sustainability

MDPI

Article

From Complementarity to Selectability: Evaluating Hybrid Retrieval for Standard-Informed Furniture Carbon Screening

Feng Ye [1], Min Yin [2,*] and Shouqian Sun [3]

[1] School of Design and Art, Communication University of Zhejiang, Hangzhou 314500, China; yefeng@cuz.edu.cn
[2] The Innovation Center of Yangtze Delta, Zhejiang University, Jiaxing 314100, China
[3] College of Computer Science and Technology, Zhejiang University, Hangzhou 310027, China; ssq@zju.edu.cn
* Correspondence: amyin@zju.edu.cn; Tel.: +86-0573-84576639

Abstract

Concept-stage designers make carbon choices before complete product evidence exists, yet their answers must be computable and traceable to governing standards. Hybrid retrieval combines these functions: a knowledge graph can execute carbon calculations, while text retrieval recovers conditional standards guidance. We test whether complementarity, in which sources succeed on different cases, translates into selectability: query-time evidence may indicate which source to trust. A knowledge graph and a GraphRAG text index were evaluated on 100 furniture-carbon questions across five primary conditions, with oracle and routing analyses. Retrieval value depended on task class: correctness improved by 0.5600 (cluster-bootstrap 95% CI [0.4487, 0.6583]) over the closed-book condition for factual questions (factor lookups, product totals, and material lists), but by only 0.0160 (95% CI [−0.0986, 0.1420]) for standards interpretation. The stores reversed rank across categories. The observed correctness difference between concurrent fusion and the graph alone was 0.0007 (95% CI [−0.0532, 0.0628]). A cross-fitted oracle exposed recoverable headroom, with category routing examined through both a score-pool diagnostic and a separate live evaluation on the same benchmark. Concurrent fusion recovered little of the source-selection headroom. The evaluated outputs are material-stage screening estimates from author-compiled bills of materials. Designers and retrieval-system developers should prioritize structured access to exact factor fields, unit conversions, and linked inventory data for record-specific questions, using standards indexing to provide inspectable clause context for conditional interpretation.

Keywords: product carbon footprint; knowledge graph; retrieval-augmented generation; GraphRAG; design-stage decision support; furniture; life cycle assessment; hybrid retrieval architecture; retrieval evaluation; traceability

Check for updates

Received: 11 August 2026
Revised: 8 October 2026
Accepted: 20 October 2026
Published: 22 October 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Furniture designers face a timing paradox. Furniture life-cycle studies identify material production and processing as important environmental hotspots [1–3], yet the opportunity to intervene is greatest before verified bills of materials, supplier data, and matched emission factors exist. Prior work motivates integrating environmental assessment into early product development [4,5]. Carbon knowledge must therefore become useful before it becomes complete. GB/T 46486-2025 is China's recommended national standard for furniture product carbon footprints [6]. Applying it during design requires structured emission factors and product compositions for numerical comparison, as well as textual provisions on boundaries,

allocation, exclusions, and reporting for defensible interpretation. A knowledge graph appears suited to the first need and standards retrieval to the second.

The graph store and text index succeed on different question categories, yet historical concurrent fusion recovered little of their complementary performance. The category reversal creates a source-selection problem when the two returns disagree. Concurrent fusion improved correctness by only 0.0007 over the graph store, whereas a cross-fitted oracle that selects the stronger source reveals 0.1067 of recoverable headroom. The observed gap motivates examining query-time signals for selecting a source. This distinction between complementarity and selectability organizes the study.

We evaluate retrieval value using two proposed dimensions. Representational determinacy asks whether available records uniquely determine an answer; closed-book answerability asks whether a model can already produce that answer from its parameterized knowledge. A proprietary bill of materials paired with an emission factor may be determinate yet unavailable to the model, whereas a published standards provision may be both determinate and familiar from pretraining. The benchmark samples high-determinacy task groups, so the analysis compares answerability across categories; it does not independently manipulate both dimensions.

The study addresses three research questions. **RQ1:** How is furniture carbon knowledge distributed across the proposed dimensions of representational determinacy and closed-book answerability? **RQ2:** When does retrieval add value, and when does hybrid retrieval improve on a single source? **RQ2a:** Within the sampled high-determinacy region, is variation in closed-book answerability associated with retrieval gain? **RQ2b:** Do demonstrably complementary sources benefit from concurrent fusion, and how much gain is recoverable through per-query selection of the stronger source? **RQ3:** How well do the resulting answers meet decision-relevant requirements for numerical fidelity and traceability, and what failure modes can be identified?

We evaluate a structured knowledge graph and GraphRAG over a 23-document corpus of standards and supporting material across four question categories and five primary conditions, separating correctness from source attribution. Contributions include distinguishing complementarity from selectability, testing a routing criterion against a score-pool diagnostic and a separate live evaluation, and providing a traceable furniture-carbon screening artifact with 45,342 nodes and 483,246 edges. Calculations agree within 5% with a separate implementation for 20 of 24 products under shared inputs; the remaining disagreements expose limits in mapping composite materials. With author-compiled bills of materials, the artifact supports conditional material-stage comparison and screening.

2. Related Work

2.1. Carbon Footprint Assessment for Furniture: The Designer Decision Gap

Furniture life-cycle studies analyze environmental impacts and improvement options [1,2]. GB/T 46486-2025 specifies a recommended national method for quantifying furniture product carbon footprints [6]. This shifts carbon knowledge from a specialist output toward a working input for designers making concept-stage material decisions [7,8]. LCA-based tools have been studied for wooden-furniture ecodesign [9], alongside broader approaches to simplifying LCA [10]. Product-category rules provide category-specific assessment requirements [11]. Design-stage use of LCA must account for the information available at each development stage [12].

Studies have examined ways to support early design decisions and the obstacles to applying LCA and selecting materials in that setting [4,5,13].

The literature therefore calls for iterative guidance that accepts under-specification in early-design LCA [4,14,15]. This study operationalizes that requirement by linking carbon

hotspots to product structures and material alternatives. The underlying representational gap is that quantitative assessment uses deterministic structured data, whereas compliance requires interpreting conditional clauses governing functional units, system boundaries, allocation, and biogenic carbon [11,16].

A closed-book comparison is needed to interpret retrieval gains. Prior work compares question answering with and without retrieval [17,18], and domain-specific systems also include standalone-language-model baselines [19]. The closed-book condition estimates performance before retrieval is added. Section 4.3 therefore uses the closed-book comparison as a primary condition.

2.2. Knowledge Graphs for Life-Cycle Data Integration

Ontologies, semantic representations, flow mappings, and data models have been developed for reconciling heterogeneous LCA information [20–25].

Related work addresses industrial-product data models and product-life-cycle ontologies [26,27], manufacturing data and design-process integration [28,29], and process simulation and semantic modeling for copper-alloy production [30]. Recent systems investigate graph-based LCI storage, background-dataset recommendation, and linking BOM components to LCA database entries [31–33]. AutoPCF explores LLM-assisted life-cycle modeling and parameter matching for product carbon accounting [34].

Graphs may represent that a product contains a material and that the material has an impact while omitting the mass assignments, emission factors, and allocation rules needed for parametric comparison. Reported matching results use different tasks and metrics: Peng et al. report flow-recommendation Precision@10 of 79.52% [32], whereas Castle et al. report Hits@1 of 0.24 and Hits@5 of 0.48 for their full pipeline on 21 components [33]. A typical mismatch is a BOM entry such as "oak plywood panel" versus an LCI process indexed as "plywood production, hardwood, at plant". The matcher must bridge product- and process-oriented classifications. Complete-BOM assumptions also fail when grade or supplier is unknown [4,14].

These limitations become sharper when standards impose conditional interpretation. For example, GB/T 46486-2025 permits omitting a link whose carbon contribution is less than 1%, provided that all omitted links together do not exceed 5% of the total footprint [6]. A schema must represent both thresholds and their scope; storing isolated values is insufficient for this check. The present architecture uses text retrieval to supply the surrounding clause context.

2.3. Retrieval-Augmented Generation for Technical Information

Retrieval-augmented generation gives natural-language access to technical documentation whose meaning depends on context, cross-references, and conditional logic [35–37]. Retrieved passages make supporting material inspectable, although citation correctness alone does not establish that the model used the cited evidence [38]. Related studies investigate graph-guided and hybrid retrieval across several question-answering and technical domains [19,39–47].

Prior studies examine incomplete retrieval and insufficient context [48,49], failures to use available evidence [50–52], and citation unfaithfulness [38]. Engineering and legal-compliance evaluations provide further domain-specific tests [53–55]. Numerical and unit errors form an additional diagnostic category in the present study. Diagnostic frameworks accordingly separate citation quality from retrieval, answer, and grounding errors [56]. We use these four categories to organize the diagnostic taxonomy for furniture-carbon retrieval in Supplementary Material S3.

Whether retrieval was needed at all is a question logically prior to these failure modes. Closed-book studies show that models answer a substantial share of factual questions from parameters alone [17], with retrieval helping most for infrequent entities and sometimes degrading answers for popular ones [18]. Some question-answering tasks mark cases where the supplied page contains no answer [57], and the joint failure rate of member systems limits what any selection policy can recover [58–60]. A circular-economy study compares its graph-based system with standalone-LLM and naive-RAG baselines [19]. The present study additionally examines task-category contrasts and a member-selection bound. Furniture-carbon questions expose both properties before a system is built, which Section 2.4 develops into a criterion.

These failure modes reveal a tension between interpretive coverage and numerical precision. Comparative studies and benchmarks examine task and configuration effects in graph-based retrieval [61–64]. In the present implementation, the graph provides factor lookup and calculation, while text retrieval supplies standards context (Section 3). Design-stage carbon work needs both, but dividing labor between two stores presumes that the system can identify which source to trust.

2.4. Task Differences in Retrieval Evaluation

Natural Questions includes cases where the supplied Wikipedia page contains no annotated answer [57], whereas closed-book/open-book evaluation asks whether a model can produce that answer without retrieval [17]. These are orthogonal dimensions: representational determinacy concerns the evidence, while closed-book answerability concerns the model's parametric knowledge. Mallen et al. link retrieval benefits to entity popularity in a general corpus [18]. A standards-governed design domain introduces a different contrast: a supplier's factor may be proprietary, whereas a national standard is deliberately published. Domain studies already evaluate standalone models or analyze evidence sufficiency [19,50]. Here, the proposed distinction organizes task-category analysis within the furniture benchmark.

The distinction matters because aggregate scores average task classes whose preferred sources may point in opposite directions. This is the hidden-stratification or aggregation-bias problem documented in clinical imaging and machine-learning fairness [65,66], distinct from benchmark saturation, which additionally requires performance near an empirical ceiling rather than statistical indistinguishability alone [67]. Small category subsets also make effect estimates and rankings uncertain [68]. We therefore assess precision using the present study's cluster-bootstrap intervals and conditional sensitivity analysis.

The present study therefore operationalizes the two dimensions in a standards-governed design domain and examines their architectural consequence. Within the high-determinacy groups discussed in Section 5.1, retrieval is most justified when closed-book answerability is low. It may contribute little, or displace a correct parametric answer, when answerability is high [18]. This association organizes routing within the benchmark's sampled region.

3. Materials and Methods

We construct a hybrid knowledge service for carbon-related queries in furniture design. The service comprises three components. A Neo4j 5.x knowledge graph (Neo4j, Inc., San Mateo, CA, USA) reorganizes emission factors at product-component-material granularity, a GraphRAG index covers the full document corpus, and a routing layer synthesizes responses from both according to task-specific epistemic requirements.

3.1. System Architecture and Design Rationale

Furniture-carbon queries differ in the information needed to answer them. In this study, numerical comparisons require traceable emission-factor records, units, and calculation assumptions. Standards provisions require interpretation of system boundaries and cut-off conditions [6]. Design decisions require what-if exploration, since designers ask counterfactual questions and expect numerical comparisons synthesized with trade-offs in cost, weight, and structural performance [7,8]. The benchmark examines how the two stores perform on queries with these differing information requirements.

The system separates storage, indexing, and service into three tiers, each matched to one of those demands (Figure 1). A data tier supplies precise numerical facts, an indexing tier preserves interpretive document context, and a service tier classifies each query so that retrieval and synthesis can draw on both. The data tier is the Neo4j 5.x knowledge graph of Section 3.2, integrating emission factors, bills of materials, and regulatory metadata. The indexing tier is a Microsoft GraphRAG v3.1.0 (Microsoft Corporation, Redmond, WA, USA) [35] index over 11,087 indexed records: 3452 standards chunks, 7611 process records, and 24 bills of materials, re-segmented into 11,364 text units. The service tier classifies each query, queries both backends, and synthesizes their returns, with GPT-5.4 (OpenAI, San Francisco, CA, USA) performing the synthesis at temperature 0.1. The model family, access period, and sampling configuration are summarized in Supplementary Material S1. Classification conditions what the graph store looks for rather than which backend is consulted, as Section 3.4 sets out.

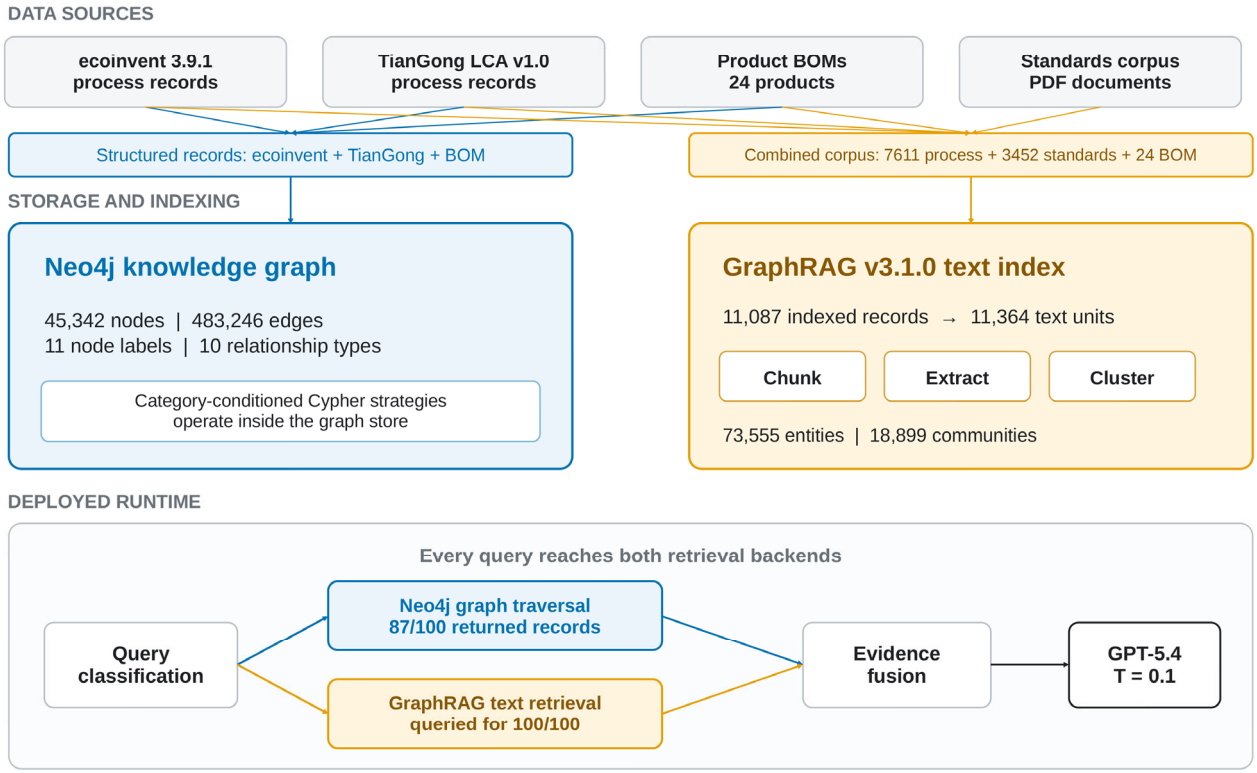

Figure 1. System architecture and information flow across the three tiers. Data sources feed a Neo4j knowledge graph and a GraphRAG text index over a combined 11,087 indexed records. The runtime path shows the deployed retrieval behavior: every query reaches both backends in parallel, and their returns are fused into one synthesis prompt. Category conditioning acts inside the graph store, selecting which Cypher strategies may run, and on the synthesis prompt; it does not select between the two sources. The graph store returned records on 87 of 100 questions; on the other 13, every strategy its category admitted came back empty and synthesis proceeded from text alone.

3.2. Knowledge Graph Construction

LCA databases organize emission factors at the process level, using identifiers such as market for particleboard, uncoated, RoW [69]. Designers reason at component level instead, weighing a frame panel in particleboard against solid pine on cost, weight, and carbon together [4]. The graph bridges that vocabulary gap by reorganizing the data on the designer's structure.

The graph integrates four data sources, with aggregate schema statistics reported in Supplementary Table S1. The export contains 21,240 ecoinvent 3.9.1 process records and 1778 TianGong process records (23,018 combined) with database-of-origin labels [69,70]. The two worked ecoinvent factors use IPCC 2021 GWP100 characterization [71] applied to cutoff system-model cumulative LCI_RESULT exchanges. The two worked factors were independently reproduced from these cumulative inventories (Supplementary Table S11).

The authors compiled bills of materials for 24 IKEA products from manufacturer product pages and 25 assembly instruction PDFs. Product pages supplied article numbers, weights and material descriptions; assembly instructions supplied part inventories. Panel geometry was inferred from overall dimensions less thickness, component masses were calculated as density times volume, and hardware masses were assigned from domain knowledge. Data-quality grades distinguish verified source fields from inferred values.

Twenty-three standards and supporting documents, among them GB/T 46486-2025 [6], ISO 14067 [72], EN 15804+A2 [11], and the GHG Protocol Product Standard [73], supply regulatory requirements. These 23 documents are the unit of the indexed text corpus. The graph separately holds 10 standard metadata records carrying identifiers and clause numbers for the subset referenced by traversal queries; the text index supplies the provisions.

The implementation organizes records in a three-level Product-Component-Material hierarchy. Particleboard maps to the Rest-of-World market process, giving 388.33 kg CO_2e/m^3; division by the assumed density of 650 kg/m^3 gives the stored 0.5974 kg CO_2e/kg. The Global raw-softwood market process gives 40.76 kg CO_2e/m^3; division by 500 kg/m^3 gives the stored 0.0815 kg CO_2e/kg. These market averages include the upstream activities and market transport represented by their source inventories. Supplementary Table S11 gives the process identifiers and full-precision calculations.

The deployed schema defines eleven node types connected by ten relationship types, shown complete in Figure 2. Four types represent products, components, materials, and processes, while four represent emission factors, global warming potential factors, flows, and lifecycle stages. Locations, categories, and standards provide the attributes used for query filtering. The highlighted traversal begins with a 13.65 kg wardrobe side panel and follows it to particleboard and its volumetric emission-factor record in the held ground-truth ledger. Because source databases report factors on the process reference flow, the traversal divides by material density before multiplication by component mass. The inset in Figure 2 gives the worked conversion to a mass basis.

Cross-database entity alignment was implemented to connect records from the two source databases. In this implementation, candidates were scored by multi-feature TF-IDF similarity over process names, reference flows, and category paths, with the score retained on each accepted edge. Equivalences were accepted only when reference flow and unit agreed. The alignment audit inspected the 30 lowest-confidence pairs and a random sample of 20 additional pairs. None of the Cypher strategies in Section 3.4 traverse the accepted equivalence edges.

Implementation consistency of the carbon pathway was assessed by comparing graph traversal with a separately implemented spreadsheet calculation. Both implementations used the same 24 bills of materials and ecoinvent 3.9.1 factors, isolating differences in implementation from differences in input data. Agreement within 5% was defined in

advance and mean absolute percentage error summarized the full set. Section 4.5 reports the alignment and calculation checks.

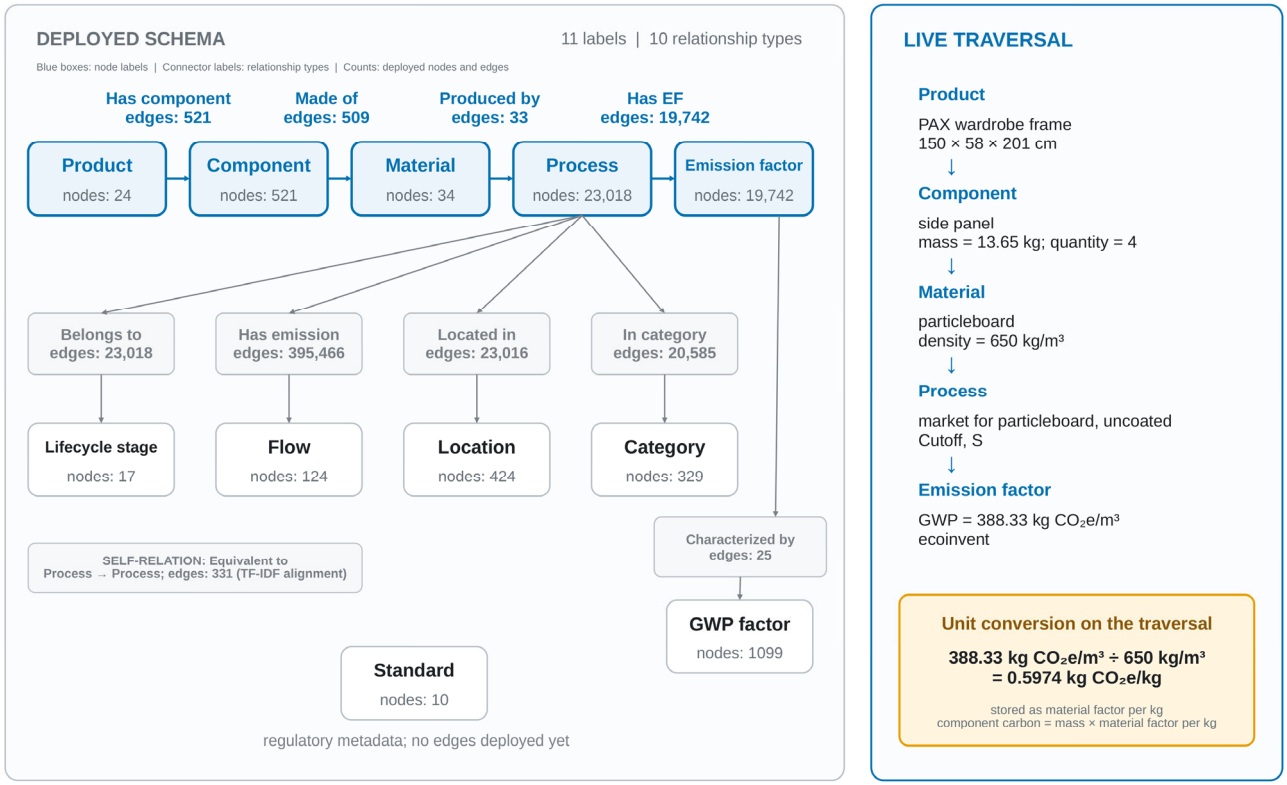

Figure 2. Knowledge graph schema with eleven node labels and ten relationship types, with node counts shown beneath each label. Highlighted nodes and edges trace the Product-to-Emission Factor path for a wardrobe side panel; the inset records the density-based unit conversion that places the emission factor on a mass basis.

3.3. Standards Corpus Indexing with GraphRAG

The graph stores factor records, while standards documents contain conditional provisions and cross-references [37]. Clause 6.3.2 of GB/T 46486-2025 [6] illustrates the difficulty: it permits omitting links whose individual carbon contribution is less than 1%, provided the omitted set totals no more than 5% of the footprint. The rule combines an individual threshold with an aggregate constraint, so admissibility also depends on the other omitted links. A graph representation would need to encode both the individual and aggregate conditions. The implementation therefore retains clause context and cross-references when retrieving provisions, with full-text retrieval supplying that context.

The standards corpus comprises 23 documents, among them GB/T 46486-2025 [6], GB/T 24067-2024 [74], ISO 14067:2018 [72], EN 15804+A2 [11], PAS 2050:2008 [75], the GHG Protocol Product Standard [73], and IPCC AR6 WGIII 2022 [76]. They total approximately 1.03 million tokens after PDF-to-Markdown conversion. Section-aware chunking respects clause boundaries and retains section numbering, clause identifiers, and cross-references to enable citation tracking [42], yielding 3452 chunks.

The GraphRAG pipeline applies entity extraction, Leiden community detection [77], and summarization [35] across five entity types: standard identifiers, method terms, requirements, boundary definitions, and impact categories. Extraction yielded 73,555 entities with confidence scores. Leiden resolution at three levels produced 18,899 communities, of which 6258 received GPT-5.4 summaries. Chunks are embedded with text-embedding-3-small in

1536 dimensions. Supplementary Material S1 reports the full indexing configuration and response constraints.

Standards queries use GraphRAG local search exclusively, because designer queries about provisions are clause-focused rather than corpus-wide. Global search traverses all community summaries, so its cost scales with community count rather than question specificity, which pilot runs found impractical at this corpus scale. Local search retrieves the top 10 entities by query embedding, expands to their 1-hop neighborhood, and returns the top 15 text chunks for synthesis with clause citations. Retrieval quality is reported per category in Section 4.2.

3.4. Concurrent Dual Retrieval and Synthesis

The graph retrieves recorded emission factors and performs deterministic calculations. Its standards records contain identifiers and clause numbers; interpretive provisions and applicability conditions come from text retrieval. Numerical errors in text-based answers include approximate factors in place of recorded values, inconsistent units, and incorrect aggregation over dispersed passages. Concurrent fusion concatenates both returns into one synthesis prompt, leaving source selection to the generator when they conflict. Supplementary Material S3 summarizes the diagnostic failure modes of each path and the fused configuration.

Supplementary Table S2 sets out the proposed source-selection policy: factual and multi-hop queries would use the graph store, standards queries the text index, and design-optimization queries both. The historical evaluation used fixed concurrent fusion, querying both sources for each question.

Category conditioning acts within each retrieval backend rather than between sources. The query category selects among disjoint Cypher strategies in the graph store: material emission-factor lookup and bill-of-materials traversal serve factual, multi-hop, and design-optimization queries; substitution calculation is restricted to design optimization; standards-scope lookup serves standards and design-optimization queries; and a general emission-factor fallback is restricted to factual queries. The system shortens the text-index timeout for multi-hop queries when the graph already returns a component ranking, and a strict numerical prompt is selected for factual graph returns. Supplementary Table S3 reports the graph store return count for each category.

The graph store returned records on 87 of 100 questions, identically in all three runs. On the remaining 13, every admitted strategy came back empty and synthesis proceeded from retrieved text alone. Eleven of the 13 are the items on which the graph store returns nothing when queried alone; the remaining two return nothing only in the fused pipeline.

Three implementation properties bear on Section 4. Failed graph queries are logged by omission rather than by a status field, so an absent record and an empty return are indistinguishable in the execution log without the separate single-source runs. The quality gate rejects a result lacking text retrieval but not one lacking graph retrieval, so text-only answers are tolerated while graph-only answers are treated as failures. The multi-hop timeout reduction applies to one backend in one category, so any multi-hop comparison between the two stores inside the fused system is not resource-matched. The first two are asymmetric in the text index's favor; the third favors the graph store.

Historical synthesis records report GPT-5.4 with temperature 0.1; Supplementary Material S1 summarizes the response specifications. The inspected implementation uses category-specific prompts. Factual prompts request numerical values, units and source labels; standards prompts request standard identifiers and descriptions grounded in the supplied text. Multi-hop prompts request ranked components and their recorded quantities. The source-absence branch returns an empty result when no backend content is available.

A separate post-generation diagnostic extracts identifier patterns from the answer text; its output is recorded alongside the returned answer. Section 4.6 reports the extraction results.

Supplementary Materials S1–S8 document the system specification, scoring protocol, failure taxonomy, and sensitivity analyses.

3.5. Evaluation Framework

The evaluation compares retrieval architectures across task types and examines answer quality under different information conditions. A no-retrieval baseline measures answerability without external context; an oracle-context condition assesses answer quality when reference information is supplied. These conditions provide diagnostic comparisons for interpreting the retrieval results.

3.5.1. Question Set

In this study, a category denotes a predefined question type: factual, standards, multi-hop, or design optimization. Category assignment was fixed before evaluation, and each category contains 25 questions. The evaluation set therefore holds 100 questions in four categories. Factual questions request a specific emission factor or product footprint. Standards questions ask what a published provision requires. Multi-hop questions require component-level aggregation across a bill of materials. Design optimization questions ask for a material substitution and its carbon consequence.

Questions were authored in Chinese and English because the corpus is bilingual and the deployment setting is Chinese manufacturing with export obligations. The four categories were chosen because they differ in two properties that Section 5.1 treats as explanatory: whether the held corpus uniquely fixes an answer, and whether the answer already sits in a language model's weights. Category assignment was fixed before any condition was run, and no question moved category during the revisions of Section 3.6. Table 1 records each category's epistemic demand, ground-truth source, and correctness criterion.

Table 1. Evaluation dataset structure and ground-truth sources.

Success Criteria	Ground-Truth Source	Epistemic Demand	Count	Category
Stated numerical criterion: value within ±10%, unit correct, source database cited. Material-list questions are also included; all items use the generic correctness rubric (Section 3.5.3).	ecoinvent 3.9.1 and TianGong Data Platform v1.0 factor records, read directly from the source databases	Numerical precision, database provenance	25	Factual
Clause citation matches published text, interpretation accurate	The 23-document corpus of standards and supporting material, with each target linked to its source and documented derivation (Section 3.6); the ISO 14068-1 target uses a secondary briefing (Section 4.4); 6 of 25 targets rewritten in the audit	Textual interpretation, clause citation	25	Standards
Top-3 components identified correctly, carbon footprint within ±10%	Neo4j pre-computation, manual spreadsheet verification	Graph traversal, numerical aggregation	25	Multi-hop
Carbon reduction within ±10%, minimum 2 trade-off dimensions addressed	Recomputed from BOM mass and emission-factor records held in the graph; 17 of 25 targets rewritten and 8 replaced through seed enumeration in the audit (Section 3.6)	Numerical comparison, contextual trade-offs	25	Design optimization

3.5.2. Evaluation Conditions

The evaluation comprises five primary conditions, one reference-informed diagnostic condition, one dense-retrieval sensitivity check, and two routing policy analyses. The five primary conditions differ in what they can reach.

- **Hybrid (proposed):** Category-conditioned concurrent fusion of graph traversal and text retrieval.
- **Neo4j-only:** The knowledge graph without text retrieval, isolating structured precision.
- **Local-only:** GraphRAG local search without graph access, isolating text coverage.
- **BM25-RAG:** Okapi BM25 lexical retrieval over the flat corpus, as a standard sparse-retrieval reference.
- **LLM-only (closed book):** No retrieval, establishing what parametric knowledge alone supplies.
- **Oracle-context (reference-informed diagnostic):** Reference passages supplied directly to measure generation with reference-selected context.

BM25 serves as the lexical baseline because it provides transparent, reproducible sparse matching over the same flat corpus, supplying a non-neural reference point distinct from both graph traversal and GraphRAG.

The last two conditions support a descriptive budget decomposition. Closed book provides the no-retrieval reference, and oracle context supplies a reference-informed diagnostic level. Section 4.6 reports the increments between these levels and the remaining distance to a perfect score for design optimization. Oracle-context passages are selected from ground-truth provenance.

A dense-retrieval condition using bge-m3 was also run on the original target basis and reached 0.5970, but was not re-measured after ground-truth revision.

3.5.3. Scoring Protocol

Scores come from an LLM judge over four dimensions: correctness, completeness, faithfulness and relevance. The archived configuration identifies Claude Sonnet 4.6 (Anthropic, San Francisco, CA, USA), a different model family from the reported generator. The retained prompts request integer scores from 0 to 5; reported means divide the stored dimensional scores by five. Correctness and completeness use the question, reference and response; relevance uses the question and response; faithfulness uses the question, response and captured context without the reference. No category label is included in these calls. The stated reference criteria include a ±10% tolerance, units and database provenance for factual carbon-factor items, but the retained generic correctness prompt does not separately encode those rules. Supplementary Material S2 reports the actual anchors and parser behavior. Source attribution is audited separately because a correct answer can omit an identifier. Correctness is the primary metric throughout, for the reason Section 3.5.5 gives.

Three protocol decisions change how the results should be read, and Supplementary Material S4 summarizes each one. Faithfulness is undefined for closed-book generation and for a backend without directly scorable context, so nonempty-answer records retain nulls there. The blank-answer branch can instead assign zero when retrieval is enabled. Aggregate comparisons shared across conditions use the three generally defined dimensions and their available paired observations. Retaining genuine empty returns is necessary to describe failure to answer, which affects 8 of 25 standards and 3 of 25 design-optimization questions in the archived graph condition. The intended protocol distinguished retrieval absence from infrastructure failure. Subsequent inspection found that some query paths collapsed these states and that the judge maps blank, N/A, and literal timeout responses to zero on the three common dimensions. Supplementary Material S6.8 reports the recovery limits.

3.5.4. Statistical Treatment

Comparisons are gated on a Friedman test [78] across all conditions, and pairwise Wilcoxon signed-rank tests [79] follow only if the gate rejects. Holm correction [80] is applied within each declared family of comparisons, and the family is stated wherever results appear. Effect sizes use rank-biserial correlation [81,82] for paired comparisons. Two-sided Wilcoxon tests use the normal approximation with tied-rank variance adjustment, removal of zero differences, and no continuity correction. Integer or rational score differences preserve mathematical ties. Signed ranks use magnitudes of differences under the prespecified numerical score coding; the location-shift interpretation assumes symmetric difference distributions. The marginal intervals use 10,000 cluster bootstrap resamples (seed 20260803) [83], drawing the first matched product or standard cluster with all member items retained, conditional on fixed targets, judge scores, and database inputs.

Category estimates are interpreted with the study-specific cluster-bootstrap intervals in Sections 4.2 and 5.1. For example, the standards retrieval gain over closed book is 0.0160 with a 95% CI of $[-0.0986, 0.1420]$; the interval spans both a small loss and a moderate gain. Per-category ratio metrics are not reported: subset denominators are small and the interval width is itself material. A conditional simulation using the same 25 standards items and 22 clusters estimates an 80%-power minimum detectable positive mean difference of about 0.16 for the nominal two-sided 5% percentile-bootstrap rule (empirical null rejection 6.45%; Supplementary Material S6.1).

Observations are not independent: repeated runs of a question are averaged into one per-item score, while items grounded in the same product or standards document share a data-generating process. Treating those per-item scores as independent would overstate the effective sample size and understate uncertainty, so confirmatory comparisons also use the cluster-robust procedures defined below. Questions draw on 24 products and the 23-document corpus. Each question is assigned to a product cluster where its evidence names one and otherwise to a standard cluster. Items matching neither are singletons, giving 76 first-match label clusters over 100 items at a maximum cluster size of five. This partition can split aliases of one standard and leaves some shared source dependencies unmodelled; a 74-cluster alias-merged sensitivity is reported in Supplementary Material S8. Three estimators accompany each headline comparison. A cluster bootstrap resamples whole clusters with replacement 10,000 times for percentile intervals. A cluster-level sign permutation test flips each cluster's mean difference 10,000 times. A leave-one-cluster-out sweep reports the estimate's range when any single cluster is removed. Where a cluster-robust interval spans zero, the item-level test is not treated as confirmatory.

3.5.5. Judge Validation

The original judge-validation sample comprised 30 stratified items scored by one annotator (Supplementary Material S4). Two annotators independently rated all 25 standards questions under local-only and closed-book LLM-only conditions from one preselected run (Supplementary Material S6.2), using separately randomized blind packets and the frozen 0–5 correctness rubric. This yielded 50 responses per annotator and 100 rating events. Marginal 95% intervals used 10,000 question-level bootstrap resamples (seed 2026092303), retaining both conditions and both fixed annotators within each question. Results are separate from the original 30-item sample (Section 4.4 and Supplementary Material S6.2).

Correctness is the primary outcome; faithfulness and relevance are secondary. The identifier-provenance diagnostic is computed separately from judge scores. The judge-swap control on 20 responses gave a three-dimension composite difference of −0.0067 (two-sided asymptotic Wilcoxon $p = 0.3173$). The additional standards validation reports agreement patterns alongside the frozen judge scores.

3.6. Reference-Answer Construction and Verification

In a standards-governed domain, each reference answer must be traceable to a source record. Otherwise, the scoring process cannot distinguish an incorrect system answer from an unverifiable target. The former indicates a system error; the latter indicates a defect in the benchmark.

Ground-truth construction therefore followed an explicit admission rule. Every reference answer identifies the records from which it was derived, documents the calculation or interpretive step, and carries a changelog for subsequent revisions. Targets that could not be derived from the held data were excluded from the primary metric or replaced with questions supported by those data.

Revision of the target set covered 23 original targets: six standards questions were corrected on 27 July 2026, and 17 design-optimization questions were repaired on 28 July 2026. The 17 repaired items were recomputed from source records; 11 had contained unsourced numerical approximations and six had wording or constraint defects. Seed enumeration supplied replacements for a further eight design-optimization items. Stage 3a re-judged stored Hybrid and BM25 answers against the repaired targets, whereas Stage 3b generated fresh full-set runs for the other conditions and the router. A further standards correction (QA_039) followed the corpus audit on 2 August 2026. The final reporting basis was frozen on 3 August 2026, and the primary results use the final corrected and enumerated target set.

Corpus support for standards targets was assessed separately. Each target was reduced to content keys and searched first within chunks from the named standard and then across the full corpus. Items flagged as partially supported or unsupported were adjudicated manually because paraphrases can distribute the target keys across differently worded passages. Section 4.4 reports the audit outcomes, and Supplementary Material S1 summarizes the screening and adjudication outcomes.

3.7. Material-Stage Carbon Estimates

Carbon totals describe the material-stage impacts of the recorded inventory configurations.

The accounting unit is one modeled inventory configuration represented by the recorded bill of materials, including frames and incomplete furniture combinations. Material-substitution calculations hold this recorded configuration fixed. Functional equivalence requires a separate design assessment.

The reported carbon values are material-stage screening estimates. They combine author-compiled component masses with source market factors, whose cumulative inventories include upstream activities and market transport where documented. Product-specific manufacturing energy, distribution, installation, use, end-of-life, and carbon storage are outside the modeled scope.

For ecoinvent, the selected source category is IPCC 2021 GWP100 [71]. In the two inspected inventories, ordinary non-fossil CO_2 emissions and CO_2-in-air uptake have no characterization factors in that category and contribute zero to this calculation; non-fossil methane uses a factor of 27, while specified soil/biomass-stock CO_2 flows carry +1 or −1. The TianGong factor map was reproduced from four Environmental Footprint climate-method files, whose metadata cite IPCC 2013 [84]. The manuscript reports the two database-specific factor systems separately and uses the resulting calculations for material-stage screening (Supplementary Material S6.3).

Across the 24 bills of materials, product-page records, assembly instructions, geometric estimates, and assumed densities define the input rows. Independent reconstruction reproduced historical carbon-total calculations to within 0.000337 kg CO_2e (Supplementary Material S6.4). Under reconciled accounting, 35 of 50 objectives differ from the historical

convention: all 25 multi-hop and 10 design-optimization targets (QA_083, QA_084, QA_086, QA_088, QA_091, QA_092, QA_095, QA_096, QA_098, QA_100). Bare-component contributions without a source mass are encoded as null. A separate seven-product mass-gap mask (H04, H08, H13, H14, H18, H19, H21) affects 13 of 100 numerical targets (Supplementary Material S6.7). Conservative bounded-score sensitivity analysis retains all 100 items and all historical scores: each affected normalized three-run item correctness is relaxed to $[0, 1]$ while unaffected scores are held fixed, and sharp bounds are computed over the resulting cube (Supplementary Material S7.1). For design optimization, the sensitivity intervals admit different level orderings and performance-budget increments; multi-hop category means are unrestricted in $[0, 1]$ because all 25 items are affected. The factual graph-minus-local contrast remains positive, $[0.1227, 0.5227]$, under the broader 40-question union including all five factual product-total questions; these bounds quantify the specified accounting and mass dependencies. The category estimates in Section 4.2 and performance-budget values in Section 4.6 are conditional on recorded benchmark targets; the sensitivity intervals appear in Supplementary Material S7.1. The $\pm 10\%$ tolerance is a stated reference-answer criterion (Section 3.5.3). Source reconciliation and record corrections are detailed in Supplementary Material S6.4–S6.7.

Under these material-stage assumptions, the values support comparative screening between material alternatives within a recorded inventory. Section 4.5 assesses computational consistency under shared inputs.

4. Results

4.1. Aggregate Performance

Each of the five primary conditions and the archived router has three scored full-set runs against the final question set of Section 3.6. A Friedman test across the five conditions rejects equal performance on correctness (chi-square 82.36, $p = 5.50 \times 10^{-17}$).

Three of the four margins are large and survive Holm correction (Table 2). The hybrid gains 0.1140 to 0.2960 over local-only, closed-book, and BM25-RAG, all at corrected p below 0.003. Against the graph store the difference is 0.0007, which is 0.1 times the between-run standard deviation.

Table 2. Aggregate performance and pairwise comparisons with the hybrid condition.

n	*p* (Holm)	Hybrid Difference	SD	3-dim Composite	SD	Correctness	Condition
N/A	N/A	N/A	0.0077	0.7264	0.0092	0.6220	Hybrid (proposed)
66	0.477	+0.0007	0.0064	0.6780	0.0061	0.6213	Neo4j-only
86	0.00241	+0.1140	0.0092	0.6756	0.0092	0.5080	Local-only
90	<0.00001	+0.2487	0.0080	0.5051	0.0160	0.3733	LLM-only (closed book)
89	<0.00001	+0.2960	0.0099	0.5020	0.0111	0.3260	BM25-RAG

Note: $n = 100$ questions per condition, three runs each on one frozen basis. SD is between-run standard deviation, a stability estimate under re-execution. The 3-dimension composite averages correctness, completeness, and relevance, using the available defined dimensions within each run before averaging runs (one completeness and two relevance scores are missing in the 1800-record pool); faithfulness is excluded because it is undefined for LLM-only and local-only (Section 3.5.3). Differences are hybrid minus baseline correctness. Wilcoxon signed-rank tests use run-averaged per-item scores with Holm correction across four comparisons; n counts non-tied pairs.

Aggregate correctness differs by only 0.0007, yet the four equally sized categories diverge: −0.0133 for factual, +0.1360 for standards, −0.0373 for multi-hop, and −0.0827 for design optimization. The positive standards gain offsets the three negative differences, making the near equality a benchmark-specific cancelation.

Table 3 and Figure 3 report cluster contrasts for hybrid versus closed book, hybrid versus text, graph versus text, and hybrid versus graph. The first three margins remain positive, with a maximum leave-one-cluster-out shift of 0.0113. The interval for concurrent fusion minus the graph store spans zero.

Table 3. Cluster-robust inference on the aggregate comparisons: correctness.

Leave-One-Out Range	*p* (perm)	95% CI	Difference	Comparison
[+0.2418, +0.2585]	0.0001	[+0.1762, +0.3217]	+0.2487	Hybrid vs. closed book
[+0.1057, +0.1219]	0.0003	[+0.0571, +0.1772]	+0.1140	Hybrid vs. text index
[+0.1051, +0.1246]	0.0080	[+0.0327, +0.1903]	+0.1133	Graph store vs. text index
[−0.0088, +0.0088]	1.0000	[−0.0532, +0.0628]	+0.0007	Hybrid vs. graph store

Note: 76 clusters over 100 items, assigned by product when evidence names one and by standard otherwise. Cluster bootstrap with 10,000 resamples, cluster-level sign permutation with 10,000 permutations, bootstrap seed 20260803 and permutation seed 20260804. The leave-one-out range gives the estimate's extremes when any single cluster is dropped.

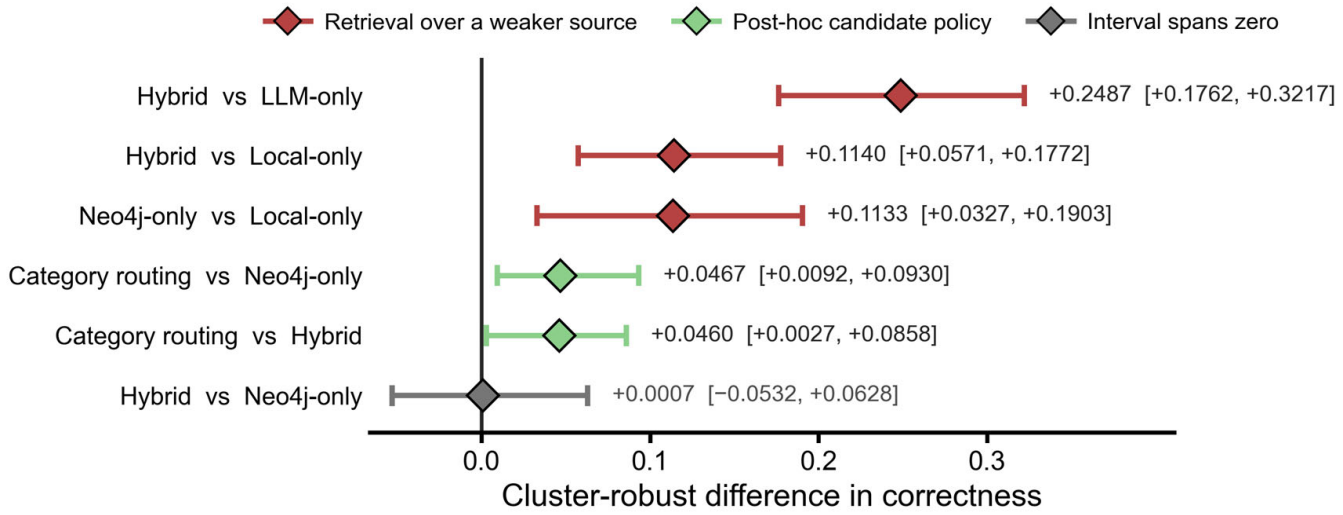

Figure 3. Cluster-bootstrap intervals for correctness contrasts, $n = 100$ questions in 76 clusters. Points are mean paired differences and bars are marginal 95% percentile intervals from resampling whole clusters. The hybrid-versus-graph interval spans zero. Green points show post hoc category-policy estimates derived from the existing score pool. Section 4.7 gives the related routing diagnostics.

Figure 4 summarizes aggregate correctness. The difference on the 3-dimension composite was not statistically significant ($\Delta = 0.0484$, Holm-adjusted $p = 0.5823$). The secondary 4-dimension composite showed a positive difference ($\Delta = 0.1006$, Holm-adjusted $p = 0.0164$), which includes faithfulness measured at kappa 0.567. Friedman repeat-run checks on correctness gave hybrid chi-square 0.34, $p = 0.845$, and BM25-RAG chi-square 0.27, $p = 0.875$. The aggregate between-run standard deviation is 0.0092 on correctness and 0.0120 on the four-dimension composite, requiring all four defined dimensions within each run (one hybrid run-score is unavailable).

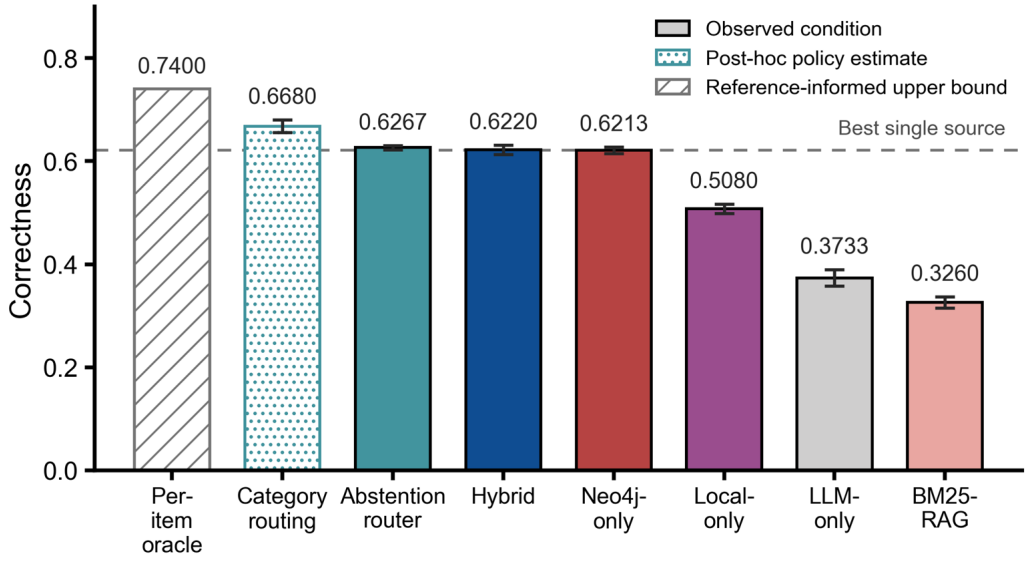

Figure 4. Aggregate correctness by condition, ordered by mean, $n = 100$ questions and three scored runs per condition. Error bars give the between-run standard deviation, a stability estimate under re-execution. The dashed line marks the best single source. The per-item oracle (0.7400) is the run-averaged per-item maximum. Colors distinguish the conditions labeled on the horizontal axis: solid bars show observed conditions, the dotted bar shows the post hoc category policy, and the hatched bar shows the reference-informed upper bound. The abstention router is the observed live policy above the best single source; category routing is shown as a post hoc policy estimate.

4.2. Category-Level Analysis

Figure 5 shows the category-specific ordering of retrieval conditions. Neo4j-only leads on factual, multi-hop, and design optimization but drops to 0.4533 on standards (Table 4). Local-only leads on standards but reaches only 0.3867 on multi-hop. The closed-book control reaches 0.6240 on standards yet only 0.2880 on factual. The hybrid leads in none of the four categories.

Table 4. Category-level correctness and paired contrasts. Each mean or difference is followed by its marginal cluster-bootstrap 95% CI. Parentheses report the recomputed raw and/or Holm-adjusted Wilcoxon *p*-values as labeled. Bootstrap intervals are unadjusted and conditional on frozen targets, original judge scores, and the first-match clusters.

Graph − Text (p Raw; p Holm)	Hybrid − Closed Book (p Holm)	Closed Book	Text Index	Graph Store	Hybrid	Category
+0.2747 (0.0036; 0.0109) 95% CI [0.1024, 0.4528]	+0.5467 (0.0000853) 95% CI [0.4361, 0.6472]	0.2880 [0.2361, 0.3444]	0.5733 [0.4453, 0.6864]	0.8480 [0.7253, 0.9444]	0.8347 [0.7304, 0.9253]	Factual
−0.1867 (0.0343; 0.0685) 95% CI [−0.3576, −0.0405]	−0.0347 (0.2951) 95% CI [−0.1385, 0.0833]	0.6240 [0.5101, 0.7262]	0.6400 [0.5462, 0.7307]	0.4533 [0.2720, 0.6190]	0.5893 [0.4939, 0.6812]	Standards
+0.2933 (0.0004; 0.0016) 95% CI [0.1364, 0.4067]	+0.3147 (0.0001322) 95% CI [0.2254, 0.4099]	0.3280 [0.2444, 0.4049]	0.3867 [0.3256, 0.4606]	0.6800 [0.5565, 0.7704]	0.6427 [0.5420, 0.7227]	Multi-hop
+0.0720 (0.4327; 0.4327) 95% CI [−0.0619, 0.2167]	+0.1680 (0.1456) 95% CI [0.0213, 0.3173]	0.2533 [0.2000, 0.3037]	0.4320 [0.3333, 0.5333]	0.5040 [0.3361, 0.6773]	0.4213 [0.2889, 0.5590]	Design optimization

Note: Correctness, $n = 25$ questions per category and three runs per condition. Contrast *p*-values come from Wilcoxon signed-rank tests on run-averaged per-item scores, with Holm correction within each four-comparison family. Full means for BM25-RAG and the abstention router, together with comparisons against the strongest category-specific baseline, appear in Supplementary Table S4.

Against the strongest observed baseline in each category, hybrid correctness is lower in all four categories, and none of these comparisons reaches significance after Holm correction (Supplementary Table S4). Table 4 reports different contrasts: graph minus text and hybrid minus closed book. Its marginal bootstrap intervals apply to those specific paired differences.

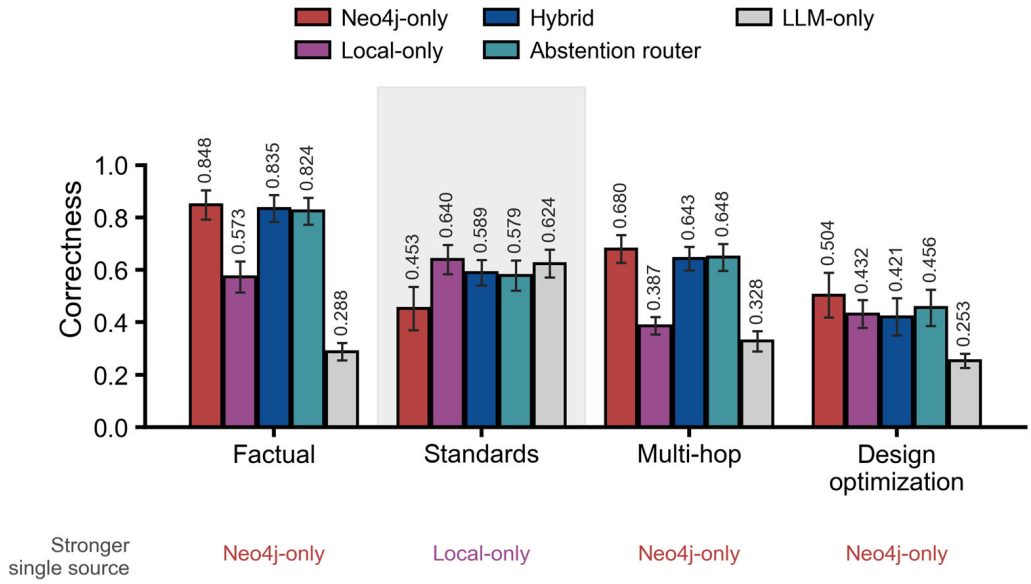

Figure 5. Category-level mean correctness by condition, $n = 25$ items per category and three runs per condition. Error bars give the descriptive item-based standard error (SD of run-averaged item correctness divided by $\sqrt{25}$); Table 4 reports the cluster-bootstrap intervals used for inference. The reversal between the two single sources is the substantive pattern: the graph store leads on factual and multi-hop queries while the text index leads on standards. BM25-RAG is omitted from this category plot; its full category means are provided in Supplementary Table S4. Differences between the hybrid and the strongest per-category baseline are not significant after Holm correction, so no significance markers are shown. The graph–text contrasts for factual and multi-hop are significant after Holm correction and are reported in Table 4 rather than encoded in this plot.

Closed-book generation reaches its lowest category score, 0.2533, on design optimization. The hybrid's own worst category is design optimization at 0.4213. The abstention-gated router of Section 4.7 reaches a worst-category score of 0.4560, the highest of any condition tested.

The hybrid's multi-hop deficit against the strongest baseline is 0.0373, significant before Holm correction ($p = 0.0331$) but not after ($p = 0.1324$). For hybrid correctness, category-specific between-run standard deviations are 1.8 to 6.3 times the aggregate between-run standard deviation.

Against the closed-book control (Table 4), factual and multi-hop gains are large and survive correction, at 0.5467 and 0.3147. Standards run slightly negative at −0.0347.

4.3. The Representational Crossover

The graph-minus-text difference varied by category. A Kruskal–Wallis test on the per-item differences indicated heterogeneity ($H = 23.39$, $p = 3.35 \times 10^{-5}$). Contrasting standards against the other three pooled gives $H = 16.28$, $p = 5.46 \times 10^{-5}$, the difference running −0.1867 on standards against +0.2133 elsewhere. Cluster bootstrap over the 76 product and standard clusters confirms this: 95% CI [−0.59, −0.23], cluster permutation $p = 0.0001$ (plus-one correction).

Table 4 decomposes that interaction. Graph minus text is +0.2747 for factual questions (95% CI [0.1024, 0.4528]; raw $p = 0.0036$, Holm-adjusted $p = 0.0109$) and +0.2933 for multi-hop questions (95% CI [0.1364, 0.4067]; raw $p = 0.0004$, Holm-adjusted $p = 0.0016$). The standards reversal is −0.1867 (95% CI [−0.3576, −0.0405]; raw $p = 0.0343$, Holm-adjusted $p = 0.0685$), and design optimization is +0.0720 (95% CI [−0.0619, 0.2167]; raw and adjusted $p = 0.4327$). The percentile-bootstrap intervals are marginal and unadjusted, reported alongside the Holm-adjusted Wilcoxon tests. The factual means of 0.8480 for graph retrieval and 0.5733 for local text retrieval measure correctness, not source coverage. A source-locator audit decomposes the 25 factual questions into 15 material-factor lookups (graph 1.0000, local 0.4578), five product-total questions (0.4933, 0.7333), and five material-list questions (0.7467, 0.7600). Their respective contributions to the 25-item graph-minus-local gap are +0.3253, −0.0480, and −0.0027, summing to +0.2747. All 25 corresponding process or product identities are present in the current text index, but the 15 factor questions refer to only seven base process records whose descriptions do not encode the characterized factors and material-specific conversions supplied to the graph. The current index is not fingerprint-bound to the historical runs, and the local logs did not independently preserve the returned source records, so historical retrieval coverage and a pure architectural effect remain unidentified.

The closed-book control adds a third contrast. Standards carry the highest closed-book score at 0.6240, against 0.6400 for the best retrieval-only condition; factual figures are 0.2880 and 0.8480. The resulting structure is a crossover interaction under hidden stratification: the ordering of conditions depends on an unmodelled property of the query.

The aggregate interaction and the factual and multi-hop differences of 0.27 to 0.29 support descriptive boundary mapping between the two representations. The standards reversal does not survive Holm adjustment ($p = 0.0685$), and the expanded standards validation shows annotator-dependent judge agreement and imprecise local-only versus closed-book differences (Section 4.4). The category pattern is hypothesis-generating within this question set.

4.4. Reference Answers and Judge Agreement

The evaluation checks addressed two potential sources of measurement error: variation in the LLM judge and defects in the reference answers. Agreement between the judge and human annotation depended on the scoring dimension (Supplementary Table S5). Quadratic-weighted kappa was 0.836 for correctness and 0.848 for completeness, compared with 0.567 for faithfulness and 0.552 for relevance. Confirmatory comparisons consequently use correctness.

The historical seven-item standards subset had quadratic-weighted kappa = 0.521, compared with 0.908 on the factual subset. In the expanded standards validation, A/B quadratic-weighted kappa was 0.709 [0.519, 0.855] (estimate [95% CI]), and the A − B mean difference was 0.54 [0.28, 0.82] points on the 0–5 scale. Agreement with the original judge differed between annotators: A/judge kappa was 0.541 [0.368, 0.701] and B/judge kappa was 0.804 [0.685, 0.882]. Local-only minus closed-book differences were 0.24 [−0.28, 0.80] points for A and 0.36 [−0.16, 0.92] for B; both intervals include zero. Supplementary Tables S13 and S14 give agreement, condition means, and paired contrasts.

The historical target audit changed the reported design-optimization scores. Of the 25 design-optimization items, 17 were repaired (11 numerical target corrections and six wording or constraint corrections) and eight were replaced through seed enumeration. Across the target-revision stages, retrieval performance rose and closed-book performance fell. Target repair, item replacement, and generation or rejudging schedules changed together, so their separate effects are confounded. This audit reuses the existing archived data;

independent candidate arithmetic has been reconstructed from the original BOM rows and frozen factor assumptions. Bounded-score sensitivity analysis over the resulting 35-of-50 affected-target mask and the 13-of-100 mass-gap mask is reported in Supplementary Material S7.1; full details of the reconstructed conditional arithmetic appear in Supplementary Material S6.4–S6.7. The frozen numerical-target dependency audit identifies six questions involving H05, four involving H06, three involving H10, and four involving H14, with seven unique questions overall (six design-optimization and one factual; Supplementary Table S12). These counts include all-product and series aggregates.

The corpus-side screen classified 18 of the 25 standards targets as fully grounded, five as partially grounded, and two as unsupported by keyword matching. Manual inspection showed that six of the seven flags were caused by paraphrases whose content keys appeared in different wording. The remaining flag identified a corpus limitation: for the definition in ISO 14068-1:2023 [85] of carbon neutrality, the index contains a BSI executive briefing [86] rather than the standard text, and the target is limited to the distinction stated in that source.

4.5. Entity Alignment and Calculation Consistency

The entity-alignment audit found no incorrect process match among the 30 lowest-confidence equivalences or the 20 randomly sampled pairs. Ambiguity was confined to co-product naming conventions for records with the same reference flow. Because the evaluation queries do not traverse these equivalence edges, the alignment result does not affect the reported architecture scores.

Carbon accounting through the graph was then compared with the separate spreadsheet implementation for 24 products. Both calculations used the author-compiled bills of materials and the same ecoinvent 3.9.1 emission factors. The comparison measures implementation consistency under shared inputs (Section 3.7).

The archived graph-to-ledger validation reports mean absolute percentage error of 6.4962% across 24 products. H05, H06, H10 and H14 have signed graph-minus-ledger differences of −33.7647%, −54.1236%, −32.2237% and −28.2888%, respectively. The other 20 products have MAPE of 0.3755%. Composite mapping is implicated in the four outliers. Conditional constituent-decomposition scenarios are reported separately in Supplementary Table S16.

4.6. Source Attribution and Design-Optimization Performance

We examined 190 responses to standards questions in 13 available hybrid evaluation exports, including the 75 responses in the current three-run analysis. Seven responses received correctness 5/5 while the legacy identifier extractor returned an empty list; all seven explicitly named a document or framework. The extractor recognizes GB/T, ISO, EN, PAS, ASTM and IEC patterns but omits identifiers such as T/CSTE and T/CECA-G. The seven named-source responses are false negatives of this pattern-based diagnostic.

Source attribution requires clause-level support and correct edition identification beyond document names alone.

Design optimization shows where remaining headroom lies. Closed book reaches 0.2533, retrieval adds 0.2507, selection between backends would add 0.0933, and oracle context adds 0.1947 to reach 0.7920. A residual of 0.2080 to a perfect score remains in the separate one-run oracle-context condition. Figure 6 shows this descriptive budget.

Section 4.7 examines backend complementarity at the level of individual questions.

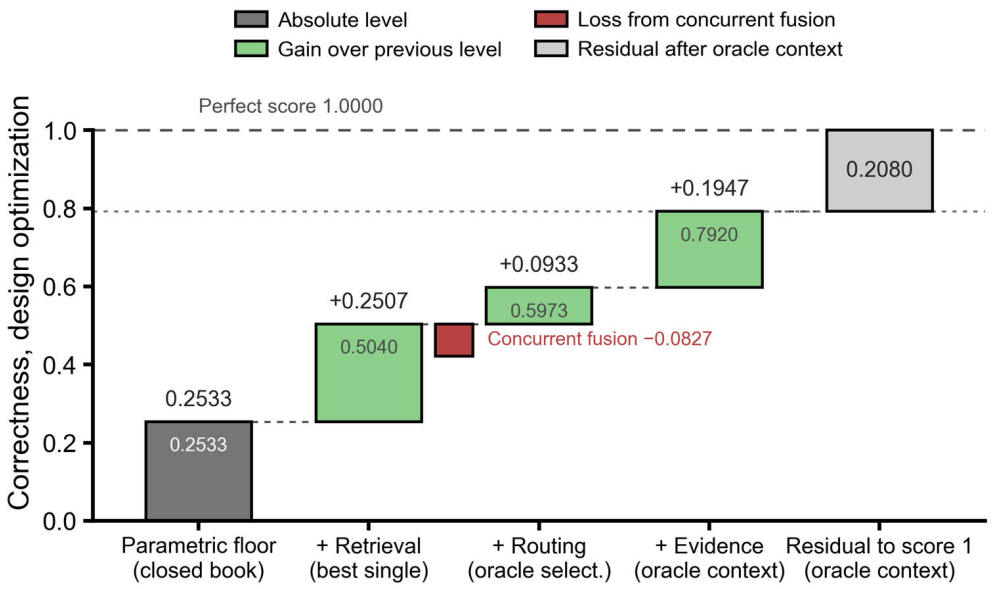

Figure 6. Performance budget for design optimization on correctness, $n = 25$ items. Gray bars are absolute levels anchored at zero and colored bars are the increments bridging each level to the next: the parametric floor supplied by closed-book generation, the increment from retrieval, the further increment a per-item oracle over the two backends would add, and the increment from supplying oracle context. The residual is the difference between the separate one-run oracle-context score and 1.0. The downward step marks concurrent fusion, which scores below the better of the two backends it combines in this category. Displayed values are conditional on recorded benchmark targets; bounded-score sensitivity intervals under the 35-of-50 affected-target relaxation are reported in Supplementary Material S7.1.

4.7. Realized and Recoverable Gains from Backend Complementarity

The two backends cross over by category (Section 4.2) and fail on different items (Figure 7). A per-item oracle choosing the better backend for each question sets the member-selection ceiling. The central question is how much of that ceiling a deployable routing policy can reach.

Cross-fitting evaluates whether source selection transfers across runs (Table 5). It selects on one run and scores on a held-out run, averaged over six ordered pairs of three runs. The result estimates cross-run headroom on the same 100 questions.

Table 5. Cross-fitted oracle and decomposition of routable headroom.

Share of Correctness Headroom	3-dim Composite	Correctness	Quantity
N/A	0.8189	0.7447	In-sample oracle
N/A	0.8068	0.7280	Cross-fitted oracle
14% of apparent headroom	+0.0121	+0.0167	Selection bias
N/A	0.6780	0.6213	Best single source
100%	+0.1288	+0.1067	Cross-run routable headroom
44%	N/A	+0.0467	Observable: graph store abstains
56%	N/A	+0.0600	Unobservable: graph store answers

Cross-fitting reduced the apparent correctness headroom by 14%, leaving an estimated headroom of 0.1067, or seventeen times the graph store's between-run standard deviation. Table 5 also separates this remainder according to whether the graph store's failure was observable at query time. The estimands differ by operation order: averaging each item over runs before selecting the better source yields 0.7400, hence 0.1187 above the graph

store; selecting per run before averaging yields 0.7446667, hence 0.1233333. The six ordered training/test-run pairs yield 0.7280 and a cross-run headroom of 0.1067, with the graph selected on training ties. This estimate characterizes source-selection transfer within the fixed 100-question benchmark. Separately, routing all 16 graph-zero items in the run-averaged pool would contribute 0.0707; the 11 abstentions contribute 0.0473 and the five non-empty zero-score answers contribute 0.0233. Supplementary Table S10 reconciles these item sets with the signal-depth analysis in Section 5.4.

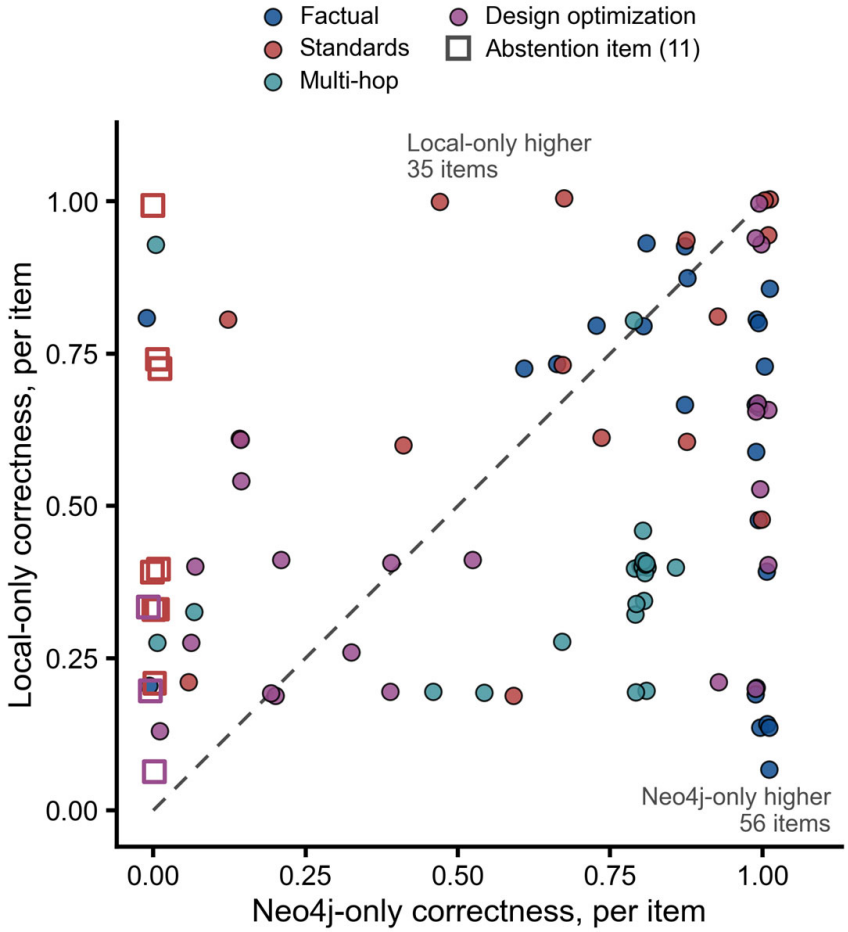

Figure 7. Per-question correctness under the graph-store backend against the text-retrieval backend, colored by category, $n = 100$. Points off the equal-performance diagonal are questions where the two backends receive different ordinal scores, and they are distributed across categories rather than concentrated in one. Hollow square markers on the left edge are the eleven questions with recorded empty graph-path returns and zero scores; marker shape identifies abstention and marker color identifies question category. These empty returns triggered the archived router's fallback, although the records do not distinguish retrieval absences from infrastructure failures. Exploratory item-level Spearman rho = 0.2285 (unadjusted asymptotic $p = 0.0222$). Points are jittered by ±0.012 to separate tied ordinal scores, so plotted positions are displaced slightly from the exact judged values.

The cross-run observable contribution gives a reference performance level of 0.6680. The deployed hybrid reached 0.6220, compared with 0.6213 for the best single source, realizing less than 1% of the cross-fitted headroom. At item level, it matched or exceeded the better backend on 39 questions, fell between the two backend scores on 30, equaled the weaker backend on 11, and scored below both on 20.

4.7.1. Recovery Through Abstention-Based Selection

The archived abstention-routing records distinguish a rerun from a paired score-pool counterfactual. The implemented wrapper queried the graph store first and fell back on a recorded empty status or blank response, without consulting ground truth. Across the three archived runs, 89 items followed the primary path and 11 the fallback. The historical trigger records the 11 fallback activations while leaving retrieval absence and possible infrastructure failures indistinguishable.

The −0.0427 shift is the mean correctness difference conditional on the 89 primary-path items, contributing −0.0380 to the 100-item mean. The 11 fallback items contributed +0.0433, leaving an overall archived rerun difference of +0.0053. Pairing the historical baseline and rerun blocks by ordinal position showed identical stored retrieval-context objects for all 267 primary-path pairs, but identical answer text for only one pair. The observed change therefore occurs downstream of the exposed context; generation variability, judge variability, hidden prompt or backend changes, and retries cannot be separated in these nonrandomized run blocks. The paired score-pool counterfactual held each graph score fixed and substituted the archived local score for the 11 fallback items only, yielding +0.0473 (between-run SD 0.0012). This conditional reuse of recorded scores isolates the fallback substitution effect within the existing score pool.

4.7.2. Recovery Through Category-Based Selection

Sending benchmark-labeled standards questions to the text index and all other questions to the graph store gives a score-pool estimate of 0.6680 (Table 6). This retrospective calculation applies the fixed benchmark category labels. A separate offline implementation applies the frozen existing question-only classifier and dispatches the selected backend. Its route agrees with the benchmark-label assignment on 95 of 100 questions; four-class category agreement is 46 of 100, and these routing rules were developed on the benchmark without a held-out evaluation. Prospective live execution of both policies under the three-run protocol is reported below and in Supplement S7.3.

Table 6. Performance of abstention- and category-based routing policies.

p (Cluster)	95% CI (Cluster)	p (Item)	SD	Difference	Policy Comparison
N/A	N/A	0.4708	0.0092	+0.0053	Live abstention rerun vs. graph store
N/A	N/A	0.0032	0.0012	+0.0473	Paired abstention counterfactual vs. graph store
0.0454	[+0.0027, +0.0858]	0.0043	N/A	+0.0460	Category routing vs. concurrent fusion
0.0240	[+0.0092, +0.0930]	0.0343	N/A	+0.0467	Category routing vs. graph store
N/A	N/A	0.8273	N/A	−0.0007	Category routing vs. abstention routing

Note: correctness, $n = 100$. Item-level p-values use Wilcoxon signed-rank tests on run-averaged scores; cluster columns use the bootstrap and permutation procedures in Section 3.5.4 over 76 clusters. The category policy was selected using category-level reference-scored performance on this benchmark. The reported counterfactual applies fixed category assignments to that same score pool. The final row compares the two policies as paired counterfactuals, the only form in which their reported scores are directly comparable.

Both intervals for the counterfactual correctness contrasts excluded zero. On the three-dimensional composite, a metric-specific post-hoc category rule (text for standards and design optimization; graph otherwise) exceeded the graph store by 0.0689 (95% CI [+0.0202, +0.1241], cluster permutation $p = 0.0079$). The two counterfactual policies disagreed on 18 questions; their aggregate difference was −0.0007 ($p = 0.8273$). The score-pool category diagnostic recovered 0.0467, or 44% of the cross-fitted headroom. These retrospective estimates use benchmark scores from the same question set. A separate three-run live evaluation used the requested gpt-5 alias, fresh single-source baselines and a 180 s per-channel limit (Supplementary Material S7.3). Complete-pair service-utility gains over the graph were 0.0785 (95% cluster CI [0.0207, 0.1463]) for abstention routing and 0.0756 [0.0287, 0.1308] for category routing, each from 270 of 300 matched run pairs. All-pair missing-score bounds were [0.0233, 0.1500] and [0.0020, 0.1320], respectively. These finite-sample bounds hold scored outcomes fixed, assign zero to non-delivery and allow every unscored success to range over [0, 1]. The sequential hybrid-minus-graph contrast was 0.0098 [−0.0261, 0.0504] from 265 complete pairs. The live evaluation used the provider configuration documented in Supplementary Material S7.3 and is separate from the historical GPT-5.4 concurrent-fusion study.

Most of the historical score-pool abstention gain came from standards questions: eight abstentions contributed 0.0413, or 87% of the full 0.0473 gain (Supplementary Table S6).

Figure 8 locates each policy within the headroom band.

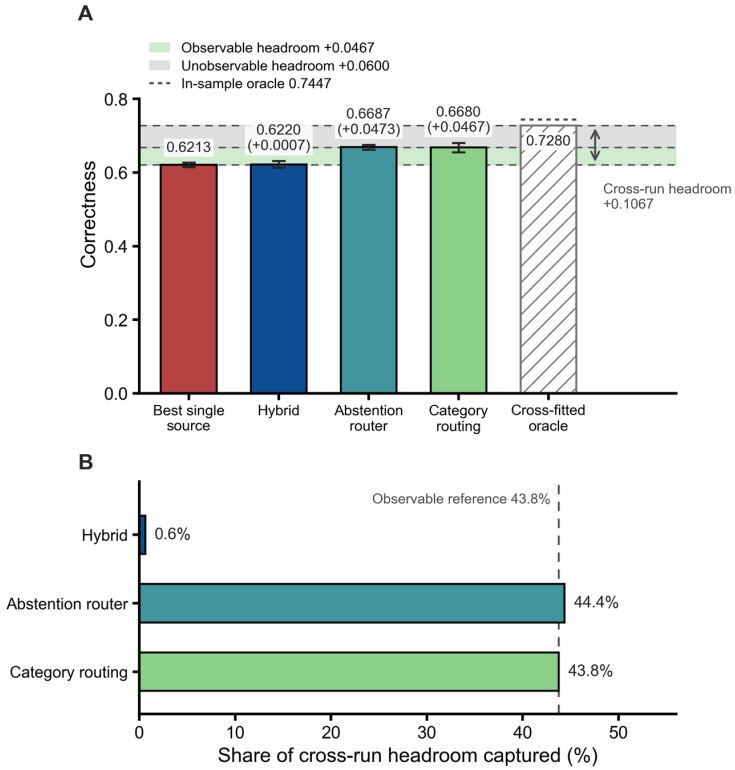

Figure 8. Comparison of routing headroom estimates, correctness, $n = 100$. Panel A gives absolute levels: the band above the best single source is shaded to separate the portion observable at query time, where the graph store abstains, from the portion where it answers but answers less well. Error bars give between-run standard deviations; the router's paired-counterfactual level is 0.6687 with SD 0.0064 (Section 4.7). The SD of its paired gain is 0.0012, as reported in Table 6. The oracle bar is open and hatched because it selects using the reference answer. Panel B expresses each policy as a share of the cross-run headroom estimate, with the observable portion marked as a cross-run reference. The 43.8% line uses the cross-run observable contribution (0.0467), whereas the router's 44.4% uses its paired counterfactual gain (0.0473); these are different estimands.

The historical live router trailed the strongest baseline in all four categories: by 0.0240 on factual, 0.0613 on standards, 0.0320 on multi-hop and 0.0480 on design optimization. The overall archived rerun difference of +0.0053 reflects both the fallback contribution on eleven items and a −0.0380 primary-path shift on the remaining 89 items.

Five-fold cross-fitting of the category rule returned a gain of 0.0467, equal to its in-sample estimate and corresponding to optimism of 0.0000. The folds partition this same benchmark. In the random-grouping control, each item was independently assigned with seed 7 to one of 10, 25, or 50 unequal groups; five folds were formed after shuffling item order with seed 20260807, using `order[j::5]`. The category choice was the training-fold mean-score argmax, with the graph selected on ties and in empty training groups; out-of-fold contributions were weighted by held-out item count. The controls produced optimism of 0.0320, 0.0427, and 0.0747, respectively, with negative out-of-fold gains.

5. Discussion

5.1. A Framework for Retrieval Value

Structured traversal led on factual and multi-hop questions, whereas text retrieval led on standards with a three-run gain of 0.0160 over closed-book generation in that category (95% CI [−0.0986, 0.1420]). Representational determinacy and empirical closed-book answerability organize these outcomes.

Representational determinacy concerns whether the available records uniquely fix an answer. Given a bill of materials and per-kilogram emission factors, a product's material-stage footprint is determined to the precision of the underlying inventory, whereas an approximate range does not provide the same constraint. Carbon assessment also contains a representational mismatch: emission-factor databases are organized by process, while designers reason over products and components [8,87]. By encoding both hierarchies and the relations between them, the graph addresses this mismatch.

Empirical closed-book answerability concerns whether the model can produce the answer without access to the retrieval corpus [17,18]. In this study, the standards are public documents and the component-level bills of materials were compiled by the authors. The estimate comes from the closed-book scores subsequently compared with retrieval performance, making it an operational measure tied to the question, corpus, model, and date. Table 7 places the four task groups in the resulting space.

Table 7. Retrieval gain by task position in the determinacy x closed-book answerability space. Gain is the best retrieval-only condition minus closed-book generation.

Retrieval Gain	Closed Book	Closed-Book Answerability	Determinacy	Task Group
+0.5600 95% CI [0.4487, 0.6583]	0.2880 95% CI [0.2361, 0.3444]	Low	High	Factual (factor lookups, product totals and material lists)
+0.3520 95% CI [0.2508, 0.4472]	0.3280 95% CI [0.2444, 0.4049]	Low	High	Multi-hop (BOM aggregation across products)
+0.2507 95% CI [0.0795, 0.4293]	0.2533 95% CI [0.2000, 0.3037]	Low	High	Design optimization, derivable targets
+0.0160 95% CI [−0.0986, 0.1420]	0.6240 95% CI [0.5101, 0.7262]	High	High	Standards (published national standards)

Note: correctness, with three runs per condition. Gain is the better of the two single-source retrieval conditions minus closed-book generation. Every row sits at high determinacy, so these four groups trace one cross-section through the space along the answerability axis. The low-determinacy quadrants are unpopulated by this design. Answerability is measured against GPT-5.4 as of July 2026. The reported 95% cluster-bootstrap intervals keep the observed winning source fixed and are unadjusted for category-wise source selection or multiple comparisons.

Across the four rows, retrieval gain varies by more than a factor of thirty without following a simple structured-versus-textual distinction. Standards and factual questions both have determinate answers but occupy opposite ends of the gain range, with adequate closed-book performance as their main measured difference. On standards, the closed-book score of 0.6240 was the model's highest category result and exceeded its factual score by 0.3360. The remaining retrieval gain of 0.0160 is consistent with redundant access to knowledge already represented in the model, although retrieval quality and corpus coverage remain alternative explanations that require direct tests of answerability.

The literature considers recall–precision relationships in information retrieval [88] and risk–coverage trade-offs in selective classification [89]. In this study, these trade-offs depend on source availability and the preservation of clause qualifiers: GB/T 46486-2025 combines computable thresholds with conditional provisions [6]. Although the graph reproduced the structured elements, it returned no standards record for eight questions; text retrieval made the clause context available to the generator, but none of the three local-only or three hybrid responses to the core stage question explicitly preserved its mandatory/recommended distinction (Supplementary Material S7.4). Representing a provision and benefiting from its retrieval are therefore separate issues, consistent with prior closed-book analyses [17,18].

The historical design-optimization gain changed from 0.0560 to 0.2507: closed-book correctness fell from 0.3360 to 0.2533, while retrieval rose from 0.3920 to 0.5040. Record-derived targets replaced approximate ranges, but the comparison also spans 17 repairs (11 numerical and six wording or constraint corrections), eight replacements, and different generation and rejudging blocks; these stages changed jointly, confounding their individual contributions. For fixed member answers, co-failure $\beta = 0.480$ at correctness ≥ 0.5 gives a binary selector ceiling of $1 - \beta = 0.520$ [59]. This threshold-based ceiling differs from the graded member bound of 0.5973 and the separate one-run oracle-context score of 0.7920: the three quantities use different selection mechanisms and scoring granularity. In the descriptive budget, the oracle-context increment of 0.1947 exceeds the member-selection increment of 0.0933. Under the 35-of-50 affected-target relaxation, β spans $[0.240, 0.640]$ and the binary ceiling spans $[0.360, 0.760]$. The relaxed scores admit different orderings of Figure 6 levels (Supplementary Material S7.1).

With respect to RQ1, the sampled high-determinacy task classes differ strongly in closed-book answerability, and that variation corresponds to retrieval gain. Table 7 therefore traces retrieval gain along the answerability axis; varying determinacy independently is a separate experimental comparison.

5.2. Predictions That Distinguish the Two Properties

The proposed account predicts different retrieval gains when determinacy and closed-book answerability are varied separately. A preliminary provider diagnostic requested the provider alias gpt-5 without client-side retrieval on GB/T 46486-2025 questions; the documented cutoff for the reported model version precedes the standard. Its categorical audit found less complete identification of the standard and stage names than local-only and hybrid retrieval, but a clearer mandatory/recommended distinction in two responses (Supplementary Material S7.4). The diagnostic uses provider-reported model labels; a matched study using an authenticated earlier-cutoff model would test whether retrieval gain rises as closed-book answerability falls. Conversely, replacing exact targets with plausible ranges would reduce determinacy and should contract the retrieval gain. A precisely estimated gain remaining near 0.0160 under matched conditions would count against the proposed answerability explanation.

Section 5.4 distinguishes fixed-score selection constraints from empirical routing hypotheses. The cross-run analysis estimates headroom on the same questions. A further test

would remove the observability of empty graph returns: under the proposed selectability account, the benefit of abstention routing should then diminish. Supplementary Material S5 states these constraints and tests.

5.3. The Crossover as a Difference in Failure Type

The crossover in Section 4.3 can be read as a boundary between representational failure modes. A structured lookup requires every premise along a query path to be present. Under a simplifying independence approximation, its success probability scales as p^n, where p is per-premise availability and n is path length; coverage can therefore decline sharply as the chain grows. A statistical retriever returns ranked context for any input and usually degrades through partial relevance or incomplete evidence. The expected distinction is consequently between discrete absences from the structured store and dispersed partial credit from text retrieval.

The archived graph failures included empty returns and incorrect answers containing records (Section 4.7; Supplementary Table S10). Empty returns expose a query-time routing signal. Nonempty errors require an additional diagnostic signal or comparison with a reference answer. This distinction explains which part of the observed headroom the abstention rule can access.

Score dispersion provides a second view of the same boundary (Supplementary Table S8). The graph store's SD of run-averaged item correctness increased from 0.2812 on factual questions to 0.4123 on standards, whereas text retrieval showed the opposite category ordering.

The archived pipeline exposed return status as a routing signal. Empty returns could arise from missing retrieval or infrastructure failures, while nonempty answers included partial-credit errors. A symbolic system that supplies defaults could replace empty returns with incorrect nonempty answers. The observed failure pattern therefore depends on the implementation.

5.4. Selectability as a Condition for Realizing Complementarity

The backends crossed over by category and failed on different items, while concurrent fusion realized 0.625% (about 0.6%) of the available headroom. Diversity measures have an inconsistent relationship with ensemble gain [60]; here, the run-averaged correctness scores showed a weak positive item-level Spearman association ($\rho = 0.2285$; unadjusted asymptotic $p = 0.0222$), an exploratory description conditional on the benchmark. The cross-fitted oracle provides a cross-run opportunity estimate of 0.1067. For policies restricted to emitting one member's fixed answer, co-failure imposes the threshold-based ceiling discussed in Section 5.1 [59].

These results motivate selectability as a proposed condition for realizing complementarity: the source-selection decision must be supported by signals available at query time. Selectability is a joint property of the sources and the signals exposed by their outputs. It can also vary in degree. Some graph queries return no record, which is directly observable; others resolve to a neighboring entity and disclose the substitution in the answer text. Table 8 grades the 35 items that contribute to in-sample headroom by the depth of the signals available at query time.

Approximately two-thirds of the in-sample headroom was associated with potentially observable signals. In the historical fixed-score diagnostic across all 100 questions, the abstention rule achieved precision of 1.0000, recall of 0.3143, F1 of 0.4783, and no loss from misrouting. Its limited recall reflects errors in nonempty answers that the abstention signal does not identify (Supplementary Table S10).

Table 8. Signal depth of the routable headroom in correctness.

Detectable from	Share	Headroom	Items	Depth
status field	40%	+0.0473	11	0, abstention
answer text	25%	+0.0293	8	1, disclosed substitution
reference answer only	35%	+0.0420	16	2, silent degradation

Note: all 35 items on which the run-averaged text score exceeds the run-averaged graph score. Depth is assigned by a conservative rule-based detector requiring a majority of runs; depth 1 is therefore a lower bound. The depths partition the 0.1187 gain obtained by averaging each item over three runs before source selection. The separate 0.1233 gain selects per run before averaging. The cross-fitted gain is 0.1067 (Section 4.7). Supplementary Table S10 shows the depth-by-failure-class decomposition.

These results place source selection after the assessment of retrieval value. Cross-fitting estimates the available headroom, whereas inspection of backend failures identifies signals a routing policy can use (Supplementary Material S5).

A prospective implementation would expose, from both backends, the graph return status, the record count, a provenance completeness measure, any disclosed substitution, and calibrated confidence with an associated error profile. Observable graph abstentions would route to the text index. When those signals conflict, or when neither source appears reliable, the system would abstain or request expert review rather than emit an answer. Thresholds and the routing policy would be learned on a development set and evaluated once on an independent benchmark held out from policy development.

5.5. Implications for Design-Stage Carbon Practice

For a manufacturer with a limited data-preparation budget, the results favor structuring exact emission-factor fields, unit conversions, and linked inventory data so that record-specific queries can be answered directly. Retrieval over published standards produced the smallest marginal correctness gain while the closed-book model already performed well on that category (Section 4.2). Within the factual category, the 15 factor-lookup questions contributed +0.3253 to the 25-item graph-minus-local difference, whereas the five product-total questions favored local retrieval (Section 4.3). Standards indexing serves a different role, centered on inspectable clause context and conditional interpretation.

Correctness and source attribution require separate evaluation. The audit in Section 4.6 identified document names missed by the pattern-based extractor. A source-attribution measure needs reliable identifier coverage and clause-level support.

Section 4.5 tests whether two implementations reproduce calculations from the same inputs. For design use, a material comparison also depends on the assumed mass and composition of each alternative and on whether the substitution preserves the required furniture function. The service can support conditional screening of alternatives (Section 3.7).

5.6. Limitations and Future Work

In Supplementary Material S6.2, the 25-question, two-annotator validation covers one preselected run under local-only and closed-book conditions. Agreement varied by annotator; source locators and proposed reference corrections are retained with the audit records. The bootstrap intervals are conditioned on the two annotators and omit between-annotator uncertainty and dependence among questions about the same standard. Faithfulness and relevance remain secondary outcomes given their moderate agreement in the original sample (approximately 0.55); Section 4.6 addresses the separate identifier diagnostic.

The benchmark question set was constructed from the retrieval corpus, and the target-construction audit changed the measured position of some task classes (Supplementary Table S9). The closed-book condition uses no corpus context. A separately constructed

question set would extend the answerability analysis, and holding answerability fixed while varying determinacy would provide the complementary comparison.

Within the hybrid pipeline, multi-hop queries receive a shorter text-index timeout, so the two backends are not resource-matched in that condition; the crossover result is therefore based on the separate single-source runs. The matched timeout collection (Supplementary Material S7.3) supplies per-channel counts at 60 s and 180 s with complete-pair contrasts and intention-to-evaluate endpoint bounds. The oracle-context score of 0.7920 on design optimization uses reference information unavailable at deployment; both that score and the derived 0.1947 evidence term are reference-informed diagnostics.

External validation is required for the routing and domain conclusions. The abstention rule was specified before its effect was measured, but the category policy was derived from the same means it exploits; a held-out question set is needed to estimate its gain without selection bias. The paired counterfactual also assumes that the fallback behaves as it did in isolation. Every question concerns home furnishing under Chinese standards; testing other product categories and regulatory systems is needed to establish transfer, and a user study is needed to determine whether the screening service changes designers' decisions. Held-out estimation of the routing gain and independent variation in determinacy bear most directly on the claims advanced here.

6. Conclusions

Designers choosing materials at the concept stage need carbon figures that are both computable and traceable to a governing document. This study evaluated when retrieval can supply those figures using a test bed that combines a 45,342-node knowledge graph with a GraphRAG index over 11,087 records. The system was tested on 100 standard-informed furniture questions against four baselines and a routing condition. Retrieval gain was 0.5600 (95% CI [0.4487, 0.6583]) for factual questions (factor lookups, product totals, and material lists), but 0.0160 (95% CI [−0.0986, 0.1420]) for standards interpretation, where the closed-book model was already competitive. This contrast links retrieval value to empirical closed-book answerability and places the stronger data-engineering priority on structuring exact factor fields, unit conversions, and linked inventory data for record-specific queries.

Concurrent fusion recovered only a small fraction of the observed source-selection headroom. The graph and text stores reversed rank across categories, while the concurrent-fusion minus graph difference was 0.0007 (cluster-bootstrap 95% CI [−0.0532, 0.0628]). A cross-fitted per-item oracle exposed headroom of 0.1067. Source selection is therefore the main problem raised by this comparison: empty graph returns exposed a routing signal, while confident errors remained difficult to select. The identifier extractor also missed named sources, supporting separate evaluation of source attribution and correctness.

Selectability is proposed as an extension of selective-prediction principles [89] to heterogeneous knowledge sources, whereas provenance requires a separate evaluation axis in standards-governed use.

The artifact calculates material-stage carbon estimates from author-compiled bills of materials. Using the same mass, composition, and characterization inputs, graph traversal and a separate calculation agreed within 5% for 20 of 24 products. The four discrepancies involved composite-material mappings.

Further tests should vary representational determinacy independently of closed-book answerability and use an authenticated model cutoff, following the predictions in Section 5.2.

Supplementary Materials: The following supporting information can be downloaded at https://www.mdpi.com/article/10.3390/su182110735/s1: Section S1. System and prompt specifications; Section S2. Condensed evaluation rubric; Section S3. Failure mode summary; Section S4. Scoring protocol and judge validation; Section S5. Predictions and decision procedure; Section S6. Additional tables; Section S7. Sensitivity analyses and prospective diagnostics; Section S8. Statistical definitions and clustering sensitivity. Table S1. Knowledge graph statistics; Table S2. Designed source selection policy; Table S3. Deployed graph store returns; Table S4. Category level correctness; Table S5. Judge and human agreement; Table S6. Contribution of abstention routing; Table S7. Performance budget for design optimization; Table S8. SD of run-averaged item correctness by category (25 items per category); Table S9. Ground-truth revision audit; Table S10. Reconciliation of Table 8 depth and graph failure classes; Table S11. Source-verified named processes and factor conversion; Table S12. Frozen target dependencies on the four outlier products; Table S13. Agreement across 50 standards responses; estimates [95% CI]; Table S14. Standards condition means and paired differences; estimates [95% CI]; Table S15. Product mass and carbon under two accounting scenarios; Table S16. Conditional composite inheritance bias at fixed historical row masses; Table S17. Prospective delivery and scoring counts by timeout arm; Table S18. Paired prospective contrasts at 180 s per channel; Table S19. Failures and exact repeated responses by category; Table S20. Multi-hop timeout sensitivity, 180 s minus 60 s per channel; Table S21. Source-content indicators for the two core standards questions.

Author Contributions: Conceptualization, F.Y. and S.S.; Methodology, F.Y. and M.Y.; Software, F.Y.; Validation, F.Y. and M.Y.; Formal analysis, F.Y.; Investigation, F.Y. and M.Y.; Resources, S.S.; Data curation, F.Y.; Writing – original draft, F.Y.; Writing – review & editing, M.Y. and S.S.; Supervision, S.S.; Project administration, F.Y.; Funding acquisition, F.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by Zhejiang Provincial Natural Science Foundation of China under Grant No. LGG22G030003.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data underlying this study are not publicly available. Requests to access the datasets should be directed to corresponding author.

Acknowledgments: The authors acknowledge the ecoinvent Centre, the TianGong LCA Data Platform, publicly available IKEA product documentation, the annotators. During manuscript preparation, OpenAI and Anthropic generative AI tools assisted with language editing, formatting, and document and analysis checking. The authors reviewed and edited the outputs and take full responsibility for the content of this publication.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

BOM	Bill of Materials
CFP	Carbon Footprint of Products
CRediT	Contributor Roles Taxonomy
EPD	Environmental Product Declaration
GWP	Global Warming Potential
KG	Knowledge Graph
LCA	Life Cycle Assessment
LLM	Large Language Model
MAPE	Mean Absolute Percentage Error
PLM	Product Lifecycle Management
RAG	Retrieval-Augmented Generation
RoW	Rest of World (ecoinvent geography code)
SAC	Standardization Administration of China

References

1. Cordella, M.; Hidalgo, C. Analysis of key environmental areas in the design and labelling of furniture products: Application of a screening approach based on a literature review of LCA studies. *Sustain. Prod. Consum.* **2016**, *8*, 64–77. [CrossRef]
2. Wang, S.; Su, D.; Zhu, S. A comparative study on life cycle assessment of typical wood base furniture. In *Proceedings of the 2016 5th International Conference on Sustainable Energy and Environment Engineering (ICSEEE 2016)*; Atlantis Press: Dordrecht, The Netherlands, 2016; Volume 63, pp. 634–640. [CrossRef]
3. Liu, Y.; Wang, Y.; Wang, C.; Zhou, T.; Hu, J.; Wu, Z. Carbon Footprint Accounting and Analysis of Chinese Furniture Enterprises' Panel Cabinets. *Sustainability* **2025**, *17*, 9267. [CrossRef]
4. Chatty, T.; Moeller, B.L.; Pantelimon, I.A.; Parnell, C.D.; Khan, T.M.; Laurin, L.; Faludi, J.; Murnane, E.L. EcoSketch: Promoting Sustainable Design through Iterative Environmental Assessment during Early-Stage Product Development. *ACM J. Comput. Sustain. Soc.* **2024**, *2*, 1–29. [CrossRef] [PubMed]
5. Englhardt, Z.; Hähnlein, F.; Mei, Y.; Lin, T.; Sun, C.M.; Zhang, Z.; Patel, S.; Schulz, A.; Iyer, V. Incorporating Sustainability in Electronics Design: Obstacles and Opportunities. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*; Association for Computing Machinery: New York, NY, USA, 2025; pp. 1–17. [CrossRef]
6. *GB/T 46486-2025*; Greenhouse Gases—Quantification Requirement and Method for Carbon Footprint of Products—Furniture. Standardization Administration of China: Beijing, China, 2025.
7. Chatty, T.; Harrison, W.; Ba-Sabaa, H.H.; Faludi, J.; Murnane, E.L. Co-Creating a Framework to Integrate Sustainable Design into Product Development Practice: Case Study at an Engineering Consultancy Firm. *Sustainability* **2022**, *14*, 9740. [CrossRef]
8. Brundage, M.P.; Bernstein, W.Z.; Hoffenson, S.; Chang, Q.; Nishi, H.; Kliks, T.; Morris, K.C. Analyzing environmental sustainability methods for use earlier in the product lifecycle. *J. Clean. Prod.* **2018**, *187*, 877–892. [CrossRef] [PubMed]
9. Bianco, I.; Thiébat, F.; Carbonaro, C.; Pagliolico, S.; Blengini, G.A.; Comino, E. Life Cycle Assessment (LCA)-based tools for the eco-design of wooden furniture. *J. Clean. Prod.* **2021**, *324*, 129249. [CrossRef]
10. Kiemel, S.; Rietdorf, C.; Schutzbach, M.; Miehe, R. How to Simplify Life Cycle Assessment for Industrial Applications—A Comprehensive Review. *Sustainability* **2022**, *14*, 15704. [CrossRef]
11. *EN 15804:2012+A2:2019*; Sustainability of Construction Works-Environmental Product Declarations-Core Rules for the Product Category of Construction Products. European Committee for Standardization: Brussels, Belgium, 2019.
12. Bhander, G.S.; Hauschild, M.; McAloone, T. Implementing life cycle assessment in product development. *Environ. Prog.* **2003**, *22*, 255–267. [CrossRef]
13. Pollini, B.; Rognoli, V. Early-stage material selection based on life cycle approach: Tools, obstacles and opportunities for design. *Sustain. Prod. Consum.* **2021**, *28*, 1130–1139. [CrossRef]
14. Olivetti, E.; Patanavanich, S.; Kirchain, R. Exploring the viability of probabilistic under-specification to streamline life cycle assessment. *Environ. Sci. Technol.* **2013**, *47*, 5208–5216. [CrossRef] [PubMed]
15. Chatty, T.; Qu, Y.; Ba-Sabaa, H.H.; Murnane, E.L. Examining the user experience of life cycle assessment tools and their ability to cater to ecodesign in early-stage product development practice. In *Proceedings of the Design Society*; Cambridge University Press: Cambridge, UK, 2021; Volume 1, pp. 1441–1450. [CrossRef]
16. Wenker, J.L.; Richter, K.; Rüter, S. A Methodical Approach for Systematic Life Cycle Assessment of Wood-Based Furniture. *J. Ind. Ecol.* **2018**, *22*, 671–685. [CrossRef]
17. Roberts, A.; Raffel, C.; Shazeer, N. 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 (EMNLP 2020), Online, 16–20 November 2020*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2020; pp. 5418–5426. [CrossRef]
18. Mallen, A.; Asai, A.; Zhong, V.; Das, R.; Khashabi, D.; Hajishirzi, H. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL 2023), Toronto, ON, Canada, 9–14 July 2023*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2023; pp. 9802–9822. [CrossRef]
19. Zhao, Y.; Dai, C.; Niyato, D.; Tan, C.F.; Xiang, K.; Wang, Y.; Yeo, Z.; Loong, D.T.Z.; Zhaozhi, J.L.; Ho, E.H.Z. A Graph-Retrieval-Augmented Generation Framework Enhances Decision-Making in the Circular Economy. *arXiv* **2025**, arXiv:2506.04252. [CrossRef]
20. Janowicz, K.; Krisnadhi, A.A.; Hu, Y.; Suh, S.; Weidema, B.P.; Rivela, B.; Tivander, J.; Meyer, D.E.; Berg-Cross, G.; Hitzler, P.; et al. A Minimal Ontology Pattern for Life Cycle Assessment Data. In Proceedings of the 6th Workshop on Ontology and Semantic Web Patterns, Bethlehem, PA, USA, 11 October 2015; Volume 1461. Available online: https://ceur-ws.org/Vol-1461/WOP2015_pattern_abstract_1.pdf (accessed on 10 August 2026).
21. Ghose, A.; Lissandrini, M.; Hansen, E.R.; Weidema, B.P. A core ontology for modeling life cycle sustainability assessment on the Semantic Web. *J. Ind. Ecol.* **2022**, *26*, 731–747. [CrossRef]
22. Edelen, A.N.; Cashman, S.; Young, B.; Ingwersen, W.W. Life Cycle Data Interoperability Improvements through Implementation of the Federal LCA Commons Elementary Flow List. *Appl. Sci.* **2022**, *12*, 9687. [CrossRef] [PubMed]

23. Valente, A.; Vadenbo, C.; Fazio, S.; Shobatake, K.; Edelen, A.N.; Sonderegger, T.; Karkour, S.; Kusche, O.; Diaconu, E.; Ingwersen, W.W. Elementary flow mapping across life cycle inventory data systems: A case study for data interoperability under the Global Life Cycle Assessment Data Access (GLAD) initiative. *Int. J. Life Cycle Assess.* **2024**, *29*, 789–802. [CrossRef] [PubMed]
24. Zhang, Y.; Luo, X.; Buis, J.J.; Sutherland, J.W. LCA-oriented semantic representation for the product life cycle. *J. Clean. Prod.* **2015**, *86*, 146–162. [CrossRef]
25. Pauliuk, S.; Heeren, N.; Hasan, M.M.; Müller, D.B. A general data model for socioeconomic metabolism and its implementation in an industrial ecology data commons prototype. *J. Ind. Ecol.* **2019**, *23*, 1016–1027. [CrossRef]
26. Mandolini, M.; Marconi, M.; Rossi, M.; Favi, C.; Germani, M. A standard data model for life cycle analysis of industrial products: A support for eco-design initiatives. *Comput. Ind.* **2019**, *109*, 31–44. [CrossRef]
27. Otte, J.N.; Kiritsi, D.; Ali, M.M.; Yang, R.; Zhang, B.; Rudnicki, R.; Rai, R.; Smith, B. An ontological approach to representing the product life cycle. *Appl. Ontol.* **2019**, *14*, 179–197. [CrossRef]
28. Bernstein, W.Z.; Tamayo, C.D.; Lechevalier, D.; Brundage, M.P. Incorporating unit manufacturing process models into life cycle assessment workflows. *Procedia CIRP* **2019**, *80*, 364–369. [CrossRef]
29. Tschiltschke, T.; Wehking, S.; Riedelsheimer, T.; Lindow, K. Automated LCA through semantic integration of design and process data. *Procedia CIRP* **2025**, *136*, 407–412. [CrossRef]
30. Sajjad, M.; van den Boogaart, K.G.; Steinmeier, L.; Parvez, A.M. Environmental Impact Assessment of Copper-Alloy Production Using Process Simulation and Semantic Modeling. *Adv. Eng. Mater.* **2025**, *27*, 2401702. [CrossRef]
31. Saad, M.; Zhang, Y.; Tian, J.; Jia, J. A graph database for life cycle inventory using Neo4j. *J. Clean. Prod.* **2023**, *393*, 136344. [CrossRef]
32. Peng, T.; Gao, L.; Agbozo, R.S.K.; Xu, Y.; Svynarenko, K.; Wu, Q.; Li, C.; Tang, R. Knowledge graph-based mapping and recommendation to automate life cycle assessment. *Adv. Eng. Inform.* **2024**, *62*, 102752. [CrossRef]
33. Castle, S.; Moreno Schneider, J.; Hennig, L.; Rehm, G. Entity Linking Using LLMs for Automated Product Carbon Footprint Estimation. In *Proceedings of the 1st Workshop on Ecology, Environment, and Natural Language Processing (NLP4Ecology 2025), Tallinn, Estonia, 2 March 2025*; University of Tartu Library: Tartu, Estonia, 2025; pp. 56–60. Available online: https://aclanthology.org/2025.nlp4ecology-1.12/ (accessed on 10 August 2026).
34. Deng, Z.; Liu, J.; Luo, B.; Yuan, C.; Yang, Q.; Xiao, L.; Zhou, W.; Liu, Z. AutoPCF: Efficient Product Carbon Footprint Accounting with Large Language Models. *arXiv* **2023**, arXiv:2308.04241. [CrossRef]
35. Edge, D.; Trinh, H.; Cheng, N.; Bradley, J.; Chao, A.; Mody, A.; Truitt, S.; Larson, J. From Local to Global: A Graph RAG Approach to Query-Focused Summarization. *arXiv* **2024**, arXiv:2404.16130v1.
36. Zhu, Z.; Huang, T.; Wang, K.; Ye, J.; Chen, X.; Luo, S. Graph-Based Approaches and Functionalities in Retrieval-Augmented Generation: A Comprehensive Survey. *arXiv* **2026**, arXiv:2504.10499v2. [CrossRef]
37. Masoud, A.A.; Arazzi, M.; Germani, S.; Nocera, A. Exploring Structural Complexity in Normative RAG with Graph-based approaches: A case study on the ETSI Standards. *arXiv* **2026**, arXiv:2604.09868. [CrossRef]
38. Wallat, J.; Heuss, M.; de Rijke, M.; Anand, A. Correctness Is Not Faithfulness in Retrieval Augmented Generation Attributions. In *Proceedings of the 2025 International ACM SIGIR Conference on Innovative Concepts and Theories in Information Retrieval (ICTIR 2025), Padua, Italy, 18 July 2025*; Association for Computing Machinery: New York, NY, USA, 2025; pp. 22–32. [CrossRef]
39. Zhu, X.; Xie, Y.; Liu, Y.; Li, Y.; Hu, W. Knowledge Graph-Guided Retrieval Augmented Generation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies*; Association for Computational Linguistics: Albuquerque, NM, USA, 2025; Volume 1, pp. 8912–8924. [CrossRef]
40. Lee, M.-C.; Zhu, Q.; Mavromatis, C.; Han, Z.; Adeshina, S.; Ioannidis, V.N.; Rangwala, H.; Faloutsos, C. HybGRAG: Hybrid Retrieval-Augmented Generation on Textual and Relational Knowledge Bases. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*; Association for Computational Linguistics: Vienna, Austria, 2025; Volume 1, pp. 879–893. [CrossRef]
41. Krainovskikh, V.; Samigulin, T. Domain-Adapted Retrieval-Augmented Generation for Technical Documentation: Enhancing Reliability and Faithfulness in Technical QA. *IEEE Access* **2026**, *14*, 36016–36024. [CrossRef]
42. Ali, Ș.; Oprea, S.-V.; Bâra, A. Engineering Trustworthy Retrieval-Augmented Generation for EU Electricity Market Regulation. *Electronics* **2026**, *15*, 749. [CrossRef]
43. Bahr, L.; Wehner, C.; Wewerka, J.; Bittencourt, J.; Schmid, U.; Daub, R. Knowledge Graph Enhanced Retrieval-Augmented Generation for Failure Mode and Effects Analysis. *arXiv* **2025**, arXiv:2406.18114v3. [CrossRef]
44. Hu, Y.; Lei, Z.; Zhang, Z.; Pan, B.; Ling, C.; Zhao, L. GRAG: Graph Retrieval-Augmented Generation. In *Findings of the Association for Computational Linguistics: NAACL 2025*; Association for Computational Linguistics: Albuquerque, NM, USA, 2025; pp. 4145–4157. [CrossRef]
45. Chae, K.; Yeom, J.; Park, J.; Bae, S.; Jang, I.; Jin, H.; Jang, J.; Kim, T. Beyond Case Law: Evaluating Structure-Aware Retrieval and Safety in Statute-Centric Legal QA. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics*; Association for Computational Linguistics: San Diego, CA, USA, 2026; Volume 1, pp. 45553–45573. [CrossRef]

46. Rasyid, S.A.; Wiranata, A.; Usman, K.; Wibowo, S.A.; Hantoro, G.D.; Syamsul, M.; Prihatini, Y.I. Evidence-Grounded Hybrid Framework for ISO/IEC 17025 Laboratory Operations Audits: A Neuro-Symbolic Decision-Support Approach. *Logist. Oper. Manag. Res.* **2026**, *5*, 36–50. [CrossRef]
47. Mavromatis, C.; Karypis, G. GNN-RAG: Graph Neural Retrieval for Efficient Large Language Model Reasoning on Knowledge Graphs. In *Findings of the Association for Computational Linguistics: ACL 2025*; Association for Computational Linguistics: Vienna, Austria, 2025; pp. 16682–16699. [CrossRef]
48. Li, J.; Zhou, A. Multi-Regulation RAG for AI Product Counsel: A Legal Governance Framework for Cross-Border Digital Commerces. *Rule Law Stud. J.* **2026**, *2*, 106–123. [CrossRef]
49. Joren, H.; Zhang, J.; Ferng, C.-S.; Juan, D.-C.; Taly, A.; Rashtchian, C. Sufficient Context: A New Lens on Retrieval Augmented Generation Systems. In Proceedings of the Thirteenth International Conference on Learning Representations (ICLR 2025), Singapore, 24–28 April 2025. Available online: https://openreview.net/forum?id=Jjr2Odj8DJ (accessed on 10 August 2026).
50. Elchafei, P.; Swain, M.; Masoudian, S.; Schedl, M. Facet-Level Tracing of Evidence Uncertainty and Hallucination in RAG. *arXiv* **2026**, arXiv:2604.09174. [CrossRef]
51. Kassimi, Z.E.; Fourati, F.; Alouini, M.-S. Retrieval-Augmented Generation for Reliable Interpretation of Radio Regulations. *arXiv* **2025**, arXiv:2509.09651. [CrossRef]
52. Kim, H.; Sohn, J.; Gilson, A.; Cochran-Caggiano, N.; Applebaum, S.; Jin, H.; Park, S.; Park, Y.; Park, J.; Choi, S.; et al. Rethinking Retrieval-Augmented Generation for Medicine: A Large-Scale, Systematic Expert Evaluation and Practical Insights. *arXiv* **2025**, arXiv:2511.06738. [CrossRef]
53. Chen, Z.; Zhou, W. RegCheck-Hybrid: A Hybrid Rule-LLM Framework for Automated Multi-modal Fire Code Compliance Checking. In Proceedings of the 2025 International Conference on Mechatronic Engineering and Artificial Intelligence (MEAI), Shenyang, China, 5–7 December 2025; pp. 104–108. [CrossRef]
54. Liang, C.; Huang, Z.; Wang, H.; Chai, F.; Yu, C.; Wei, H.; Liu, Z.; Li, Y.; Wang, H.; Luo, R.; et al. AECBench: A Hierarchical Benchmark for Knowledge Evaluation of Large Language Models in the AEC Field. *Adv. Eng. Inform.* **2026**, *71*, 104314. [CrossRef]
55. Afane, M.; Hariri, E.; Ouyang, D.; Ho, D.E. Benchmarking Legal RAG: The Promise and Limits of AI Statutory Surveys. In Proceedings of the Symposium on Computer Science and Law, Berkeley, CA, USA, 3–5 March 2026; pp. 101–125. [CrossRef]
56. Murugaraj, K.; Lamsiyah, S.; Theobald, M. RAGVUE: A Diagnostic View for Explainable and Automated Evaluation of Retrieval-Augmented Generation. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics, Rabat, Morocco, March 2026*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2026; Volume 3, pp. 512–526. [CrossRef]
57. Kwiatkowski, T.; Palomaki, J.; Redfield, O.; Collins, M.; Parikh, A.; Alberti, C.; Epstein, D.; Polosukhin, I.; Devlin, J.; Lee, K.; et al. Natural questions: A benchmark for question answering research. *Trans. Assoc. Comput. Linguist.* **2019**, *7*, 453–466. [CrossRef]
58. Madhyastha, P. Task-aware evaluation and error-overlap analysis for large language models. In *Proceedings of the 1st Workshop on Confabulation, Hallucinations and Overgeneration in Multilingual and Practical Settings (CHOMPS 2025), Mumbai, India, 23 December 2025*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2025; Volume 2025, pp. 1–10. [CrossRef]
59. Chen, J. When does combining language models help? A co-failure ceiling on routing, voting, and mixture-of-agents across 67 frontier models. *arXiv* **2026**, arXiv:2606.27288. [CrossRef]
60. Kuncheva, L.I.; Whitaker, C.J. Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Mach. Learn.* **2003**, *51*, 181–207. [CrossRef]
61. Han, H.; Wang, Y.; Shomer, H.; Lei, Y.; Guo, K.; Hua, Z.; Long, B.; Liu, H.; Tang, J. RAG vs. GraphRAG: A Systematic Evaluation and Key Insights. *arXiv* **2025**, arXiv:2502.11371v1.
62. Xiao, Y.; Dong, J.; Zhou, C.; Dong, S.; Zhang, Q.-W.; Yin, D.; Sun, X.; Huang, X. GraphRAG-Bench: Challenging Domain-Specific Reasoning for Evaluating Graph Retrieval-Augmented Generation. *arXiv* **2025**, arXiv:2506.02404. [CrossRef]
63. Yuan, X.; Liu, Y.; Di, S.; Wu, S.; Zheng, L.; Meng, R.; Chen, L.; Zhou, X.; Yin, J. A Pilot Empirical Study on When and How to Use Knowledge Graphs as Retrieval Augmented Generation. *arXiv* **2025**, arXiv:2502.20854v3. [CrossRef]
64. Xiang, Z.; Wu, C.; Zhang, Q.; Chen, S.; Hong, Z.; Huang, X.; Su, J. When to Use Graphs in RAG: A Comprehensive Analysis for Graph Retrieval-Augmented Generation. *arXiv* **2026**, arXiv:2506.05690v3. [CrossRef]
65. Oakden-Rayner, L.; Dunnmon, J.; Carneiro, G.; Ré, C. Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. In *Proceedings of the ACM Conference on Health, Inference, and Learning (CHIL 2020), Toronto, ON, Canada, 2–4 April 2020*; Association for Computing Machinery: New York, NY, USA, 2020; pp. 151–159. [CrossRef] [PubMed]
66. Suresh, H.; Guttag, J. A framework for understanding sources of harm throughout the machine learning life cycle. In *Proceedings of the 1st ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO 2021), Virtual Event, 5–9 October 2021*; Association for Computing Machinery: New York, NY, USA, 2021; pp. 1–9. [CrossRef]
67. Akhtar, M.; Reuel, A.; Soni, P.; Ahuja, S.; Ammanamanchi, P.S.; Rawal, R.; Zouhar, V.; Yadav, S.; Whitehouse, C.; Ki, D.; et al. When AI benchmarks plateau: A systematic study of benchmark saturation. *arXiv* **2026**, arXiv:2602.16763.

68. Voorhees, E.M.; Buckley, C. The effect of topic set size on retrieval experiment error. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2002), Tampere, Finland, 11–15 August 2002*; Association for Computing Machinery: New York, NY, USA, 2002; pp. 316–323. [CrossRef]
69. Wernet, G.; Bauer, C.; Steubing, B.; Reinhard, J.; Moreno-Ruiz, E.; Weidema, B. The ecoinvent database version 3 (part I): Overview and methodology. *Int. J. Life Cycle Assess.* **2016**, *21*, 1218–1230. [CrossRef]
70. TianGong, L.C.A. TianGong LCA Data Platform. Available online: https://www.tiangong.earth (accessed on 25 September 2026).
71. Forster, P.; Storelvmo, T.; Armour, K.; Collins, W.; Dufresne, J.-L.; Frame, D.; Lunt, D.J.; Mauritsen, T.; Palmer, M.D.; Watanabe, M.; et al. The Earth's Energy Budget, Climate Feedbacks, and Climate Sensitivity. In *Climate Change 2021: The Physical Science Basis*; Cambridge University Press: Cambridge, UK, 2021; pp. 923–1054. [CrossRef]
72. *ISO 14067:2018*; Greenhouse Gases—Carbon Footprint of Products—Requirements and Guidelines for Quantification. International Organization for Standardization: Geneva, Switzerland, 2018.
73. World Resources Institute; World Business Council for Sustainable Development. *Product Life Cycle Accounting and Reporting Standard*; World Resources Institute: Washington, DC, USA; World Business Council for Sustainable Development: Geneva, Switzerland, 2011.
74. *GB/T 24067-2024*; Greenhouse Gases—Carbon Footprint of Products—Requirements and Guidelines for Quantification. State Administration for Market Regulation; Standardization Administration of China: Beijing, China, 2024.
75. *PAS 2050:2008*; Specification for the Assessment of the Life Cycle Greenhouse Gas Emissions of Goods and Services. British Standards Institution: London, UK, 2008.
76. Intergovernmental Panel on Climate Change. *Climate Change 2022: Mitigation of Climate Change. Contribution of Working Group III to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*; Cambridge University Press: Cambridge, UK, 2022. [CrossRef]
77. Traag, V.A.; Waltman, L.; van Eck, N.J. From Louvain to Leiden: Guaranteeing well-connected communities. *Sci. Rep.* **2019**, *9*, 5233. [CrossRef] [PubMed]
78. Friedman, M. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *J. Am. Stat. Assoc.* **1937**, *32*, 675–701. [CrossRef]
79. Wilcoxon, F. Individual comparisons by ranking methods. *Biom. Bull.* **1945**, *1*, 80–83. [CrossRef]
80. Holm, S. A simple sequentially rejective multiple test procedure. *Scand. J. Stat.* **1979**, *6*, 65–70. Available online: https://www.jstor.org/stable/4615733 (accessed on 31 July 2026).
81. Cureton, E.E. Rank-biserial correlation. *Psychometrika* **1956**, *21*, 287–290. [CrossRef]
82. Kerby, D.S. The simple difference formula: An approach to teaching nonparametric correlation. *Compr. Psychol.* **2014**, *3*, 1. [CrossRef]
83. Efron, B. Bootstrap methods: Another look at the jackknife. *Ann. Stat.* **1979**, *7*, 1–26. [CrossRef]
84. Intergovernmental Panel on Climate Change. *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*; Cambridge University Press: Cambridge, UK, 2013. [CrossRef]
85. *ISO 14068-1:2023*; Climate Change Management—Transition to Net Zero—Part 1: Carbon Neutrality. International Organization for Standardization: Geneva, Switzerland, 2023.
86. British Standards Institution. *ISO 14068-1:2023: Climate Change Management—Transition to Net Zero—Part 1: Carbon Neutrality: Executive Briefing*; British Standards Institution: London, UK, 2024.
87. Goridkov, N.; Wang, Y.; Goucher-Lambert, K. Empowering designers to create life cycle informed products: Heuristics for extracting insights from LCA reports. *Des. Sci.* **2025**, *11*, e27. [CrossRef]
88. Buckland, M.; Gey, F. The relationship between recall and precision. *J. Am. Soc. Inf. Sci.* **1994**, *45*, 12–19. [CrossRef]
89. El-Yaniv, R.; Wiener, Y. On the foundations of noise-free selective classification. *J. Mach. Learn. Res.* **2010**, *11*, 1605–1641. Available online: https://jmlr.org/papers/v11/el-yaniv10a.html (accessed on 28 July 2026).

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.