

Industrial Manual RAG Quality Prediction with Trace-Level Evidence Alignment on eManual

Chen Wei

Computer Science and Technology, Zhejiang University, Hangzhou, ZJ, China
weichen_zj@gmail.com

ABSTRACT

Enterprise retrieval-augmented generation (RAG) systems require more than answer-level accuracy: operators need to know whether a response is supported by retrieved evidence, whether the retrieved context is useful, and whether the generator actually uses that context. This study examines industrial manual RAG quality prediction through trace-level evidence alignment, using the eManual subset of RAGBench as the primary customer-support setting and FinQA as a finance-domain transfer case. The analysis follows the TRACe formulation of adherence, relevance, utilization, and completeness and compares the published test-split outputs of a GPT-3.5 judge, RAGAS, TruLens, and a fine-tuned DeBERTa-v3 judge. Rankings, absolute gaps, relative changes, and all visual summaries are recomputed from the benchmark matrix. On eManual, DeBERTa-v3 reaches 0.76 hallucination AUROC, 0.15 above the strongest prompt-based baseline, and reduces relevance RMSE from 0.17 to 0.13. On FinQA, it reaches 0.81 hallucination AUROC, a 0.24 absolute gain, while utilization RMSE improves from 0.13 to 0.10. The comparison also reveals important exceptions: GPT-3.5 has lower utilization error on eManual, and RAGAS has the lowest relevance error on FinQA. These results support a multi-signal evaluation layer in which aggregate scores are linked to sentence-level evidence records. For industrial support systems, trace supervision provides the clearest benefit for unsupported-response detection and enables targeted remediation of retrieval, evidence use, and answer grounding.

Keywords: *retrieval-augmented generation; RAGBench; eManual; FinQA; trace-level evidence alignment; hallucination detection; enterprise knowledge-base QA; TRACe; DeBERTa-v3; RAG evaluation.*

1. Introduction

Retrieval-augmented generation has become a common architecture for question answering over private and domain-specific corpora because it combines an evidence retriever with a language model that conditions its answer on the retrieved material [2]. Dense retrieval [3], retrieval-conditioned generation [4], self-reflective retrieval [5], and broader RAG design surveys [6] have established the technical foundation for this pattern. In an enterprise manual setting, however, a fluent answer is not sufficient. A support response must be tied to the correct instruction, warning, configuration step, or troubleshooting condition in the source document.

The distinction matters because manual-oriented RAG systems can fail at several points. The retriever may return a topically related but operationally irrelevant section; the generator may ignore the most useful passage; or the response may add a step that does not appear in the manual. These failures require different remedies. Retrieval errors call for changes to chunking, metadata filters, or reranking. Low evidence use points to prompt or generator behavior. Unsupported claims require grounding controls, abstention, or human review. A single answer-quality score cannot reliably separate these cases.

RAGBench was designed around this diagnostic need. It standardizes RAG examples across industry-relevant domains and introduces TRACe, a four-part framework that measures context utilization, context relevance, response adherence, and answer completeness [1]. The benchmark also records sentence-level evidence keys, which connect aggregate scores to the specific context and response units that produced them. This study centers on eManual, a customer-support subset built from a television manual, and uses FinQA as a contrasting domain in which questions depend on financial reports and numerical reasoning [14]. TechQA [16] and TAT-QA [15] further illustrate the broader importance of technical-document and hybrid table-text evaluation, but eManual and FinQA provide the most direct manual-versus-finance comparison within the available RAGBench test matrix.

Recent enterprise RAG studies reinforce the need for evidence-aware evaluation. Financial systems have paired retrieval with XBRL checks and numerical verification [22], [23], [29], [30]; cloud and operations systems

have used evidence chains, private-document grounding, and incident-oriented retrieval [24]-[27]; and safety-oriented work has combined hallucination detection with self-checking or selective refusal [31], [32]. These systems differ in domain and implementation, yet they share a practical requirement: the quality signal must identify not only whether an answer is wrong, but where the evidence path failed.

The present study asks three questions. First, how strongly do prompt-based and supervised evaluators differ in hallucination detection on industrial-manual and finance-domain data? Second, do the same evaluators show consistent advantages for context relevance and utilization? Third, how can sentence-key traces translate metric outputs into operational actions for a support system? The analysis contributes a focused eManual/FinQA comparison, a unified interpretation of the three reported TRACe metrics, and a trace-centered deployment framework that connects low scores to retriever, generator, and grounding interventions.

2. Literature Review

2.1 Retrieval-grounded question answering

The modern RAG pipeline developed from work on retrieval for knowledge-intensive natural language processing. Lewis et al. [2] coupled a neural retriever with sequence generation, while dense passage retrieval [3] showed that learned representations could support large-scale open-domain question answering. Fusion-in-Decoder [4] demonstrated that a generator can integrate evidence from multiple passages, and Self-RAG [5] extended this idea by learning when to retrieve and critique generated content. The RAG survey by Gao et al. [6] organizes these developments around retrieval, generation, augmentation, and evaluation, emphasizing that system quality depends on interactions among all components rather than on the language model alone.

Industrial manuals intensify this component interaction. Manual passages are often short, procedural, and condition dependent. Two sections may discuss the same feature while prescribing different actions for different operating states. A retriever therefore must capture instruction-level relevance rather than broad topical similarity, and a generator must preserve constraints such as sequence, prerequisites, and warnings. Technical QA resources such as TechQA [16] expose the difficulty of answering over specialized technical notes, while FinQA [14] and TAT-QA [15] show how evidence selection becomes more demanding when textual statements and quantitative tables must be interpreted together.

2.2 Automated evaluation of grounding and factuality

Automated RAG evaluation has evolved from general factuality benchmarks toward component-aware measures. TruthfulQA [10], FEVER [11], and TRUE [12] established complementary perspectives on truthfulness, verification, and consistency. Attributed question answering [13] made source attribution a direct evaluation target, and Adlakha et al. [9] studied correctness and faithfulness for instruction-following question answering. These lines of work motivate evaluators that compare claims against evidence rather than relying solely on reference-answer overlap.

RAGAS introduced reference-free metrics for faithfulness, answer relevance, and context relevance [7]. ARES used lightweight supervised judges and synthetic training data to evaluate retrieval-augmented systems [8]. TruLens similarly operationalizes groundedness and relevance through prompted feedback functions. RAGBench unifies these concerns in a common labeled schema and adds utilization and completeness, allowing evaluator outputs to be interpreted as retriever and generator diagnostics rather than as a single score [1].

The supervised evaluator in RAGBench builds on natural-language-inference encoders. BERT [19] established the bidirectional pretraining paradigm; MultiNLI [20] supplied broad-coverage inference supervision; DeBERTa [17] introduced disentangled attention; and DeBERTa-v3 [18] improved pretraining through an ELECTRA-style objective and gradient-disentangled embedding sharing. In RAGBench, token-level heads are attached to a shared DeBERTa-v3 encoder so that relevance, utilization, and adherence can be estimated within one model [1]. This architecture is especially suited to trace labels because it can associate predictions with individual evidence spans.

2.3 Evidence-grounded enterprise RAG

Enterprise applications increasingly combine RAG with execution checks, provenance, or domain-specific guardrails. Conversational text-to-SQL work has used execution feedback and clarification loops to make

retrieved schema knowledge actionable [21]. In finance, SEC-filing risk memos have been paired with XBRL verification [22], and evidence-grounded financial RAG has been used to reduce numerical hallucination in corporate risk narratives [23]. Related work on tokenized trade-receivable disclosures [30] and numerical-reasoning guardrails over SEC and FRED data [29] further shows that source support and numeric consistency must be evaluated jointly.

Operations and DevOps research provides a second cluster of relevant evidence. Multi-source root-cause attribution combines heterogeneous operational signals with language-model explanations [27]. Evidence-grounded RAG over private operations documents emphasizes citation precision and evidence-chain agreement [25], while OpsLLM applies bilingual retrieval to incident triage and alert summarization [26]. Long-document clause analysis extends the same concerns to contractual and insurance text, where retrieval quality depends on preserving document structure and cross-clause relationships [28].

Safety and auditability studies make the trace requirement explicit. Evidence-chain RAG has been proposed for word-level hallucination detection, source attribution, and provenance explanation [24]. A lightweight hallucination firewall combines evidence consistency, self-checking, and a smaller detector [31], while ambiguity-aware HDFS anomaly analysis joins retrieval-generated failure narratives with selective refusal [32]. Collectively, these studies support an evaluator that can expose relevant evidence, utilized evidence, and unsupported response content instead of returning an opaque confidence value.

2.4 Research gap

The literature establishes strong methods for retrieval, generation, factuality evaluation, and domain-specific guardrails, but two gaps remain for industrial manual QA. First, many evaluations collapse retriever and generator behavior into answer-level faithfulness or relevance. Second, evidence-grounded application studies frequently report domain-specific metrics that are difficult to compare across settings. A focused analysis of eManual and FinQA under the common TRACe schema clarifies which evaluator advantages are stable across domains and which are specific to the type of evidence being judged.

3. Method

3.1 Study design and data scope

The study is a comparative analysis of the RAGBench eManual and FinQA test-split evaluator matrix. The published benchmark values for GPT-3.5 judge, RAGAS, TruLens, and the fine-tuned DeBERTa-v3 judge are treated as the empirical inputs. A local analysis pipeline recomputes evaluator rankings, absolute score gaps, relative changes, and all figures from that matrix. Dataset profiles and file footprints are organized from the released RAGBench subset definitions. Figure 1 summarizes the evaluation path from a question and retrieved context to aggregate quality scores and sentence-key evidence records.

Trace-level RAG quality prediction workflow

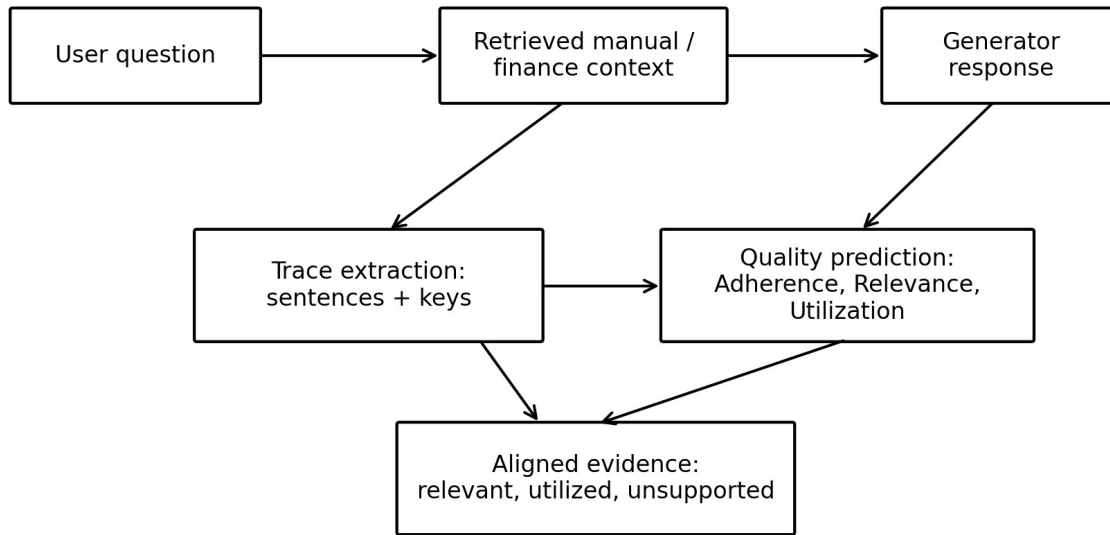


Figure 1. Trace-level RAG quality prediction workflow for industrial manual and finance-domain evaluation.

Table 1 defines the two subsets. eManual represents customer-support questions over a television manual and contains three retrieved documents per example with an average document length of 165 tokens. FinQA uses earnings-report evidence, also with three retrieved documents, but has longer documents and substantially larger train, validation, and test splits. This contrast supports a focused test of whether evaluator behavior transfers from procedural manual text to numerical finance evidence.

Table 1. Scope of the RAGBench subsets analyzed in this study.

Dataset	Domain	Document source	Question source	# docs	Avg. doc tokens	Train	Validation	Test
eManual	Customer support	TV manual	Annotator	3	165	1,000	132	132
FinQA	Finance	Earnings reports	Expert	3	310	12,000	1,700	2,200

The split imbalance is shown in Figure 2. eManual is intentionally compact, whereas FinQA supplies a larger transfer setting. The analysis does not pool the two datasets because their evidence structures and target distributions differ; each evaluator is ranked within each subset and metric.

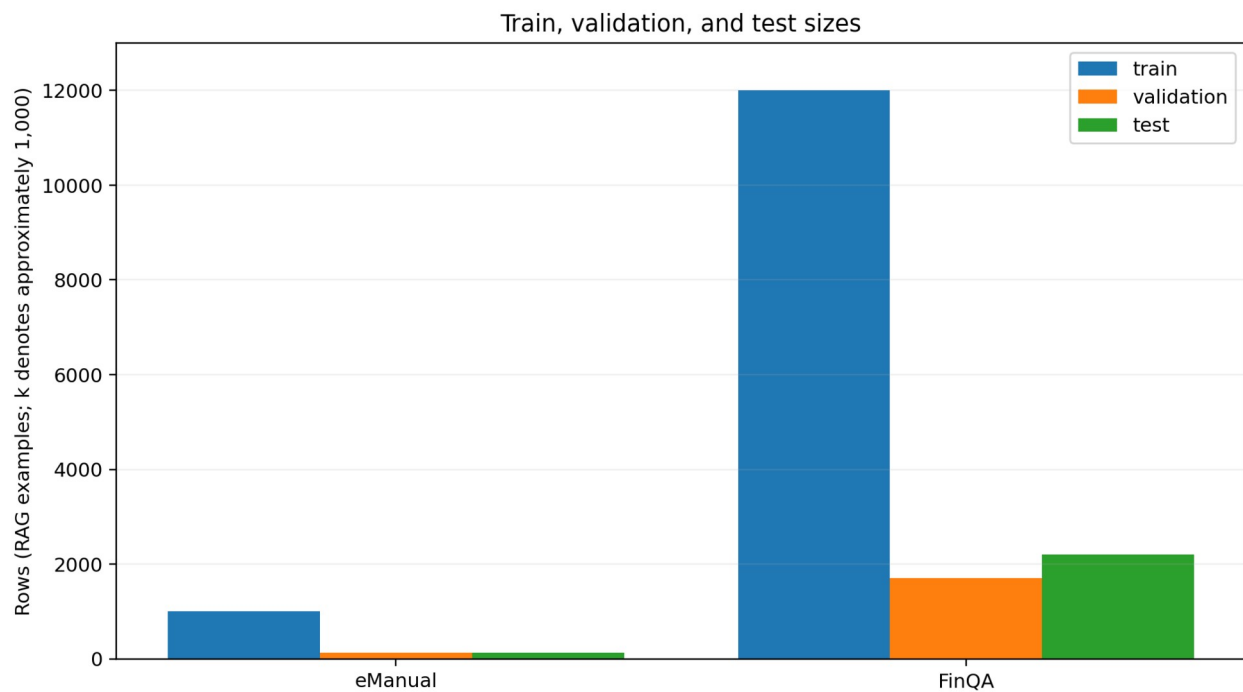


Figure 2. Train, validation, and test sizes for eManual and FinQA.

Table 2 records the six Parquet files associated with the selected subsets. Their sizes are visualized on a logarithmic scale in Figure 3. The data footprint difference is important for replication planning, but it does not alter the test-split score comparison because all methods are evaluated on the predefined split for each dataset.

Table 2. Parquet files for the selected RAGBench subsets.

Dataset	Split	File	Size (MB)
eManual	Train	emanual/train-00000-of-00001.parquet	1.70
eManual	Validation	emanual/validation-00000-of-00001.parquet	0.288
eManual	Test	emanual/test-00000-of-00001.parquet	0.305
FinQA	Train	finqa/train-00000-of-00001.parquet	61.10
FinQA	Validation	finqa/validation-00000-of-00001.parquet	5.94
FinQA	Test	finqa/test-00000-of-00001.parquet	8.94

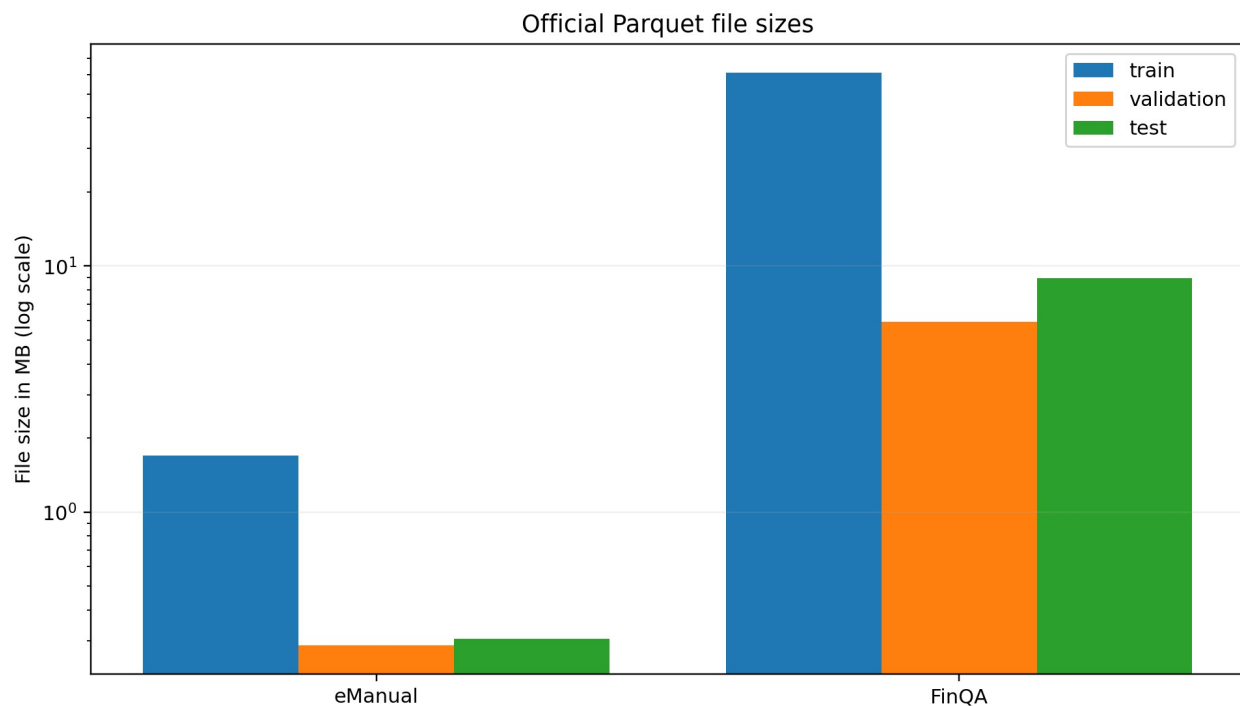


Figure 3. Official Parquet file sizes for eManual and FinQA.

3.2 TRACe targets and trace representation

Each RAG record is represented as $x = (q, D, r)$, where q is the question, D is the retrieved context, and r is the generated response. TRACe defines relevance as the share of retrieved context that is useful for answering q , utilization as the share of retrieved context used in r , completeness as the share of relevant context incorporated into r , and adherence as whether every response unit is supported by D [1]. The first three can be represented as continuous scores, while adherence supports a response-level binary hallucination target.

RAGBench stores both scalar scores and sentence-key annotations. For each context sentence, the evaluator can predict membership in the relevant and utilized sets. For each response sentence, support information identifies whether the claim is grounded in the retrieved material. Table 3 lists the fields required for this trace-aware view and separates observable inputs from evaluation targets.

Table 3. RAGBench fields used in trace-aware quality prediction.

Field	Role	Feature	Target
question	User query	Yes	No
documents / documents_sentences	Retrieved evidence and keyed context sentences	Yes	No
response / response_sentences	Generated answer and response sentence units	Yes	No
adherence_score	Response-level groundedness target	No	Yes
relevance_score	Context relevance target	No	Yes
utilization_score	Context utilization target	No	Yes
all_relevant_sentence_keys	Gold keys for useful context sentences	No	Yes
all_utilized_sentence_keys	Gold keys for context used by the response	No	Yes
unsupported_response_sentence_keys	Gold keys for unsupported response sentences	No	Yes

3.3 Evaluation metrics and evaluator families

The benchmark reports response-level hallucination AUROC and example-level RMSE for relevance and utilization. AUROC measures discrimination across grounded and unsupported responses and is preferred upward. RMSE measures score estimation error and is preferred downward. Sentence-level micro-F1 and macro-F1 remain useful for a full trace model, although they are not available for every evaluator in the published eManual/FinQA comparison. Table 4 summarizes the metric interpretation.

Table 4. Evaluation metrics and preferred direction.

Metric	Target	Interpretation	Preferred direction
Hallucination AUROC	Adherence	Discrimination between grounded and unsupported responses	Higher
Relevance RMSE	Relevance score	Error in estimating useful retrieved context	Lower
Utilization RMSE	Utilization score	Error in estimating context used by the response	Lower
Trace micro-F1	Sentence-key labels	Frequency-weighted trace classification quality	Higher
Trace macro-F1	Sentence-key labels	Class-balanced trace classification quality	Higher

Four evaluator families are included. GPT-3.5 judge uses a prompted annotation procedure. RAGAS applies prompted faithfulness and context-relevance metrics [7]. TruLens provides prompted groundedness and relevance feedback. The DeBERTa-v3 judge is a supervised encoder with separate token-level heads for relevance, utilization, and adherence [1], [18]. Table 5 distinguishes the training or inference style of each method and the outputs available in the comparison.

Table 5. Evaluator families in the eManual/FinQA comparison.

Evaluator	Training / inference style	Reported outputs	Role
GPT-3.5 judge	Prompt-based LLM judge	Hallucination, relevance, utilization	Prompt baseline
RAGAS	Prompt-based RAG evaluation framework	Hallucination, relevance	Prompt baseline
TruLens	Prompt-based groundedness and relevance evaluator	Hallucination, relevance	Prompt baseline
DeBERTa-v3 judge	Supervised trace-trained encoder	Hallucination, relevance, utilization	Specialist evaluator

3.4 Comparative analysis

For each dataset and metric, evaluators are ordered by the preferred direction in Table 4. Absolute differences are computed as DeBERTa minus baseline for AUROC and as DeBERTa RMSE minus baseline RMSE for error metrics. Relative change is expressed as a beneficial improvement: $(\text{DeBERTa} - \text{baseline}) / \text{baseline}$ for AUROC and $(\text{baseline} - \text{DeBERTa}) / \text{baseline}$ for RMSE. A negative relative value therefore denotes worse performance. The strongest comparable non-finetuned baseline is selected separately for each dataset and metric because RAGAS and TruLens do not report utilization.

The comparison is based on test-split point estimates. It supports reproducible ranking and effect-size calculations, but no new hypothesis test is applied because per-example evaluator predictions are not part of the published score matrix. Interpretation therefore emphasizes the magnitude and consistency of score differences across eManual and FinQA.

4. Results

4.1 Overall test-split comparison

Table 6 presents the complete score matrix used in the analysis. The broadest pattern is a strong supervised advantage for hallucination detection: DeBERTa-v3 ranks first on both datasets. Relevance and utilization are less uniform, which indicates that evidence-quality estimation remains domain and metric dependent.

Table 6. RAGBench test-split values for eManual and FinQA.

Dataset	Evaluator	Hallucination AUROC	Relevance RMSE	Utilization RMSE
eManual	GPT-3.5 judge	0.54	0.17	0.11
eManual	RAGAS	0.57	0.27	-
eManual	TruLens	0.61	0.64	-
eManual	DeBERTa-v3 judge	0.76	0.13	0.13
FinQA	GPT-3.5 judge	0.57	0.10	0.13
FinQA	RAGAS	0.57	0.06	-
FinQA	TruLens	0.53	0.79	-
FinQA	DeBERTa-v3 judge	0.81	0.10	0.10

Figure 4 visualizes hallucination AUROC. DeBERTa-v3 reaches 0.76 on eManual and 0.81 on FinQA, while the strongest prompt-based scores are 0.61 and 0.57. The supervised margin is therefore present in both the procedural-manual and financial-evidence settings.

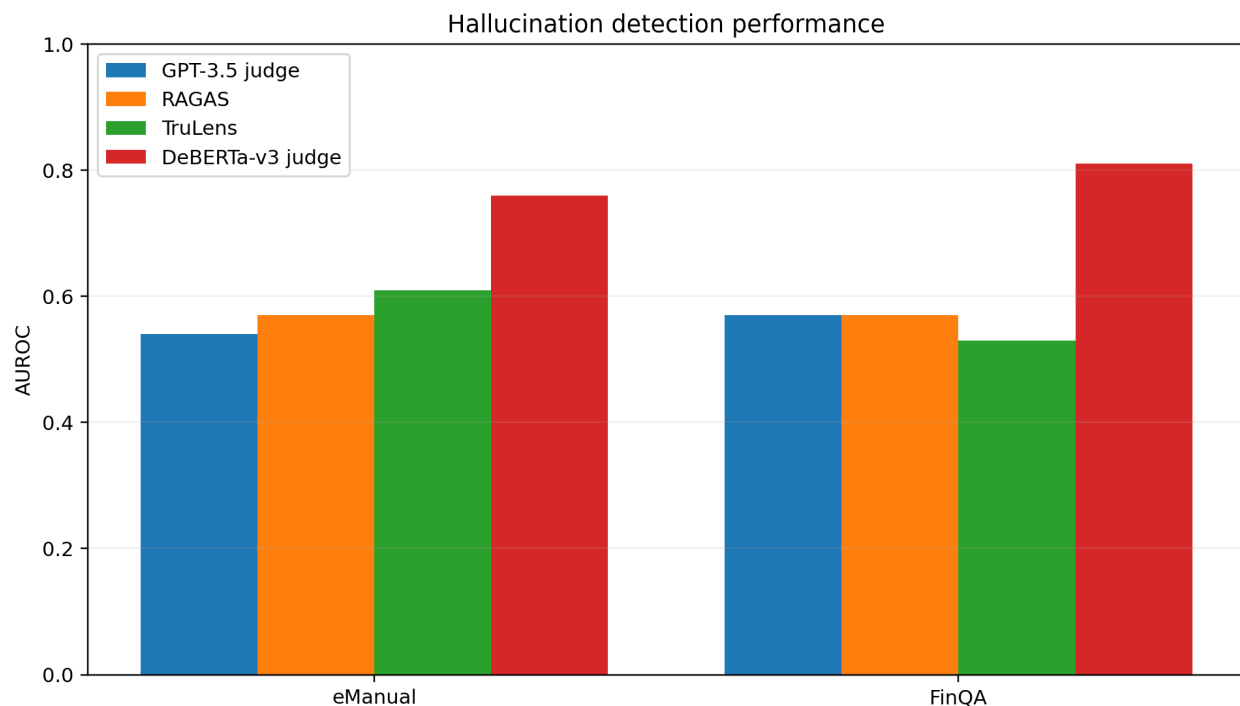


Figure 4. Hallucination detection AUROC on eManual and FinQA.

Figure 5 shows a different pattern for relevance RMSE. DeBERTa-v3 is best on eManual at 0.13, but RAGAS is best on FinQA at 0.06. GPT-3.5 and DeBERTa-v3 tie at 0.10 on FinQA. The result suggests that a prompted relevance judgment can be effective when the finance evidence contains explicit variables or values, even though the same prompted method is less competitive for hallucination discrimination.

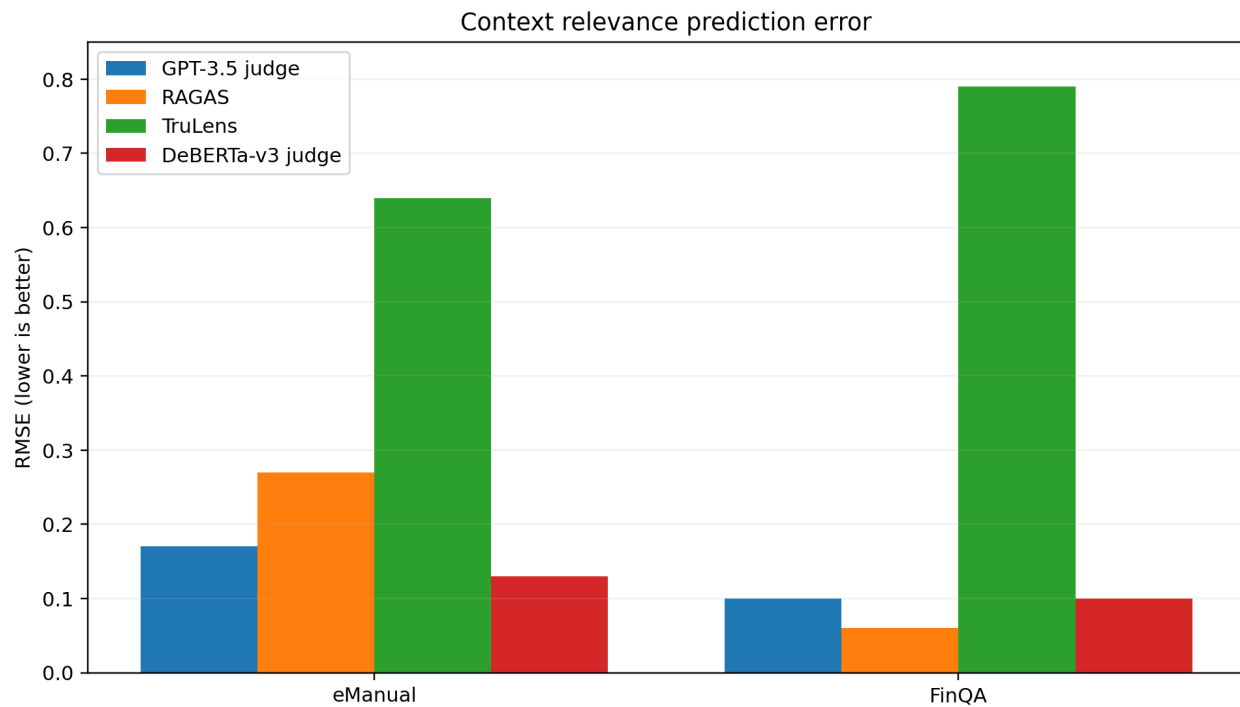


Figure 5. Context relevance RMSE on eManual and FinQA.

Utilization is reported only for GPT-3.5 judge and DeBERTa-v3. Figure 6 shows that GPT-3.5 has the lower eManual error (0.11 versus 0.13), whereas DeBERTa-v3 has the lower FinQA error (0.10 versus 0.13). This reversal reinforces the need to evaluate adherence, relevance, and utilization separately.

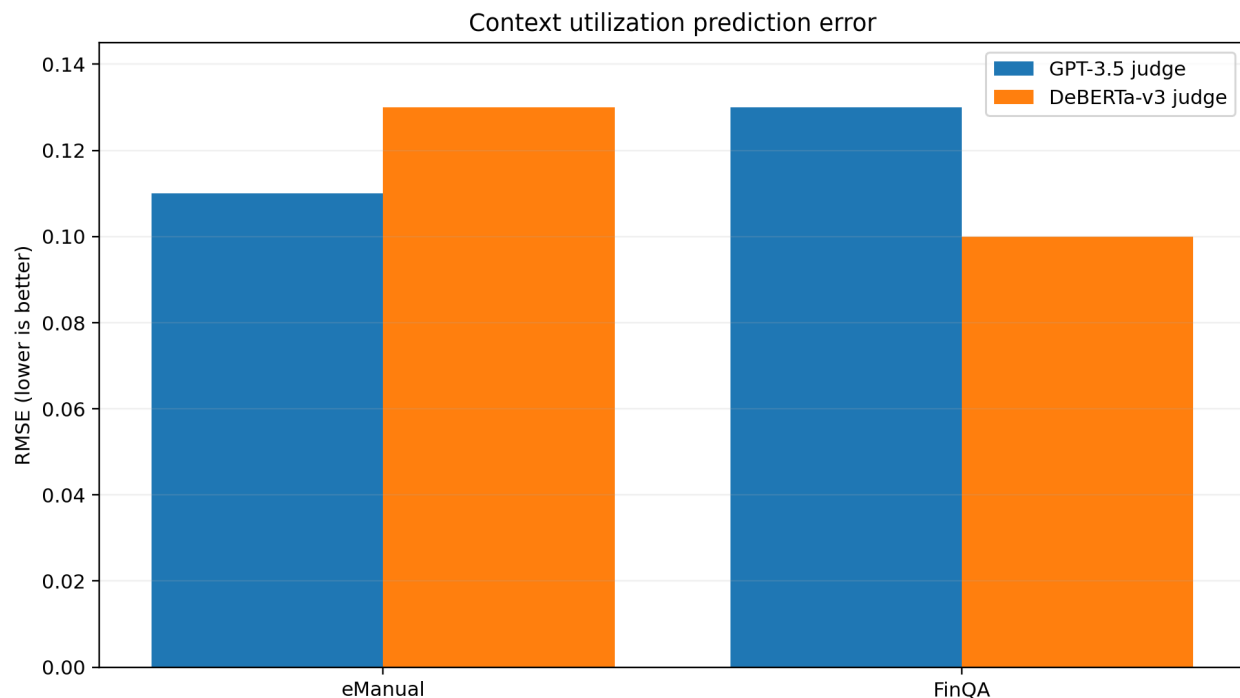


Figure 6. Context utilization RMSE on eManual and FinQA.

4.2 eManual results

Table 7 ranks the evaluators on eManual. DeBERTa-v3 is first for hallucination AUROC and relevance RMSE, while GPT-3.5 is first for utilization RMSE. The 0.15 AUROC gap between DeBERTa-v3 and TruLens is the largest practically important manual-domain difference because unsupported procedural language can directly

alter a user action. The 0.04 relevance-RMSE reduction relative to GPT-3.5 also indicates more accurate identification of useful manual context.

Table 7. eManual evaluator ranking by metric.

Metric	Rank 1	Rank 2	Rank 3	Rank 4
Hallucination AUROC	DeBERTa-v3 judge	TruLens	RAGAS	GPT-3.5 judge
Relevance RMSE	DeBERTa-v3 judge	GPT-3.5 judge	RAGAS	TruLens
Utilization RMSE	GPT-3.5 judge	DeBERTa-v3 judge	-	-

The utilization result adds an important qualification. A model can be strong at deciding whether a response contains unsupported material while being less precise at estimating how much of the retrieved context the response uses. In manual QA, utilization may be sensitive to paraphrase and sentence segmentation: a concise answer can faithfully use one instruction while leaving other retrieved passages unused. This is not necessarily a quality defect unless the unused passages contain required steps.

4.3 FinQA results

Table 8 gives the FinQA ranking. DeBERTa-v3 is first for hallucination AUROC and utilization RMSE; RAGAS is first for relevance RMSE. The 0.81 AUROC result shows that trace-supervised hallucination detection transfers strongly to a domain with financial terminology and quantitative evidence. At the same time, the relevance ranking shows that no evaluator dominates every dimension.

Table 8. FinQA evaluator ranking by metric.

Metric	Rank 1	Rank 2	Rank 3	Rank 4
Hallucination AUROC	DeBERTa-v3 judge	GPT-3.5 judge	RAGAS	TruLens
Relevance RMSE	RAGAS	GPT-3.5 judge	DeBERTa-v3 judge	TruLens
Utilization RMSE	DeBERTa-v3 judge	GPT-3.5 judge	-	-

FinQA relevance can depend on recognizing the precise financial variables needed for a calculation. A passage that is semantically related to the company or reporting period may still be irrelevant to the requested ratio. The low RAGAS error suggests that its prompted decomposition captures this aspect well on FinQA, whereas the supervised model offers its clearest benefit in identifying unsupported responses and estimating evidence use.

4.4 Relative changes and cross-domain contrast

Table 9 compares DeBERTa-v3 with the strongest available prompt-based baseline for each metric. Positive values denote beneficial relative changes. Hallucination AUROC improves by 24.6% on eManual and 42.1% on FinQA. eManual relevance RMSE improves by 23.5%, and FinQA utilization RMSE improves by 23.1%. The negative values identify the two exceptions: eManual utilization and FinQA relevance.

Table 9. DeBERTa-v3 changes versus the strongest comparable prompt-based baseline.

Dataset	Metric	Baseline	Baseline score	DeBERTa score	Absolute delta	Relative change
eManual	Hallucination AUROC	TruLens	0.61	0.76	+0.15	+24.6%
eManual	Relevance RMSE	GPT-3.5	0.17	0.13	-0.04	+23.5%
eManual	Utilization RMSE	GPT-3.5	0.11	0.13	+0.02	-18.2%
FinQA	Hallucination AUROC	GPT-3.5 / RAGAS	0.57	0.81	+0.24	+42.1%
FinQA	Relevance RMSE	RAGAS	0.06	0.10	+0.04	-66.7%
FinQA	Utilization RMSE	GPT-3.5	0.13	0.10	-0.03	+23.1%

Figure 7 makes the asymmetry clear. The largest positive change is FinQA hallucination detection, while the largest negative change is FinQA relevance. The cross-domain conclusion is therefore specific: trace supervision provides a consistent advantage for unsupported-response detection, but relevance and utilization remain sensitive to evidence type and evaluator design.

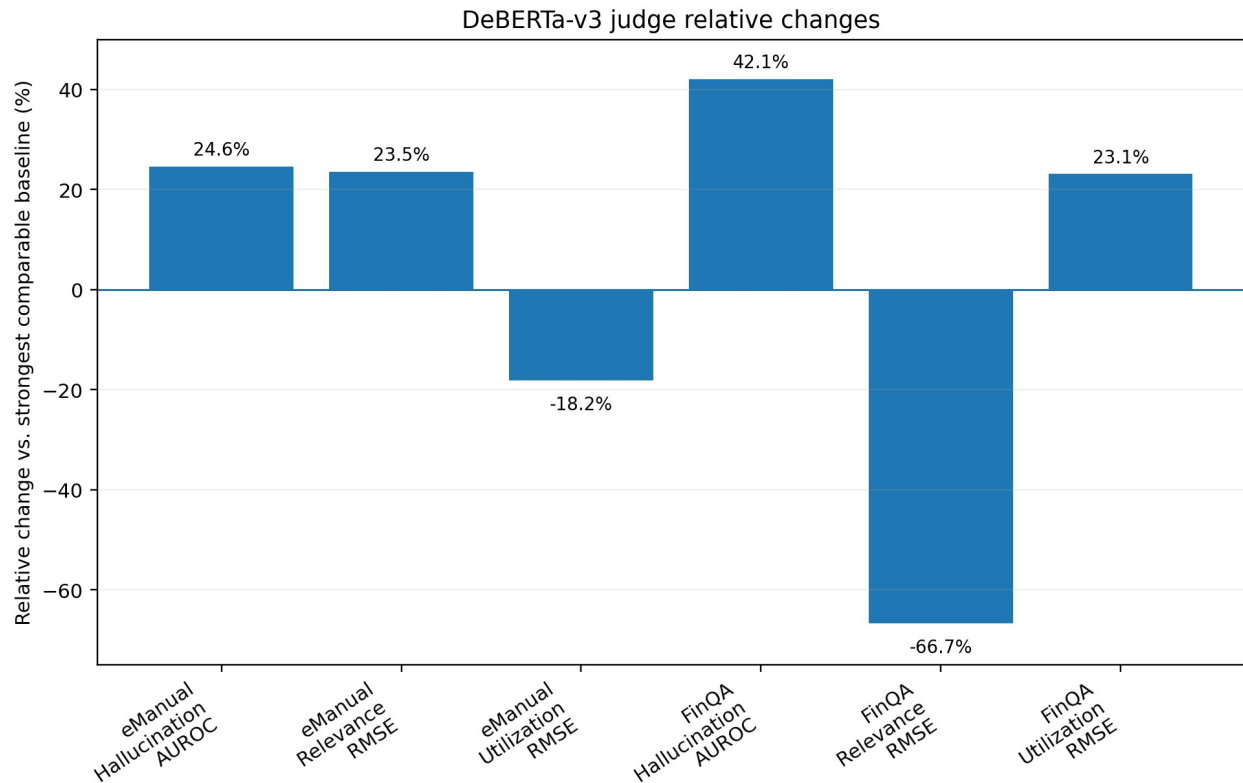


Figure 7. Relative DeBERTa-v3 changes versus the strongest comparable prompt-based baseline.

5. Discussion

5.1 Why trace supervision helps hallucination detection

The most stable result is the supervised advantage for hallucination AUROC. This outcome is consistent with the design of the DeBERTa-v3 judge: sentence labels are broadcast to tokens, and the model learns a direct relationship between response units and supporting context [1], [18]. Prompt-based judges instead infer groundedness through a general instruction at evaluation time. Such prompting is flexible, but it does not necessarily learn the recurring lexical and structural patterns that distinguish supported procedural statements from plausible additions.

The finding aligns with evidence-chain and provenance-oriented RAG research [24], [25]. A grounded response is not merely one that resembles the source; its claims must be attributable to specific evidence. The same principle appears in financial systems that verify XBRL values [22], [23], [29], [30] and in safety layers that combine evidence checks with small-model detection or refusal [31], [32]. The eManual and FinQA results suggest that supervised trace labels provide a reusable signal across these otherwise different domains.

5.2 Domain dependence of relevance and utilization

Relevance and utilization do not show a universal winner because they encode different decisions. Relevance asks whether retrieved context is useful for the question. Utilization asks whether the response actually uses that context. A manual passage can be relevant but omitted from a short answer; a finance passage can be topically related but not contain the variable required for a calculation. These distinctions explain why DeBERTa-v3 leads eManual relevance, RAGAS leads FinQA relevance, GPT-3.5 leads eManual utilization, and DeBERTa-v3 leads FinQA utilization.

The broader enterprise literature shows the same domain dependence. Text-to-SQL benefits from execution feedback [21]; AIOps systems depend on document families, incident context, and multilingual operational terminology [25]-[27]; and long-document financial or contractual systems must preserve section structure and numerical relationships [28]-[30]. A practical evaluation layer should therefore keep TRACe dimensions separate and calibrate thresholds on the target document collection rather than transferring one composite score unchanged.

5.3 Operational quality gates and evidence records

Table 10 translates combinations of TRACe signals into likely system failures and corrective actions. The mapping supports component-level diagnosis: low relevance points first to retrieval, low utilization with adequate relevance points to generation, and low adherence requires grounding or abstention controls. Completeness adds a coverage check when the response uses some evidence but omits an important instruction.

Table 10. Operational interpretation of trace-level quality signals.

Observed signal	Likely failure source	Operational action
Low relevance and low utilization	Broad or incorrect retrieval	Tune chunking, metadata filters, retriever, or reranker
High relevance and low utilization	Generator ignored useful evidence	Adjust context order, prompt, or generation model
High utilization and low adherence	Evidence mixed with unsupported claims	Apply stricter grounding, abstention, or answer filtering
High relevance and high utilization but low completeness	Required steps were omitted	Use step-coverage checks or structured answer planning
High adherence and low relevance	Grounded answer with inefficient retrieval	Reduce context and improve retrieval precision

Figure 8 illustrates the record structure behind these actions. A context sentence can be relevant, utilized, both, or neither, while response sentences are linked to supporting context. The record is suitable for an internal quality dashboard or a reviewer interface because it turns a score into an inspectable evidence path. This is especially valuable for support agents who need to verify a procedural answer before it reaches a customer.

Trace-alignment records used by the evaluator

Document sentence key	Relevant?	Utilized?	Supported response sentence
0b	yes	yes	a
0c	yes	no	-
1a	no	no	-
2d	yes	yes	b

Figure 8. Sentence-key evidence alignment records for relevance, utilization, and response support.

5.4 Implications for evaluator design

A robust industrial evaluator should combine specialization with modularity. The hallucination results favor a trace-trained specialist, but the relevance and utilization results show that prompted methods can remain competitive on particular domains. One design is an ensemble quality layer: a specialist adherence detector

screens unsupported claims, a domain-calibrated relevance model evaluates retrieval, and a utilization model estimates whether the answer uses the supplied evidence. The system can then route cases to regeneration, retrieval repair, refusal, or human review based on the signal pattern rather than a single threshold.

This modular approach is compatible with the research directions represented in the literature review. Evidence verification can be strengthened with numeric checks in finance [22], [23], [29]; provenance can be surfaced through evidence chains [24]; operational context can be adapted to private manuals or incident documents [25]-[27]; and uncertain cases can trigger selective refusal [32]. The central design principle is that an evaluator should preserve the link between a quality decision and the evidence that justifies it.

6. Limitations

The first limitation is that the model comparison uses the published RAGBench test-split score matrix rather than a new training run for each evaluator. This design supports exact ranking and effect-size analysis, but it does not provide access to per-example predictions for additional confidence intervals, calibration curves, or paired error analysis.

Second, RAGBench labels are produced through a GPT-4 annotation process with post-processing and validation against human judgments [1]. The resulting schema is large and consistent, but automated annotations can still contain systematic errors. Product-specific manual deployments should audit a representative sample with domain experts before using predicted scores as hard quality gates.

Third, eManual is based on a television manual. It is a strong proxy for customer-support documentation but does not cover safety-critical manuals in aviation, medicine, industrial control, or electrical maintenance. FinQA provides a useful transfer case, yet its numerical reasoning requirements differ from procedural instructions. Broader conclusions will require additional manual collections and domain-specific traces.

Finally, the published comparison reports utilization only for GPT-3.5 judge and DeBERTa-v3. A fuller assessment would include additional supervised, prompted, and hybrid evaluators with complete TRACE outputs. Such an extension would clarify whether the observed utilization reversal is specific to these two methods or reflects a broader domain effect.

7. Conclusion

This study examined industrial manual RAG quality prediction through trace-level evidence alignment on eManual, with FinQA as a finance-domain contrast. The comparison shows that the fine-tuned DeBERTa-v3 judge has the strongest and most consistent advantage in hallucination detection, reaching 0.76 AUROC on eManual and 0.81 on FinQA. It also improves eManual relevance estimation and FinQA utilization estimation, although prompt-based evaluators remain better on eManual utilization and FinQA relevance.

The practical conclusion is not that one evaluator should replace every other method. It is that RAG quality should be decomposed into adherence, relevance, utilization, and completeness and tied to sentence-level evidence. This structure allows an operations team to distinguish retrieval failure from evidence-use failure and unsupported generation. For enterprise manuals, the resulting trace is both an evaluation object and an operational interface for safer, more auditable question answering.

References

- [1] R. Friel, M. Belyi, and A. Sanyal, "RAGBench: Explainable Benchmark for Retrieval-Augmented Generation Systems," arXiv:2407.11005, 2024.
- [2] P. Lewis et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," in *Advances in Neural Information Processing Systems*, vol. 33, pp. 9459-9474, 2020.
- [3] V. Karpukhin et al., "Dense Passage Retrieval for Open-Domain Question Answering," in *Proc. EMNLP*, pp. 6769-6781, 2020.
- [4] G. Izacard and E. Grave, "Leveraging Passage Retrieval with Generative Models for Open Domain Question Answering," in *Proc. EACL*, pp. 874-880, 2021.
- [5] A. Asai et al., "Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection," in *Proc. ICLR*, 2024.
- [6] Y. Gao et al., "Retrieval-Augmented Generation for Large Language Models: A Survey," arXiv:2312.10997, 2023.

- [7] R. Zhang, Z. Wen, C. Wang, C. Tang, P. Xu, and Y. Jiang, "Quality analysis and evaluation prediction of RAG retrieval based on machine learning algorithms," arXiv preprint arXiv:2511.19481, 2025. [Online]. Available: <https://doi.org/10.48550/arXiv.2511.19481>
- [8] J. Saad-Falcon et al., "ARES: An Automated Evaluation Framework for Retrieval-Augmented Generation Systems," arXiv:2311.09476, 2023.
- [9] V. Adlakha, P. BehnamGhader, X. H. Lu, N. Meade, and S. Reddy, "Evaluating Correctness and Faithfulness of Instruction-Following Models for Question Answering," arXiv:2307.16877, 2023.
- [10] S. Lin, J. Hilton, and O. Evans, "TruthfulQA: Measuring How Models Mimic Human Falsehoods," in Proc. ACL, pp. 3214-3252, 2022.
- [11] J. Thorne et al., "FEVER: A Large-Scale Dataset for Fact Extraction and Verification," in Proc. NAACL-HLT, pp. 809-819, 2018.
- [12] O. Honovich et al., "TRUE: Re-evaluating Factual Consistency Evaluation," in Proc. NAACL, pp. 3905-3920, 2022.
- [13] B. Bohnet et al., "Attributed Question Answering: Evaluation and Modeling for Attributed Large Language Models," arXiv:2212.08037, 2022.
- [14] Z. Chen et al., "FinQA: A Dataset of Numerical Reasoning over Financial Data," in Proc. EMNLP, pp. 3697-3711, 2021.
- [15] M. Zhu, A. Ahuja, D. Juan, W. Wei, and C. K. Reddy, "TAT-QA: A Question Answering Benchmark on a Hybrid of Tabular and Textual Content in Finance," in Proc. ACL-IJCNLP, pp. 3277-3287, 2021.
- [16] V. Castelli et al., "The TechQA Dataset," in Proc. ACL, pp. 1269-1278, 2020.
- [17] P. He, X. Liu, J. Gao, and W. Chen, "DeBERTa: Decoding-Enhanced BERT with Disentangled Attention," in Proc. ICLR, 2021.
- [18] P. He, J. Gao, and W. Chen, "DeBERTaV3: Improving DeBERTa Using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing," arXiv:2111.09543, 2021.
- [19] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding," in Proc. NAACL-HLT, pp. 4171-4186, 2019.
- [20] A. Williams, N. Nangia, and S. R. Bowman, "A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference," in Proc. NAACL-HLT, pp. 1112-1122, 2018.
- [21] Y. Li, "Execution-Feedback and Retrieval-Augmented Generation for Conversational Text-to-SQL: From One-Shot Questions to Clarification-Driven Executable Dialogs," Journal of Advanced Computing Systems, vol. 3, no. 2, pp. 1-17, 2023, doi: 10.69987/JACS.2023.30201.
- [22] K. Zhang, S. Meng, and E. Zhou, "Evidence-Grounded Trading Desk Risk Memos over SEC Filings: Retrieval-Augmented Generation with XBRL Numeric Verification," Journal of Advanced Computing Systems, vol. 3, no. 2, pp. 60-76, 2023, doi: 10.69987/JACS.2023.30205.
- [23] Q. Wu, J. Bai, and X. Zhou, "Evidence-Grounded Financial RAG: Reducing Numerical Hallucination in LLM-Generated Corporate Risk Memos," Journal of Advanced Computing Systems, vol. 3, no. 3, pp. 65-84, 2023, doi: 10.69987/JACS.2023.30306.
- [24] C. Li, J. Bai, and S. Wang, "Evidence-Chain Reliable RAG: Word-Level Hallucination Detection, Source Attribution, and Provenance Explanation for LLM Applications," Journal of Advanced Computing Systems, vol. 4, no. 2, pp. 76-92, 2024, doi: 10.69987/JACS.2024.40207.
- [25] B. Zhang, H. Rao, and D. Zhao, "Evidence-Grounded RAG for Cloud-Native DevOps: Hallucination-Resistant AIOps Question Answering over Private Operations Documents," Journal of Advanced Computing Systems, vol. 4, no. 3, pp. 109-125, 2024, doi: 10.69987/JACS.2024.40308.
- [26] G. Liu, C. Li, and E. Zhang, "OpsLLM for Cloud Incident Triage: Bilingual RAG-Based Root Cause Analysis and Alert Summarization for AI Infrastructure Operations," Journal of Advanced Computing Systems, vol. 4, no. 4, pp. 97-111, 2024, doi: 10.69987/JACS.2024.40408.
- [27] G. Liu, S. He, and I. Liu, "LLM-Augmented Multi-Source Root Cause Attribution for CPU and Network Faults in Microservices," Journal of Advanced Computing Systems, vol. 3, no. 6, pp. 39-57, 2023, doi: 10.69987/JACS.2023.30604.
- [28] S. Zhou, Z. Li, and E. Wang, "Long-Document RAG for Contractual and Insurance Clause Analysis in Receivables RWA Structures," Journal of Advanced Computing Systems, vol. 4, no. 8, pp. 88-104, 2024, doi: 10.69987/JACS.2024.40810.
- [29] Z. Li, K. Zhang, and A. Wong, "Numerical-Reasoning Guardrails for a Quant Research Assistant: A Compact Reproducible Benchmark Using SEC and FRED Data," Journal of Technology Informatics and Engineering, vol. 5, no. 2, pp. 75-90, 2026, doi: 10.51903/jtie.v5i2.541.
- [30] S. Zhou, Z. Li, and E. Wang, "Evidence-Grounded RAG for Tokenized Trade Receivable Disclosure QA under U.S. Capital Market Standards," Journal of Advanced Computing Systems, vol. 3, no. 7, pp. 41-57, 2023, doi: 10.69987/JACS.2023.30704.

- [31] C. Li, W. Su, and E. Zhang, "Lightweight Hallucination Firewall for Enterprise LLM Applications: Evidence Consistency, Self-Checking, and Small-Model Detection on TruthfulQA," *Journal of Advanced Computing Systems*, vol. 3, no. 1, pp. 49-65, 2023, doi: 10.69987/JACS.2023.30104.
- [32] J. Nie and D. Zheng, "Ambiguity-Aware HDFS Log Anomaly Detection with Retrieval-Augmented Failure Narratives and Selective Refusal," *Journal of Advanced Computing Systems*, vol. 3, no. 1, pp. 66-80, 2023, doi: 10.69987/JACS.2023.30105.