ENTERPRISERAG-BENCH: A REALISTIC BENCHMARK FOR CROSS-DOCUMENT AND MULTIMODAL RAG OVER MEDIUM-SCALE CORPORA

Anonymous authors
Paper under double-blind review

ABSTRACT

Retrieval-augmented generation (RAG) evaluation has focused heavily on short-context question answering, synthetic retrieval tasks, and single-document reasoning. These settings underrepresent realistic enterprise workloads, where systems must retrieve, preserve, and synthesize evidence across long documents, noisy scans, audio, presentations, and scientific visual media. We introduce **EnterpriseRAG-Bench**, a medium-scale benchmark for end-to-end evaluation over one heterogeneous corpus. The current release contains 325 scored questions in five 65-question suites: factual, comparison, analytical, multi-hop, and multimodal. Of these items, 303 are answerable and 22 are unsupported-by-design abstention probes. The multimodal suite contributes 45 evidence-supported questions over presentation video and slide content, scanned handwriting, audio, and scientific video/frame analysis, plus 20 abstention probes. The benchmark combines human-verified gold answers with semantic judging, deterministic factual-anchor checks, source annotations, and media-specific audit metadata. Eight systems now have results across all five suites, including three fixed-pipeline Open WebUI baselines. The results expose substantial variation by task and access regime: VaultIQ has the highest five-suite mean at 67, Claude Opus 5 has the highest simple-factual score at 88, VaultIQ and Codex 5.6 Sol tie on multi-hop at 62, and VaultIQ has the highest multimodal score at 72. The fixed-pipeline Open WebUI study also reproduces prior findings that generator scale alone is not a reliable proxy for end-to-end RAG performance. We release the benchmark specification, scoring rubric, source manifest, and evaluation harness to support reproducible extension.

Keywords: Retrieval-augmented generation; RAG evaluation; multi-hop reasoning; multimodal RAG; cross-document question answering; enterprise search; benchmark design; grounded generation

1 INTRODUCTION

Retrieval-augmented generation (Lewis et al., 2020) has become a dominant pattern for grounding large language model outputs in external knowledge, yet evaluation remains fragmented despite rapid progress in retrieval, reranking, and long-context generation. Many commonly cited RAG benchmarks (Thakur et al., 2021; Wadden et al., 2020; Dasigi et al., 2021; Jin et al., 2019) emphasize small corpora, short passages, synthetic queries, single-hop QA, or claim verification against narrow evidence windows—settings that underrepresent the deployment environments in which enterprise RAG systems actually succeed or fail.

In realistic knowledge-work settings, a question often cannot be answered from a single passage or document: the system must identify evidence scattered across heterogeneous materials, preserve it through fusion and reranking, and synthesize it into a grounded answer. This introduces failure modes that existing benchmarks weakly capture: retrieving the right document family but collapsing onto one source before the answer stage; producing a plausible but one-sided comparative answer; losing critical evidence spans during context packing despite sufficient retrieval; or answering fluently despite incomplete or unsupported evidence.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

These are not edge cases: they arise repeatedly in real product evaluations of enterprise RAG systems, yet current public benchmarks rarely isolate them, and product-grade systems such as Google NotebookLM (Google, 2024) are often evaluated only in domain-specific or non-reproducible settings.

We present EnterpriseRAG-Bench, a benchmark for measuring RAG performance under realistic medium-scale, cross-document conditions. It is motivated by a concrete observation: product-grade evaluation requires measuring not just whether a relevant chunk was retrieved, but whether the system preserved and used the right evidence across the full retrieval-to-answer pipeline.

Our contributions are: (1) a 325-question benchmark organized into five 65-question suites spanning direct factual lookup, comparison, analytical synthesis, cross-document multi-hop reasoning, and multimodal evidence use; (2) a merged heterogeneous corpus whose document backbone spans 3,121 pages and whose multimodal extension includes presentation/slide content, scanned handwriting, audio, and scientific video/frame analysis; (3) an evidence-aware evaluation protocol that combines semantic answer judging with deterministic factual-anchor checks, source requirements, unsupported-information abstention, and media-specific audit metadata; (4) a shared evaluation harness in which the multimodal suite uses the same runner, evaluator, judge, comparison framework, and reporting path as the document suites rather than a separate benchmark stack; (5) a cross-system evaluation spanning managed RAG products, a self-hosted RAG application, direct-access agentic baselines, and an independently executed fixed-pipeline Open WebUI study; and (6) a released benchmark specification, source manifest, scoring rubric, and evaluation harness to support reproducible follow-on work.

2 Related Work

Existing RAG benchmarks cover open-domain QA, claim verification, and grounded answer generation. BEIR (Thakur et al., 2021) has been influential for retrieval evaluation, while RAGBench (Friel et al., 2024) and related frameworks incorporate answer generation and grounding metrics, but both emphasize passage retrieval or relatively clean corpora over cross-document synthesis. Scientific QA datasets such as SciFact (Wadden et al., 2020), Qasper (Dasigi et al., 2021), and PubMedQA (Jin et al., 2019) examine reasoning over scientific text but typically operate within a single paper or narrow domain ontology, which does not fully capture cross-document synthesis under heterogeneous corpora. Product systems such as NotebookLM (Google, 2024), Perplexity, and enterprise copilots support document-grounded questions, but hidden retrieval, orchestration, and lack of programmatic APIs make controlled, reproducible evaluation difficult. EnterpriseRAG-Bench complements, rather than replaces, these efforts by addressing a realistic deployment regime that remains underrepresented: medium-scale, heterogeneous, cross-document, multi-hop RAG.

LLM-based judging is also an established approach for scalable evaluation of open-ended model outputs. Zheng et al. (Zheng et al., 2023) report high agreement between strong LLM judges and human preferences while also documenting position, verbosity, self-enhancement, and reasoning biases. We use this work as methodological precedent, not as evidence that human-agreement results for a different judge model transfer to ours. Accordingly, our LLM judge is only one component of the protocol: gold answers, factual anchors, and required sources are human-verified, and deterministic factual-anchor and evidence-coverage checks reduce reliance on a single semantic judge score.

Prior RAG work also shows that generator scale alone does not determine end-to-end quality. Ram et al. (Ram et al., 2023) find substantial gains from retrieval and ranking across model sizes, and Guinet et al. (Guinet et al., 2024) report that retrieval-algorithm choice can produce larger gains than simply using a larger language model. More recent work explicitly studies corpus-generator tradeoffs and finds that smaller or mid-sized generators with stronger retrieval resources can rival larger generators (Ning et al., 2026). We therefore do not present non-monotonic performance with generator size as a new general scaling law. Instead, our independently executed Open WebUI experiment tests whether that established system-level behavior is reproduced under the EnterpriseRAG-Bench workload.

3 BENCHMARK DESIGN GOALS

We designed EnterpriseRAG-Bench around the following goals. **Realistic corpus scale:** a medium-scale corpus large enough that retrieval, reranking, and context selection matter, without reducing the task to web-scale retrieval. **Cross-document reasoning:** questions often require evidence from multiple documents or regions, such that a single retrieved chunk is insufficient. **Multi-hop structure:** questions require at least two reasoning steps, such as locating evidence in multiple sources and then comparing or synthesizing them. **Modality heterogeneity:** the same benchmark should exercise long-form text, technical documentation, scanned/handwritten pages, presentation slides, audio transcripts, and scientific visual media rather than treating each modality as a separate benchmark. **Realistic failure modes:** the benchmark surfaces evidence collapse, one-sided comparison, partial coverage, synthesis failure after correct retrieval, unsupported answer drift, and overconfident answering when the corpus does not support a response. **End-to-end accountability:** evaluation spans retrieval success, evidence preservation, answer correctness, and grounding/support quality, so failures are not collapsed into a single end score.

4 METHOD

The current EnterpriseRAG-Bench release runs against one merged heterogeneous corpus. Its original document backbone comprises six source works totaling 3,121 pages, covering literature and history (Tolstoy, 1869), medical informatics (Röhrig et al., 2021; Mantas et al., 2022), enterprise IT (Veritas Technologies LLC, 2025), scanned lab notes (Lang, n.d.), and postcolonial studies (Singh & Maheshwari, 2024). The source manifest extends that corpus with five additional media sources: one presentation video, one audio recording, and three scientific videos. The scanned laboratory notebook is reused as the handwriting source and is therefore counted once in the corpus rather than duplicated. The resulting benchmark exercises text, OCR, handwriting, audio-derived evidence, slide content, and scientific visual analysis within one corpus.

The benchmark contains 325 scored questions divided evenly across five 65-question suites: simple factual, comparison, analytical, multi-hop, and multimodal. Across the full benchmark, 303 items are answerable and 22 are unsupported-by-design abstention probes: 20 in the multimodal suite and 2 in the simple-factual suite. The 65-question multi-hop suite operationalizes cross-document reasoning directly: each item requires evidence from exactly two documents and a reasoning bridge that connects them. The multimodal suite contains 45 supported questions and 20 abstention probes. Supported cases cover presentation/slide reading, scanned handwriting, audio, and scientific video/frame analysis. Multimodal questions can include source IDs, modality metadata, question subtypes, and expected timestamp- or page-level citations. Three items that explicitly request a visualization also carry a deterministic `visual_expectation` contract. Additional repository folders support engineering diagnostics and regression testing and are not counted in the 325-question publication benchmark.

Each question includes a gold-answer rubric, factual anchors represented through `match_keywords` and optional concept aliases, and expected supporting sources. Human review verifies these annotations against the corpus at authoring time. The primary cross-system score combines a local semantic judge with deterministic factual-anchor coverage through the hybrid rule described below. Source coverage is reported separately when an answer artifact includes source evidence. For multimodal items, `expected_citations` records the timestamp or page and a short evidence description for audit; the current harness does not include those location annotations in the primary score. The narrower `visual_expectation` check is scored only on the three questions that explicitly request a structured visualization.

4.1 CURRENT BENCHMARK COMPOSITION

Table 1 summarizes the current five-suite release. The multimodal suite is an integrated fifth suite, not a separate benchmark: it uses the same execution, scoring, comparison, and reporting infrastructure as the four document-oriented suites. The full set contains 303 answerable items and 22 unsupported-by-design abstention probes.

Table 1: Current EnterpriseRAG-Bench composition. All five suites contain 65 scored items, for 325 total; 303 are answerable and 22 are unsupported-by-design abstention probes.

Suite	Questions	Primary evaluation focus
Simple factual	65	Direct grounded lookup + 2 abstention probes
Comparison	65	Balanced evidence across compared entities/sources
Analytical	65	Evidence-backed synthesis and interpretation
Multi-hop	65	Two-document relational reasoning and evidence preservation
Multimodal	65	45 supported media questions + 20 unsupported/abstention probes

4.2 Evaluation Protocol

4.2.1 End-to-End System Setting

Each evaluated system receives access to the benchmark corpus and a query and returns an answer. The current benchmark contains five 65-question suites, for 325 scored items. VaultIQ, Google NotebookLM, AnythingLLM, Claude Opus 5, Codex 5.6 Sol, and three Open WebUI baselines have answer artifacts for all five suites.

The system set spans two access regimes. VaultIQ, NotebookLM, AnythingLLM, and the Open WebUI baselines answer from retrieval indexes. Claude Opus 5 and Codex 5.6 Sol answer through command-line agents with direct filesystem access to the corpus and their own search, OCR, and frame-extraction tools. We therefore treat the latter as direct-access agentic baselines rather than like-for-like RAG pipeline comparisons. This distinction matters especially for simple factual lookup and multimodal tasks.

4.2.2 Scoring Dimensions

The benchmark records four evaluation dimensions plus one narrow multimodal contract. **Answer Correctness** measures semantic agreement with the human-verified gold answer through a local LLM judge. **Factual Anchor Coverage** measures deterministic coverage of required `match_keywords` or concept aliases after normalization. **Evidence Coverage** checks expected supporting sources when the answer artifact includes source evidence. **Grounding/Support Quality** captures whether the returned evidence supports the answer rather than merely appearing topically related; this remains partly diagnostic and cannot be compared uniformly when external answer artifacts lack equivalent evidence blocks. For three explicitly visual multimodal items, a deterministic **Visual Output Check** validates the requested structured visualization.

The primary score reported in the cross-system tables is the hybrid semantic/anchor score below. Evidence coverage, grounding diagnostics, failure labels, and multimodal location annotations remain available for audit and system diagnosis but do not silently change the primary score.

4.2.3 Hybrid Scoring Rule

The benchmark uses the LLM judge as the primary semantic score and applies a deterministic factual-anchor floor:

```
anchor_pct = matched_anchors / total_anchors * 100
combined_score = (llm_judge_pct if anchor_pct >= 50
                  else min(llm_judge_pct, 49))
```

The `50%` floor is applied consistently in the internal VaultIQ evaluator and the offline multi-system comparison path. Its purpose is narrow: a fluent response that omits most required anchors cannot receive a high primary score solely because the semantic judge is permissive. For unsupported-by-design items, the expected behavior is explicit abstention. An answer that relies on outside knowledge rather than the benchmark corpus does not satisfy the benchmark even when its factual content is correct.

Grading thresholds are:

Combined Score	Grade
$\geq 85\%$	Pass
$\geq 60\%$ and $< 85\%$	Marginal
$< 60\%$	Fail

4.2.4 MULTIMODAL EVIDENCE METADATA

Supported multimodal questions can carry `expected_citations` with timestamp- or page-level evidence. These annotations let a reviewer distinguish citing the correct media file from locating the correct moment or page. The current primary scoring path does not consume `expected_citations`; they are released as audit metadata. Three multimodal questions additionally carry `visual_expectation`, which is actively scored by checking the requested structured visualization against the declared chart type and values.

4.2.5 LLM JUDGE

The current evaluation harness uses a local Ollama-hosted judge model, "gemma3:12b-it-qat", with temperature set to 0. The judge receives the benchmark question, gold answer, and system answer. LLM-as-a-judge has prior empirical support as a scalable semantic-evaluation method, but published work also documents systematic biases (Zheng et al., 2023). We therefore pair semantic judging with deterministic anchors and preserve judge rationales for audit. The judge rubric is:

LLM Judge Score	Interpretation
95–100%	All facts match the gold answer
85–94%	Accurate with minor missing details
70–84%	Mostly accurate
50–69%	Partially correct
20–49%	Mostly wrong
0–19%	Completely wrong or fabricated

4.2.6 HUMAN AUDIT AND AMBIGUOUS CASES

Human review remains useful for ambiguous cases, near misses, OCR variants, and semantically correct answers that use unexpected wording. Benchmark construction verifies each gold answer, factual anchor, expected source, and unsupported probe against the corpus. Evaluation artifacts preserve the gold answer, system answer, anchor matches, judge rationale, expected sources, and failure labels so automated decisions can be audited. We treat calibration of this specific 12B judge against independent human ratings as future validation work.

4.2.7 EVALUATED SYSTEMS AND ACCESS REGIMES

The comparison intentionally spans different deployment regimes rather than attempting a controlled ranking of language models. VaultIQ represents an end-to-end enterprise RAG system (VaultIQ, 2026); NotebookLM represents a managed document-grounded product; AnythingLLM represents a self-hosted RAG application. Claude Opus 5 and Codex 5.6 Sol provide direct-access agentic baselines over the same corpus. Their filesystem access, tool use, and exhaustive search capability differ from retrieval-index-based RAG, so their scores should not be interpreted as isolated generator-model effects.

A separate external evaluation was executed by the HPC SYS AI Operations team at Los Alamos National Laboratory (LANL) in an independently configured Open WebUI RAG environment. The team used LANL AI Portal-hosted instruction-tuned models and generated benchmark-compatible answer artifacts for three generators in the current five-suite run: `mistral_7b`, `gemma-4-31b`, and `gpt-oss-120b`. Within this slice, retrieval and orchestration are held fixed while the generator varies. We use the completed five-suite Open WebUI results as an external benchmark-validation experiment.

5 RESULTS

The current benchmark contains 325 scored questions across five suites. Eight systems have audited answer artifacts for all five suites: VaultIQ, Google NotebookLM, AnythingLLM, Claude Opus 5, Codex 5.6 Sol, and three Open WebUI baselines. Because the evaluation includes different access regimes and does not yet provide matched repeated trials for every external system, all values below are descriptive point estimates rather than claims of statistical significance.

Table 2 reports the current suite scores and each system's five-suite mean. Figures 1 and 2 show the same latest result set as graph images.

Table 2: Primary combined scores for the current EnterpriseRAG-Bench evaluation. All rows cover all five suites, and the overall value is the five-suite mean.

System	Analytical	Multi-hop	Comparison	Factual	Multimodal	Overall	Suites
VaultIQ – `Qwen3_8b`	65	62	58	79	72	67	5
Claude Opus 5	53	59	54	88	49	61	5
Codex 5.6 Sol	59	62	54	73	44	58	5
Google NotebookLM	52	54	48	72	46	54	5
AnythingLLM – `Qwen3_8b`	61	46	55	46	27	47	5
Open WebUI – `gemma-4-31b`	49	47	45	49	40	46	5
Open WebUI – `gpt-oss-120b`	50	44	42	39	34	42	5
Open WebUI – `mistral_7b`	53	41	39	39	33	41	5

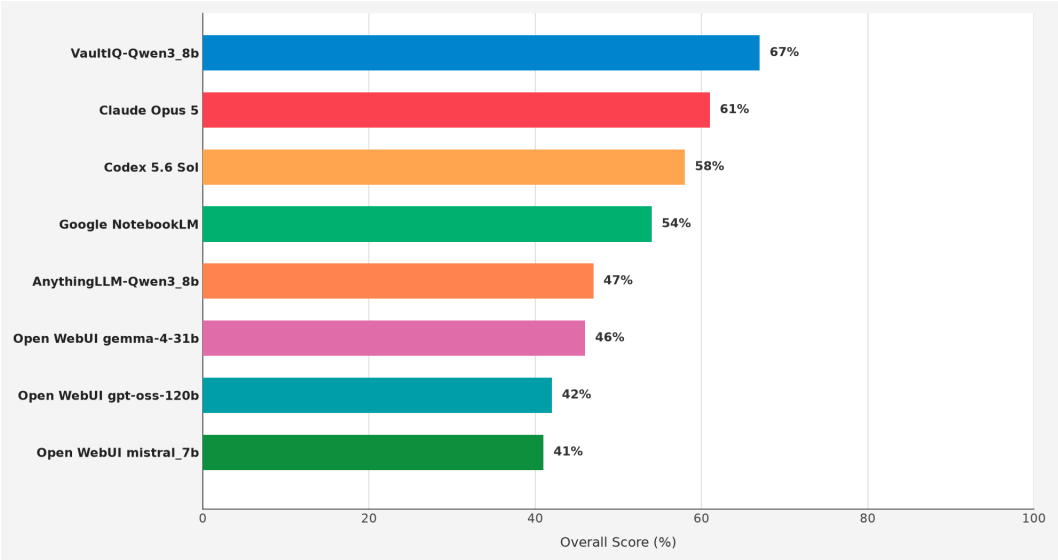

Figure 1: Overall five-suite mean combined score for the current EnterpriseRAG-Bench result set.

Among the eight systems evaluated on every suite, VaultIQ has the highest five-suite mean at 67. Claude Opus 5 follows at 61, Codex 5.6 Sol at 58, NotebookLM at 54, AnythingLLM at 47, Open WebUI `gemma-4-31b` at 46, Open WebUI `gpt-oss-120b` at 42, and Open WebUI `mistral_7b` at 41. The suite-level results are less uniform than the aggregate. VaultIQ has the highest analytical score at 65, the highest comparison score at 58, and the highest multimodal score at 72. VaultIQ and Codex 5.6 Sol tie on multi-hop at 62. Claude Opus 5 has the highest simple-factual score at 88, followed by VaultIQ at 79.

These differences show why EnterpriseRAG-Bench reports per-suite behavior rather than relying only on an overall mean. AnythingLLM is close to VaultIQ on analytical and comparison tasks but falls to 27 on multimodal. Claude Opus 5 is strongest on simple factual lookup but scores 49 on multimodal. Codex 5.6 Sol matches VaultIQ on multi-hop while remaining lower on multimodal. The access-regime distinction is important for interpreting these results: Claude Opus 5 and Codex 5.6 Sol search the corpus directly with agent tools rather than through a retrieval index.

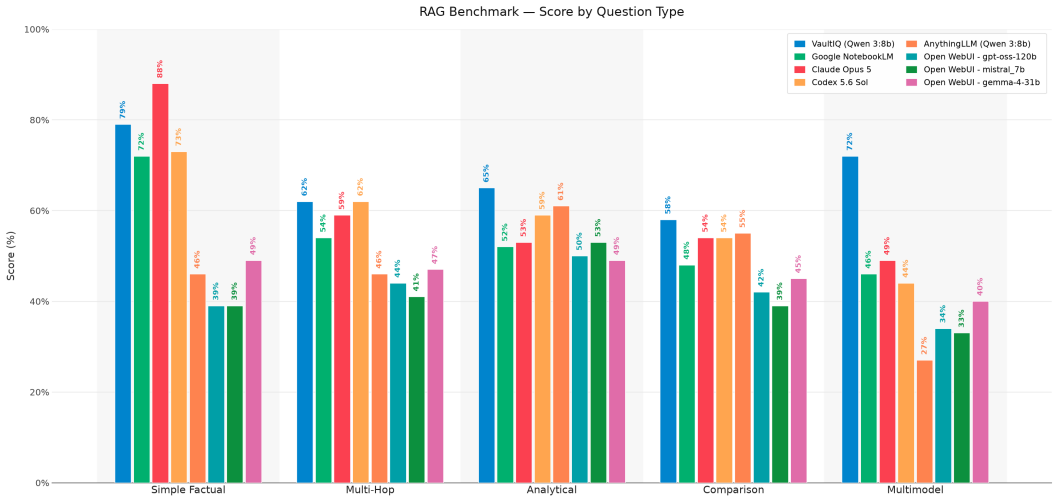

Figure 2: Primary combined score by question type for the current EnterpriseRAG-Bench result set.

The multimodal suite has the lowest mean score across the eight systems, but that observation does not establish that multimodal reasoning is intrinsically harder. Twenty of its 65 items are abstention probes, compared with two in simple factual and none in the other suites, and the evaluated systems do not all access media in the same way. One concrete limitation affects Claude Opus 5: its benchmark environment had no reachable speech-to-text model, and its saved answers report that limitation on 8 of the 9 audio-source items. We therefore treat its multimodal score as a lower bound for that evaluation environment rather than as a clean measurement of audio reasoning.

The Open WebUI comparison provides a separate controlled slice. Under fixed Open WebUI retrieval and orchestration, generator parameter count is not a monotonic predictor of benchmark score: `gemma-4-31b` has the highest five-suite mean among the three at 46, while the larger `gpt-oss-120b` scores 42 and `mistral_7b` scores 41. Prior work has already shown that retrieval and ranking choices can matter as much as, or more than, generator scale (Ram et al., 2023; Guinet et al., 2024; Ning et al., 2026). We therefore interpret the LANL result as an external replication of that qualitative behavior rather than as a new scaling law.

The multimodal answers also expose a failure mode that text-only evaluation cannot isolate cleanly: *text-channel substitution*. A system can return a fluent answer drawn from nearby or unrelated text without reading the requested slide, frame, handwritten page, or audio segment. Timestamp- and page-level annotations in the multimodal manifest make these cases auditable even though the current primary score does not yet include location-level citation correctness.

6 Discussion

EnterpriseRAG-Bench exposes failures that aggregate answer scores and retrieval metrics can hide. Performance changes substantially across direct lookup, analytical synthesis, explicit comparison, cross-document multi-hop reasoning, and multimodal evidence use. The current results also show that access regime matters: retrieval-index systems and direct-access agents solve the same questions under materially different search constraints. We therefore treat the benchmark as a system-level workload rather than as a pure language-model comparison.

The multimodal expansion extends the same end-to-end accountability across ingestion and evidence representation. A system must preserve source identity and usable evidence across slide content, handwriting, transcript-derived audio, and frame-level scientific visual analysis. The observed text-channel-substitution failures motivate location-aware evidence auditing: naming the correct media file is not sufficient when the answer depends on one slide, frame, page, or spoken segment.

Retrieval alone does not address all failures. The benchmark's design and error analysis target downstream evidence preservation, fusion, reranking, context construction, answer synthesis, and

abstention. The current cross-system table illustrates this directly: a system that performs well on direct factual lookup can still lose substantial accuracy on comparison or multimodal questions, and a system that performs well on analytical questions can fail when the relevant evidence lives outside extractable text.

The controlled Open WebUI comparison remains consistent with earlier findings that generator scale alone is not a reliable proxy for RAG quality (Ram et al., 2023; Guinet et al., 2024; Ning et al., 2026). Its value here is external benchmark validation: the same five-suite workload reproduces a known system-level behavior in an independently configured RAG stack while retrieval and orchestration are held fixed.

7 LIMITATIONS

EnterpriseRAG-Bench has several validity considerations.

Corpus scale and representativeness. The benchmark uses six long text documents plus five additional media sources, with the scanned notebook reused across text and multimodal evaluation. This remains modest relative to large enterprise repositories or web-scale retrieval, and results may not generalize to substantially larger or more repetitive collections.

Benchmark construction bias. The taxonomy and failure modes were informed by real RAG development workflows observed by the designers. This improves realism but can bias the benchmark toward failure modes visible in familiar systems. External annotators, additional corpora, and independent benchmark construction would strengthen validity.

Access-regime asymmetry. VaultIQ, NotebookLM, AnythingLLM, and Open WebUI answer from retrieval indexes, while Claude Opus 5 and Codex 5.6 Sol use direct filesystem access through command-line agents with their own search, OCR, and frame-extraction tools. Their results measure agentic corpus access rather than the same RAG pipeline setting, so direct numerical comparisons do not isolate model or retrieval quality.

Open WebUI generator coverage. The current Open WebUI slice includes three completed five-suite generators. This supports the fixed-pipeline validation analysis, but it does not exhaust the available model family or deployment settings.

Multimodal access limitations. The Claude Opus 5 benchmark environment had no reachable speech-to-text model and explicitly reports this limitation on 8 of the 9 audio-source items. Its multimodal score therefore understates what a configuration with audio transcription might achieve.

Abstention imbalance across suites. The multimodal suite contains 20 unsupported-by-design items, or 31% of the suite, while simple factual contains 2 and the other suites contain none. Cross-suite score differences therefore mix reasoning difficulty with different abstention burdens.

Media redistribution and reconstruction. The repository records source identifiers, hashes, categories, local paths, and evidence locations, but not every media source currently has a resolved public reconstruction URL. The multimodal suite therefore cannot yet be reconstructed from the repository alone. Third-party media must also satisfy redistribution or reference-by-URL requirements before a turnkey public release.

Judge-model bias. The primary score uses a local `gemma3:12b-it-qat` semantic judge. Deterministic decoding improves repeatability, not correctness. The factual-anchor floor, preserved rationales, source annotations, and human audit path reduce dependence on the judge, but a dedicated human-agreement calibration study remains future work.

Factual-anchor limitations. Anchor and concept matching helps prevent fluent answers that omit required facts from receiving high scores, but deterministic matching can penalize valid paraphrases or reward surface matches without full semantic support. We therefore use it as a guardrail on semantic judging rather than as a complete correctness metric.

Location-level multimodal grounding is not yet in the primary score. The benchmark records expected timestamps and page locations for multimodal evidence, but the current primary scoring path does not consume those annotations. They support audit today and leave room for a future location-aware grounding metric.

Run-level uncertainty. The current cross-system table combines the latest internal benchmark summaries with saved external answer artifacts and does not provide matched repeated trials for every system. We therefore report descriptive point estimates and avoid claims of statistical significance until repeated-run dispersion is available across the full system set.

Domain and modality imbalance. The benchmark spans heterogeneous domains and several modalities, but coverage is not balanced. The multimodal suite concentrates on one presentation, one scanned notebook, one audio source, and three scientific videos. Future versions should add more independent sources per modality and include native images, structured tables, simulation outputs, and additional audio/video domains.

8 CONCLUSION

We introduced the current EnterpriseRAG-Bench release, a 325-question benchmark for end-to-end evaluation over a medium-scale heterogeneous corpus. Five 65-question suites cover simple factual lookup, comparison, analytical synthesis, two-document multi-hop reasoning, and multimodal evidence use. Across the full benchmark, 303 items are answerable and 22 are unsupported-by-design abstention probes. The multimodal suite extends the same benchmark framework to presentation/slide content, scanned handwriting, audio, and scientific video/frame analysis while preserving the same scoring and reporting infrastructure.

The current five-suite results show substantial task-specific variation. VaultIQ has the highest mean across the eight systems with complete suite coverage at 67, leads analytical, comparison, and multimodal, and ties Codex 5.6 Sol on multi-hop. Claude Opus 5 leads simple factual lookup. These values describe system behavior under different access regimes rather than a controlled language-model ranking: six systems use retrieval indexes, while the two agentic baselines search the corpus directly with filesystem tools.

The independently executed LANL Open WebUI experiment provides a complementary validation signal. In its current five-suite run, retrieval and orchestration remain fixed while the generator varies. Generator size is not a monotonic predictor of benchmark score, consistent with prior work on the importance of retrieval and pipeline design (Ram et al., 2023; Guinet et al., 2024; Ning et al., 2026).

By combining cross-document reasoning, unsupported-information tests, multimodal evidence, an evidence-aware annotation scheme, and a shared end-to-end harness, EnterpriseRAG-Bench provides a testbed for studying whether systems preserve and use the evidence required for realistic enterprise questions. We release the benchmark specification, source manifest, scoring rubric, and evaluation harness to support independent replication and extension.

AI USE STATEMENT

In this work, we used generative AI tools to assist with copy-editing and language refinement of the manuscript text, and as an evaluated subject of study (the LLM-based systems and the local LLM judge described in Section 4.2). All AI-assisted writing was reviewed and edited by the authors. The benchmark corpus, question taxonomy, gold-answer rubrics, scoring protocol, and experimental analysis were designed, executed, and verified by the authors. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI. See the ICLR 2027 AI Policy for Authors for more details.

REPRODUCIBILITY STATEMENT

Section 4.2 documents the current five-suite, 325-question benchmark; the 303 answerable and 22 unsupported-by-design items; the primary semantic-plus-anchor scoring rule; the local judge and deterministic anchor floor; the distinction between retrieval-index and direct-access agentic baselines; the current five-suite Open WebUI results; and the multimodal audit metadata. The repository contains benchmark-compatible answer artifacts for the eight systems currently evaluated on all suites. The multimodal manifest records source identifiers, categories, hashes, and expected timestamp- or page-level evidence. These location annotations support audit but are not part of the current primary

score. Full third-party reconstruction of the multimodal suite remains subject to the media-source URL and redistribution constraints described in the Limitations section.

Ethics statement

The benchmark does not involve human subjects or intentionally include personally identifiable information. The document corpus and source metadata retain provenance information. Some multimodal sources require publication or redistribution verification, so the repository stores source metadata rather than committing all media bytes. Product and agentic baselines use externally accessible systems or locally executed tools and are reported with their access-regime differences. We note construction, judge-model, source-availability, and access-regime biases in the Limitations section and encourage external annotation and audit.

References

Pradeep Dasigi, Kyle Lo, Iz Beltagy, Arman Cohan, Noah A. Smith, and Matt Gardner. A dataset of information-seeking questions and answers anchored in research papers. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4599–4610. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.naacl-main.365.

Robert Friel, Masha Belyi, and Atindriyo Sanyal. RAGBench: Explainable benchmark for retrieval-augmented generation systems. *arXiv preprint arXiv:2407.11005*, 2024. doi: 10.48550/arXiv.2407.11005.

Google. NotebookLM goes global with slides support and better ways to fact-check. `https://blog.google/innovation-and-ai/products/notebooklm-goes-global-support-for-websites-slides-fact-check/`, 2024. Published 2024-06-06; accessed 2026-09-18.

Gauthier Guinet, Behrooz Omidvar-Tehrani, Anoop Deoras, and Laurent Callot. Automated evaluation of retrieval-augmented language models with task-specific exam generation. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 16773–16801. PMLR, 2024.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua 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)*, pp. 2567–2577. Association for Computational Linguistics, 2019. doi: 10.18653/v1/D19-1259.

Gregory I. Lang. Evolution 9: Bulk-fitness assays. Lang Lab Laboratory Notebooks, `https://www.glanglab.com/notebooks.html`, n.d. Public laboratory notebook; companion to Buskirk et al. (2017); accessed 2026-09-18.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, pp. 9459–9474. Curran Associates, Inc., 2020.

John Mantas, Arie Hasman, and Reinhold Haux (eds.). *Achievements, Milestones and Challenges in Biomedical and Health Informatics*, volume 300 of *Studies in Health Technology and Informatics*. IOS Press, 2022. doi: 10.3233/SHTI300.

Jingjie Ning, Yibo Kong, Yunfan Long, and Jamie Callan. Less LLM, more documents: Searching for improved RAG. In *Advances in Information Retrieval: 48th European Conference on Information Retrieval, ECIR 2026*, volume 16483 of *Lecture Notes in Computer Science*, pp. 598–613. Springer, 2026. doi: 10.1007/978-3-032-21289-4_38.

Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. In-context retrieval-augmented language models. *Transactions of the Association for Computational Linguistics*, 11:1316–1331, 2023. doi: 10.1162/tacl_a_00605.

Rainer Röhrig, Tim Beißbarth, Werner Brannath, Hans-Ulrich Prokosch, Irene Schmidtmann, Susanne Stolpe, and Antonia Zapf (eds.). *German Medical Data Sciences: Bringing Data to Life*, volume 278 of *Studies in Health Technology and Informatics*. IOS Press, 2021. doi: 10.3233/SHTI278.

Kundan Singh and Krishna Maheshwari. *Colonial Discourse and the Suffering of Indian American Children: A Francophone Postcolonial Analysis*. Palgrave Macmillan, Cham, 2024. doi: 10.1007/978-3-031-57627-0.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021.

Leo Tolstoy. War and peace. Public-domain electronic edition, Project Gutenberg, `https://www.gutenberg.org/ebooks/2600`, 1869. First published in book form in 1869.

VaultIQ. VaultIQ: Turn organizational knowledge into trusted answers. `https://www.vaultiq.ai/`, 2026. Accessed 2026-09-17.

Veritas Technologies LLC. *Backup Exec 25 Administrator's Guide*. Veritas Technologies LLC, 2025. Last published 2025-05-23; accessed 2026-09-18.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. Fact or fiction: Verifying scientific claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 7534–7550. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.emnlp-main.609.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems*, volume 36, 2023.