provenance annotation, and a unified metric.

TABLE II: Benchmark comparison.

Benchmark	Domains	Size	Num.	Prov.	Met.
RAGAS [2]	Generic	—	×	×	✓
RAGChecker [3]	10 gen.	4,000	×	×	✓
MedRAG [7]	Medical	3,560	×	×	×
LegalBench [8]	Legal	53k	×	×	×
FinanceBench [9]	Finance	150	✓	✓	×
FaithJudge [6]	Generic	2,000	×	×	✓
TRIDOMRAG-BENCH	**3**	**1,200**	✓	✓	✓

VI. DOMAIN HALLUCINATION SCORE (DHS)

A. Motivation

Standard faithfulness metrics break answers into atomic claims. They check each claim against the retrieved context using NLI. This fails for two reasons.

First, numerical answers. "$19.88B" produces zero NLI-decomposable sentence pairs. RAGAS returns no score. This is a documented problem for finance [9].

Second, terminological hallucination. An answer can be semantically faithful but use the wrong domain vocabulary. "Profit" instead of "EBITDA margin." "Heart attack" instead of "acute myocardial infarction." Generic metrics accept this. Professionals do not.

B. DHS Formulation

DHS is a weighted combination of three sub-scores:

$$\mathrm{DHS}(a, C, a^*, d) = \alpha_d \cdot F(a, C) + \beta_d \cdot A(a, C) + \gamma_d \cdot T(a, a^*, d) \quad (1)$$

Here a is the generated answer, C is the retrieved context, a^* is the gold answer, and d is the domain. F is faithfulness. A is attribution. T is terminology precision. The weights sum to 1 for each domain (Table III).

The weighting choices are based on a simple question: what kind of error is most professionally consequential in the domain? The most dangerous failure in healthcare is a generated answer that states something unsupported by the retrieved evidence; a hallucinated drug dosage by a clinician can cause direct harm. Faithfulness is therefore the most important factor ($\alpha = 0.40$). In legal reasoning, a valid answer must be traceable to a specific statute or precedent; a lawyer who cannot point to the provision they rely on has no argument. Thus attribution dominates ($\beta = 0.45$). In finance, analysts speak a common technical language, and a near-synonym is a material error. Precision of terminology is the most important ($\gamma = 0.45$). The empirical validity of these choices is demonstrated in Section XIII, where the configuration rankings are unaffected by ± 0.05 perturbations.

TABLE III: Domain-specific DHS component weights.

Domain	α_d (F)	β_d (A)	γ_d (T)
Healthcare	0.40	0.30	0.30
Legal	0.35	0.45	0.20
Finance	0.30	0.25	0.45

C. Faithfulness Score (F)

For Type-T answers, we use an NLI cross-encoder (`cross-encoder/nli-deberta-v3-base`). It scores entailment probability. This is stricter and more reliable than token-overlap F1.

For Type-N answers, we use numerical proximity:

$$F_N(a, a^*) = \max\left(0, 1 - \frac{|v_a - v_{a^*}|}{\epsilon_d \cdot v_{a^*}}\right) \quad (2)$$

Here v_a and v_{a^*} are the parsed numbers. ϵ_d is a domain tolerance: 0.05 for healthcare (clinical rounding), 0.01 for legal (exactness), 0.02 for finance (financial reporting). For Type-E answers, we use normalised string match after entity canonicalisation.

D. Attribution Score (A)

$$A(a, C) = \max_{c_i \in C} \text{BM25-sim}(a, c_i) \cdot \mathbb{1}[\text{src}(c_i) \neq \emptyset] \quad (3)$$

The indicator $\mathbb{1}[\text{src}(c_i) \neq \emptyset]$ penalizes answers where the retrieved context has no traceable source. This directly measures the F4 (attribution gap) failure mode.

E. Terminology Precision (T)

We keep a small terminology lexicon L_d for each domain. SNOMED-CT (104 terms) for healthcare. Black's Law Dictionary (89 terms) for legal. XBRL taxonomy (105 terms) for finance.

$$T(a, a^*, d) = \frac{|T(a) \cap T(a^*)|}{\max(|T(a)|, |T(a^*)|) + \epsilon} \quad (4)$$

Here $T(x)$ is the set of domain terms from L_d present in the text x. ϵ is a smoothing constant.

F. DHS Computation Algorithm

The entire computation is summarised in Algorithm 1. The CLASSIFY step is a deterministic rule applied to the gold answer a^*: answers with only numeric tokens are classified Type-N; answers matching named entity patterns are categorised Type-E; all others are Type-T. No trained model is required, and the rule is completely reproducible.

Algorithm 1 DHS Computation

Require: generated answer a, context C, gold answer a^*, domain d
Ensure: DHS $\in [0, 1]$
$\tau \leftarrow$ CLASSIFY(a^*) ▷ Type-T, N, or E
if τ = Type-T **then**
  $F \leftarrow$ NLI-ENTAILMENT(a, C)
else if τ = Type-N **then**
  $F \leftarrow F_N(a, a^*)$ ▷ Eq. (2)
else
  $F \leftarrow$ ENTITYMATCH(a, a^*)
end if
$A \leftarrow \max_{c_i \in C} \text{BM25-SIM}(a, c_i) \cdot \mathbb{1}[\text{src}(c_i) \neq \emptyset]$
$T \leftarrow |T(a) \cap T(a^*)| \,/\, \max(|T(a)|, |T(a^*)|)$
return $\alpha_d \cdot F + \beta_d \cdot A + \gamma_d \cdot T$

VII. EXPERIMENTAL SETUP

A. RAG Configurations

We tested six configurations. Two design factors.

Retrieval strategy: (i) BM25 [26] (sparse, $k_1 = 0.9$, $b = 0.4$); (ii) Dense using `BAAI/bge-base-en-v1.5` [25] (768-dim bi-encoder, FAISS index); (iii) Hybrid (linear combination of BM25 and dense, $\lambda = 0.5$).

Retrieval granularity:

(i) Chunk – Instead of feeding the whole document to the retriever, it cuts each document into overlapping 256-word pieces, where each new piece starts 32 words before the previous one ended. So if a document is 1000 words, you get many small overlapping windows. The overlap (stride) ensures that information sitting at the boundary between two chunks isn't lost.

(ii) Document – Here we fed the complete document without splitting; the full source document is indexed and retrieved as one unit. Simpler, but the retrieved context is much longer and may contain irrelevant content alongside the useful part.

The reason we test both is to see which granularity works better per domain. Chunk-level retrieval achieves the best overall average (DHS 0.585 for *dense_chunk* vs 0.580 for *dense_document*), but legal is the exception; document-level is competitive there because legal documents have cross-references between sections that get broken when you split them into chunks.

This gives six configurations: *bm25_chunk*, *bm25_document*, *dense_chunk*, *dense_document*, *hybrid_chunk*, *and hybrid_document*. We retrieve top-k = 5 passages per query.



Fig. 2: TriDomRAG-Bench evaluation pipeline. Five stages from dataset ingestion through DHS scoring to output analysis. Dense chunk retrieval (⋆) achieves the best overall DHS of 0.585.

B. Generators

We used two generators to test generator-agnosticism.

GPT-4o-mini (`openai/gpt-4o-mini`) is our primary generator. It runs on all six configurations (1,200 samples each).

Llama-3.3-70B-Instruct (meta-llama/llama-3.3-70B-instruct) runs only on *dense_chunk* (1,200 samples). This is enough to check if the rankings stay the same.

Both generators get the same prompt. It tells them to answer using only the provided context. If the answer is not there, respond "INSUFFICIENT CONTEXT." Temperature is 0 for reproducibility.

C. Implementation Details

The dense retriever loads once and is shared across all configurations. The full run uses 400 samples × 3 domains × 6 configs = 7,200 API calls. Runtime is about 3.4 hours. The NLI cross-encoder runs on GPU. All tests pass 20/20 automated unit tests before each run. Results are saved to `results/experiment_results.json`.

VIII. RESULTS

A. Main Results

Table IV shows DHS for all six configurations across all three domains. GPT-4o-mini generator. NLI faithfulness. 1,200 samples.

A few patterns stand out.

Dense beats BM25. dense_chunk has the best average DHS at 0.585. It beats bm25_chunk (0.569) by 1.6 points. The advantage is largest in finance (0.665 vs 0.638). Semantic retrieval helps with numerical financial context.

Chunk granularity is slightly better overall. dense_chunk (0.585) beats dense_document (0.580). But legal is different. hybrid_document (0.570) ties dense_chunk (0.569). Full documents help with long-range legal reasoning.

Finance has the highest DHS. Finance averages 0.653 across all configurations. It peaks at 0.665 with dense_chunk. This makes sense. SEC 10-K filings are well-structured and predictable.

Healthcare is the most consistent. Healthcare DHS ranges only 0.501 to 0.520. Variance is less than 0.02. This makes it a stable sub-benchmark for comparing systems.

Faithfulness is the bottleneck. F ranges from 0.209 to 0.470. Attribution (0.643–0.686) and terminology (0.742–0.829) are much stronger. Since Recall@5 is above 0.877 in all cases, the problem is generation faithfulness, not retrieval.

B. DHS Sub-Component Analysis

Table V breaks DHS down for the best configuration in each domain.

Faithfulness is the weakest in every domain. Attribution and terminology are strong. This means retrieval is working. The generator is the problem. It does not consistently present retrieved information faithfully.

Finance has the strongest faithfulness (0.470) because 10-K filings are structured. Healthcare has the weakest (0.233) because biomedical abstracts require complex inference.

IX. RETRIEVAL ANALYSIS

Table VI shows Recall@1, Recall@5, and MRR for all configurations. These are retrieval-only runs. No LLM calls.

Three observations.

First, all configurations achieve Recall@5 above 0.877. Yet DHS scores stay between 0.50 and 0.66. This gap means the bottleneck is generation, not retrieval.

Second, finance has the lowest retrieval scores. Recall@5 is 0.877–0.905. Healthcare is nearly perfect at 0.988–0.990. Legal is in between. Finance questions are semantically complex and often numerical.

Third, dense retrieval helps finance. It improves Recall@1 by 5.5 points over BM25 chunk (0.795 vs 0.740). This explains its DHS advantage in that domain.

X. GENERATOR-AGNOSTICISM VALIDATION

Does DHS work the same with different generators? We tested this by running Llama-3.3-70B on dense_chunk and comparing it to GPT-4o-mini. Table VII shows the results.

The scores are very close. The largest difference is 0.013 in healthcare. More importantly, the domain ranking is identical for both generators: Finance > Healthcare > Legal. This confirms that DHS captures domain-level patterns, not generator-specific artifacts.

We clarify the scope of this claim: we assert that the *domain ordering* is consistent across generators, not that DHS values are model-invariant constants. The absolute DHS values differ slightly between GPT-4o-mini (0.554) and Llama-3.3-70B (0.560), which is expected given different model architectures and training. The key finding is that which domain is hardest (Legal) and which is easiest (Finance) remains unchanged regardless of the generator choice.

Note that these NLI-based scores (0.554–0.560) are lower than the token-overlap DHS in Table IV (0.585). That is expected. NLI entailment is stricter. It gives a more conservative estimate of faithfulness.

XI. ABLATION STUDY

Does every DHS component actually matter? We tested this by removing one component at a time and renormalizing the remaining weights. Table VIII shows the results for dense_chunk.

Removing terminology costs 9.0 points on average. This is the largest drop. Domain vocabulary precision is not redundant with faithfulness or attribution.

Legal shows the biggest drop from removing attribution: 9.1 points. This makes sense. Statutory reasoning requires explicit source traceability.

Using faithfulness alone gives an average DHS of 0.328. F alone is not enough.

The w/o Faithfulness row scores higher (0.728) than the full metric. Why? Because faithfulness is the weakest sub-score. When you remove it and renormalize attribution and terminology, the score goes up. This confirms that improving faithfulness is the highest-leverage direction for future work.

TABLE IV: Full DHS results across all RAG configurations and domains.

		Healthcare			Legal			Finance			
Config	**Gran.**	DHS	F	Fail	DHS	F	Fail	DHS	F	Fail	**Avg DHS**
BM25	Chunk	0.509	0.217	193	0.561	0.265	147	0.638	0.414	180	0.569
BM25	Document	0.501	0.209	199	0.564	0.260	142	0.636	0.420	177	0.567
Dense	Chunk	**0.520**	**0.233**	191	**0.569**	**0.283**	141	**0.665**	**0.470**	153	**0.585**
Dense	Document	0.513	0.221	196	0.564	0.269	142	0.662	0.463	152	0.580
Hybrid	Chunk	0.514	0.223	191	0.559	0.268	147	0.657	0.455	161	0.577
Hybrid	Document	0.512	0.218	194	0.570	0.276	142	0.662	0.454	157	0.581

TABLE V: DHS sub-component scores for *dense_chunk*.

Domain	*F*	*A*	*T*	**DHS**
Healthcare	0.233	0.674	0.749	0.520
Legal	0.283	0.680	0.821	0.569
Finance	0.470	0.661	0.797	0.665

XII. STATISTICAL ANALYSIS

Table IX shows 95% confidence intervals for dense_chunk. We used the standard error approximation ($n = 400$ per domain, $\sigma \approx 0.15$).

The ranking of dense_chunk above all BM25 configurations is consistent across all confidence intervals. This suggests that the difference is not just sampling noise. We emphasize that these are *estimated* 95% confidence intervals using the standard error approximation ($\sigma \approx 0.15$, $n = 400$ per domain). This method assumes independence and approximate normality of the bounded DHS scores. Per-sample bootstrap confidence intervals would provide stronger statistical guaranties and are explicitly identified as future work in Section XIV.

XIII. FAILURE MODE ANALYSIS

We classify low-DHS evaluations into four failure types.

F1 (wrong chunk): The retrieved passage is irrelevant. All sub-scores are low.

F2 (numerical hallucination): The correct passage is retrieved, but the generator outputs the wrong number.

F3 (terminology mismatch): The answer uses vocabulary outside the domain lexicon.

F4 (attribution gap): The answer is plausible but cannot be traced to any retrieved source ($A < 0.2$).

Table X shows the counts for dense_chunk.

Attribution failure (F4) dominates healthcare (119 cases) and legal (92 cases). Even when the right context is retrieved, the generator adds unsupported claims.

Numerical hallucination (F2) dominates finance (95 cases). The model finds the right table but picks the wrong number.

Terminology drift (F3) is most common in healthcare (57 cases). Biomedical vocabulary is dense and precise.

A. Qualitative Failure Examples

Table XI shows concrete examples of the three most common failure modes.

XIV. DHS WEIGHT SENSITIVITY ANALYSIS

Are our conclusions sensitive to the specific weight choices? We tested this by shifting ± 0.05 between components while keeping the sum at 1. Table XI shows the average DHS rankings under each variant.

The ranking (dense_chunk > hybrid_document > dense_document > hybrid_chunk > bm25_chunk > bm25_document) is preserved. Our conclusions do not depend on the exact weights.

XV. DISCUSSION

Retrieval is not the bottleneck. All six configurations achieve Recall@5 above 0.877. Yet DHS scores stay between 0.50 and 0.66. This is our central finding. The gap is generation faithfulness. Improving generator instruction-following through fine-tuning or structured prompting will likely raise DHS more than any retrieval change.

No single configuration wins everywhere. dense_chunk has the best overall average (0.585). But hybrid_document ties for legal (0.570). A setup optimised for finance will behave differently in healthcare. Single-domain benchmarks cannot select configurations for multi-domain systems. This is why TRIDOMRAG-BENCHmatters.

Corpus authenticity is not optional for legal. Switching from synthetic to real Australian legal data improved attribution by 67.8 points. That is the strongest methodological finding in this study. Any prior legal RAG evaluation on synthetic data has likely underestimated attribution capabilities.

Finance is near a retrieval ceiling. Finance has the highest DHS (0.665) despite having the most numerical answers. SEC 10-K filings are structured and predictable. Further gains in finance will come from better generator faithfulness, not better retrieval.

Generator-agnosticism confirmed. Domain rankings are identical for GPT-4o-mini and Llama-3.3-70B. DHS captures domain properties, not generator artifacts. This is important for benchmark validity.

XVI. PRACTICAL GUIDELINES

Based on our results, here is what we recommend.

G1 : Prefer dense or hybrid over BM25. Dense retrieval beats BM25 by 1.6 to 2.8 DHS points across all domains. Make it your default choice.

G2 : Use chunk-level by default; use document-level for legal. Chunk-level gives the best overall average (0.585).

TABLE VI: Retrieval metrics across all configurations.

		Healthcare			Legal			Finance		
Config	**Gran.**	R@1	R@5	MRR	R@1	R@5	MRR	R@1	R@5	MRR
BM25	Chunk	0.965	0.988	0.973	0.868	0.970	0.907	0.740	0.877	0.797
BM25	Document	0.965	0.988	0.973	0.880	0.975	0.916	0.743	0.887	0.801
Dense	Chunk	**0.973**	**0.990**	**0.980**	0.875	0.955	0.907	**0.795**	0.890	**0.829**
Dense	Document	0.975	0.990	0.981	**0.887**	**0.963**	**0.914**	0.795	0.892	0.831
Hybrid	Chunk	0.968	0.990	0.977	0.850	0.960	0.895	0.745	**0.905**	0.813
Hybrid	Document	0.965	0.990	0.975	0.860	0.965	0.905	0.750	0.905	0.814

TABLE VII: Generator comparison on *dense_chunk* with NLI faithfulness.

Domain	**GPT-4o-mini**	**Llama-3.3-70B**	**Δ**
Healthcare	0.532	0.545	0.013
Legal	0.521	0.527	0.006
Finance	0.610	0.609	0.001
Overall	**0.554**	**0.560**	**0.006**

TABLE VIII: DHS ablation study for *dense_chun* .

Variant	**Healthcare**	**Legal**	**Finance**	**Avg**
Full DHS (F+A+T)	0.520	0.569	0.665	0.585
w/o Terminology (−T)	0.422	0.506	0.556	0.495
w/o Attribution (−A)	0.454	0.478	0.666	0.533
w/o Faithfulness (−F)	0.711	0.724	0.749	0.728
F only	0.233	0.282	0.469	0.328
A only	0.673	0.680	0.661	0.671

TABLE IX: 95% confidence intervals for *dense_chunk*.

Domain	**Mean**	**CI Low**	**CI High**
Healthcare	0.520	0.505	0.534
Legal	0.569	0.555	0.584
Finance	0.665	0.650	0.679

TABLE X: Failure mode counts for *dense_chunk*.

Domain	**F1**	**F2**	**F3**	**F4**
Healthcare	12	3	57	119
Legal	16	26	7	92
Finance	0	95	16	42
Total	**28**	**124**	**80**	**253**

Legal is the exception. Document-level retrieval is competitive there (0.570) because splitting legal documents destroys cross-references.

G3 : Do not use RAGAS as your only metric for finance. RAGAS cannot score numerical answers. Numerical answers are 57.8% of finance samples. Any RAGAS score on finance data only reflects the minority of sentence answers. It is not representative.

G4 : Enforce explicit attribution in legal RAG. F4 attribution failures accounts for 65.2% of all legal failures. Adding an instruction to cite the relevant statute section is cheap and likely effective.

G5 : Always test on all target domains before deployment. Single-domain results do not predict cross-domain performance. The variance across domains in this study proves this clearly.

XVII. LIMITATIONS

Single legal jurisdiction: Our legal split uses Australian law. US common law or EU civil law may have given different results.

Static terminology lexicons: SNOMED-CT, Black's Law Dictionary, and XBRL do not update automatically. New drugs, new laws, new accounting rules require manual updates.

Finance split composition: The finance split combines two sub-corpora with different question styles. We treated them uniformly. Sub-corpus analysis is future work.

Confidence interval method: We used standard error approximation, not per-sample bootstrap. The main experiment stores only domain-level averages. Storing per-sample scores would enable stronger statistical tests.

XVIII. FUTURE WORK

Multi-generator evaluation at scale. We ran Llama-3.3-70B only on dense_chunk. A full run across all six configurations with multiple generators (Mistral, Gemini) would show how faithfulness varies across model families.

Bootstrap confidence intervals. Storing per-sample DHS scores would enable bootstrap resampling and paired significance tests. This would give stronger evidence for the dense-vs-BM25 gap.

Dynamic lexicon maintenance. An automated pipeline to track SNOMED-CT updates, legislative amendments, and XBRL changes would keep terminology scoring valid without manual work.

Scale and multilingual extension. Expanding to 5,000+ samples and adding Hindi, German, and French corpora would make the benchmark useful for non-English professional NLP.

XIX. CONCLUSION

We built TRIDOMRAG-BENCH, the first benchmark that evaluates RAG systems across healthcare, legal, and finance under one protocol. The Domain Hallucination Score fixes two problems with existing metrics. It can score numerical answers, which RAGAS cannot. It checks domain-terminology precision, which generic metrics ignore.

We tested six retrieval configurations with two generators on 1,200 samples. Dense chunk retrieval works best, with an overall DHS of 0.585. Domain rankings are the same for GPT-4o-mini and Llama-3.3-70B, so the metric is generator-agnostic.

TABLE XI: Qualitative examples of failure modes.

Failure Type	Example
F2: Numerical Hallucination (Finance)	*Query:* "What was Microsoft's total revenue in FY2022?" *Retrieved context:* "Microsoft reported revenue of $198.3 billion for FY2022..." *Model output:* "$185.2 billion" *Diagnosis:* Correct document retrieved, wrong number extracted.
F4: Attribution Gap (Healthcare)	*Query:* "What are the side effects of Lisinopril?" *Retrieved context:* "Lisinopril may cause dizziness, cough, and headache..." *Model output:* "Lisinopril may cause dizziness, nausea, and fatigue." *Diagnosis:* Model added "nausea and fatigue" not in context.
F3: Terminology Mismatch (Healthcare)	*Query:* "What is myocardial infarction?" *Retrieved context:* "Acute myocardial infarction occurs when blood flow decreases..." *Model output:* "Heart attack occurs when blood flow decreases..." *Diagnosis:* Correct meaning but lay term instead of clinical term.

TABLE XII: Config rankings under weight perturbations.

Variant	d_ck	h_dc	d_dc	h_ck	b_ck	b_dc
Original	0.585	0.581	0.580	0.577	0.569	0.567
Faith+0.05	0.565	0.561	0.560	0.557	0.550	0.547
Attr+0.05	0.593	0.588	0.588	0.584	0.576	0.574
Term+0.05	0.601	0.599	0.596	0.595	0.588	0.585
Ranking is identical across all variants						

All three DHS components matter. Removing terminology costs 9.0 points on average.

The central finding is that retrieval is not the bottleneck. Every configuration achieves Recall@5 above 0.877. But generation faithfulness scores stay between 0.22 and 0.47. The problem is how the generator uses what it retrieves. Improving faithfulness through fine-tuning or structured prompting is the most promising direction for raising DHS.

ACKNOWLEDGEMENT

We thank NIMS University Rajasthan, Jaipur, for supporting this work. We are also grateful to the creators of PubMedQA, Open Australian Legal QA, FinanceBench, and Financial QA-10K for making their datasets publicly available. All experiments, data analysis, results, interpretations, and conclusions are entirely original work by the authors. The authors have thoroughly reviewed, verified, and take full responsibility for all content in this manuscript. AI-assisted tools were used for languages and grammar checking. All benchmark data, evaluation code, baseline results, and experimental configurations are publicly available under the MIT license at https://github.com/abhi6579/tridomrag-bench.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflict of interest. This research received no external funding.

FUNDING STATEMENT

This research was conducted as part of an M.Tech degree requirement at NIMS University Rajasthan, Jaipur, India. No external funding was received. All computational costs were covered by the authors personally.

REFERENCES

[1] P. Lewis *et al.*, "Retrieval-augmented generation for knowledge-intensive NLP tasks," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, 2020, pp. 9459–9474.

[2] S. Es, J. James, L. Espinosa-Anke, and S. Schockaert, "RAGAS: Automated evaluation of retrieval augmented generation," in *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (EACL): System Demonstrations*, St. Julians, Malta, 2024, pp. 150–158.

[3] D. Ru *et al.*, "RAGChecker: A fine-grained framework for diagnosing retrieval-augmented generation," in *Advances in Neural Information Processing Systems (NeurIPS 2024), Datasets and Benchmarks Track*, 2024.

[4] TruEra, "TruLens: Evaluation and tracking for LLM experiments," 2023. [Online]. Available: https://trulens.org

[5] J. Saad-Falcon, O. Khattab, C. Potts, and M. Zaharia, "ARES: An automated evaluation framework for retrieval-augmented generation systems," in *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, Mexico City, Mexico, 2024, pp. 338–354.

[6] M. S. Tamber, F. S. Bao, C. Xu, G. Luo, S. Kazi, M. Bae, M. Li, O. Mendelevitch, R. Qu, and J. Lin, "Benchmarking LLM faithfulness in RAG with evolving leaderboards," in *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP): Industry Track*, Suzhou, China, 2025, pp. 799–811.

[7] G. Xiong, Q. Jin, Z. Lu, and A. Zhang, "Benchmarking retrieval-augmented generation for medicine," *arXiv preprint arXiv:2402.13178*, 2024.

[8] N. Guha *et al.*, "LegalBench: A collaboratively built benchmark for measuring legal reasoning in large language models," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[9] P. Islam *et al.*, "FinanceBench: A new benchmark for financial question answering," *arXiv preprint arXiv:2311.11944*, 2023.

[10] I. Srinivasan, "Open Australian Legal QA," Hugging Face Datasets, 2024. [Online]. Available: https://huggingface.co/datasets/isaacus/open-australian-legal-qa

[11] virattt, "Financial QA 10K," Hugging Face Datasets, 2023. [Online]. Available: https://huggingface.co/datasets/virattt/financial-qa-10K

[12] Q. Jin, B. Dhingra, Z. Liu, W. Cohen, and X. Lu, "PubMedQA: A dataset for biomedical research question answering," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Hong Kong, China, 2019, pp. 2567–2577.

[13] T. Zhang *et al.*, "BERTScore: Evaluating text generation with BERT," in *Proceedings of the 8th International Conference on Learning Representations (ICLR)*, 2020.

[14] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Proceedings of the ACL Workshop: Text Summarization Branches Out*, Barcelona, Spain, July 2004, pp. 74–81.

[15] W. Shi *et al.*, "REPLUG: Retrieval-augmented black-box language models," in *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, Mexico City, Mexico, 2024, pp. 8371–8384.

[16] G. Izacard and E. Grave, "Leveraging passage retrieval with generative models for open domain question answering," in *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, 2021, pp. 874–880.

[17] S. Borgeaud *et al.*, "Improving language models by retrieving from trillions of tokens," in *Proceedings of the 39th International Conference on Machine Learning (ICML)*, vol. 162, 2022, pp. 2206–2240.

[18] Z. Ji *et al.*, “Survey of hallucination in natural language generation,” *ACM Computing Surveys*, vol. 55, no. 12, pp. 1–38, 2023.
[19] S. Min *et al.*, “FActScore: Fine-grained atomic evaluation of factual precision in long form text generation,” in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Singapore, 2023, pp. 12076–12100.
[20] H. Yang *et al.*, “FinGPT: Open-source financial large language models,” *arXiv preprint arXiv:2306.06031*, 2023.
[21] S. Wu *et al.*, “BloombergGPT: A large language model for finance,” *arXiv preprint arXiv:2303.17564*, 2023.
[22] I. Chalkidis *et al.*, “LexGLUE: A benchmark dataset for legal language understanding in English,” in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, Dublin, Ireland, 2022, pp. 4310–4330.
[23] A. Louis and G. Spanakis, “A statutory article retrieval dataset in French,” in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, Dublin, Ireland, 2022, pp. 6789–6803.
[24] N. Mündler, J. He, S. Jenko, and M. Vechev, “Self-contradictory hallucinations of large language models: Evaluation, detection and mitigation,” *arXiv preprint arXiv:2305.15852*, 2023.
[25] S. Xiao *et al.*, “C-Pack: Packaged resources to advance general Chinese embedding,” *arXiv preprint arXiv:2309.07597*, 2023.
[26] 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.
[27] V. Karpukhin, B. Og˘uz, S. Min, P. Lewis, L. Wu, S. Edunov, D. Chen, and W.-T. Yih, “Dense passage retrieval for open-domain question answering,” in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020, pp. 6769–6781.