

# Ontology-Guided Query Expansion for Evidence Retrieval in ESRS E1 Climate Disclosures

Tatsuya Ishisaka

ASUENE Inc.

ishisaka.tatsuya@asuene.com

## Abstract

Auditable analysis of corporate climate disclosures requires *retrieving*, for each regulatory disclosure requirement, the report page or section that contains the evidence. We study **ontology-aware query expansion** (OAE) for **ESRS E1** (Climate Change) requirements over a new page/section-level benchmark: **five companies**, 1,177 real-artifact chunks, and 114 human-reviewed gold labels, plus a held-out ESRS-native company (64 labels) unseen during ontology construction. Against lexical, dense, hybrid, and pseudo-relevance-feedback (PRF) baselines, a lightweight ontology built from the standard’s requirements improves retrieval: the best configuration reaches Recall@10 **0.842** vs. BM25’s 0.702 (McNemar  $p=.003$ ). Generic PRF expansion *hurts* retrieval while ontology-guided expansion helps; a corpus-blind LLM query expander is a far stronger control and ties with the ontology in-domain, but does not transfer to the held-out company, where OAE stays ahead of lexical, dense, and LLM-expansion baselines while being statistically indistinguishable from cross-encoder reranking — a comparison we treat as *diagnostic* rather than conclusive. We provide an error analysis of the residual misses. “Gold” denotes retrieval ground truth only; we make no company scoring or compliance judgement.

## 1 Introduction

Under the EU Corporate Sustainability Reporting Directive, companies report against the European Sustainability Reporting Standards; ESRS E1 (Climate Change) defines concrete disclosure requirements — transition plans, climate targets with base and target years, energy consumption and renewable shares, and gross GHG emissions by scope (European Financial Reporting Advisory Group (EFRAG) and European Commission, 2023). ESRS E1 is, in effect, a checklist of disclosure requirements a company must address.

Company sustainability reports, however, are free-form PDFs of 100–200 pages, and their layout, structure, and level of detail vary by company. This holds even for ESRS-native reports, since the standard does not prescribe where each item must appear or under which heading, so where a given checklist item is stated still varies by company. In addition, the regulatory wording of the checklist and the wording a company actually uses in its report often do not match, so using the regulatory wording directly as a search query does not reliably find the relevant passage. The problem therefore combines two factors: free-form layout and vocabulary mismatch.

For analysts, auditors, and regulators, accurately and quickly locating the passage that addresses each checklist item is essential. Constructing such a *disclosure matrix* (requirement  $\times$  company, populated with evidence pointers) thus begins with a retrieval problem: rank a company’s report pages/sections so the evidence surfaces at the top for human verification.

This vocabulary and structural mismatch makes the retrieval problem harder than generic ad-hoc search: sustainability reports are saturated with climate-adjacent boilerplate, which makes topically “close” pages adversarial distractors. Classical pseudo-relevance feedback, which expands the query with terms from top-ranked documents (Lavrenko and Croft, 2001; Carpineto and Romano, 2012), is prone to *query drift* in exactly this setting. We instead expand queries from a *disclosure ontology*: requirement labels, curated evidence-query phrases, field names, synonyms, units, and cross-standard mappings.

**RQ1.** *Does ontology-aware query expansion improve evidence retrieval for disclosure requirements? Is the benefit attributable to the ontology’s curated content, rather than to query expansion in general, to score fusion, or to what a general-*

*purpose LLM would propose unprompted by the standard? And how does query-side expansion compare with rerank-stage alternatives?*

**Contributions.** (1) A **human-reviewed page/section-level ESRS E1 evidence-retrieval benchmark**: 5 companies, 12 requirements, 114 gold labels over 1,177 real-artifact chunks, each subset checked by an independent blinded verification pass ( $\kappa=0.67$  and  $0.94$ ) and protected by an evaluation-contamination guard, plus a fair-pooled **held-out ESRS-native company** (64 labels). (2) **Ontology-aware query expansion** from a lightweight, auditable regulation ontology, with LITE/FIELD/FULL field tiers evaluated across three independent scorers. (3) A **controlled decomposition** of where the gain comes from — matched fusion, PRF, and LLM-expansion controls, held-out transfer, and paired significance tests — showing that curated ontology content, not expansion per se, is the operative factor, and identifying where page-level reranking and unconstrained LLM expansion fall short. We scope this paper to retrieval; extraction-stage validation is future work so that retrieval is established as stable first.

## 2 Related Work

**Climate disclosure retrieval.** We build on a recent line studying retrieval over corporate climate disclosures and expert-informed query reformulation, but focus on ESRS E1 requirement-grounded evidence retrieval and isolate the effect of curated ontology content through matched controls. Schimanski et al. (2024) release CLIMRETRIEVE, an expert-annotated IR benchmark over 30 sustainability reports with 16 analyst questions and graded relevance on  $\sim 350$ -word paragraphs, and show that progressively replacing the raw question with longer expert-informed explanations improves embedding retrieval. Ahmed et al. (2025) build ESG-CID by mining companies’ own disclosure content indexes as weak supervision for requirement-to-section retrieval across GRI and ESRS, with an ESRS evaluation split, and show fine-tuned BERT encoders can beat commercial embeddings. Neither displaces the present contribution: relative to CLIMRETRIEVE, our queries are ESRS E1 *requirements* rather than analyst questions and our expansion content is authored from the standard rather than distilled from labelled passages, tested with a budget-matched

PRF control, an LLM-expansion control, replication across three scorers, and a held-out ESRS-native company; relative to ESG-CID, our E1 labels are human-reviewed and adjudicated rather than content-index weak supervision, and we isolate *which* ontology-derived query fields cause the gain rather than fine-tuning an encoder.

**Query expansion and retrieval baselines.** Automatic query expansion spans pseudo-relevance feedback (Lavrenko and Croft, 2001; Carpineto and Romano, 2012) and ontology/knowledge-base methods (Bhagal et al., 2007); we compare both, plus an LLM expander, under an identical scorer. We situate OAE against BM25 (Robertson and Zaragoza, 2009), dense sentence encoders (Reimers and Gurevych, 2019; Karpukhin et al., 2020), rank fusion (Cormack et al., 2009), and BERT cross-encoder reranking (Nogueira and Cho, 2019); full configurations are given in §4.

## 3 Task and Data

**Task.** For a disclosure requirement  $r$  and a company’s reporting corpus, order chunks so the gold evidence chunk(s) rank highly. The 12 ESRS E1 evaluation requirements span transition plan (2), targets (4), energy (2), GHG emissions (3), and gross/removals separation (1).

**Corpus (real artifacts).** Official reporting artifacts from five companies, parsed to **1,177 chunks** (per-company breakdown in Appendix D): PDF pages and HTML sections spanning US and EU/CSRD reporters (one report is ESRS-native), so requirement phrasing is evaluated against both native and non-native disclosures.

**Gold annotation and verification.** All gold is **human-reviewed** over ranked candidate chunks per (requirement, document), yielding **114 labels** (92 direct, 22 partial) covering all 12 requirements and all 5 companies. Reliability was measured by *independent blinded* verification passes (given only the requirement, its codebook notes, and the chunk text; no access to the gold). On Companies B–D the blinded pass agreed at 84.3% (Cohen’s  $\kappa=0.67$ ); all 13 disagreements were adjudicated with document-level verification (5 codebook-violating labels removed, 7 best-available-evidence labels relabelled *partial*). On the two EU reporters, where the human reviewed AI-suggested candidates, a same-protocol pass agreed at 97.4% ( $\kappa=0.94$ ). We treat  $\kappa=0.67$  as

the headline reliability, since the 0.94 figure is likely optimistic (AI-assisted review, and the verifier shares the model family that produced the candidates). A validation guard rejects labels referencing nonexistent chunks, low-confidence silver labels, local-snapshot artifacts, or document-level smoke results. Our **annotation protocol** defines **“gold” as retrieval ground truth only**: the labels record where document-grounded evidence is located, and support no claim about company quality, disclosure adequacy, or compliance.

**Held-out company.** To check generalisation beyond the five development companies, we add **Company F** (2025 Sustainability Statement, 158 pages, ESRS-native), used in no ontology authoring, tuning, or pooling decision. Candidates were pooled as the union of every compared method’s top- $k$  (BM25, Hyb-SBERT, OAE-lite, CE-rerank), blind-labelled, and re-checked for completeness as each method was added — yielding **64 gold labels** (12/12 requirements) with a pool neutral to any one system’s query vocabulary.

## 4 Methods

All methods rank the same 1,177 chunks (restricted to the label’s company at evaluation time) over the concatenation of section title and chunk text.

**Baselines.** BM25: Okapi BM25 ( $k_1=1.2$ ,  $b=0.75$ ) on the raw `requirement_text`. LSA: TF-IDF (uni+bigrams, 8k features) reduced by truncated SVD to 64 dimensions, cosine. SBERT: all-MiniLM-L6-v2 sentence embeddings (Reimers and Gurevych, 2019), cosine. BGE: bge-small-en-v1.5 retrieval embeddings with its query prefix. HYBRID: equal-weight sum of min-max-normalised BM25 and dense scores (HYB-LSA, HYB-SBERT, HYB-BGE). The three standalone dense scorers all underperform their hybrids and are omitted from Table 1 for space. PRF-RM3: relevance-model expansion (Lavrenko and Croft, 2001) — top-20 terms from the top-10 BM25 chunks (weighted by normalised score  $\times$  relative  $\text{tf} \times \text{IDF}$ ), scored by the same HYB-LSA scorer as the OAE variants. LLM-QE: an LLM query expander, the natural low-cost competitor to a hand-curated ontology. A fixed prompt (Appendix C) shows a current-generation instruction-tuned LLM *only* the standard name and the `requirement_text` — no corpus, no gold

labels, no ontology — and asks for at most 20 expansion terms: exactly PRF-RM3’s budget, inside the 13–24 terms OAE-lite adds. The generated expansions were frozen before retrieval evaluation, making all reported runs deterministic. Unlike an LLM *reranker*, an expander never sees the corpus and so does not inherit the circularity risk of our partly AI-assisted labels. We score it with HYB-LSA and HYB-SBERT, matching the two OAE-lite configurations. CE-RERANK: a BERT cross-encoder (Nogueira and Cho, 2019) (ms-marco-MiniLM-L-6-v2) reranks the company-restricted top-50 of a seed ranking (CE-RERANK(BM25) or CE-RERANK(HYB-SBERT)) by scoring (`requirement_text`, page text, truncated to 2k chars) pairs; chunks beyond rank 50 keep their seed order. Base retrievers already reach Recall@50 of .99–1.00, so the reranker has near-full candidate coverage.

**Ontology construction.** We designed the ontology as a lightweight, auditable retrieval artifact rather than a full OWL/SHACL implementation. Each ESRS E1 requirement node stores a label, evidence-query phrases, fields, units/values, and validation notes, from which the LITE/FIELD/FULL query variants are generated. Expansion terms are authored from the standard and general domain vocabulary, never from gold pages; Appendix B details the workflow and leakage guardrails.

**Ontology-aware expansion (OAE).** The query is the requirement text plus ontology fields: OAE-LITE adds label + curated evidence-query phrases; OAE-FIELD adds required field names; OAE-FULL adds synonyms, units/values, and cross-standard mappings. Expanded queries are scored by HYB-LSA; to test scorer-independence we also score the lite expansion with HYB-SBERT and HYB-BGE.

**Metrics and protocol.** Recall@5, Recall@10, MRR, and page-level accuracy; primary metric Recall@10 (and MRR for expansion contrasts). Significance: exact McNemar on paired hit@ $k$  and sign-permutation ( $10^5$  resamples) on reciprocal ranks; Holm correction within the expansion family; fixed seeds. Only the human+LLM `manual_page_gold` set is used; document-level smoke and snapshot-based results are excluded by a guard.

| method                               | R@5          | R@10         | MRR          | p.acc        |
|--------------------------------------|--------------|--------------|--------------|--------------|
| <i>Baselines</i>                     |              |              |              |              |
| BM25                                 | 0.456        | 0.702        | 0.305        | 0.140        |
| Hyb-LSA                              | 0.465        | 0.693        | 0.296        | 0.132        |
| Hyb-SBERT                            | 0.491        | 0.693        | 0.297        | 0.123        |
| Hyb-BGE                              | 0.500        | 0.719        | 0.287        | 0.114        |
| <i>Generic expansion / reranking</i> |              |              |              |              |
| PRF-RM3                              | 0.518        | 0.632        | 0.305        | 0.140        |
| LLM-QE                               | <b>0.614</b> | 0.807        | <b>0.365</b> | <b>0.184</b> |
| LLM-QE <sub>SBERT</sub>              | 0.596        | 0.798        | 0.345        | 0.167        |
| CE-rerank(BM25)                      | 0.430        | 0.693        | 0.276        | 0.123        |
| CE-rerank(Hyb-SBERT)                 | 0.412        | 0.658        | 0.277        | 0.132        |
| <i>Ontology expansion</i>            |              |              |              |              |
| OAE-lite                             | <b>0.614</b> | 0.807        | 0.358        | 0.149        |
| OAE-lite <sub>BGE</sub>              | <b>0.614</b> | 0.816        | 0.353        | 0.149        |
| OAE-lite <sub>SBERT</sub>            | 0.605        | <b>0.842</b> | 0.353        | 0.149        |

Table 1: Retrieval on manual\_page\_gold ( $n=114$ , 5 companies, human-reviewed). Best per column in bold (ties bolded); additional variants are omitted for space (§4). In-domain, LLM-QE is indistinguishable from OAE-lite (Appendix Table 3); the two separate on the held-out company (Appendix Table 2).

## 5 Results

**Main result.** The best system, OAE-LITE<sub>SBERT</sub>, reaches Recall@10 **0.842** vs. 0.702 for BM25 (McNemar  $p=.0025$ ; R@5  $p=.008$ ); full pairwise  $p$ -values are in Appendix A.

**Where does the gain come from?** Ontology expansion improves over every baseline on all three scorers, while PRF degrades its own scorer. (i) *Not fusion:* Hyb-LSA over BM25 is not significant ( $p \geq .65$ ). (ii) *The expansion itself:* lite expansion is significant on MRR for all three scorers (LSA  $p=.014$ , SBERT  $p=.042$ , BGE  $p=.0015$ , Holm-corrected) and on SBERT Recall@10 ( $p=.0002$ , 19 discordant pairs vs. 2). (iii) *Ontology content, not expansion per se:* PRF-RM3 drops Recall@10 to 0.632, below its own unexpanded scorer, and loses to OAE-lite ( $p=.0002$ ). (iv) *But a corpus-blind LLM expander is a far stronger control:* it clears PRF-RM3 and BM25 (R@10  $p<.0001/.029$ ) and ties the ontology in-domain ( $p=1.0/.69$ ), taking the best MRR/page-accuracy in Table 1 — substitutable in-domain, not out of domain (below).

**Robustness and ablation.** The effect holds regardless of company set (US  $n=52$   $p=.002$ ; EU/CSRD  $n=62$  ontology-vs-PRF  $p=.001$ ); across OAE tiers, lite captures the benefit — field/full do not beat it ( $p \geq .29$ ) and cost more to maintain.

**Cross-encoder reranking (diagnostic).** We treat this comparison as diagnostic rather than conclusive: both CE-RERANK variants (Table 1) degrade their own seed rankings — misapplied rather than outmatched, likely because the tested model is trained for short passages and truncates long report pages. A page-length-matched reranker is the appropriate comparison and remains future work.

**Zero-shot generalisation (held-out company).** On the held-out ESRS-native Company F, OAE-lite continues to transfer: it improves over BM25 (R@10 .812 vs. .594,  $p=.009$ ) and separates from LLM-QE (R@10 .812 vs. .656,  $p=.041$ ), although the latter comparison is suggestive at  $n=64$  and uncorrected. The cross-encoder closes the in-domain gap (.734/.750 R@10), and OAE-vs-CE is not significant ( $p \geq .75$ ). We therefore claim zero-shot generalisation against lexical, dense, PRF, and LLM-expansion baselines, not superiority over a modern reranker; the full held-out table is in Appendix Table 2.

## 6 Error Analysis

The remaining errors are concentrated in target-family requirements: 13 of the 18 missed gold labels @10 involve target statements, target years, or base years. Most are ranking inversions among topically correct pages in target-dense reports, where reporting-period and target/base years compete. Appendix E shows which expansions help and hurt: OAE helps when regulatory terms and report headings diverge (gross/removals separation, target year), but can over-focus when multiple pages are valid evidence for one requirement.

## 7 Conclusion

We introduced an ESRS E1 page/section-level evidence-retrieval benchmark and showed that ontology-aware query expansion improves requirement-grounded retrieval beyond BM25, hybrid retrieval, and generic PRF, with the effect replicated across scorers and company subsets. A corpus-blind LLM expander is a strong in-domain control, but OAE transfers better to a held-out ESRS-native company while remaining statistically tied with the tested cross-encoder reranker. These results support lightweight, auditable regulation ontologies as practical retrieval guardrails; page-length-matched reranking and extraction-stage validation are the next steps.

## Limitations

**Small benchmark.** Our benchmark remains small: five development companies and one held-out ESRS-native company, 12 requirements, 114 + 64 labels, English-language reports. The results should be read as evidence from a focused pilot benchmark, not as broad generalisation across sectors, jurisdictions, and reporting styles. Expanding the company and requirement coverage is the single most useful extension.

**Single primary annotator.** Gold labels rely primarily on one human reviewer with independent *blinded* verification passes, not full independent double annotation of every item; the verifier is an LLM rather than a second human, so human-human agreement is unmeasured. We report the conservative of our two agreement estimates ( $\kappa=0.67$ ) as the headline and describe the annotation protocol in §3 to make the labelling contract explicit.

**Retrieval-only evaluation.** This work evaluates the retrieval step that precedes extraction and verification. We make no claim of end-to-end improvement in disclosure drafting, assurance, or regulatory analysis, and we do not evaluate whether better-ranked evidence yields better extracted values — that is exactly the next stage of this line of work.

**Cross-encoder comparison is diagnostic.** The rerankers we test are trained for short passages and truncate long page-level inputs, which is the most likely reason they can degrade their own seed rankings here. This makes them an informative diagnostic about granularity but a poor basis for any claim about neural reranking in general; a reranker matched to page-length inputs (or a passage-level decomposition of each page) is needed for a fair comparison and is left to future work.

**LLM query expansion.** Our LLM-QE baseline uses one frozen generation per requirement from a single model and a single fixed prompt. We did not sweep prompts, models, sampling temperature, or multi-query expansion, all of which could move it in either direction; its in-domain parity with the ontology should therefore not be read as an upper bound on LLM expansion. Our gold is also partly AI-assisted by a related model family — an expander is far less exposed to this than a reranker, since it never sees the corpus, but

the risk is not zero. For the same reason we exclude an LLM-as-reranker baseline, and a comparison against direct LLM prompting (asking a model to locate evidence with no retrieval step) remains open; our retrieval-based design is motivated by deterministic, inspectable evidence pointers rather than a demonstrated accuracy or cost advantage over LLM-direct-answering, which independently annotated labels would be needed to evaluate fairly.

**Metric and tuning caveats.** Page-level accuracy is low for all methods ( $\leq 0.19$ ): adjacent pages share vocabulary, so exact-page top-1 is harsh — we treat Recall@10 and MRR as primary. The main benchmark’s candidate-pool seed vocabulary overlaps more with OAE’s evidence-query terms than with BM25’s (Jaccard 0.87 vs. 0.79), so the absolute margin on this benchmark (Table 1) may be somewhat optimistic; the fair-pooled Company F held-out check (Table 2, §5), which is neutral to any one system’s query vocabulary, shows the qualitative conclusion is unaffected — if anything, the margin there is larger. That check is itself limited to one company ( $n=64$ ), so the comparisons made on it — OAE vs. CE-reranker, and OAE vs. LLM-QE — are underpowered (§5). We did not tune fusion weights, PRF depth, LLM-QE budget, or BM25/expansion hyperparameters; all comparisons share untuned defaults.

**Ethics and evaluation safeguards.** We analyse only publicly released sustainability reports and make **no company-scoring, compliance, weakness-detection, or “not disclosed” judgement**; “gold” denotes retrieval ground truth only. We refer to companies by anonymized labels to avoid the appearance of assessing specific companies’ disclosure quality. Randomised procedures use fixed seeds, and the LLM-QE generations were frozen before evaluation. Paper results use only the `manual_page_gold` set; snapshot-based and document-level smoke results are excluded.

## References

Shafiuddin Rehan Ahmed, Ankit Shah, Quan Hung Tran, Vivek Khetan, Sukryool Kang, Ankit Mehta, Yujia Bao, and Wei Wei. 2025. [Enhancing retrieval for ESGLLM via ESG-CID: A disclosure content index finetuning dataset for mapping GRI and ESRS](#). In *Proceedings of the 2nd Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2025)*, pages 1–16, Vienna, Austria. Association for Computational Linguistics.

Jagdev Bhogal, Andrew Macfarlane, and Peter W. H. Smith. 2007. [A review of ontology based query expansion](#). *Information Processing & Management*, 43(4):866–886.

Claudio Carpineto and Giovanni Romano. 2012. [A survey of automatic query expansion in information retrieval](#). *ACM Computing Surveys*, 44(1):1:1–1:50.

Gordon V. Cormack, Charles L. A. Clarke, and Stefan Büttcher. 2009. [Reciprocal rank fusion outperforms Condorcet and individual rank learning methods](#). In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pages 758–759. ACM.

European Financial Reporting Advisory Group (EFRAG) and European Commission. 2023. [ESRS E1 climate change — european sustainability reporting standards, annex i to commission delegated regulation \(eu\) 2023/2772](#). Official Journal of the European Union. [https://eur-lex.europa.eu/eli/reg\\_del/2023/2772/oj](https://eur-lex.europa.eu/eli/reg_del/2023/2772/oj).

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. [Dense passage retrieval for open-domain question answering](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781. Association for Computational Linguistics.

Victor Lavrenko and W. Bruce Croft. 2001. [Relevance-based language models](#). In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pages 120–127. ACM.

Rodrigo Nogueira and Kyunghyun Cho. 2019. [Passage re-ranking with BERT](#). *arXiv preprint arXiv:1901.04085*.

Nils Reimers and Iryna Gurevych. 2019. [Sentence-BERT: Sentence embeddings using siamese BERT-networks](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992. Association for Computational Linguistics.

Stephen Robertson and Hugo Zaragoza. 2009. [The probabilistic relevance framework: BM25 and beyond](#). *Foundations and Trends in Information Retrieval*, 3(4):333–389.

Tobias Schimanski, Jingwei Ni, Roberto Spacey Martín, Nicola Ranger, and Markus Leippold. 2024. [ClimRetrieve: A benchmarking dataset for information retrieval from corporate climate disclosures](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 17509–17524, Miami, Florida, USA. Association for Computational Linguistics.

## A Held-out Results and Significance Tests

Table 2 gives the full held-out recall/MRR numbers summarised in §5; Table 3 gives the paired significance tests for the main benchmark and the held-out company.

| method                  | R@5          | R@10         | MRR          |
|-------------------------|--------------|--------------|--------------|
| BM25                    | 0.406        | 0.594        | 0.269        |
| LLM-QE                  | 0.422        | 0.656        | 0.291        |
| LLM-QE <sub>SBERT</sub> | 0.438        | 0.641        | 0.310        |
| CE-rerank(BM25)         | <b>0.547</b> | 0.734        | 0.334        |
| CE-rerank(Hyb-SBERT)    | 0.531        | 0.750        | 0.333        |
| OAE-lite(LSA)           | 0.516        | <b>0.812</b> | 0.333        |
| OAE-lite(SBERT)         | 0.500        | 0.766        | <b>0.349</b> |

Table 2: Held-out zero-shot retrieval on Company F ( $n=64$ ), unseen in ontology construction or pooling design. Best per column in bold.

| comparison                                    | R@5          | R@10             | MRR          |
|-----------------------------------------------|--------------|------------------|--------------|
| OAE-lite vs. Hyb-LSA                          | <b>.0009</b> | <b>.019</b>      | <b>.014</b>  |
| OAE-lite <sub>S</sub> vs. Hyb-SBERT           | <b>.015</b>  | <b>.0002</b>     | <b>.042</b>  |
| OAE-lite <sub>B</sub> vs. Hyb-BGE             | <b>.019</b>  | <b>.043</b>      | <b>.0015</b> |
| OAE-lite vs. PRF-RM3                          | .071         | <b>.0002</b>     | .072         |
| OAE-lite <sub>S</sub> vs. BM25                | <b>.008</b>  | <b>.0025</b>     | .151         |
| OAE-lite <sub>S</sub> vs. CE(BM25)            | <b>.0003</b> | <b>.0023</b>     | <b>.006</b>  |
| OAE-lite <sub>S</sub> vs. CE(Hyb-SBERT)       | <b>.0001</b> | <b>.0002</b>     | <b>.009</b>  |
| <i>Controls (outside the Holm family)</i>     |              |                  |              |
| Hyb-LSA vs. BM25 (fusion only)                | 1.0          | 1.0              | .65          |
| LLM-QE vs. BM25                               | <b>.008</b>  | <b>.029</b>      | .062         |
| LLM-QE vs. PRF-RM3                            | .071         | <b>&lt;.0001</b> | <b>.030</b>  |
| OAE-lite vs. LLM-QE                           | 1.0          | 1.0              | .69          |
| OAE-lite <sub>S</sub> vs. LLM-QE <sub>S</sub> | 1.0          | .30              | .75          |
| <i>Held-out Company F (<math>n=64</math>)</i> |              |                  |              |
| OAE-lite(LSA) vs. BM25                        | .144         | <b>.0094</b>     | <b>.0499</b> |
| OAE-lite(S) vs. BM25                          | .238         | <b>.0266</b>     | <b>.0272</b> |
| LLM-QE vs. BM25                               | 1.0          | .455             | .459         |
| OAE-lite(LSA) vs. LLM-QE                      | .109         | <b>.041</b>      | .234         |
| OAE-lite(S) vs. LLM-QE <sub>S</sub>           | .388         | .057             | .099         |
| OAE-lite(S) vs. CE(BM25)                      | .607         | .824             | .749         |
| OAE-lite(S) vs. CE(Hyb-SBERT)                 | .791         | 1.0              | .741         |

Table 3:  $p$ -values: exact McNemar (R@5, R@10) and sign-permutation on reciprocal ranks ( $10^5$  resamples, fixed seed; MRR). Bold:  $p < .05$ . In the first block, all five MRR comparisons except best-vs-BM25 (the three scorer-wise expansion tests and the two CE tests) survive Holm correction as a combined family; the control and held-out blocks are reported uncorrected and are not part of that family.

## B Ontology Construction

We designed the ontology as a lightweight, auditable retrieval artifact rather than a full OWL/SHACL implementation. Each node stores requirement labels, expected evidence phrases,

fields, units, and validation notes used to generate query variants; the validation rules are SHACL-like and implemented in code.

**Source materials.** (i) The ESRS E1 requirements and their implementation guidance (European Financial Reporting Advisory Group (EFRAG) and European Commission, 2023), including the datapoint field names. (ii) IFRS S2 and SSBJ climate mappings, held as *reference-only* pointers for cross-standard transfer; they contribute terms only to OAE-FULL. (iii) General domain vocabulary (GHG Protocol scope terminology, common units). Company reports were *not* used to author the core ontology terms, and the held-out company was excluded entirely.

**Workflow.** (1) Select the 12 ESRS E1 evaluation requirements spanning transition plan (2), targets (4), energy (2), GHG emissions (3), and gross/removals separation (1). (2) Normalise each requirement into one node with: label, topic, required\_fields, optional\_fields, allowed\_units, allowed\_values, synonyms, evidence\_queries, standard\_mappings, and validation\_notes. The evidence\_queries are 2–3 short phrases naming how the evidence is typically *titled or tabulated* in a report (e.g. for gross Scope 2: “Scope 2 market-based location-based emissions”; “Scope 2 GHG emissions tCO<sub>2</sub>e”; “purchased electricity emissions Scope 2”). (3) Derive the three expansion tiers: LITE = label + evidence\_queries; FIELD = lite + required/optional field names; FULL = field + synonyms + units/values + cross-standard mappings. Adding 13–24 terms, LITE sits inside the same budget as PRF-RM3’s 20 feedback terms and LLM-QE’s 20-term cap.

**Guardrails against leakage.** No gold label was consulted while writing expansion terms; terms are grounded in the standard and in general domain vocabulary, never in the content of a target page. The held-out company took no part in ontology authoring, hyperparameter choices, or candidate pooling, and its candidate pool is the union of every compared method’s top-*k*. We report the residual risk that the main benchmark’s pool vocabulary overlaps OAE terms more than BM25’s (Jaccard 0.87 vs. 0.79) in Limitations.

**Maintenance.** Because every field is explicit and versioned, a standards change can be implemented as a localised edit and audited through

version-controlled diffs. Governance of human-reviewed additions — who may add an evidence phrase, and on what basis — remains future work.

## C LLM-QE: Prompt and Protocol

The expander is prompted once per requirement, with no corpus, gold labels, or ontology content in scope. Exact prompt:

Standard: ESRS E1 (Climate Change).  
Disclosure requirement: {requirement\_text}  
Task: produce search-query expansion terms that would help a keyword-and-embedding retrieval system find the page or section of a corporate sustainability report that contains the evidence for this requirement. Output at most 20 whitespace-separated terms or short phrases. Output only the terms, no explanation, no punctuation other than what appears inside a term. The retrieval query is requirement\_text followed by the returned terms (truncated to 20), scored by the same HYB-LSA and HYB-SBERT scorers used for OAE-lite. The generated expansions were frozen before retrieval evaluation, so the reported runs use deterministic queries. Example output for gross Scope 2: Scope-2 gross indirect emissions market-based location-based purchased-electricity purchased-heat steam cooling tCO<sub>2</sub>e dual-reporting grid-emission-factor residual-mix contractual-instruments . . .

## D Corpus Composition

| company             | subset  | chunks | artifacts |
|---------------------|---------|--------|-----------|
| Company A           | EU/CSRD | 451    | PDF       |
| Company B           | US      | 128    | PDF, HTML |
| Company C           | US      | 262    | PDF       |
| Company D           | US      | 212    | PDF       |
| Company E           | EU/CSRD | 124    | PDF       |
| Total (5 companies) |         | 1,177  |           |

Table 4: Real-artifact corpus. Company F, held out from ontology construction and tuning, adds 158 ESRS-native pages.

## E Which Expansions Help, and Which Hurt

This section reports per-requirement Recall@10 for BM25 and OAE-lite on the 5-company benchmark and explains the two largest gains and the two weakest cases. The pattern is consistent: ontology phrases help where the requirement’s regulatory wording and the report’s own table/section wording genuinely diverge, and hurt where a requirement has *several* gold pages that the sharpened query collapses onto one.

**Success: carbon credits / gross separation** ( $n=5$ , **R@10** .60  $\rightarrow$  1.00). Helpful terms: gross emissions, carbon credits, offsets, removals, retired credits, net emissions. Retrieved evidence: credits table, methodology note. The requirement asks whether gross figures are kept *separate* from removals and credits. Companies title these pages “Carbon credits” or “Beyond value chain mitigation”; the raw requirement text matches the main GHG table instead. Ontology credit/removal vocabulary surfaces the credits page in its own right.

**Success: target year** ( $n=19$ , **R@10** .47  $\rightarrow$  .84). Helpful terms: target year, milestone year, 2030, 2050, net zero by 2050. Retrieved evidence: target table, target narrative. “Extract the target year or milestone year” shares almost no surface form with report prose, which writes “by 2030” or “2050 net zero”. Explicit milestone tokens bridge the mismatch — the largest single-requirement gain in the benchmark.

**Failure: target statement** ( $n=16$ , **R@10** .94  $\rightarrow$  .75). Harmful terms: net zero target Scope 1 Scope 2 Scope 3, science-based target. Retrieved evidence: target narrative (wrong sibling page). The only regression. Where a company states targets across several pages (all of them gold), the added target vocabulary concentrates the ranking on one canonical target-summary page: for Company C the gold pages fall from rank 5 and 6 to 17 and 16 while an adjacent target page takes rank 1. Expansion sharpens precision at the cost of multi-gold coverage.

**Weakest case: target base year** ( $n=17$ , **R@10** .53  $\rightarrow$  .65). Ambiguous terms: base year, base-line year (year tokens carry little discriminative signal). Retrieved evidence: ESG data table (wrong: reporting-period years). Improves but stays the weakest requirement. Year-heavy queries

retrieve year-dense data tables whose columns are *reporting* periods rather than the page tying a base year to a target — the “reporting-period year vs. base/target year” type in the error taxonomy (§6). Year terms need target context to disambiguate.

The residual “transition plan vs. generic sustainability strategy” type also persists for the LSA scorer (TP existence .50  $\rightarrow$  .67) but is resolved by the SBERT scorer ( $\rightarrow$  1.00), indicating a scorer rather than a vocabulary limitation.