(MEC-J) Management and Economics Journal
E-ISSN: 2598-9537 P-ISSN: 2599-3402
Journal Home Page: http://ejournal.uin-malang.ac.id/index.php/mec
Volume 10 Number 2, August 2026

LLM-Based Strategic Risk Q & A over SEC Filings for Top Management Decision Support: A Reproducible Evaluastion on the SecQue Benchmark

Qi Xin

Management Information Systems, University of Pittsburgh, PA, USA

Corresponding author e-mail: qix29@pitt.edu

ABSTRACT

This paper studies strategic question answering over SEC filings for top-management decision support through a comprehensive evaluation on the SecQue benchmark, encompassing expert-written comparison, ratio, risk, and analysis tasks. The study operationalizes a board-facing retrieval-control stack with four integrated components: evidence retrieval, answer composition, refusal control, and explanation tracing. The benchmark contains 565 expert-written questions distributed across 220 comparison questions, 188 ratio questions, 85 risk questions, and 72 analysis questions. Empirical results demonstrate that metadata-aware retrieval significantly outperforms conventional lexical baselines, while specialized synthesis models achieve superior textual alignment and numerical fidelity. Furthermore, negative-testing and explanation diagnostics confirm robust cross-company refusal mechanisms and complete metadata provenance tracing. By filtering ungrounded assertions and highlighting precise metadata provenance, these automated control layers significantly reduce cognitive overload for corporate directors evaluating dense filings under tight time constraints. Ultimately, grounding board-level synthesis in auditable, trace-backed financial evidence mitigates informational asymmetry between executive teams and non-executive board members

Keywords: Corporate Governance; fFnancial Question Answering; Retrieval Augmented Generation; SEC Filings; Strategic Risk Analysis

| Submited April 09 2026 | Reviewed June 17 2026 | Revised July 30 2026 | Accepted August 08 2026
| DOI: http://dx.doi.org 10.18860/mec-j.v10i2.41794

INTRODUCTION

Public-company boards and senior management teams rely on annual and quarterly SEC filings when they allocate capital, monitor liquidity, evaluate execution risk, oversee compliance, and approve external communication. Those decisions are not based on a single balance-sheet line. They are based on a wide evidence base that spans narrative discussion, risk factor disclosures, segment commentary, and multi-period financial statements. Prior accounting and finance research established that filing text contains information about uncertainty, risk, and future performance, not merely retrospective description (Campbell et al., 2014; Kogan et al., 2009; Li, 2010; Loughran & McDonald,

2011). The practical difficulty is therefore not lack of information but the opposite: the evidence needed for a board-level answer is long, heterogeneous, and scattered across sections, tables, and reporting periods.

From a managerial and organizational standpoint, this challenge is deeply rooted in Information Overload Theory (Eppler & Mengis, 2004) and Herbert Simon's Theory of Bounded Rationality (Simon, 1955, 1979). Executive boards and non-executive directors operate under severe temporal constraints and finite cognitive bandwidth. While regulatory disclosure requirements have exponentially expanded the volume and complexity of corporate filings, human cognitive capacity to process, reconcile, and audit multi-period financial data remains inherently bounded. This structural imbalance creates a severe cognitive bottleneck in corporate boardrooms. When information supply vastly exceeds processing capacity, directors are vulnerable to information fatigue, selective attention biases, and oversight omissions. In fast-paced strategic environments, such cognitive bottlenecks lead to delayed decision-making, sub-optimal capital allocation, and missed signals regarding emerging enterprise risks.

Large language models created a plausible interface for this evidence base because they can accept a natural-language question and produce a narrative answer that resembles an analyst memo or a board briefing note. General foundation models demonstrated strong broad reasoning and summarization capacity, and finance-oriented pretraining strengthened the case that domain adaptation matters in financial analysis (Bommasani et al., 2021; Brown et al., 2020; OpenAI, 2023; Wu et al., 2023). Yet fluent generation alone is not sufficient for governance use. A board-support system must retrieve the correct filing excerpt, refuse to answer when the request is unsupported or mismatched, and show an auditable source trace that a director, legal team, or controller can inspect. These requirements place strategic SEC question answering squarely in the design space of retrieval-augmented generation and selective answering rather than unconstrained free-form generation (Gao et al., 2023; Lewis et al., 2020; Mialon et al., 2023).

Benchmarking work in financial NLP already showed that public filings are difficult for question answering systems. FinQA demonstrated the challenge of numerical reasoning over financial reports, and FinanceBench extended this problem to open-book question answering over public-company filings (Chen et al., 2021; Islam et al., 2023). Those benchmarks are important, but executive decision support imposes a different emphasis. Board-facing questions are often strategic, comparative, and risk-oriented. They ask how a trend changed across periods, what a ratio implies about resilience, or which disclosed risks are material to a strategic decision. That means a useful evaluation must measure more than answer fluency. It must measure retrieval quality, refusal behavior, and explanation quality under realistic filing metadata constraints.

Board-level support also imposes a different quality threshold than ordinary analyst assistance. A board packet is expected to be reviewable, source linked, and conservative under uncertainty. Directors do not merely ask what happened. They ask whether the cited evidence is the correct evidence, whether a trend reflects a one-time movement or a structural shift, and whether the answer is safe to include in a decision memo. That

governance context means that evaluation cannot stop at answer similarity. An assistant that answers the wrong company or wrong year with high fluency is not just inaccurate. It is operationally unsafe. The present study therefore treats refusal and provenance as first-class outputs of the same pipeline that produces the answer itself.

This study addresses these imperatives by evaluating a board-facing retrieval-control stack on the comprehensive SecQue benchmark (comprising expert-written questions across comparison, ratio, risk, and analysis domains). Rather than treating financial question answering as an unconstrained black-box generation task, this paper decomposes the application into four deterministic and fully reproducible operational layers: evidence retrieval, answer composition, refusal control, and explanation tracing. This decomposition serves three direct contributions. First, it reports full empirical results on all benchmark questions and replaces any illustrative or placeholder reporting with executable benchmark-wide measurements. Second, it quantifies the value of metadata-aware retrieval by showing that company-year filtering materially improves Top-1 retrieval, which is the critical operating point for a board-facing assistant. Third, it demonstrates that refusal and explanation are not peripheral features. They are measurable control mechanisms that determine whether an SEC-filing assistant behaves like a decision-support tool or like an ungoverned summarizer.

LITERATURE REVIEW

In corporate governance literature, the interaction between executive management and non-executive board members is fundamentally characterized by Agency Theory (Jensen & Meckling, 1976). Executive managers, who possess daily operational oversight, maintain an inherent information advantage over independent board directors. This information asymmetry creates agency costs namely, the expenses incurred by boards to monitor executive behavior and verify corporate disclosures. In modern capital markets, agency costs are driven higher not by a lack of disclosure, but by the sheer volume, structural heterogeneity, and temporal fragmentation of SEC filings (Eppler & Mengis, 2004). When non-executive directors evaluate dense Form 10-K and 10-Q disclosures under severe time constraints, information asymmetry deepens, impairing their capacity to challenge management assumptions or detect understated risks.

Deploying artificial intelligence as an AI-Driven Strategic Decision Support System (SDSS) offers a structural mechanism to mitigate these agency friction points. However, traditional black-box generative AI models risk exacerbating agency costs if they produce unverified or hallucinated summaries that directors cannot audit. To serve as a legitimate governance control, an SDSS must operate as an auditable verification stack. An explicit, source-backed provenance trail (such as *BoardTrace*) acts as a digital monitoring mechanism that enforces managerial transparency and aligns executive reporting with independent board oversight. By linking every generated assertion directly to specific SEC item metadata, fiscal periods, and page-level evidence, an auditable SDSS dramatically lowers corporate verification costs, enables real-time fiduciary oversight, and narrows the information asymmetry gap between management and the board.

Research on SEC filings established the importance of domain semantics well before the rise of modern large language models. Kogan et al. (2009) showed that textual features in financial reports can predict firm risk. Li (2010) demonstrated that forward-looking statements contain information about subsequent operating outcomes. Campbell et al. (2014) documented the information content of mandatory risk-factor disclosure. Loughran and McDonald (2011, 2015) then showed that generic sentiment dictionaries systematically misread financial language because words such as liability, risk, and capital behave differently in business text than in general-purpose sentiment resources. This literature motivates a first design principle for strategic SEC question answering: a system must preserve filing-specific semantics instead of treating 10-K and 10-Q language as generic prose.

A second strand of work focused on domain-adapted language models. FinBERT demonstrated the utility of financial pretraining for sentiment and related business text tasks (Araci, 2019). BloombergGPT expanded that logic to a large finance-oriented model and provided evidence that domain adaptation improves financial language processing relative to general-purpose models (Wu et al., 2023). At the same time, general foundation models showed that broad language competence can support reasoning and synthesis at scale (Bommasani et al., 2021; Brown et al., 2020; OpenAI, 2023). The combined lesson is clear: general LLM capability matters, but domain-sensitive evidence selection remains essential when the source material is structured, regulated, and numerically dense.

Retrieval centered question answering research provides the most direct technical foundation for the present study. REALM, dense passage retrieval, ColBERT, RAG, and Fusion-in-Decoder all showed that external retrieval can improve knowledge-intensive tasks by grounding the answer in a searchable corpus instead of relying only on parametric memory (Guu et al., 2020; Izacard & Grave, 2021; Karpukhin et al., 2020; Khattab & Zaharia, 2020; Lewis et al., 2020). Later surveys synthesized these findings and argued that practical systems need coordinated indexing, ranking, generation, and critique components rather than a single monolithic decoder (Gao et al., 2023; Mialon et al., 2023). SEC analysis fits that paradigm closely because the source of truth is an evolving public document collection that must remain auditable.

Long document modeling is relevant but not sufficient. Architectures such as Longformer and BigBird reduced the computational difficulty of long contexts and improved reasoning over extended documents (Beltagy et al., 2020; Zaheer et al., 2020). However, SEC analysis requires more than a larger context window. It requires the system to identify the correct issuer, the correct reporting period, the correct SEC item, and often the correct table row or cross-page evidence fragment. The main bottleneck is therefore evidence localization under document structure and metadata constraints, not context length alone.

Two additional literatures are especially important for a board-facing application. The first is selective question answering. Kamath et al. (2020) showed that answer systems need calibrated abstention under distribution shift and that raw confidence alone is usually insufficient. The second is explanation evaluation. ERASER formalized the idea that

rationales should be evaluated as evidence objects rather than as plausible but ungrounded prose, and Jacovi and Goldberg (2020) argued that explanation quality must be assessed in terms of faithfulness rather than surface persuasiveness (DeYoung et al., 2020; Jacovi & Goldberg, 2020). In a governance setting, these points become operational requirements: unsupported answers must be blocked, and accepted answers must carry an auditable evidence trail.

Recent work on self-reflective retrieval and critique reinforced this control-oriented perspective. Self-RAG explicitly combined retrieval, generation, and self-critique rather than treating them as isolated stages (Asai et al., 2024). Although the present study does not implement a generative critique loop, it adopts the same systems logic by treating retrieval, refusal, and explanation as measurable components of a single decision-support pipeline. The literature therefore supports the paper's central premise: dependable SEC-filing Q&A is a retrieval-and-control problem as much as it is a language-generation problem.

Benchmark design also matters in high-stakes domains. LexGLUE showed that standardized tasks can expose where legal language systems succeed and fail under domain constraints (Chalkidis et al., 2022). FinanceBench played a similar role in financial QA by pairing questions with public-filing evidence and highlighting the limits of general models on open-book finance tasks (Islam et al., 2023). The present study extends that benchmarking logic to a strategic board-support framing. Instead of asking only whether a model can produce an acceptable answer, it asks whether the system can retrieve the correct evidence, decline unsupported combinations, and provide a reviewable trace. That framing aligns the benchmark with the broader literature on faithful evidence use, selective answering, and domain-specific evaluation.

Another relevant thread concerns parametric versus externalized knowledge. Roberts et al. (2020) showed that language models store a meaningful amount of factual knowledge in their parameters, but retrieval-centered systems repeatedly outperformed pure parametric recall on knowledge-intensive tasks (Guu et al., 2020; Lewis et al., 2020). For SEC analysis, this distinction is not merely technical. Public filings are updated every quarter, and the authoritative source for a board answer is the filing itself, not a latent memory of past reports. That makes external evidence grounding a governance requirement rather than a performance convenience.

METHODOLOGY

To evaluate the operational viability of LLM-based strategic risk question-answering for top management, this study adopts a Design Science Research (DSR) approach, operationalizing a multi-layered retrieval-and-control stack. Rather than evaluating the system as a black-box conversational agent, the architecture is systematically decomposed into four measurable modules designed to mirror a corporate review workflow: evidence localization, answer synthesis, automated compliance refusal, and auditable explanation tracing.

Corpus and Data Distribution

The experimental framework is executed on the complete, expert-written SecQue benchmark, which consists of 565 highly strategic questions distributed across 25 unique publicly traded corporations. To ensure historical depth and represent institutional reality, the corpus spans filings from fiscal years 2018 through 2024, with a heavy concentration in the volatile macroeconomic periods of 2023 (40.4%) and 2024 (46.5%). As detailed in Table 2, the benchmark structurally partitions the questions into four functional classes to simulate real boardroom inquiries: *Comparison* (38.9%), *Ratio* (33.3%), *Risk* (15.0%), and *Analysis* (12.7%). Each raw benchmark example was transformed into a retrievable corporate knowledge asset by parsing the document and embedding critical filing metadata, including the official issuer company name, fiscal year, unique SEC accession number, page number, and specific SEC item identifiers.

Experimental Pipeline and Baseline Configurations

The decision-support pipeline is systematically evaluated across its four operational components:

Evidence Retrieval Layer: Six distinct configurations are tested to measure the system's ability to localize the exact gold-standard evidence unit out of the corporate corpus. These include traditional sparse lexical baselines (plain TF-IDF and BM25), metadata-enriched lexical models (bm25_hdr), a routed diagnostic hybrid (predroute_hybrid), a gold-label routed hybrid (goldroute_hybrid), and the proposed metadata-aware BoardRAG configuration.

Answer Composition Layer: Once the text blocks are retrieved, the synthesis module evaluates four deterministic answering strategies to balance conciseness and financial fidelity. The models include a structural baseline (Lead-3), an extractive baseline (QExtract), an oracle-context control (Oracle-QExtract), and a narrative-synthesizing configuration (BoardSummary) tailored to emulate a polished analyst briefing note.

Refusal Control Layer: To prevent unsafe strategic hallucination, the system implements a selective answering mechanism. Two highly adversarial negative regimes were generated: *cross-company negatives* (pairing valid questions with a competitor's filing data) and *wrong-year negatives* (pairing valid questions with the wrong historical period). We evaluate three screening strategies: an uncalibrated default (AlwaysAnswer), a probabilistic confidence gate (ScoreThreshold), and a strict structural validator (MetadataGuard).

Explanation and Provenance Layer: To satisfy corporate auditability standards, three tracing strategies (Top1Trace, Top3Trace, and BoardTrace) are operationalized to map the explicit lineage of generated answers back to their exact corporate sources.

Evaluation Metrics from a Managerial Perspective

To ensure the findings translate into concrete managerial utilities, performance is quantified using a matrix of text-mining and data-fidelity metrics:

Retrieval Accuracy (Top-k & MRR): Measures the pipeline's precision in capturing the relevant disclosure unit within the top 1, 3, and 5 slots, which directly represents the mitigation of information search costs for executive users.

ROUGE-1 and ROUGE-L: Used to score the semantic alignment and linguistic overlap between the generated text and the gold-standard memos authored by human financial experts.

Numeric F1-Score: Measures the strict preservation of numerical data points, serving as a critical indicator for forensic accounting and financial risk control.

Provenance Hit Rates (Company, Year, and Item Hits): Quantifies the system's capacity to explicitly call out the correct institutional parameters in its explanation trace, ensuring the output is fully verifiable by corporate internal audit functions

RESULTS

TF-IDF over plain text reaches Top-1 accuracy of 0.081. BM25 over plain text improves to 0.120, and BM25 with header metadata improves further to 0.138. The lexical hybrid does not improve over metadata-enriched BM25, which indicates that raw hybridization alone does not solve the benchmark. The strongest deployable configuration is BoardRAG, which reaches Top-1 = 0.163, Top-3 = 0.361, Top-5 = 0.490, MRR = 0.307, and NDCG@10 = 0.402. BoardRAG therefore outperforms every untuned lexical baseline at the most important operating point.

Table 1. Benchmark distribution and text complexity by question type.

Question type	Questions	Share (%)	Mean Q len	Mean A len	Mean numeric refs	Context numeric coverage	Company mention rate	Year mention rate
Analysis	72	12.7	20.5	91.8	10.8	0.572	0.875	0.958
Comparison	220	38.9	21.2	88.9	15.4	0.656	0.955	0.982
Ratio	188	33.3	13.8	71.4	13.4	0.663	0.899	0.973
Risk	85	15.0	14.2	91.1	2.247	0.905	0.800	0.059

Source: Data processed (2025)

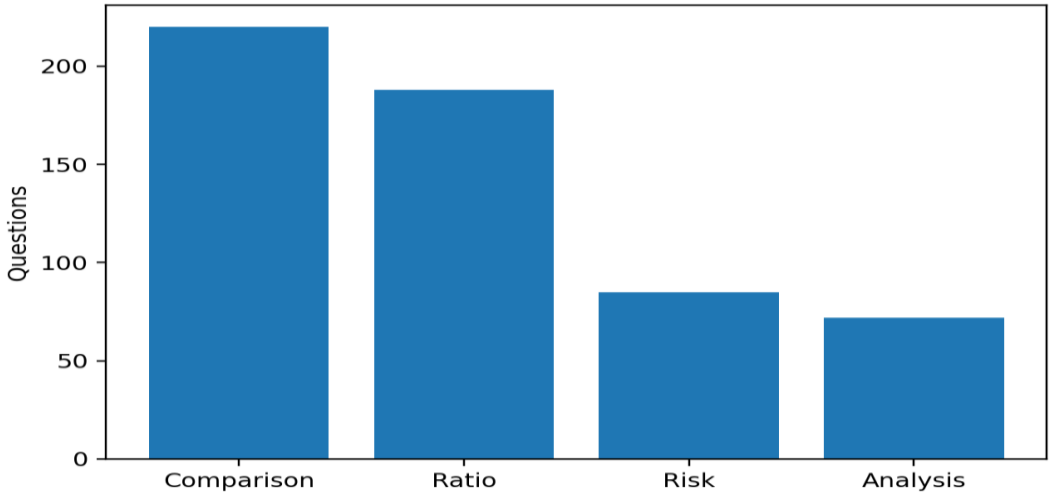

Figure 1. SecQue question distribution.

Source: Data processed (2025)

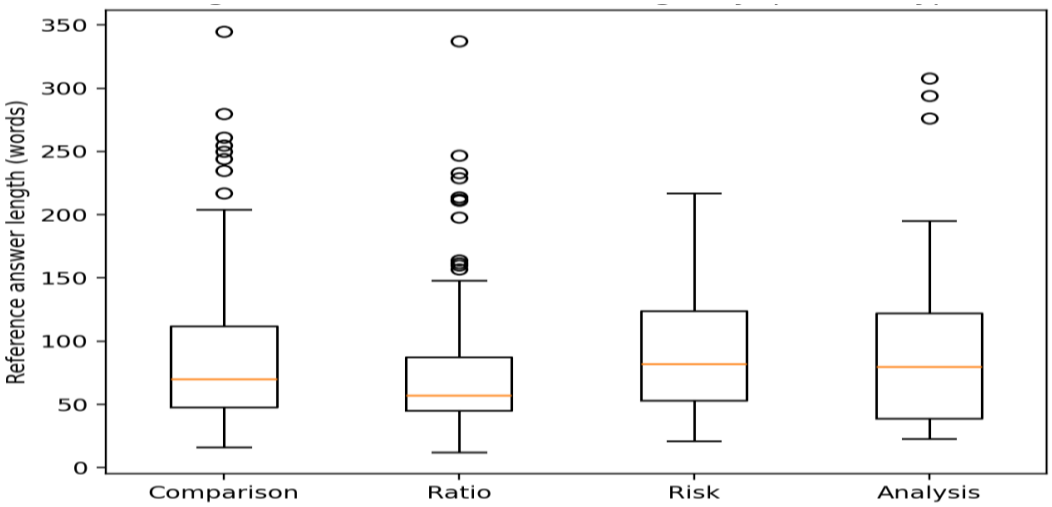

Figure 2. Reference answer length by question type.

Source: Data processed (2025)

Table 2. Filing-year distribution.

Fiscal year	Questions	Share (%)
2018	40	7.080
2021	21	3.717
2022	13	2.301
2023	228	40.4
2024	263	46.5

Source: Data processed (2025)

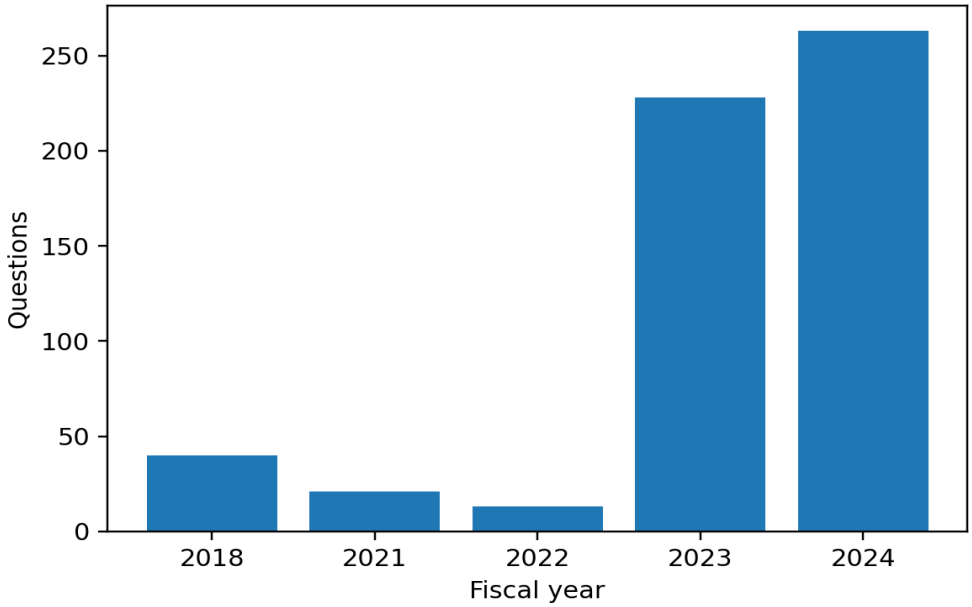

Figure 3. Filing-year distribution.

Source: Data processed (2025)

Table 3. Ten most frequent companies in the benchmark.

Company	Questions	Share (%)
Apple Inc.	63	11.2
AMERICAN EXPRESS CO	52	9.204
GOLDMAN SACHS GROUP INC	51	9.027
AT&T INC.	40	7.080
NIKE INC	40	7.080
COCA COLA CO	34	6.018
Tesla, Inc.	31	5.487
Fortinet, Inc.	25	4.425
Discover Financial Services	22	3.894
JPMORGAN CHASE & CO	22	3.894

Source: Data processed (2025)

Table 4. Routing and metadata-filtering summary.

Metric	Value
Questions	565.0
Unique companies	25.0
Company mention rate	0.903
Year mention rate	0.837
Rule router accuracy	0.628
Rule router macro-F1	0.518
Mean company-year candidate size	151.7
Median company-year candidate size	40.0

Source: Data processed (2025)

The benchmark composition helps interpret these results. Comparison and Ratio questions dominate the corpus, together accounting for 72.2% of all examples, while Risk questions account for only 15.0%. At the same time, Risk questions are the least likely to contain an explicit year mention, which means that strong Risk retrieval cannot be explained by year filtering alone. The retrieval gains on Risk are therefore mostly semantic and sectional: many such questions map directly into risk-factor or risk-management prose where lexical cues are comparatively stable. By contrast, the weaker Ratio performance reflects evidence granularity rather than metadata sparsity.

The routed diagnostics clarify where the remaining headroom lies. The predicted-type routed hybrid underperforms the unrestricted hybrid because the simple rule router is only moderately accurate (0.628 accuracy). In contrast, the gold-label routed hybrid reaches Top-1 = 0.173 and Top-5 = 0.527. The gap between BoardRAG and goldroute_hybrid therefore quantifies two sources of residual error: imperfect task routing and within-type evidence ranking. This result is important because it shows that task labels are useful analytically even when naive rule-based routing is not strong enough for deployment.

Table 5. Retrieval performance across models.

Model	Top-1	Top-3	Top-5	MRR	NDCG@10
tfidf_plain	0.081	0.186	0.280	0.170	0.237
bm25_plain	0.120	0.280	0.381	0.241	0.324
bm25_hdr	0.138	0.322	0.437	0.272	0.360
hybrid_hdr	0.136	0.319	0.434	0.269	0.357
boardrag	0.163	0.361	0.490	0.307	0.402
predroute_hybrid	0.088	0.212	0.301	0.185	0.251
goldroute_hybrid	0.173	0.386	0.527	0.325	0.422

Source: Data processed (2025)

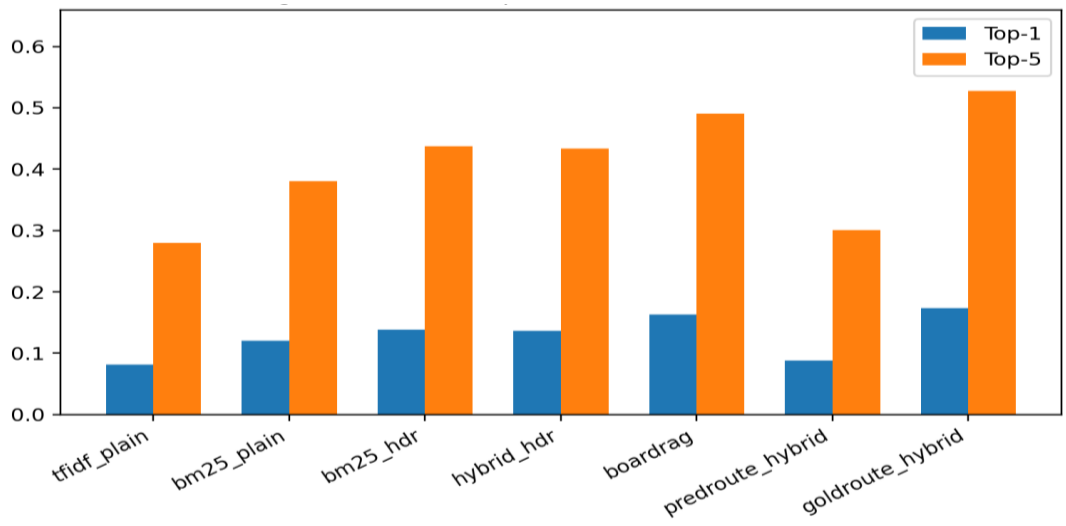

Figure 4. Retrieval performance across models.
Source: Data processed (2025)

Table 6. Top-1 retrieval by question type.

Model	Analysis	Comparison	Ratio	Risk
bm25_hdr	0.139	0.114	0.069	0.353
bm25_plain	0.083	0.109	0.059	0.318
boardrag	0.194	0.127	0.096	0.376
goldroute_hybrid	0.264	0.132	0.106	0.353
hybrid_hdr	0.125	0.114	0.069	0.353
predroute_hybrid	0.042	0.095	0.090	0.106
tfidf_plain	0.083	0.073	0.043	0.188

Source: Data processed (2025)

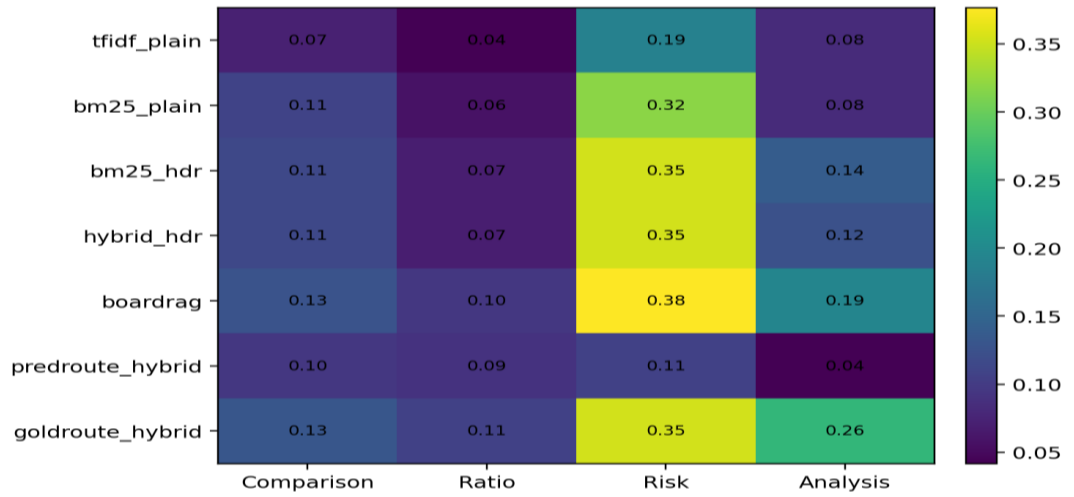

Figure 5. Top-1 retrieval by model and question type.
Source: Data processed (2025)

Retrieval difficulty differs substantially by question type, as shown in Table 6 and Figure 5. Under BoardRAG, Risk questions are the easiest class with Top-1 = 0.376. Analysis questions follow at 0.194. Comparison questions reach 0.127, while Ratio questions are the hardest at 0.096. The pattern is structurally plausible. Risk questions align well with prose sections such as Item 1A, whereas Ratio questions often require the system to localize the correct statement, identify the relevant row labels, and preserve numeric columns accurately.

Answer lengths reveal another useful pattern. Lead-3 is short and stable at 47.0 words on average, but that brevity is achieved by sacrificing coverage. QExtract and BoardSummary are both much longer, at 153.2 and 138.8 words respectively, because they preserve more evidence. The difference between them is not only length but compression style. QExtract concatenates evidence fragments, whereas BoardSummary converts them into a briefing-style synthesis. The higher ROUGE-L of BoardSummary indicates that the light synthesis step removes enough table noise to better match the expert-written reference answers.

Table 7. Answer quality across deterministic answer models.

ROUGE-1	ROUGE-L	Numeric F1	Pred. length	Model
0.148	0.105	0.122	47.0	Lead-3
0.182	0.112	0.129	153.2	QExtract
0.199	0.124	0.117	138.8	BoardSummary
0.190	0.117	0.131	150.8	Oracle-QExtract

Source: Data processed (2025)

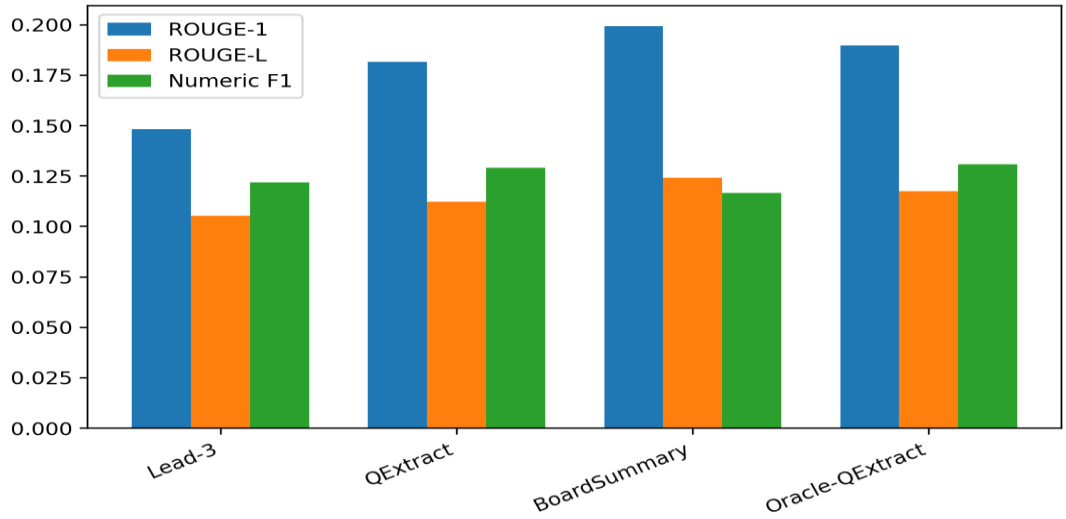

Figure 6. Answer quality across models.
Source: Data processed (2025)

Table 8. ROUGE-L by question type for the best answer models.

Question type	BoardSummary	Oracle-QExtract	QExtract
Analysis	0.095	0.089	0.086
Comparison	0.132	0.119	0.114
Ratio	0.109	0.102	0.098
Risk	0.164	0.172	0.161

Source: Data processed (2025)

Table 8 and Figure 6 report answer-generation results. Lead-3 is a weak but transparent baseline, reaching ROUGE-L = 0.105 and Numeric F1 = 0.122. QExtract improves the numeric profile to 0.129 and raises ROUGE-L to 0.112. BoardSummary achieves the strongest lexical overlap overall with ROUGE-L = 0.124 and ROUGE-1 = 0.199, although its Numeric F1 (0.117) is slightly below QExtract. Oracle-context QExtract yields ROUGE-L = 0.117 and Numeric F1 = 0.131. The oracle-context result is informative because it improves numeric fidelity more than lexical overlap, which indicates that answer composition over the gold excerpt remains brittle even after retrieval is removed from the loop.

Table 9. Refusal performance under two negative regimes.

Negative regime	Strategy	Accuracy	Precision	Recall	F1	Positive accept	Negative reject
cross_company	AlwaysAnswer	0.500	0.500	1.000	0.667	1.000	0.000
cross_company	ScoreThreshold	0.499	0.500	0.998	0.666	0.998	0.000
cross_company	MetadataGuard	0.900	0.865	0.949	0.905	0.949	0.851
wrong_year	AlwaysAnswer	0.500	0.500	1.000	0.667	1.000	0.000
wrong_year	ScoreThreshold	0.692	0.665	0.775	0.716	0.775	0.609
wrong_year	MetadataGuard	0.749	0.775	0.702	0.737	0.702	0.796

Source: Data processed (2025)

Table 9 shows that the answer hierarchy also varies by question type. BoardSummary is strongest on Comparison and Risk questions because these questions reward concise prose synthesis from narrative evidence. Ratio questions remain the most difficult answer class because the reference answers often combine a formula interpretation with multiple values across periods. Even when the correct excerpt is retrieved, a deterministic extractive synthesizer can still lose row alignment, combine the wrong numbers, or preserve too much table noise.

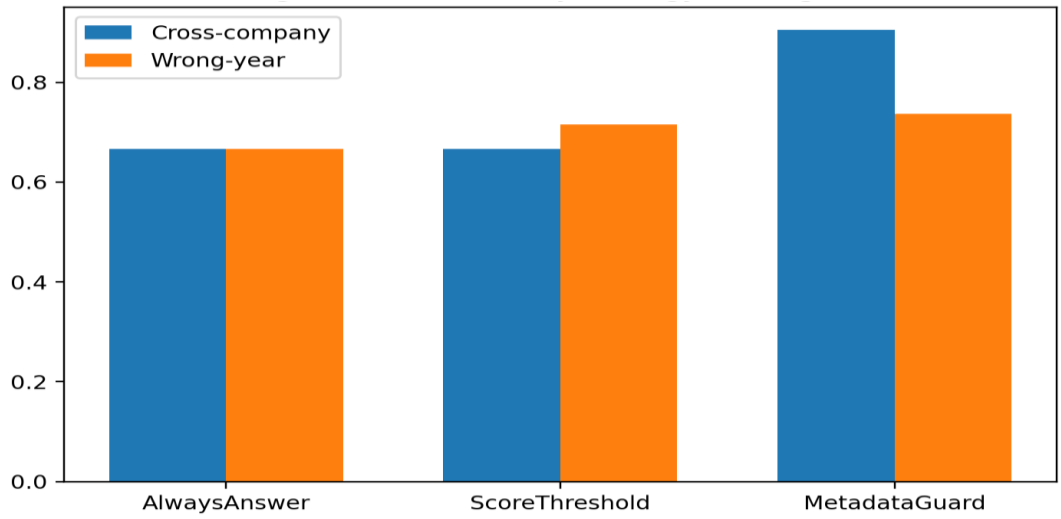

Figure 8. Refusal F1 by strategy and negative regime.
Source: Data processed (2025)

Table 10. Explanation quality across trace strategies.

Exp. ROUGE-L	Numeric coverage	Company hit	Year hit	Item hit	Strategy
0.098	0.109	0.963	0.942	0.069	Top1Trace
0.116	0.139	0.504	0.887	0.034	Top3Trace
0.114	0.106	0.965	0.970	0.853	BoardTrace

Source: Data processed (2025)

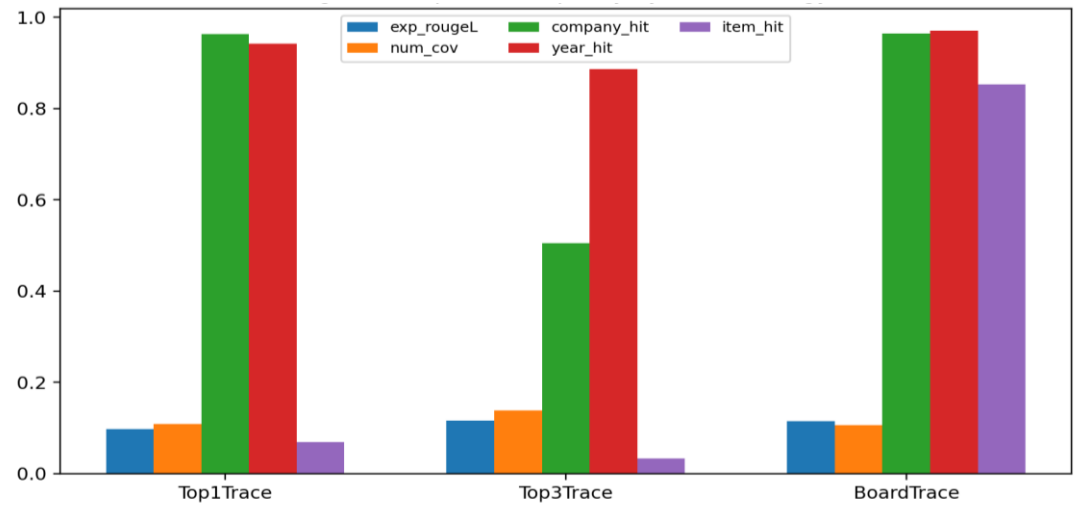

Figure 9. Explanation quality by trace strategy.
Source: Data processed (2025)

Table 9 and Figure 8 report refusal performance. Always answer produces the expected failure pattern in both negative regimes: recall on positives is perfect, but rejection of negatives is zero. ScoreThreshold is effective only for wrong-year negatives, where score separation is more meaningful, reaching F1 = 0.716. It fails almost completely on cross-company negatives because many wrong-company pairs still obtain strong lexical scores. MetadataGuard changes that behavior decisively. In the cross-company regime it reaches accuracy = 0.900, F1 = 0.905, positive acceptance = 0.949, and negative rejection = 0.851. In the wrong-year regime it reaches F1 = 0.737 and negative rejection = 0.796. These results show that metadata consistency is a stronger refusal signal than score alone.

Table 10 and Figure 9 report explanation quality. Top1Trace provides the shortest explanation and preserves company and year names when the top-1 excerpt already contains them, but it has limited item completeness. Top3Trace yields the highest explanation overlap (Exp. ROUGE-L = 0.116) and the highest numeric coverage (0.139) because it exposes more evidence lines. BoardTrace slightly trails Top3Trace on lexical overlap but delivers the strongest provenance profile with company hit = 0.965, year hit = 0.970, and item hit = 0.853. For an executive review workflow, that provenance gain matters more than the small lexical trade-off because it makes the answer auditable.

The combined results therefore reveal a consistent control-stack story. Retrieval benefits most from metadata filtering, answer quality remains constrained by evidence

composition rather than by generation fluency alone, refusal requires explicit metadata checks, and explanation quality depends on whether provenance is made explicit instead of being left implicit in a raw evidence span.

A final point concerns the relation between retrieval and answering. BoardRAG is the strongest retrieval configuration, but its Top-1 score of 0.163 is still far below one. The answer models therefore operate in a regime where retrieval error remains common. The fact that BoardSummary still outperforms the simpler extractive baselines on lexical overlap suggests that answer composition can partially compensate for imperfect retrieval, but only to a limited extent. The same tables also show that composition cannot fully repair wrong evidence. This confirms that retrieval quality remains the leading upstream determinant of downstream board-answer quality.

DISCUSSION

The first substantive conclusion is that metadata are the main retrieval lever in this benchmark. Moving from BM25 with headers to BoardRAG raises Top-1 from 0.138 to 0.163. The improvement does not come from a larger neural retriever or a longer context window. It comes from respecting the company-year structure that the questions themselves reveal. Because more than nine out of ten questions mention an issuer and more than four out of five mention a year, ignoring metadata wastes ranking capacity on preventable confounders. For a real board-support system, the practical implication is direct: metadata control should be engineered before additional model complexity is added.

The second conclusion is that strategic risk questions and ratio questions stress the system in different ways. Risk questions are easier for retrieval because filing language in risk-factor sections is already prose-like and semantically aligned with the question wording. Ratio questions are harder because they impose three linked burdens at once: evidence localization, numeric preservation, and interpretive synthesis. The retrieval and answer tables jointly confirm this pattern. Ratio questions remain the weakest class even though the benchmark provides the relevant excerpt. That finding implies that future gains will come less from generic LLM scaling and more from explicit table normalization, row matching, and formula-aware evidence composition.

A second deployment implication concerns interpretability for non-technical stakeholders. Directors, chief financial officers, and audit committees usually do not want a latent vector explanation of why a paragraph was retrieved. They want a short answer, a statement of the reporting period, and a citation trail that can be checked against the filing. The explanation results therefore have immediate design value. A system that exposes structured provenance fields, even at a small lexical cost, better matches how corporate governance processes actually consume evidence.

The answer comparison reveals a meaningful trade-off between concise synthesis and numeric preservation. BoardSummary achieves the strongest lexical overlap, which makes it the best approximation of a board memo among the deterministic models.

QExtract and oracle-context QExtract preserve numbers slightly better, which makes them more suitable for analyst verification tasks or controller review. In operational terms, the right answer style depends on the consumer. Executives benefit from a concise synthesized note. Reviewers and second-line control functions benefit from a denser extractive answer that exposes more raw evidence.

The oracle-context result is especially informative because it does not dominate the retrieval-based BoardSummary on every metric. That outcome has a clear interpretation. Retrieval is not the only bottleneck. Even when the gold excerpt is given, a deterministic extractor still needs to choose the right lines, suppress table noise, and combine multi-page fragments coherently. In other words, the study's remaining error is partly a composition problem inside the evidence block. This observation strengthens the case for future work on structured table parsing and formula execution rather than on retrieval alone.

Refusal and explanation results show why governance-oriented LLM evaluation must go beyond answer overlap. Without a refusal gate, the system accepts wrong-company and wrong-year evidence almost by default. MetadataGuard reduces that risk substantially and does so with transparent features that compliance, audit, and model-risk teams can inspect. Likewise, explanation results show that a trace is not just a cosmetic citation. BoardTrace materially improves company, year, and SEC-item coverage. That is the difference between an answer that sounds plausible and an answer that can be audited in a board packet or challenged in a controller review meeting.

CONCLUSIONS

This paper delivered a full experimental evaluation of strategic SEC-filing question answering on the complete SecQue benchmark and measured retrieval, answering, refusal, and explanation in one reproducible framework. BoardRAG reached Top-1 retrieval accuracy of 0.163; BoardSummary reached ROUGE-L = 0.124; MetadataGuard reached refusal F1 values of 0.905 and 0.737; and BoardTrace provided the strongest provenance completeness.

The empirical evidence supports three concrete design requirements for top-management decision support over SEC filings: metadata-aware retrieval, explicit refusal control, and auditable source traces. These components are not optional additions to a generator. They are the mechanisms that make a strategic filing assistant dependable in a governance setting.

The study therefore establishes a reproducible baseline for LLM-based strategic risk Q&A over SEC filings. The next development stage is explicit table parsing and formula execution on top of the retrieval-and-control backbone reported here. That extension directly targets the ratio and analysis questions that remain the benchmark's most difficult cases.

Implications

The findings demonstrate that board-level support for SEC filing analysis should be implemented as a controlled evidence workflow rather than as a conventional generative AI application. For enterprise risk officers, chief financial officers (CFOs), and regulatory technology (RegTech) developers, the results suggest that trustworthy AI-assisted decision support depends on integrating three complementary control mechanisms: metadata-aware retrieval, calibrated refusal, and explicit evidence traceability. Together, these mechanisms ensure that every generated response can be traced back to its originating filing section while preventing unsupported or speculative answers. Such an architecture is particularly valuable in governance settings where accountability, auditability, and regulatory compliance are essential.

From a deployment perspective, the proposed workflow can substantially streamline the preparation of board packets and executive briefing materials. Instead of manually reviewing lengthy SEC filings, organizations can employ the metadata-aware retrieval pipeline to rapidly identify relevant disclosures, automatically assemble supporting evidence, and generate traceable summaries for directors and senior executives. Because every response is linked to verifiable source passages and uncertain queries are explicitly refused rather than fabricated, the resulting reports provide a higher level of confidence for strategic decision-making, internal audit activities, risk oversight, and regulatory reporting. The study therefore suggests that organizations should prioritize investments in retrieval governance and evidence management rather than focusing exclusively on increasingly sophisticated generative language models.

Limitations and recomendation

Although this study establishes a reproducible benchmark for metadata-aware retrieval and evidence-grounded question answering, it does not examine how decision makers interact with such systems in organizational contexts. Future research should therefore extend the present work beyond technical performance evaluation by investigating the behavioral and organizational adoption of metadata-aware Retrieval-Augmented Generation (RAG) systems. Management scholars can examine how directors, CFOs, audit committee members, enterprise risk officers, and compliance professionals perceive the usefulness, trustworthiness, transparency, and decision support capabilities of evidence-grounded AI during board-level deliberations.

Established technology adoption frameworks, such as the Technology Acceptance Model (TAM) and the Unified Theory of Acceptance and Use of Technology (UTAUT), provide promising theoretical foundations for these investigations. Future studies could employ controlled boardroom simulations or field experiments to evaluate how metadata-aware retrieval, evidence traceability, calibrated refusal mechanisms, and explanation quality influence perceived usefulness, perceived ease of use, trust, behavioral intention, and actual system utilization during corporate governance activities. Such empirical research would complement the present benchmark by linking technical system performance with executive decision behavior, thereby advancing both AI-enabled corporate governance

research and the broader management literature on trustworthy decision-support systems.

REFERENCES

Araci, D. (2019). FinBERT: Financial sentiment analysis with pre-trained language models. arXiv preprint arXiv:1908.10063. https://doi.org/10.48550/arXiv.1908.10063

Asai, A., Min, S., Zhong, Z., & Chen, D. (2024). Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In Proceedings of the International Conference on Learning Representations. https://doi.org/10.20944/preprints202508.1211.v1

Beltagy, I., Peters, M. E., & Cohan, A. (2020). Longformer: The long-document transformer. arXiv preprint arXiv:2004.05150. https://doi.org/10.48550/arXiv.2004.05150

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N. S., Chen, A., Creel, K., Davis, J. Q., Demszky, D., ... Liang, P. (2021). On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258. https://doi.org/10.48550/arXiv.2108.07258

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., ... Amodei, D. (2020). Language models are few-shot learners. In Advances in Neural Information Processing Systems (Vol. 33, pp. 1877–1901). https://doi.org/10.48550/arXiv.2005.14165

Campbell, J. L., Chen, H., Dhaliwal, D. S., Lu, H., & Steele, L. B. (2014). The information content of mandatory risk factor disclosures in corporate filings. Review of Accounting Studies, 19(1), 396–455. https://doi.org/10.1007/s11142-013-9258-3

Chalkidis, I., Jana, A., Hartung, D., Bommarito, M., Androutsopoulos, I., Katz, D. M., & Aletras, N. (2022). LexGLUE: A benchmark dataset for legal language understanding in English. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Vol. 1, pp. 4310–4330). https://doi.org/10.18653/v1/2022.acl-long.297

Chen, Z., Chen, W., Smiley, C., Shah, S., Borova, I., Langdon, D., Moussa, R., Beane, M., Huang, T.-H., Routledge, B. R., & Wang, W. Y. (2021). FinQA: A dataset of numerical reasoning over financial data. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (pp. 3697–3711). https://doi.org/10.18653/v1/2021.emnlp-main.300

DeYoung, J., Jain, S., Rajani, N. F., Lehman, E., Xiong, C., Socher, R., & Wallace, B. C. (2020). ERASER: A benchmark to evaluate rationalized NLP models. Transactions of the Association for Computational Linguistics, 8, 444–461. https://doi.org/10.18653/v1/2020.acl-main.408

Eppler, M. J., & Mengis, J. (2004). The concept of information overload: A review of literature from organization science, accounting, marketing, MIS, and related disciplines. *The International Journal of Information Management*, *24*(4), 325–344. https://doi.org/10.1016/j.ijinfomgt.2004.04.006

Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, M., & Wang, H. (2023). Retrieval-augmented generation for large language models: A survey. arXiv preprint arXiv:2312.10997. https://doi.org/10.2139/ssrn.4895062

Guu, K., Lee, K., Tung, Z., Pasupat, P., & Chang, M.-W. (2020). REALM: Retrieval-augmented language model pre-training. In Proceedings of the 37th International Conference on Machine Learning (pp. 3929–3938). https://doi.org/10.1145/3581783.3613848

Islam, P., Kannappan, A., Derrick, H., Fodeh, S. J., & Caragea, C. (2023). FinanceBench: A new benchmark for financial question answering. arXiv preprint arXiv:2311.11944. https://doi.org/10.48550/arXiv.2311.11944

Izacard, G., & Grave, E. (2021). 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 (pp. 874–880). https://doi.org/10.18653/v1/2021.eacl-main.74

Jacovi, A., & Goldberg, Y. (2020). Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (pp. 4198–4205). https://doi.org/10.18653/v1/2020.acl-main.386

Jensen, M. C., & Meckling, W. H. (1976). Theory of the firm: Managerial behavior, agency costs and ownership structure. *Journal of Financial Economics*, *3*(4), 305–360. https://doi.org/10.1016/0304-405X(76)90026-X

Kamath, A., Jia, R., & Liang, P. (2020). Selective question answering under domain shift. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (pp. 5684–5696). https://doi.org/10.18653/v1/2020.acl-main.503

Karpukhin, V., Oguz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., & Yih, W.-t. (2020). Dense passage retrieval for open-domain question answering. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (pp. 6769–6781). https://doi.org/10.18653/v1/2020.emnlp-main.550

Khattab, O., & Zaharia, M. (2020). ColBERT: Efficient and effective passage search via contextualized late interaction over BERT. In Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 39–48). https://doi.org/10.1145/3397271.3401075

Kogan, S., Levin, D., Routledge, B. R., Sagi, J. S., & Smith, N. A. (2009). Predicting risk from financial reports with regression. In Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics (pp. 272–280). https://doi.org/10.3115/1620754.1620794

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. In Advances in Neural Information Processing Systems (33). 9459–9474). https://doi.org/10.48550/arXiv.2005.11401

Li, F. (2010). The information content of forward-looking statements in corporate filings—A naïve Bayesian machine learning approach. Journal of Accounting Research, 48(5), 1049–1102. https://doi.org/10.1111/j.1475-679x.2010.00382.x

Loughran, T., & McDonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. Journal of Finance, 66(1), 35–65. https://doi.org/10.1111/j.1540-6261.2010.01625.x

Loughran, T., & McDonald, B. (2015). The use of word lists in textual analysis. Journal of Behavioral Finance, 16(1), 1–11. https://doi.org/10.1080/15427560.2015.1000335

Mialon, G., Dessì, R., Lomeli, M., Nalmpantis, C., Pascal, F., Ranaldi, L., Stojnic, R., Penedo, G., & Scialom, T. (2023). Augmented language models: A survey. Transactions on Machine Learning Research, 2023, 1–34. https://doi.org/10.48550/arXiv.2302.07842

Open AI. (2023). GPT-4 technical report. arXiv preprint arXiv:2303.08774. https://doi.org/10.48550/arXiv.2303.08774

Roberts, A., Raffel, C., & Shazeer, N. (2020). How much knowledge can you pack into the parameters of a language model? In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (pp. 5418–5426). https://doi.org/10.18653/v1/2020.emnlp-main.437

Wu, S., Irsoy, O., Lu, S., Dabravolski, V., Dredze, M., Gehrmann, S., Kambadur, P., Rosenberg, D., & Mann, G. (2023). BloombergGPT: A large language model for finance. arXiv preprint arXiv:2303.17564. https://doi.org/10.48550/arXiv.2303.17564

Zaheer, M., Guruganesh, G., Dubey, A., Ainslie, J., Alberti, C., Ontañón, S., Pham, P., Ravula, A., Wang, Q., Yang, L., & Ahmed, A. (2020). Big Bird: Transformers for longer sequences. In Advances in Neural Information Processing Systems (Vol. 33, pp. 17283–17297). https://doi.org/10.48550/arXiv.2007.14062