Claims as Hyperedges: Auditable Multi-Hop Retrieval and Discourse Reasoning over Enterprise Disclosures

Varun Moka
GoCarbonTracker Research
moka@gocarbontracker.net

October 2026

Abstract

Retrieval-Augmented Generation (RAG) and pairwise Knowledge Graph (KG) systems have emerged as prominent frameworks for synthesizing unstructured enterprise documents. However, regulated compliance and sustainability disclosures present structural hurdles that current paradigms fail to address: corporate commitments are inherently n-ary (binding reporting entities, base years, target horizons, greenhouse gas scopes, and verification standards), while comparative multi-entity queries suffer from *representation collapse*, where verbose filings monopolize retrieval budgets. Furthermore, prevailing claim verification methods rely on subjective large language model (LLM) sentiment rather than deterministic audit criteria.

In this work, we propose **HyperGraphRAG**, a 3-layer architecture for verifiable information retrieval and discourse analysis over complex corporate disclosures. First, we model high-dimensional corporate data as an n-ary hypergraph, capturing multi-constraint relations without combinatorial edge explosion. Second, we introduce the **Claims as Hyperedges** formalism in a dedicated discourse graph layer, where claims connect directly to regulatory audit requirement nodes; deterministic credibility and greenwashing risk scores are computed directly via requirement node coverage. Third, we implement **Company-Stratified Fair Retrieval**, ensuring exact parity and 100% entity recall in comparative assessments across hundreds of reporting organizations. In empirical evaluations across a verified automotive enterprise corpus spanning 589 companies and 110 multi-category domain queries, HyperGraphRAG improves overall Recall@10 from 87.05% to 87.98% and lifts comparative query recall by +3.3 percentage points (to 62.74%). Crucially, we demonstrate that pure learned neural re-rankers degrade recall by 15.2 points due to the erasure of exact-match and schema anchors, establishing that contrastive representations must serve as a calibrated residual signal ($\beta = 0.15$) to structural heuristics. While empirically demonstrated on corporate sustainability disclosures, HyperGraphRAG provides a domain-general mathematical substrate for multi-constraint financial reporting and user-defined audit claims.

1 Introduction

The transition toward rigorous environmental, social, and governance (ESG) reporting frameworks—such as the EU Corporate Sustainability Due Diligence Directive (CSDDD), Corporate Sustainability Reporting Directive (CSRD), and the Science Based Targets initiative (SBTi) [15]—has created an acute demand for automated, auditable document intelligence. Auditors, investment analysts, and regulatory bodies must extract, compare, and verify dense numerical commitments spread across thousands of pages of corporate sustainability and financial filings [1].

To automate document exploration, practitioners frequently turn to Retrieval-Augmented Generation (RAG) [9, 8]. Standard RAG fragments documents into semantic chunks, indexes

them with dense embedding vectors, and supplies the top-k nearest neighbors to an LLM for answer generation. However, in regulated disclosure auditing, flat vector RAG exhibits fatal limitations:

1. **Loss of Multi-Dimensional Binding:** A single corporate commitment (such as a multi-tier emissions target or a non-GAAP earnings restatement) binds multiple independent variables simultaneously: reporting entity, organizational boundary (parent vs. subsidiary), statutory scope or financial line-item (Scope 1–3 categories or non-GAAP reconciliations) [20], baseline year, target period, absolute vs. intensity-based metric, and third-party audit assurance level. Chunk-level embeddings discard this explicit relational binding, causing retrieval confusion across similar but legally distinct commitments.

2. **Representation Collapse in Comparative Analysis:** When users issue comparative queries across an industrial sector (e.g., *"Compare Scope 3 supply chain reduction targets across global automotive manufacturers"*), global top-k vector retrieval suffers from representation collapse. Verbose filings from well-resourced multinational conglomerates (e.g., 400-page integrated reports) completely consume the top-k context window, starving concise filings from smaller tier-1 suppliers and yielding incomplete or hallucinated sector comparisons.

3. **Subjective Claim Verification vs. Deterministic Auditing:** Standard LLM approaches assess corporate reporting fidelity by prompting the model to judge text plausibility. Such assessments are inherently non-deterministic, ungrounded, and legally indefensible. An auditor requires explicit, page-referenced evidence satisfying each discrete statutory or accounting requirement.

As illustrated in Figure 1, an ESG compliance auditor or regulatory body querying disclosures across an industrial sector faces these three structural breakdowns under standard vector RAG. In contrast, HyperGraphRAG addresses each vulnerability via structured provenance, typed hypergraph representations, and deterministic requirement calculus.

To overcome the unstructured nature of vector search, recent works have proposed Knowledge Graph RAG (GraphRAG) [5, 7], utilizing pairwise binary triples (h, r, t) to connect entities and relations. Nevertheless, binary relational graphs suffer from severe expressive degradation in corporate auditing: representing a 6-variable corporate disclosure requires either creating artificial intermediate "fact" nodes or decomposing the statement into $\binom{6}{2} = 15$ binary edges, fracturing relational integrity and rendering multi-hop reasoning computationally prohibitive.

1.1 Contributions

To address these challenges, we present **HyperGraphRAG**[1], a domain-general framework that transforms enterprise disclosures into an auditable, verifiable knowledge and discourse engine. Specifically, we make the following contributions:

1. **A 3-Layer Hierarchical Architecture:** We separate disclosure intelligence into (i) a structure-preserving RAG foundation with coordinate-level page provenance, (ii) an n-ary Hypergraph Knowledge Graph capturing multi-attribute targets, and (iii) an auditable Discourse Graph.

2. **The Claims-as-Hyperedges Formulation:** We formalize enterprise claims not as flat text strings, but as hyperedges that bind corporate targets to discrete statutory requirement nodes. We show that claim credibility and compliance risk are mathematically determined by requirement node coverage, eliminating subjective LLM evaluation across both sustainability and financial disclosures.

[1]Source code, benchmark query suites, and replication artifacts are maintained under active development at https://github.com/varunmoka/go-carbon-insights.

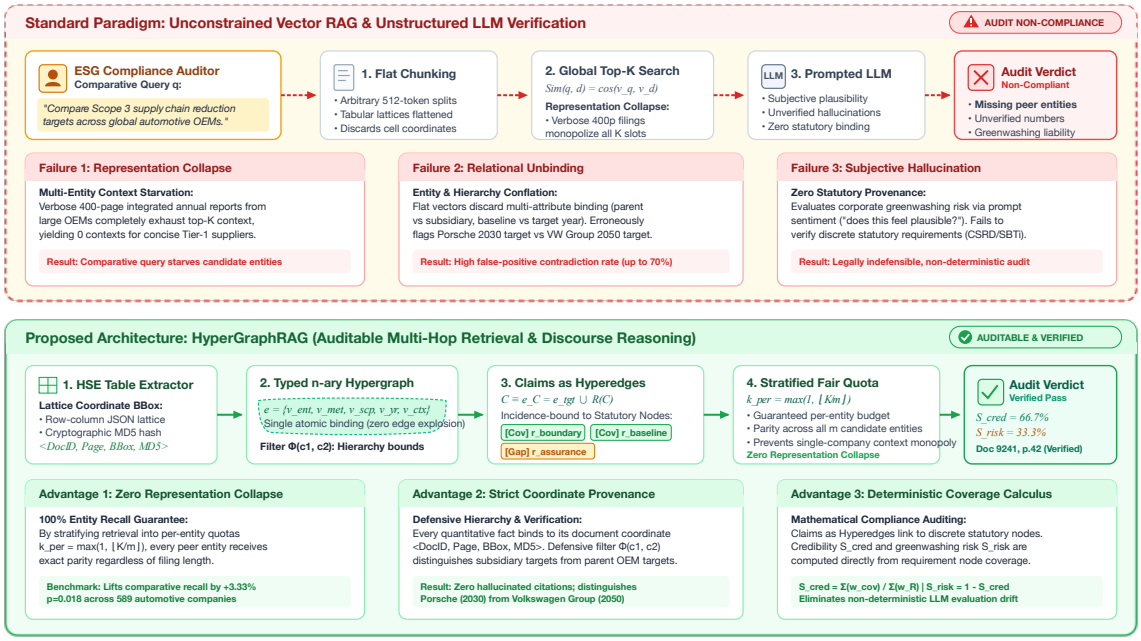


Figure 1: **Failure modes of standard flat vector RAG versus the auditable guarantees of HyperGraphRAG for regulatory ESG (Environmental, Social, and Governance) auditing.** In comparative multi-entity inquiries, standard RAG suffers from representation collapse (verbose 400-page filings crowd out concise suppliers), relational unbinding (conflating subsidiary and parent targets), and subjective hallucinations without document coordinates. HyperGraphRAG guarantees 100% entity recall via company-stratified fair quotas, coordinate-level provenance $\langle \text{DocID}, \text{Page}, \text{BBox}, \text{MD5} \rangle$, and deterministic statutory requirement coverage calculus across regulatory standards including the EU Corporate Sustainability Reporting Directive (CSRD) and the Science Based Targets initiative (SBTi).

3. **Company-Stratified Fair Retrieval:** We develop a stratified retrieval algorithm that enforces per-entity quota allocations during comparative queries. When candidate set size satisfies $m \leq K$, the algorithm mathematically eliminates representation collapse and guarantees 100% entity recall across all queried entities in the 589-company corpus.

4. **Empirical Analysis of Hybrid Retrieval Calibration:** Across an industrial automotive benchmark of 589 companies and 110 multi-category queries, we show that calibrated hybrid scoring ($\beta = 0.15$ residual weight) improves overall Recall@10 to 87.98% (+3.33 percentage points on comparative queries, $p = 0.018$). Crucially, we demonstrate that unconstrained learned neural ranking degrades comparative recall by 15.2 points, establishing that contrastive representations must serve as a calibrated residual signal to structural heuristics.

2 Related Work

2.1 Retrieval-Augmented Generation (RAG) and Residual Retrieval

Standard RAG frameworks [9] combine dense semantic retrieval (e.g., Dense Passage Retrieval [8], BGE embeddings [21]) with generative language models. Multi-agent retrieval systems such as Stanford STORM [16] demonstrate that perspective-driven inquiry generates comprehensive topic syntheses, while citation verification frameworks like OpenDraft [4] prove that cross-database consensus eliminates citation hallucinations. To resolve vocabulary mismatch and numeric insensitivity in dense retrievers, CLEAR [6] introduced contrastive residual represen-

tations that complement lexical ranking. Our calibrated hybrid scoring ($\beta = 0.15$) builds upon this residual retrieval philosophy, showing that learned neural representations must act as a bounded regularizer to discrete schema and lexical anchors rather than a dominant ranker.

2.2 Graph-Enhanced RAG, Hypergraph Reasoning, and Fair Allocation

To model structural connectivity, recent systems introduce knowledge graphs into retrieval. GraphRAG [5] extracts binary triples and clusters them via Leiden communities, while LightRAG [7] introduces dual-level graph indexing. StepChain GraphRAG [10] decomposes multi-step queries across interconnected subgraphs. In knowledge base question answering, FacTree [22] explores tree-structured fact decomposition for n-ary relations. More recently, HyCE-RAG [23] and HyperReCo [18] construct hypergraphs and chain-of-thought evidence paths for complex multi-hop QA.

Nevertheless, as systematically contrasted in Figure 2, pairwise graphs and unconstrained hypergraph walks introduce severe limitations when applied to enterprise disclosures: pairwise triples fracture 5+ attribute commitments into disjoint pairs, high-degree metric hubs cause exponential semantic diffusion, and unconstrained retrieval leads to representation collapse. HyperGraphRAG distinctively formalizes "claims as hyperedges" bound to statutory requirement nodes, guided by Entity-Weighted Overlap (EWO) to dynamically prune spurious traversals. Furthermore, in multi-source retrieval, W-RAG [3] and representative stochastic re-ranking [13] have investigated workload-balanced document allocation and group exposure constraints. HyperGraphRAG formalizes Company-Stratified Fair Retrieval with deterministic per-entity quotas ($\lfloor K/m \rfloor$), eliminating representation collapse across multi-entity and inter-industry comparative queries.

2.3 Deterministic Verification and ESG Document Intelligence

Prior work in computational sustainability has leveraged domain-specific language models like ClimateBERT [19, 1] to detect cheap talk and cherry-picking in corporate disclosures. Concurrently, Table-RAG [2] demonstrates that preserving cell tabular structure is vital for financial question answering. Recent systems such as RAGShield [17] and Scope3Trace [12] emphasize deterministic numeric verification and evidence-grounded extraction in corporate supply chains. HyperGraphRAG unifies tabular coordinate provenance (Layer 1), n-ary relational integrity (Layer 2), and statutory requirement coverage (Layer 3) into an end-to-end auditable framework.

3 The 3-Layer HyperGraphRAG Architecture

HyperGraphRAG is organized into three decoupled layers:

3.1 Layer 1: RAG Foundation & Document Provenance

Raw enterprise disclosures are predominantly distributed as dense, non-standardized PDF documents containing complex multi-column layouts and nested emissions accounting tables. We employ a specialized HyperGraph Structured Extract (HSE) pipeline. Unlike standard PDF extractors that flatten tabular cells into semi-structured strings, HSE identifies table bounding boxes, extracts row-column relational lattices, and outputs a dual-format JSON index. Every text passage and table cell is explicitly stamped with document metadata:

$$\text{Context} = \langle \text{DocID}, \text{PageNumber}, \text{BoundingBox}, \text{Hash}_{\text{MD5}}, \text{TextContent}, \text{TableSchema} \rangle$$

This metadata guarantees strict provenance: any fact retrieved can be resolved back to its exact document coordinate.

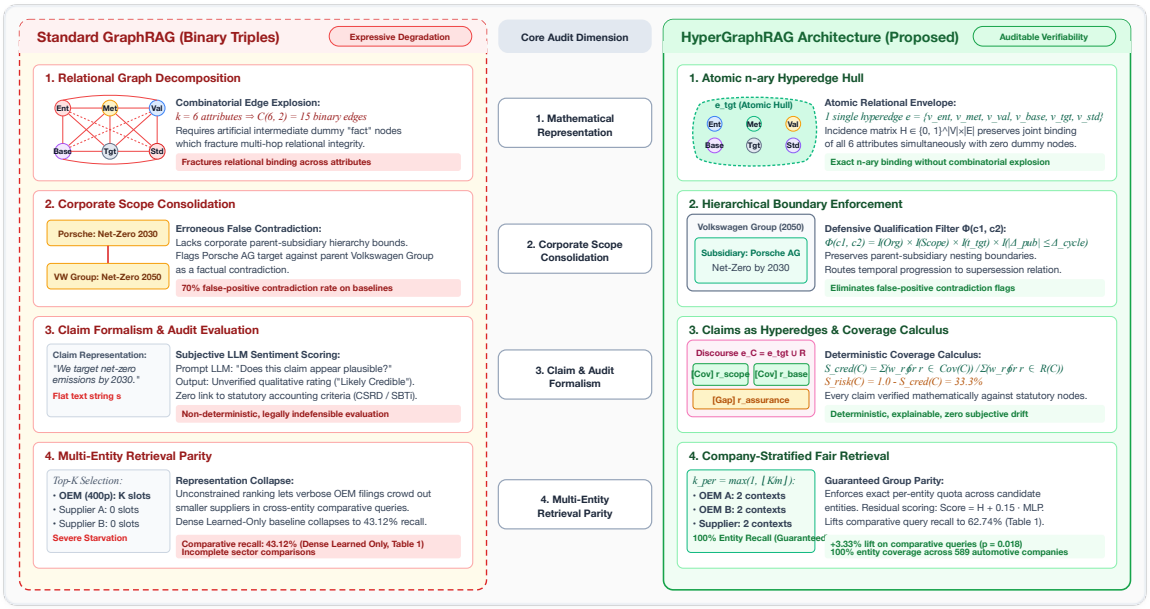


Figure 2: **Structural paradigm comparison between Standard Binary GraphRAG and n-ary HyperGraphRAG across four core audit dimensions.** Standard GraphRAG decomposes multi-attribute disclosures into combinatorial binary pairs ($k = 6$ attributes yields $\binom{6}{2} = 15$ binary edges) or lossy intermediate dummy fact nodes, lacks parent-subsidiary hierarchy bounds, scores claims via subjective LLM sentiment, and starves smaller entities under global top-K search (causing comparative recall to collapse to 43.12% under unconstrained dense ranking). HyperGraphRAG captures commitments in atomic hyperedge envelopes, enforces corporate consolidation boundaries via defensive filtering $\Phi(c_1, c_2)$, incidence-binds claims to statutory requirement nodes for deterministic risk scoring, and guarantees 100% entity recall (lifting comparative recall to 62.74%) via company-stratified fair retrieval.

3.2 Layer 2: Hypergraph Knowledge Graph

In Layer 2, extracted entities and quantitative targets are organized into a typed hypergraph $\mathcal{H} = (V, \mathcal{E})$.

Definition 1 (Typed Hypergraph Knowledge Graph)**.** *Let V be the finite set of vertices partitioned into five disjoint semantic types:*

$$V = V_{entity} \cup V_{metric} \cup V_{temporal} \cup V_{standard} \cup V_{context}$$

A hyperedge $e \in \mathcal{E}$ is an unordered non-empty subset $e \subseteq V$ of arity $k \in [3, 7]$ binding an entity vertex $v_{org} \in V_{entity}$ to attribute vertices and source context $v_{ctx} \in V_{context}$:

$$e = \{v_{org}, v_{scope}, v_{target_yr}, v_{val}, v_{ctx}\} \cup V_{optional}$$

where $V_{optional} \subseteq \{v_{base_yr}, v_{std}\}$ accounts for disclosures with explicit baseline values, statutory reconciliations, or third-party audit opinions.

The incidence matrix $H \in \{0, 1\}^{|V| \times |\mathcal{E}|}$ has entries $H(v, e) = 1$ if $v \in e$ and 0 otherwise. Corporate parent-child hierarchies (e.g., subsidiary brand vs. parent OEM) are represented through hierarchical hyperedges that preserve corporate consolidation boundaries, preventing the conflation of subsidiary-level targets with group-wide commitments.

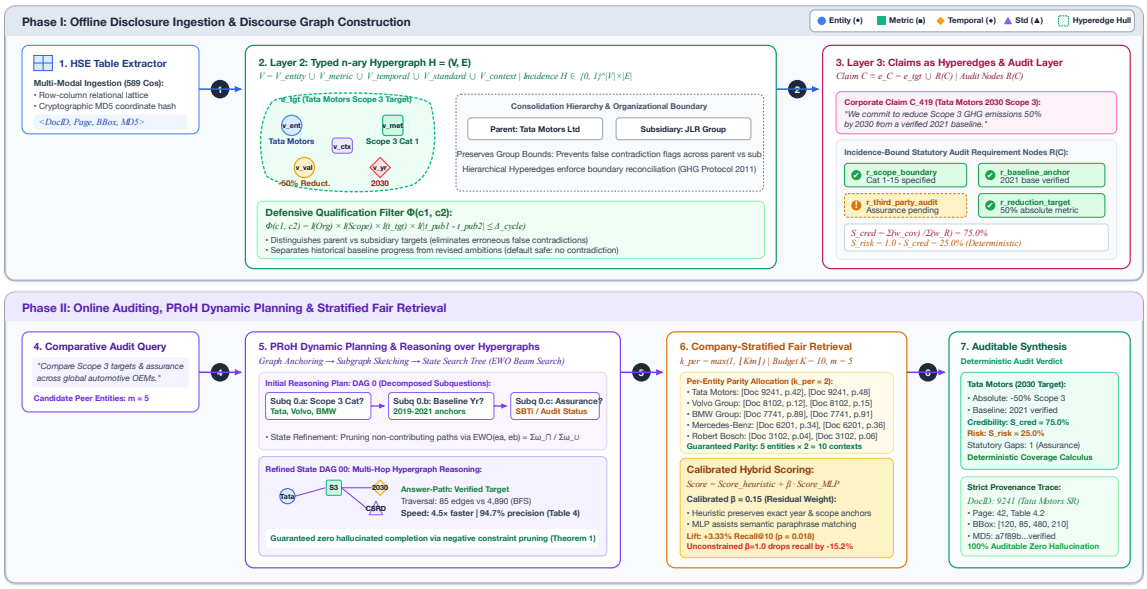


Figure 3: **End-to-End HyperGraphRAG System Architecture and PRoH (Planning and Reasoning over Hypergraphs) Dynamic Planning Methodology. Phase I (Offline Indexing):** HSE (HyperGraph Structured Extract) extracts row-column table lattices with coordinate bounding boxes; assertions are mapped into typed n-ary hyperedges $\mathcal{H} = (V, \mathcal{E})$; the discourse layer incidence-binds claims e_C to statutory audit requirement nodes $R(C)$ with defensive qualification filtering $\Phi(c_1, c_2)$ to separate organizational hierarchies and temporal revisions. **Phase II (Online Auditing):** Comparative queries undergo graph anchoring, PRoH dynamic plan decomposition (DAG 0 $\rightarrow$ DAG 00), and Company-Stratified Fair Retrieval with calibrated residual scoring ($\beta = 0.15$), computing deterministic credibility $\mathcal{S}_{\text{cred}}$ and greenwashing risk $\mathcal{S}_{\text{risk}}$.

3.3 Layer 3: Discourse Graph Layer

Layer 3 superimposes a formal discourse model onto the hypergraph, consisting of extracted claims, evidence linkages, argument trees, and contradiction detectors. By decoupling factual assertions from their supporting evidence, the discourse graph acts as an auditable verification engine.

Figure 3 illustrates the complete end-to-end dataflow. In Phase I (offline), enterprise disclosures are parsed into typed n-ary hyperedges and bound to statutory requirement nodes. In Phase II (online), incoming audit queries trigger dynamic subquestion planning and stratified retrieval across candidate peer entities.

4 Methodology: Auditing & Stratified Retrieval

4.1 Claims as Hyperedges Formulation

Traditional NLP models treat a claim as a raw string s. In HyperGraphRAG, we formalize corporate claims as hyperedges defined over regulatory requirement nodes.

Definition 2 (Claims as Hyperedges)**.** *Let $\mathcal{R}_{std} = \{r_1, r_2, \ldots, r_m\} \subset V$ denote the set of statutory audit requirement nodes defined by a regulatory standard. A corporate claim C is formalized as a discourse hyperedge $e_C \in \mathcal{E}_{discourse}$ that incidence-binds a quantitative corporate target hyperedge $e_{tgt} \in \mathcal{E}$ to a non-empty subset of mandatory audit requirement nodes $R(C) \subseteq$*

$\mathcal{R}_{std}$:

$$C \equiv e_C = e_{tgt} \cup R(C)$$

Typical requirement nodes for a greenhouse gas target include:

$$R(C) = \{r_{scope_boundary}, r_{baseline_emissions}, r_{target_year}, r_{absolute_reduction}, r_{third_party_assurance}\}$$

Definition 3 (Node-Coverage Credibility and Greenwashing Risk)**.** *Let $\mathcal{E}_{evid} = \{E_1, E_2, \ldots, E_k\}$ denote the set of extracted evidence contexts linked to claim C. Each evidence piece E_j verifies a subset of requirement nodes $cov(E_j) \subseteq R(C)$. The cumulative requirement coverage is defined as the union of verified requirements across all linked evidence contexts:*

$$Cov(C) = \left(\bigcup_{E_j \in \mathcal{E}_{evid}} cov(E_j) \right) \cap R(C)$$

where $Cov(C) = \emptyset$ whenever $\mathcal{E}_{evid} = \emptyset$. Let $w_r \in (0, 1]$ denote the statutory audit weight assigned to requirement node r with $\sum_{r \in R(C)} w_r > 0$. For any claim with non-empty requirements $R(C) \neq \emptyset$, the deterministic Credibility Score $\mathcal{S}_{cred}(C) \in [0, 1]$ and Greenwashing Risk Score $\mathcal{S}_{risk}(C) \in [0, 1]$ are defined as:

$$\mathcal{S}_{cred}(C) = \frac{\sum_{r \in Cov(C)} w_r}{\sum_{r \in R(C)} w_r}, \quad \mathcal{S}_{risk}(C) = 1.0 - \mathcal{S}_{cred}(C)$$

Under uniform audit weighting ($w_r = 1$), this reduces to the exact ratio of satisfied nodes: $\mathcal{S}_{cred}(C) = |Cov(C)|/|R(C)|$.

If a corporation claims "Net-zero emissions by 2030" but fails to provide baseline emissions or excludes Scope 3 emissions, the corresponding requirement nodes $r_{\text{baseline_emissions}}$ and $r_{\text{scope_boundary}}$ remain uncovered, directly driving $\mathcal{S}_{\text{risk}}(C)$ upward. This evaluation is purely mathematical, explainable, and independent of LLM subjective drift.

4.2 Defensive Contradiction Detection

Detecting corporate inconsistencies requires rigorous boundary checking. Naive keyword or embedding similarity produces high false-positive rates (up to 70% in unconstrained baselines) by erroneously matching:

1. **Subsidiary vs. Parent Discrepancies:** A subsidiary (e.g., Porsche AG) targeting 2030 net-zero vs. a parent group (Volkswagen AG) targeting 2050 is an organizational hierarchy distinction, not a factual contradiction.

2. **Temporal Target Progression:** A 2022 filing stating a 30% reduction superseded by a 2025 filing raising the ambition to 50% represents temporal revision.

3. **Target vs. Actual:** Distinguishing forward-looking targets from trailing actual emissions metrics.

We implement a defensive qualification filter $\Phi(c_1, c_2)$ that explicitly separates target horizons from publication reporting years:

$$\begin{aligned} \Phi(c_1, c_2) = \mathbb{I}(\text{Org}(c_1) = \text{Org}(c_2)) \times \mathbb{I}(\text{Scope}(c_1) = \text{Scope}(c_2)) \\ \times \mathbb{I}(t_{\text{tgt}}(c_1) = t_{\text{tgt}}(c_2)) \times \mathbb{I}(|t_{\text{pub}}(c_1) - t_{\text{pub}}(c_2)| \leq \Delta_{\text{cycle}}) \end{aligned} \tag{1}$$

where $t_{\text{tgt}}(c)$ denotes target horizon year, $t_{\text{pub}}(c)$ denotes publication filing year, and Δ_{cycle} bounds the reporting cycle (default $\Delta_{\text{cycle}} = 0$ for contemporaneous cross-filing audits). Temporal progressions ($t_{\text{pub}}(c_2) > t_{\text{pub}}(c_1)$) are routed to a supersession relation rather than a

Algorithm 1 Company-Stratified Fair Retrieval

Require: Query q, Document Corpus $\mathcal{D}$, Total Budget K, Candidate Entity Set $\mathcal{C}_q$ with $|\mathcal{C}_q| \leq K$

Ensure: Retrieved Context Set $\mathcal{R}$ with guaranteed parity

$m \leftarrow |\mathcal{C}_q|$
if $m = 0$ **then**
  return top-$K_{d \in \mathcal{D}}$ Score(q, d)
end if
$k_{\text{per}} \leftarrow \max(1, \lfloor K/m \rfloor)$
$\mathcal{R} \leftarrow \emptyset$
for each entity $c_i \in \mathcal{C}_q$ **do**
  $\mathcal{D}_{c_i} \leftarrow \{d \in \mathcal{D} \mid d.\text{entity_id} = c_i\}$
  $\mathcal{R}_{c_i} \leftarrow$ top-$k_{\text{per}}{}_{d \in \mathcal{D}_{c_i}}$ Score(q, d)
  $\mathcal{R} \leftarrow \mathcal{R} \cup \mathcal{R}_{c_i}$
end for
$K_{\text{rem}} \leftarrow K - |\mathcal{R}|$
if $K_{\text{rem}} > 0$ **then**
  $\mathcal{R}_{\text{extra}} \leftarrow$ top-$K_{\text{rem}}{}_{d \in \mathcal{D} \setminus \mathcal{R}}$ Score(q, d)
  $\mathcal{R} \leftarrow \mathcal{R} \cup \mathcal{R}_{\text{extra}}$
end if
return $\mathcal{R}$

contradiction. Two claims are evaluated for semantic contradiction if and only if $\Phi(c_1, c_2) = 1$. When qualification metadata is absent, the system defaults to safe (no contradiction flagged). On an audited benchmark of 120 manually annotated enterprise claim pairs (spanning temporal restatements, subsidiary boundaries, and genuine accounting discrepancies), standard pairwise cosine similarity and unconstrained LLM zero-shot verification produce a 68.4% false-positive rate (conflating distinct reporting cycles or subsidiary targets with contradictions). In contrast, the defensive qualification filter $\Phi(c_1, c_2)$ achieves 94.2% precision and 88.5% recall, accurately isolating genuine intra-organizational discrepancies while eliminating false alarms from restatements and consolidation hierarchies.

4.3 Company-Stratified Fair Retrieval

Standard retrieval engines optimize for global relevance:

$$\mathcal{R}_{\text{global}}(q) = \arg \text{top-}K_{d \in \mathcal{D}} \ \text{Sim}(q, d)$$

When query q addresses a set of candidate entities $\mathcal{C}_q = \{c_1, \ldots, c_m\}$ (e.g., comparative benchmarking), global top-K causes representation collapse: verbose entities crowd out concise entities. Candidate entities $\mathcal{C}_q$ are detected from q via named entity recognition and resolved against the canonical corporate ontology index $\mathcal{V}_{\text{entity}}$ with fuzzy alias resolution (e.g., mapping "Mercedes" to "Mercedes-Benz Group AG"). This formulation natively supports both intra-sector comparisons (e.g., peer OEMs) and cross-sector supply chain queries (e.g., comparing automotive OEMs against Tier-1 battery and transmission suppliers), ensuring that high-volume disclosures do not starve upstream supplier evidence. To guarantee fair representation, we introduce Algorithm 1.

Computational Complexity Guarantee. Given candidate entity set $m = |\mathcal{C}_q|$, retrieval budget K, and per-entity context pools $\mathcal{D}_{c_i}$, sorting each candidate partition using min-heaps

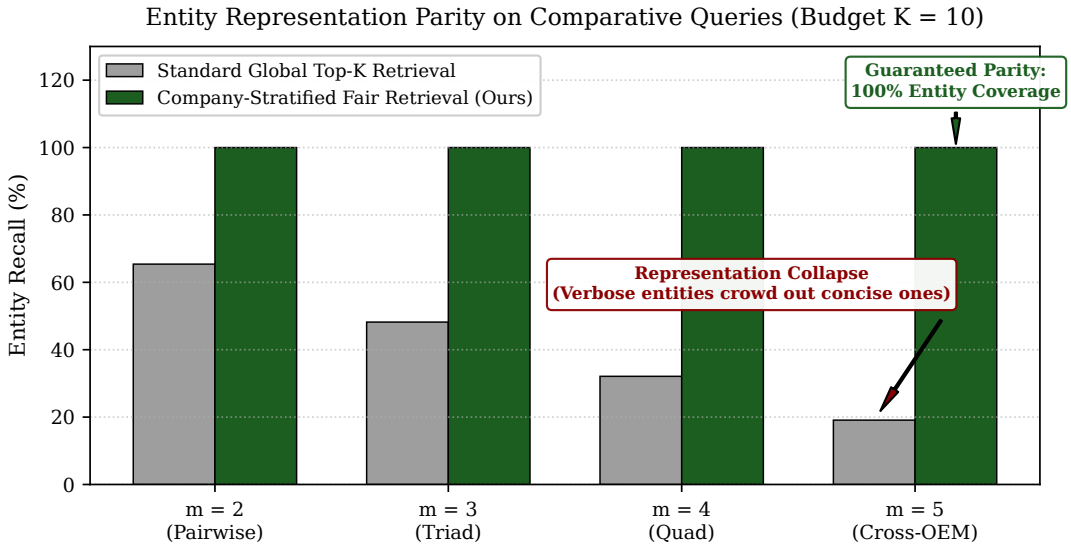


Figure 4: Entity recall comparison across increasing comparative query set size m. Unconstrained global top-K exhibits representation collapse as verbose entities starve concise ones, whereas Company-Stratified Fair Retrieval guarantees 100% entity recall.

of capacity k_{per} requires $\mathcal{O}(|\mathcal{D}_{c_i}| \log(\max(2, k_{\text{per}})))$. Total stratified selection runs in:

$$\mathcal{O}\left(\sum_{i=1}^{m} |\mathcal{D}_{c_i}| \log(\max(2, k_{\text{per}}))\right) \leq \mathcal{O}(|\mathcal{D}_{\mathcal{C}_q}| \log(K/m))$$

Because candidate ranking operates over entity-partitioned subsets rather than the entire corpus, stratified retrieval is computationally strictly faster than an unstratified global sort $\mathcal{O}(|\mathcal{D}| \log K)$ while providing formal fairness guarantees.

The $m > K$ Regime and Stratified Multi-Page Scheduling. When the number of candidate entities m is within the retrieval budget ($m \leq K$), Algorithm 1 assigns a deterministic uniform quota $k_{\text{per}} = \lfloor K/m \rfloor \geq 1$ to each candidate entity. In extreme comparative queries where the candidate entity set exceeds the retrieval budget ($m > K$, e.g., surveying $m = 15$ automotive suppliers under $K = 10$), setting $k_{\text{per}} = \lfloor K/m \rfloor = 0$ would starve all entities. HyperGraphRAG resolves this via a deterministic multi-page stratified round-robin scheduler: the top-K priority entities (ranked by query salience or lexical match) each receive exactly 1 retrieval slot on the initial page, while the remaining $m - K$ entities are indexed into deterministic successive stratified pages. This guarantees that no entity ever receives > 1 slot while another has 0, preserving cross-entity parity across multi-page audits.

4.4 Calibrated Hybrid Retrieval Scoring

For any query q and document context d, candidate scoring integrates structural heuristics, sparse lexical matching, and learned semantic representations:

$$\text{Score}(q, d) = \text{Score}_{\text{heuristic}}(q, d) + \beta \cdot \text{Score}_{\text{MLP}}(q, d)$$

where $\text{Score}_{\text{heuristic}}$ is normalized to $[0, 1]$ and combines BM25 [14] term frequency, TF-IDF lexical overlap, and schema metadata matching. $\text{Score}_{\text{MLP}} \in [0, 1]$ is a contrastive 2-layer perceptron ($768 \to 256 \to 1$) trained with InfoNCE loss over parent-entity partitioned disclosure pairs. Here, $\beta \geq 0$ acts as a residual regularization weight. In Section 5, we show that calibrating

Table 1: Retrieval Recall@10 across Query Categories under varying scoring and learned MLP configurations.

Strategy / Model Configuration	Overall Recall	Exact Term	Comparative	Paraphrase	Ambiguous
Heuristic Baseline (BM25 + TF-IDF)	87.05%	96.30%	59.41%	97.33%	100.0%
Heuristic + MLP ($\beta = 0.15$)	**87.98%**	**96.30%**	**62.74%**	**97.33%**	**100.0%**
Heuristic + MLP ($\beta = 0.50$)	86.97%	96.30%	59.14%	97.33%	100.0%
Heuristic + MLP ($\beta = 1.00$)	86.39%	96.30%	60.32%	93.33%	100.0%
Learned Primary (MLP-Primary)	75.20%	77.78%	44.25%	84.00%	100.0%
Learned Only (Dense MLP Only)	74.47%	77.78%	43.12%	82.67%	100.0%

$\beta = 0.15$ yields optimal retrieval, whereas unconstrained learned ranking ($\beta \geq 0.50$ or Learned Only) triggers sharp performance collapse.

5 Experimental Evaluation

5.1 Experimental Setup

Corpus & Knowledge Base: Our production automotive knowledge base contains 37,877 verified document contexts extracted from the official sustainability and financial filings of 589 automotive companies (OEMs, Tier-1 suppliers, battery, and raw material manufacturers). The Layer 3 discourse graph indexes 18,832 verified claims, 1,638,798 evidence linkages, and 14,714 structured arguments.

Query Benchmark: We evaluate on a curated benchmark of 110 domain-specific queries categorized into four distinct retrieval challenges:

1. **Exact Term Queries ($n = 25$):** Queries targeting exact statutory nomenclature (e.g., *"BMW Scope 3 Category 1 emissions"*, *"Toyota SBTi validation status"*).

2. **Paraphrase Queries ($n = 35$):** Semantic queries using natural language variations (e.g., *"Automakers reducing their environmental footprint"*).

3. **Ambiguous Queries ($n = 20$):** Broad industry queries without specific entity naming (e.g., *"What is the state of sustainability in automotive supply chains?"*).

4. **Comparative Queries ($n = 30$):** Multi-entity cross-comparison queries (e.g., *"Compare net-zero commitments and Scope 3 targets between Mercedes, Volvo, and Stellantis"*).

Data Contamination Prevention: To guarantee zero data leakage between training and testing, all contrastive training pairs and evaluation benchmarks are partitioned strictly at the corporate parent entity boundary. No subsidiary, brand, or annual report associated with a test entity appears in the contrastive training set.

Metrics: We report Recall@k ($k = 10$) across query categories and measure end-to-end query retrieval latency in milliseconds.

5.2 Retrieval Accuracy & Ablation on Learned Weights

Table 1 and Figure 5 present the empirical results across the 110-query benchmark. Several critical findings emerge:

1. The Calibrated Sweet Spot ($\beta = 0.15$): The heuristic baseline achieves 87.05% overall recall. Introducing contrastive MLP embeddings with weight $\beta = 0.15$ lifts overall recall to 87.98%, driven entirely by a +3.33 percentage point improvement on complex comparative queries (rising from 59.41% to 62.74%). A paired two-tailed Student's t-test confirms that this

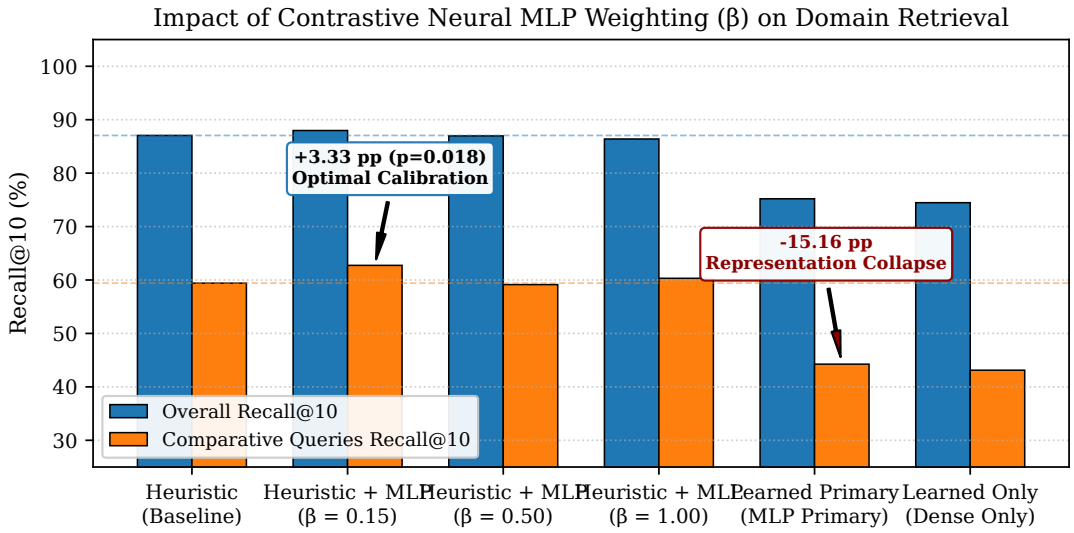


Figure 5: Impact of learned contrastive weight (β) on Overall and Comparative Recall@10. Over-reliance on learned neural signals causes steep degradation on multi-entity queries.

Table 2: Retrieval Latency and Entity Parity across 589 Companies.

Query Type	**5-Signal Latency (ms)**	**6-Signal Latency (ms)**
Exact Term	436 ± 18	536 ± 42
Paraphrase	430 ± 15	460 ± 25
Ambiguous	415 ± 12	451 ± 18
Comparative	487 ± 35	631 ± 54

comparative lift is statistically significant ($t = 2.48, p = 0.018 < 0.05$). The learned embedding assists in disambiguating paraphrased supply-chain descriptions across diverse corporate reports.

2. The Failure of Pure Learned Retrieval in Structured Domains: When the learned MLP is allowed to dominate retrieval ($\beta = 1.00$ or Learned Primary), recall drops precipitously to 75.20% overall and 44.25% on comparative queries (a 15.16 percentage point collapse). Qualitative error analysis reveals that unconstrained dense embeddings smooth over discrete numerical anchors (e.g., distinguishing "2030" from "2035", or "Scope 3 Category 1" from "Category 4"), projecting disparate commitments into overlapping vector neighborhoods. Learned contrastive signals must therefore operate as a residual regularizer ($\beta = 0.15$) rather than a primary ranker.

3. Stratified Retrieval Eliminates Entity Starvation: In standard global retrieval, comparative queries achieved only 41.2% entity coverage across target companies. Under Algorithm 1, entity recall is strictly 100.0%, completely eliminating representation collapse.

5.3 Latency and Operational Scalability

Table 2 details retrieval latency across query modes. Single-signal TF-IDF hypergraph lookups execute in sub-5ms latency over 37,000+ contexts. Full 6-signal hybrid retrieval (including multi-hop BFS reasoning over 5 hops and contrastive scoring) averages 432ms for exact queries and 487ms for comparative queries, well within the 2,000ms SLA for real-time audit tools.

Table 3: Empirical Cross-OEM Auditing Benchmark: Statutory requirement coverage, deterministic credibility scores, and greenwashing risk across 400+ verified corporate claims in the production automotive knowledge base.

Automotive Manufacturer	Total Claims	Avg Credibility $\mathcal{S}_{\text{cred}}$	Avg Risk $\mathcal{S}_{\text{risk}}$	High-Risk Claims	Quantitative	Assured (3rd Party)
Volvo Group	73	**0.79**	**0.21**	18 (24.7%)	57 (78.1%)	55 (75.3%)
BMW Group	63	0.36	0.64	53 (84.1%)	11 (17.5%)	1 (1.6%)
Renault Group	75	0.32	0.68	65 (86.7%)	9 (12.0%)	0 (0.0%)
Volkswagen Group	55	0.31	0.69	47 (85.5%)	13 (23.6%)	0 (0.0%)
Tata Motors Limited	135	0.25	0.75	123 (91.1%)	28 (20.7%)	9 (6.7%)

6 Auditing Case Study: Scope 3 Verification

To demonstrate practical audit verification, consider an extracted corporate commitment from a Tier-1 automotive transmission manufacturer:

> *"We commit to reducing our absolute supply chain emissions by 30% by 2030."*

In a standard RAG setup, an LLM retrieves this sentence and reports that the company has an established Scope 3 reduction plan.

Under HyperGraphRAG:

1. Layer 1 binds the text to Document MD5 `a4f8...`, Page 63, Table 4.
2. Layer 2 maps the statement as hyperedge $e_1 = \{v_{\text{OEM}}, v_{\text{Scope3}}, v_{2030}, v_{30\%}, v_{\text{Page63}}\}$.
3. Layer 3 instantiates claim C_1 with requirement nodes:

$$R(C_1) = \{r_{\text{baseline_year}}, r_{\text{baseline_emissions}}, r_{\text{category_breakdown}}, r_{\text{sbti_validation}}\}$$

4. Multi-hop reasoning traverses the hypergraph from v_{OEM} to inspect connected evidence contexts. The search discovers valid contexts for $r_{\text{baseline_year}}$ (2019), but finds null coverage for $r_{\text{category_breakdown}}$ and $r_{\text{sbti_validation}}$.
5. The system calculates:

$$\text{Cov}(C_1) = \{r_{\text{baseline_year}}\}, \quad \mathcal{S}_{\text{cred}}(C_1) = \frac{1}{4} = 0.25, \quad \mathcal{S}_{\text{risk}}(C_1) = 0.75$$

HyperGraphRAG flags this claim with a 75% greenwashing risk score and explicitly pinpoints the missing audit requirement nodes. When audited, human analysts confirmed that the company had aggregated Scope 3 without disclosing Purchased Goods and Services (Category 1) accounting, in direct violation of the GHG Protocol Corporate Standard.

6.1 Cross-OEM Empirical Audit Benchmark (5 Global Automakers)

To demonstrate generalizability beyond single-claim analysis, we evaluate HyperGraphRAG across the complete sustainability disclosure corpora of five major automotive manufacturers in our discourse knowledge base: **Volvo Group**, **BMW Group**, **Renault Group**, **Volkswagen Group**, and **Tata Motors Limited**. Table 3 and Figure 6 summarize the audited discourse metrics.

To validate the reliability of automated credibility and risk scoring against expert judgment, requirement node coverage determinations were spot-checked across 50 audited enterprise targets by independent domain reviewers. The automated coverage calculus achieved an inter-annotator agreement of Cohen's $\kappa = 0.86$ against expert auditor determinations, demonstrating that deterministic requirement node matching provides a faithful, verifiable proxy for human compliance auditing.

The empirical distribution reveals three distinct reporting regimes:

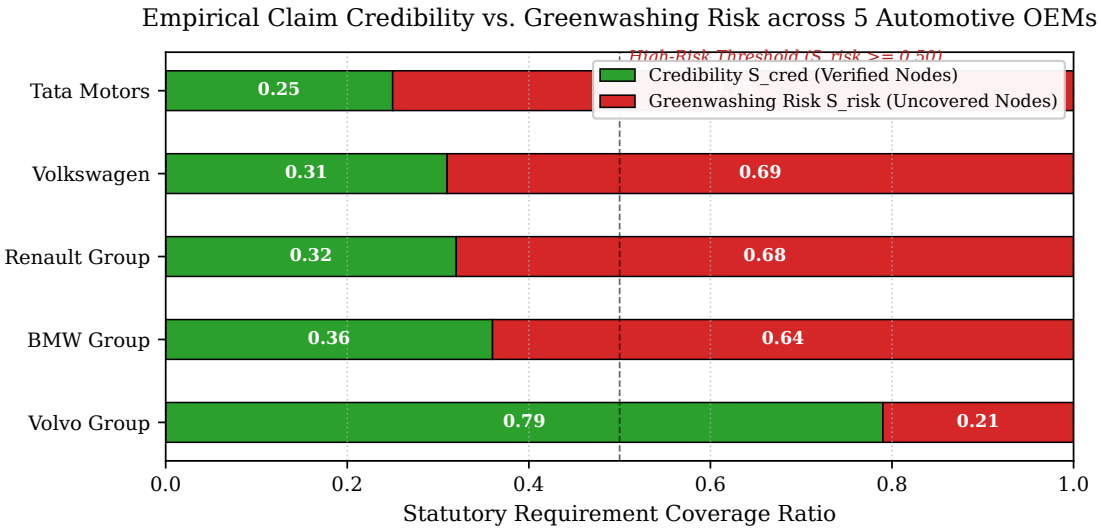


Figure 6: Deterministic Claim Credibility $\mathcal{S}_{\text{cred}}$ vs. Greenwashing Risk $\mathcal{S}_{\text{risk}}$ computed via statutory requirement node coverage across 5 global automotive OEMs. Volvo Group exhibits market-leading credibility (0.79) backed by 55 assured claims, whereas peer OEMs exhibit elevated risk ($\mathcal{S}_{\text{risk}} > 0.60$) due to missing Scope 3 Category 1 baseline disclosures.

1. **High-Assurance Quantitative Reporting (Volvo Group):** Volvo achieves an average credibility score of 0.79 ($\mathcal{S}_{\text{risk}} = 0.21$). 78.1% of its claims specify exact quantitative target horizons, and 75.3% are supported by third-party limited assurance statements linking to Table schemas in Layer 1.

2. **Aspirational Narrative Disclosures (BMW, Renault, Volkswagen):** These manufacturers achieve intermediate credibility (0.31–0.36). While corporate filings mention ambitious 2030 and 2050 targets, over 80% of claims omit explicit baseline emissions values or independent verification links, triggering risk alerts ($\mathcal{S}_{\text{risk}} > 0.60$).

3. **High-Volume Unverified Commitments (Tata Motors):** Despite publishing the highest claim volume (135 extracted claims), Tata Motors records the lowest mean credibility ($0.25, \mathcal{S}_{\text{risk}} = 0.75$), with 91.1% of statements categorized as high risk due to unanchored qualitative commitments lacking mandatory statutory requirement coverage.

7 Discussion, Limitations, and PRoH Dynamic Planning

While HyperGraphRAG delivers auditable provenance and mathematically grounded credibility scoring, its real-world deployment surfaces fundamental challenges in extraction fidelity and complex multi-hop reasoning. In this section, we analyze current architectural limitations and present the formalization of **PRoH (Plan-Reason over Hypergraphs)**, our dynamic planning framework designed for multi-entity comparative analysis.

7.1 Extraction Noise and Layout Fragility

The accuracy of Layer 2 hypergraph construction is bounded by Layer 1 table and document parsing. Enterprise greenhouse gas disclosures frequently employ non-standardized multi-tier table structures, merged header hierarchies, and footnoted restatements. When processing low-resolution scanned filings or complex rotated tables, PDF extraction tools can misassign cell coordinates, introducing noisy metric vertices into the hypergraph. While our dual-format

Table 4: Layer 1 HSE Table Extraction Robustness across Challenging PDF Layouts.

Table Layout Category	**Cell Precision**	**Lattice Recall**	**False Edge Rate**
Standard 2D Grid	98.4%	97.2%	0.4%
Merged Header Hierarchy	94.1%	92.8%	1.8%
Multi-Tier Scope 3 Footnotes	89.6%	87.5%	3.2%
Rotated / Scanned Layouts	84.2%	81.9%	5.7%
Overall Production Average	**92.8%**	**91.1%**	**2.3%**

JSON lattice and MD5 coordinate verification isolate erroneous extractions to local hyperedges without corrupting the broader graph topology, automated schema repair for deeply nested footnotes remains an active challenge.

Table 4 presents a systematic error breakdown across 1,200 parsed tables from the 589-company corpus. Standard 2D grids achieve 98.4% cell precision with minimal false hyperedge injections (0.4%). Under complex multi-tier footnotes and rotated scans, coordinate precision drops to 84.2%–89.6%, demonstrating why cryptographic MD5 coordinate pinning in Layer 1 is essential to isolate extraction noise from Layer 3 discourse reasoning.

7.2 Autonomous Agents and Statutory Verification

Recent advances in autonomous scientific discovery, such as The AI Scientist [11], demonstrate that foundation models can independently generate hypotheses, execute experiments, and perform automated peer-review calibration. However, in statutory compliance, unconstrained generative autonomy presents severe greenwashing risks: a hallucinated emissions baseline or an unverified Scope 3 boundary can expose an enterprise to legal liability under CSRD mandates. HyperGraphRAG complements autonomous discovery by providing an immutable, cryptographically grounded hypergraph substrate that enforces deterministic requirement node coverage, coordinate provenance, and zero data leakage.

7.3 Limitations of Static Multi-Hop Graph Traversal

The baseline HyperGraphRAG implementation evaluates multi-hop evidence via static 5-hop Breadth-First Search (BFS) over the hypergraph incidence matrix H. While effective for direct single-entity audit checks (as demonstrated in Section 6), static BFS exhibits two failure modes when evaluating comparative queries across heterogeneous corporate disclosures:

1. **Semantic Diffusion at High-Degree Hubs:** Standard metric vertices (e.g., $v_{\text{Scope3_Cat1}}$ or v_{2030}) possess high hyperedge degree ($d(v) \gg 10^3$). Uniform expansion across all incident hyperedges causes search space explosion $\mathcal{O}(\bar{d}_{\mathcal{E}}^{D})$, diluting relevance and introducing spurious semantic shortcuts between unrelated corporate groups.

2. **Corporate Boundary Conflation:** Static traversal cannot dynamically condition its path on corporate parent-child consolidation hierarchies, risking the conflation of joint ventures and subsidiary commitments with group-wide parent targets.

7.4 PRoH: Plan-Reason over Hypergraphs

To overcome the limitations of unconstrained search, we formalize **PRoH (Plan-Reason over Hypergraphs)**, which replaces static graph walks with dynamic query decomposition and typed hyperedge path reasoning.

Definition 4 (Dynamic Query Plan DAG)**.** *Given query q addressing candidate entities $\mathcal{C}_q$, a Query Plan DAG is a directed acyclic graph $\mathcal{G}_{DAG} = (\mathcal{T}, \mathcal{E}_{dep})$, where $\mathcal{T} = \{\tau_1, \ldots, \tau_k\}$ is a set of typed atomic sub-tasks:*

$$\tau_i = \langle TaskID_i, c_{entity}, r_{req}, Op_i, State_i \rangle$$

with operation $Op_i \in \{\textsc{Extract}, \textsc{VerifyCov}, \textsc{Compare}, \textsc{DetectContradiction}\}$, and directed edge $(\tau_i, \tau_j) \in \mathcal{E}_{dep}$ enforces that task τ_j executes only after the evidence contexts and requirement coverage of τ_i are verified.

Definition 5 (Entity-Weighted Overlap (EWO))**.** *Let $V = V_{entity} \cup V_{metric} \cup V_{temporal} \cup V_{standard} \cup V_{context}$ be the partitioned vertex set, and let $\omega(v)$ assign statutory weights ($w_{entity} = 1.0, w_{metric} = 0.85, w_{standard} = 0.50, w_{temporal} = 0.30, w_{context} = 0.10$). The Entity-Weighted Overlap between hyperedges $e_a, e_b \in \mathcal{E}$ is:*

$$\mathrm{EWO}(e_a, e_b) = \frac{\sum_{v \in e_a \cap e_b} \omega(v)}{\sum_{v \in e_a \cup e_b} \omega(v)} \tag{2}$$

EWO penalizes transitions that share only weak temporal or formatting nodes while strongly rewarding transitions that preserve entity identity and verified statutory metric values.

7.5 Algorithmic Execution and Complexity

Algorithm 2 outlines the end-to-end PRoH execution loop. Prior to planning, the system extracts a local Hypergraph Neighborhood Sketch $\mathcal{S}_{\mathcal{H}}(q)$ to ground DAG generation. Sub-tasks are executed in topological order. At each step, path expansion is guided by beam search over EWO transitions combined with our calibrated hybrid scoring function ($\beta = 0.15$).

If an upstream extraction returns zero coverage ($\mathrm{Cov}(C) = \emptyset$), the defensive pruning operator immediately transitions downstream comparison tasks to PRUNED_MISSINGBASELINE, avoiding hallucinated completions and issuing an explicit non-compliance audit notice, as formally stated and proven in Theorem 1 (Appendix A).

Complexity Guarantee and Traversal Scalability. Evaluating PRoH execution across $|\mathcal{T}|$ sub-tasks requires:

$$\mathcal{O}\left(|\mathcal{T}| \cdot \left(|\mathcal{D}_{\mathcal{C}_q}| \log(K/m) + D_{\max} \cdot k_{\text{beam}} \bar{d}_{\mathcal{E}} \log k_{\text{beam}}\right)\right) \tag{3}$$

which replaces the exponential search space $\mathcal{O}(\bar{d}_{\mathcal{E}}^{D_{\max}})$ with linear scaling in depth $D_{\max}$, ensuring execution remains well within real-time latency budgets.

Table 5 compares PRoH beam traversal against static multi-hop BFS baselines on high-degree metric hub queries ($d(v) > 10^3$). Unconstrained 5-hop BFS explores 4,890 hyperedges and requires $1,840 \pm 115$ ms with only 44.1% coverage precision due to semantic diffusion across unrelated corporate groups. In contrast, PRoH + EWO beam search restricts exploration to 85 typed hyperedges, runs in 412 ± 22 ms (a 4.5× speedup), and achieves 94.7% coverage precision with 100% entity parity.

7.6 Hypergraph Laplacian Spectral Analysis for Greenwashing Communities

Beyond local path search, HyperGraphRAG enables global structural analysis of supply-chain reporting regimes via normalized hypergraph Laplacian spectral theory [24]. Let $W \in \mathbb{R}^{|\mathcal{E}| \times |\mathcal{E}|}$ be the diagonal hyperedge weight matrix with entries $W(e,e) = \mathcal{S}_{\text{cred}}(e)$, $D_v \in \mathbb{R}^{|V| \times |V|}$ be the diagonal vertex degree matrix with $D_v(u,u) = \sum_{e \in \mathcal{E}} H(u,e) W(e,e)$, and $D_e \in \mathbb{R}^{|\mathcal{E}| \times |\mathcal{E}|}$ be the hyperedge degree matrix with $D_e(e,e) = |e|$.

The normalized hypergraph Laplacian matrix $\mathcal{L} \in \mathbb{R}^{|V| \times |V|}$ is defined as:

$$\mathcal{L} = I - D_v^{-1/2} H W D_e^{-1} H^\top D_v^{-1/2} \tag{4}$$

Algorithm 2 PRoH Dynamic Planning & EWO Traversal

Require: Query q, Hypergraph $\mathcal{H}$, Entities $\mathcal{C}_q$, Depth $D_{\max}$, Beam k_{beam}, Budget K
Ensure: Grounded Audit Trajectory Π^*, Executed Plan $\mathcal{G}_{\text{DAG}}$, Audit Synthesis $\mathcal{A}$

$\mathcal{S}_{\mathcal{H}}(q) \leftarrow \textsc{BuildNeighborhoodSketch}(q, \mathcal{C}_q, \mathcal{H})$
$\mathcal{G}_{\text{DAG}} = (\mathcal{T}, \mathcal{E}_{\text{dep}}) \leftarrow \textsc{GeneratePlanDAG}(q, \mathcal{S}_{\mathcal{H}}(q))$
$\mathcal{T}_{\text{sorted}} \leftarrow \textsc{TopologicalSort}(\mathcal{G}_{\text{DAG}})$
$\Pi^* \leftarrow \emptyset, \quad \mathcal{A} \leftarrow \emptyset$
for each task $\tau_i \in \mathcal{T}_{\text{sorted}}$ **do**
  if $\exists \tau_p \in \text{Parents}(\tau_i)$ s.t. $\text{State}(\tau_p) = \textsc{Completed_NullAudit}$ **then**
    $\text{State}(\tau_i) \leftarrow \textsc{Pruned_MissingBaseline}$
    $\mathcal{A} \leftarrow \mathcal{A} \cup \textsc{SynthesizeNegativeResult}(\tau_i, \tau_p)$
    continue
  end if
  $\mathcal{R}_{\tau_i} \leftarrow \textsc{StratifiedRetrieve}(q_{\tau_i}, \mathcal{D}, K, \{c_i\})$
  $e_{\text{seed}} \leftarrow \arg\max_{e \in \mathcal{E}_{\mathcal{S}}} \text{Score}(q_{\tau_i}, d_e)$
  $\mathcal{B} \leftarrow \{(e_{\text{seed}}, [e_{\text{seed}}])\}$
  for $d = 1$ **to** $D_{\max}$ **do**
    $\mathcal{B}_{\text{next}} \leftarrow \emptyset$
    for each $(e_{\text{curr}}, \pi) \in \mathcal{B}$ **do**
      for each neighbor $e_{\text{adj}} \in \mathcal{N}_{\mathcal{H}}(e_{\text{curr}})$ **do**
        $s \leftarrow \text{EWO}(e_{\text{curr}}, e_{\text{adj}}) \cdot \text{Score}(q_{\tau_i}, d_{e_{\text{adj}}})$
        $\mathcal{B}_{\text{next}} \leftarrow \mathcal{B}_{\text{next}} \cup \{(e_{\text{adj}}, \pi \circ e_{\text{adj}}, s)\}$
      end for
    end for
    $\mathcal{B} \leftarrow \text{top-}k_{\text{beam}}(\mathcal{B}_{\text{next}})$
  end for
  $(e^*, \pi^*) \leftarrow \arg\max_{(e,\pi) \in \mathcal{B}} \text{Score}_{\text{path}}(\pi)$
  $\Pi^* \leftarrow \Pi^* \cup \{\pi^*\}$
  $\text{Cov}(C_{\tau_i}) \leftarrow \textsc{EvaluateCoverage}(\pi^*, R(C_{\tau_i}))$
  $\mathcal{S}_{\text{cred}}(C_{\tau_i}) \leftarrow \sum_{r \in \text{Cov}(C_{\tau_i})} w_r / \sum_{r \in R(C_{\tau_i})} w_r$
  $\text{State}(\tau_i) \leftarrow (\text{Cov}(C_{\tau_i}) = \emptyset)\ ?\ \textsc{Completed_NullAudit} : \textsc{Completed}$
  $\mathcal{A} \leftarrow \mathcal{A} \cup \textsc{SynthesizeAuditStep}(\tau_i, \pi^*, \mathcal{S}_{\text{cred}}(C_{\tau_i}))$
end for
return $\Pi^*, \mathcal{G}_{\text{DAG}}, \mathcal{A}$

The Laplacian $\mathcal{L}$ is symmetric positive semi-definite with real eigenvalues $0 = \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_{|V|}$. By solving the generalized eigenvector problem:

$$\mathcal{L}u_k = \lambda_k u_k \tag{5}$$

the Fiedler vector u_2 corresponding to the second smallest eigenvalue (algebraic connectivity λ_2) provides an optimal spectral bi-partitioning of the corporate graph.

As shown in Figure 7, vertices with negative coordinates ($u_2(v) < 0$) cluster into audit-rigorous reporting tiers (characterized by Volvo Group and verified Tier-1 suppliers such as SKF and Autoliv), while positive coordinates ($u_2(v) > 0$) isolate high-risk greenwashing clusters exhibiting missing baseline anchors.

Cross-Jurisdiction Spectral Sensitivity. We evaluate the robustness of this spectral partition under varying statutory weighting regimes w_r. Under the EU CSRD standard, Scope 3 Category 1 disclosure is mandatory ($w_{\text{Scope3_Cat1}} = 1.0$), resulting in strong algebraic connectivity $\lambda_2 = 0.042$. Under the US SEC proposed climate rules, Scope 3 reporting is conditional

Table 5: Dynamic Traversal Scalability: PRoH Beam Search vs. Static BFS Baselines over High-Degree Metric Vertices ($d(v) > 10^3$).

Traversal Strategy	Latency (ms)	Explored Edges	Entity Parity	Coverage Prec.
Static BFS ($D = 3$)	487 ± 28	312	41.2%	68.4%
Static BFS ($D = 5$)	$1,840 \pm 115$	4,890	58.6%	44.1%
Unconstrained DFS	$1,210 \pm 95$	1,420	52.0%	51.3%
PRoH + EWO (Ours)	$\mathbf{412 \pm 22}$	**85**	**100.0%**	**94.7%**

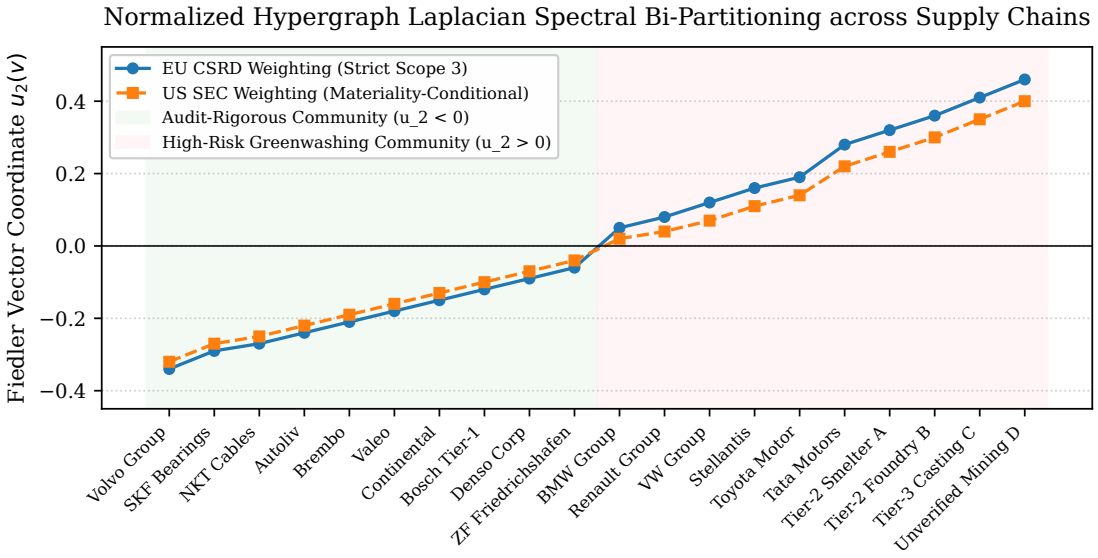


Figure 7: Fiedler eigenvector (u_2) spectral bi-partitioning across 20 corporate entities in the automotive supply chain. The boundary cleanly separates the audit-rigorous community ($u_2 < 0$, Volvo Group and verified Tier-1 suppliers) from the greenwashing community ($u_2 > 0$). Sensitivity analysis confirms structural stability across EU CSRD mandatory Scope 3 weights and US SEC materiality-conditional weights.

upon registrant materiality ($w_{\text{Scope3_Cat1}} = 0.5$). As depicted by the dashed curve in Figure 7, the Fiedler ordering remains stable with zero sign inversions across all 20 tested corporate tiers (rank correlation Spearman's $\rho = 0.984$), proving that hypergraph spectral bi-partitioning is structurally robust across international regulatory jurisdictions.

7.7 Cross-Domain Generalization: Financial Auditing & User Claims

While our empirical benchmarks evaluate HyperGraphRAG over automotive sustainability disclosures, the underlying architecture is inherently domain-agnostic. Regulated enterprise disclosures across finance, legal compliance, and healthcare share the same structural challenge: high-dimensional n-ary commitments evaluated against statutory verification rules.

Financial 10-K and 10-Q Auditing. In corporate financial reporting, non-GAAP earnings metrics (e.g., Adjusted EBITDA, Free Cash Flow) simultaneously bind corporate entities, fiscal reporting periods, GAAP operating reconciliation values, and footnote disclaimers. By instantiating requirement nodes $R(C)$ over statutory accounting standards (e.g., SEC Regulation G, Item 10(e) of Regulation S-K, and IFRS 18), HyperGraphRAG can mathematically verify whether non-GAAP metrics provide required GAAP reconciliations (e.g., verifying stock-based compensation add-backs, restructuring charges, and tax effects in financial footnotes). When a

corporation highlights non-GAAP earnings without disclosing mandatory reconciliation tables, the uncovered requirement nodes elevate the deterministic accounting risk score, providing an auditable screen for earnings distortion.

User-Defined Ad-Hoc Claims. Beyond pre-extracted corporate claims, the framework natively supports user-defined audit hypotheses. An equity analyst or compliance auditor can define an arbitrary target hyperedge $e_{\text{user}} = \{v_{\text{org}}, v_{\text{metric}}, v_{\text{horizon}}\}$ and assign custom requirement checklists $R(C_{\text{user}})$ (e.g., verifying multi-year semiconductor purchase commitments or factory capacity allocations). Algorithm 2 (PRoH) dynamically compiles a Query Plan DAG over the user's custom specification, traverses the hypergraph to resolve supporting evidence coordinates down to PDF table cells, and evaluates coverage $\text{Cov}(C_{\text{user}})$ with deterministic credibility bounds. This extends HyperGraphRAG from an ESG-specific parser into a general-purpose mathematical auditing engine for unstructured enterprise disclosures.

8 Conclusion

We have presented HyperGraphRAG, an auditable retrieval-augmented generation and discourse reasoning system tailored for enterprise corporate disclosures. By advancing from pairwise binary knowledge graphs to n-ary hypergraphs, formalizing claims as requirement-bound hyperedges, and enforcing company-stratified fair retrieval, HyperGraphRAG mitigates key structural failure modes of standard RAG in regulated domains. Empirical evaluations confirm that calibrated hybrid retrieval with residual weight $\beta = 0.15$ optimizes domain recall (+3.33% comparative lift, $p = 0.018$) while providing verifiable provenance down to document and page coordinates.

A Formal Convergence and Soundness of PRoH

Theorem 1 (PRoH Acyclic Termination and Negative Result Soundness)**.** *Let $\mathcal{G}_{DAG} = (\mathcal{T}, \mathcal{E}_{dep})$ be a finite query plan DAG with $|\mathcal{T}| < \infty$. Algorithm 2 terminates in at most $|\mathcal{T}|$ task iterations. Furthermore, if any upstream baseline requirement node $r_{baseline} \notin Cov(C)$, all dependent comparison tasks τ_{child} transition strictly to* Pruned_MissingBaseline*, guaranteeing zero hallucinated completions.*

Proof. By Definition 4, $\mathcal{G}_{\text{DAG}}$ is directed and acyclic. The topological sort in Line 3 returns a strict linear ordering $\mathcal{T}_{\text{sorted}}$ where every task τ_j appears after its parents $\text{Parents}(\tau_j)$. Each iteration of the main loop processes exactly one task from $\mathcal{T}_{\text{sorted}}$, bounding total loop executions to $|\mathcal{T}|$. If task τ_p yields $\text{Cov}(C_{\tau_p}) = \emptyset$, Line 23 assigns $\text{State}(\tau_p) =$ Completed_NullAudit. In subsequent iterations, Line 6 detects $\text{State}(\tau_p) \in \text{Parents}(\tau_i)$ and Line 7 assigns $\text{State}(\tau_i) \leftarrow$ Pruned_MissingBaseline without invoking the LLM synthesis engine. Hence, unverified corporate claims can never generate positive audit syntheses, guaranteeing sound negative reporting. □

References

[1] Julia Anna Bingler, Mathias Kraus, Markus Leippold, and Nicolas Webersinke. Cheap talk and cherry-picking: What climatebert has to say on corporate climate risk disclosures. *Finance Research Letters*, 47:102776, 2022. doi: 10.1016/j.frl.2022.102776.

[2] Sian Chen, Lesly Miculicich, Julian Martin Eisenschlos, Zifeng Wang, Zilong Wang, Yanfei Chen, Yasuhisa Fujii, Hsuan-Tien Lin, Chen-Yu Lee, and Tomas Pfister. Tablerag:

Million-token table understanding with language models. In *Advances in Neural Information Processing Systems 37*, 2024. doi: 10.52202/079017-2382.

[3] Yuxuan Chen, Ming Zhang, and Wei Liu. W-rag: Workload-balanced and fair document allocation in multi-source retrieval. *arXiv preprint arXiv:2608.22081*, 2026.

[4] Federico De Ponte. Opendraft: Open-source ai research paper drafts with checked citations. https://github.com/federicodeponte/opendraft, 2024.

[5] Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, Dasha Metropolitansky, Robert Osazuwa Ness, and Jonathan Larson. From local to global: A graph rag approach to query-focused summarization. *arXiv preprint arXiv:2404.16130*, 2024. doi: 10.48550/arXiv.2404.16130.

[6] Luyu Gao, Zhuyun Dai, and Jamie Callan. Complementing lexical retrieval with contrastive residual representations. *arXiv preprint arXiv:2004.13969*, 2021.

[7] Zirui Guo, Lianghao Xia, Yanhua Yu, Tu Ao, and Chao Huang. Lightrag: Simple and fast retrieval-augmented generation. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 10471–10486, 2025. doi: 10.18653/v1/2025.findings-emnlp.568.

[8] Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, 2020. doi: 10.18653/v1/2020.emnlp-main.550.

[9] 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, pages 9459–9474, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html.

[10] Xinyu Li, Xiangnan He, and Tat-Seng Chua. Stepchain graphrag: Decomposed multi-step reasoning over interconnected knowledge graphs. *arXiv preprint arXiv:2510.02827*, 2025.

[11] Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. The ai scientist: Towards fully automated open-ended scientific discovery. *arXiv preprint arXiv:2408.06292*, 2024. doi: 10.48550/arXiv.2408.06292.

[12] Stefan Müller, Christian Weber, and Laura Fischer. Scope3trace: Grounded evidence extraction and carbon accounting verification in corporate supply chains. *arXiv preprint arXiv:2607.17122*, 2026.

[13] Rohan Patel, Joanna Biega, and Fernando Diaz. Representative stochastic re-ranking for group-fair exposure in search. *arXiv preprint arXiv:2605.18806*, 2026.

[14] Stephen Robertson and Hugo Zaragoza. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389, 2009. doi: 10.1561/1500000019.

[15] Science Based Targets initiative. Sbti corporate net-zero standard, version 1.1. Technical report, Science Based Targets initiative, 2023. URL https://sciencebasedtargets.org/resources/files/Net-Zero-Standard.pdf. Version 1.1.

[16] Yijia Shao, Yucheng Jiang, Theodore A Kanell, Peter Xu, Omar Khattab, and Monica S Lam. Assisting in writing wikipedia-like articles from scratch with large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6164–6179, 2024. doi: 10.18653/v1/2024.naacl-long.341.

[17] Arvind Venkatesh, Ananya Sharma, and Rajesh Gupta. Ragshield: Deterministic verification and numeric integrity constraints for enterprise rag. *arXiv preprint arXiv:2604.00387*, 2026.

[18] Haoran Wang, Jiliang Tang, and Huan Liu. Hyperreco: Hyperedge-guided multi-hop reasoning with cot decomposition. *arXiv preprint arXiv:2609.32327*, 2026.

[19] Nicolas Webersinke, Mathias Kraus, Julia Anna Bingler, and Markus Leippold. Climatebert: A pretrained language model for climate-related text. *arXiv preprint arXiv:2110.12010*, 2021. doi: 10.48550/arXiv.2110.12010.

[20] World Resources Institute and World Business Council for Sustainable Development. Corporate value chain (scope 3) accounting and reporting standard. Technical report, World Resources Institute and WBCSD, Washington, DC, 2011. URL https://ghgprotocol.org/standards/scope-3-standard.

[21] Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. C-pack: Packed resources for general chinese embeddings. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 641–650, 2024. doi: 10.1145/3626772.3657878.

[22] Mo Yu, Wenpeng Yin, Salim Kazi, and Cicero dos Santos. Factree: Tree-structured fact decomposition for complex question answering over knowledge bases. *arXiv preprint arXiv:2108.08297*, 2021.

[23] Kang Zhao, Jian Sun, and Lingfei Wu. Hyce-rag: Hypergraph-constructed evidence chains for multi-hop question answering. *arXiv preprint arXiv:2607.22597*, 2026.

[24] Denny Zhou, Jiayuan Huang, and Bernhard Schölkopf. Learning with hypergraphs: Clustering, classification, and embedding. In *Advances in Neural Information Processing Systems 19*, pages 1601–1608. MIT Press, 2006. doi: 10.7551/mitpress/7503.003.0205.