Source-Aware Memory Retrieval across Engineering Standards

Sung Tae Yoo and Heyoung Yang
Applied Agent Research Center
Korea Institute of Science and Technology Information
Seoul, Republic of Korea
yst1213@kisti.re.kr, hyyang@kisti.re.kr

Abstract

Engineering design and quality review often require evidence from technical standards issued by multiple organizations. In a limited context window, partial evidence is often included, and relevance-based retrieval results can become biased toward sources with a large volume of documents, leaving other necessary sources unrepresented. We implemented a deterministic source-aware retrieval layer that identifies required source groups, retrieves candidates through source-conditioned channels, and prioritizes one candidate from each channel. We also evaluated a configuration that uses within-channel cross-encoder reranking to select that candidate. Evaluation used 15 technical standards, 72 evaluation questions, and 162 annotated evidence locations, with k set to 5 for all methods. Required-source coverage was 61.5% for dense retrieval and 78.6% for cross-encoder reranking, whereas both source-aware methods achieved 100%. Source-aware reranked channel quota achieved an evidence recall of 46.0% and retrieved all required evidence for 27.8% of questions, compared with 41.8% and 23.6% for cross-encoder reranking and 32.1% and 15.3% for dense retrieval. Within the source-aware design, within-channel reranking increased mean reciprocal rank from 0.445 to 0.545. These results support separating required-source coverage from evidence ranking in bounded-context retrieval for domain-specific agent memory.

1 Introduction

In engineering design and quality review, evaluating a product or component requires verifying requirements specified in technical standards issued by multiple organizations (Luttmer et al., 2023; Schuk et al., 2022). This study arises from an industrial valve design and review workflow in which engineers manually check the relevant clauses and tables in each standard. Because these requirements are expressed using different clause numbering, table formats, units, and terminology across organizations, engineers often need to cross-check relevant clauses and tables in multiple documents. If a required source is missing from the retrieved results, the judgment can remain incomplete even when evidence from the other sources is accurate.

These retrieval requirements also arise in domain-specific memory architectures for agents. The standards corpus can be treated as external memory, with the retrieval layer determining which evidence is brought into the limited context window (Lewis et al., 2020; Packer et al., 2024). Because the full standards corpus cannot fit within a limited context window, a small number of retrieval results are included in the context. Relevance-oriented retrieval can produce unequal document exposure, making some documents more retrievable than others (Azzopardi and Vinay, 2008; Wilkie and Azzopardi, 2017). When a question requires evidence from standards issued by multiple organizations, such unequal exposure may leave some required sources unrepresented in the limited context window.

This study examines whether identifying the source groups required by a query and retrieving through source-conditioned channels can improve coverage of the required sources. It then evaluates whether within-channel reranking can further improve evidence selection. We use a deterministic source-aware retrieval layer to separate source coverage from evidence selection. Source-aware channel quota prioritizes one candidate from each channel before filling the remaining results by dense score, while source-aware reranked channel quota applies cross-encoder reranking within each channel before selecting that candidate. Both methods remain deterministic and differ only in how that candidate is selected.

The evaluation separately considers whether the required sources are represented in the retrieved results and whether the specific evidence needed for a judgment is retrieved. Evidence is evaluated at

Proceedings of the 3rd Workshop on KNOwledge GRaphs & Agentic Systems Interplay (NORA 2026), pages 77–83
November 10, 2026 ©2026 Association for Computational Linguistics

the clause and table level rather than by page overlap, so these two aspects can be assessed separately. This distinction provides a practical way to evaluate multi-source evidence retrieval for technical standards and supports the source-aware retrieval design used in this study.

The contributions of this study are twofold. First, we distinguish required-source coverage from evidence selection as two complementary objectives in multi-source retrieval over technical standards. Second, we operationalize this distinction through a deterministic source-aware retrieval layer that uses source-group structure as a retrieval condition and prioritizes required-source representation before within-source evidence selection.

2 Related Work

External memory and retrieval have become important components of large language model (LLM)-based agent architectures. Retrieval-augmented generation (RAG) retrieves relevant passages from external memory and uses them during generation, giving the model access to information beyond what it already contains (Lewis et al., 2020). MemGPT uses an operating-system-inspired memory architecture to transfer information between the context window and external storage, allowing an LLM-based agent to access more information than the context window can hold (Packer et al., 2024). RET-LLM adds explicit read and write operations over external memory (Modarressi et al., 2024). Agent memory mechanisms also treat retrieval as a memory operation distinct from writing and management (Zhang et al., 2025).

Retrieval quality depends not only on relevance but also on whether the retrieved results cover the required sources. Maximal marginal relevance combines query relevance with information novelty to reduce redundancy in the selected results (Carbonell and Goldstein, 1998), but it does not use source identity as part of the selection criterion. Azzopardi and Vinay (2008) introduced retrievability as a document-centric measure of how easily individual documents can be retrieved within a collection across a set of queries, and Wilkie and Azzopardi (2017) later examined whether better-performing systems make relevant documents more retrievable than non-relevant ones. Retrievability characterizes differences in document exposure across a collection and a set of queries, rather than the source composition of a single result set. Retrieval methods span lexical ranking (Robertson and Zaragoza, 2009), rank fusion (Cormack et al., 2009), cross-encoder reranking (Nogueira and Cho, 2020), and dense retrieval (Karpukhin et al., 2020). These methods improve candidate ranking or combination, but do not account for the source groups required by a query when selecting the final results.

Retrieval of required evidence has also been studied in applications involving regulatory documents and technical standards. Gokhan and Briscoe (2025) report lower retrieval performance when regulatory questions require multiple passages. KG2QA (Luo et al., 2025) combines a domain knowledge graph with RAG for question answering over communication standards, while Ju and Lee (2026) cast regulatory compliance question answering as retrieving a closed set of provisions linked by explicit citations. These approaches use passage ranking, citation links, or graph structure to retrieve related evidence. When the required evidence spans standards issued by multiple organizations, however, coverage of the required source groups also becomes important.

3 Method and Evaluation

3.1 Task, Corpus, and Evaluation Set

The retrieval task is to identify the evidence needed for each evaluation question from a collection of technical standards. The evaluation corpus consists of 15 technical standards from four source groups, A to D. The standard documents are divided into 1,287 retrieval units in Korean and English. The four source groups account for 58.5%, 23.7%, 14.5%, and 3.3% of the total retrieval units, respectively. The three documents in the largest source group account for more than half of the corpus. Each retrieval unit contains source-group information used for source-conditioned retrieval.

The evaluation set consists of 72 questions linked to 162 evidence locations annotated by clause or table. Each question requires one to four evidence locations. Among the 72 questions, 15 require a single evidence location, while 12 require two or more evidence locations from the same source group. The 45 multi-source questions consist of 22 questions requiring evidence from two source groups, 15 questions requiring evidence from three source groups, and 8 questions requiring evidence from four source groups.

The evaluation uses five retrieved results per question ($k = 5$). Each evaluation question re-

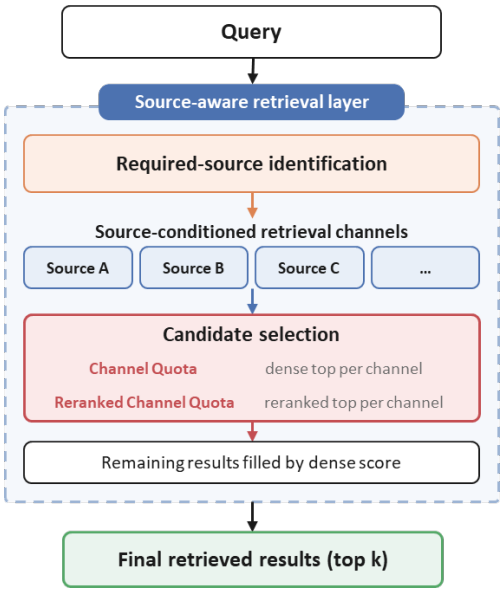


Figure 1: The source-aware retrieval layer. The two methods share source-group identification, source-conditioned channels, candidate pooling, and dense-score filling of the remaining results, and differ only in per-channel candidate selection.

quires evidence from one to four source groups. Five retrieved results can therefore cover all required source groups while leaving one to four additional results for evidence selection. This study is limited to evidence retrieval and does not include subsequent answer generation. Appendix A provides example evaluation questions, and Appendix B reports additional results for different values of k.

3.2 Source-Aware Retrieval

Figure 1 shows the source-aware retrieval layer and its retrieval flow. The retrieval process first identifies the source groups required by a query by rule-based matching between query terms and the scope of each source group. Candidates are then retrieved through source-conditioned channels for the identified groups, with candidates filtered by source-group labels. One candidate from each channel is prioritized. The remaining results are then filled by dense score from the union of channel candidates.

The two methods differ at this prioritization step. Source-aware channel quota prioritizes the highest-ranked unused candidate from each channel under dense retrieval. Source-aware reranked channel quota reranks up to 50 candidates within each channel with a cross-encoder before prioritizing the highest-ranked unused candidate. The remaining retrieval steps are shared. Both methods are deterministic and do not use an LLM during retrieval.

3.3 Evaluation Protocol

The source-aware retrieval layer is evaluated against three relevance-based baselines under the same protocol. Dense retrieval serves as the semantic retrieval baseline. Reciprocal rank fusion (RRF) (Cormack et al., 2009) is used in Hybrid RRF to combine dense and lexical rankings. Cross-encoder rerank applies reranking to an unconditioned candidate pool. BGE-M3 (Chen et al., 2024) provides the dense ranking used in dense retrieval and the source-aware methods, while BGE-reranker-v2-m3 is used for cross-encoder reranking. The comparison examines the effect of source-conditioned retrieval and the additional effect of within-channel reranking.

Scoring requires mapping annotated evidence locations to retrieval units. The mapping uses the parsed clause identifier, heading, and positional cue stored in each annotation. An annotated evidence location maps to 1.15 retrieval units on average, with a median of one, and 87.0% of annotated evidence locations map to a single unit. When clause information is unusable, positional cues are used and the mapping is marked low confidence. The mapping rules were fixed before system scoring.

We use five evaluation metrics. Evidence recall is the share of required evidence retrieved, averaged across questions. Any evidence is the percentage of questions retrieving at least one required evidence location and is reported as a diagnostic measure. All required evidence is the percentage of questions for which every required evidence location is retrieved. Source coverage measures the share of required source groups covered by the retrieved results. Mean reciprocal rank (MRR) is the reciprocal rank of the first correct retrieval, averaged across questions (Voorhees and Tice, 2000).

4 Results

4.1 Source-Aware Retrieval Performance

Table 1 compares the retrieval methods with five retrieved results per question ($k = 5$). The clearest difference appears in source coverage. Dense retrieval and hybrid RRF reach 61.5% and 52.0%, respectively, while cross-encoder rerank reaches 78.6%. Both source-aware channel quota and source-aware reranked channel quota reach 100%. The source-aware methods therefore include all re-

quired source groups in the final retrieved results, whereas the relevance-based baselines can still miss required sources.

The other evidence-level measures show different results. Source-aware channel quota retrieves all required evidence for 25.0% of questions, with source coverage remaining at 100%. Its evidence recall is 40.5%, slightly below 41.8% for cross-encoder rerank. MRR is also slightly lower at 0.445, compared with 0.501 for cross-encoder rerank. This comparison separates required-source coverage from evidence selection within those sources. Source-aware reranked channel quota improves the evidence-level results after within-channel reranking, while source coverage remains at 100%. Evidence recall rises to 46.0%, all required evidence to 27.8%, and MRR to 0.545. Source-aware reranked channel quota shows higher overall evidence-level results than the other methods, while both source-aware methods cover every required source group. Appendix B shows that the percentage of questions with all required evidence increases as k increases across all methods, while both source-aware methods continue to cover all required source groups. Reranked channel quota also has the highest evidence recall, any evidence, and MRR at each tested k.

Figure 2 shows that the effect of within-channel reranking differs by question type, with cross-encoder rerank included as a reference. For single-evidence questions, the all-required-evidence rate increases from 86.7% to 100%. Same-source multi-evidence questions change less. All required evidence remains at 16.7% and evidence recall at 41.7%, although MRR rises from 0.496 to 0.537.

For two-source questions, reranked channel quota improves all three measures over channel quota. Three-source questions show a different pattern. Evidence recall increases from 17.8% to 22.2% and MRR from 0.263 to 0.467, but all required evidence decreases from one question to none. For four-source questions, evidence recall increases, but MRR decreases slightly. These results suggest that better ranking does not always lead to retrieval of all required evidence.

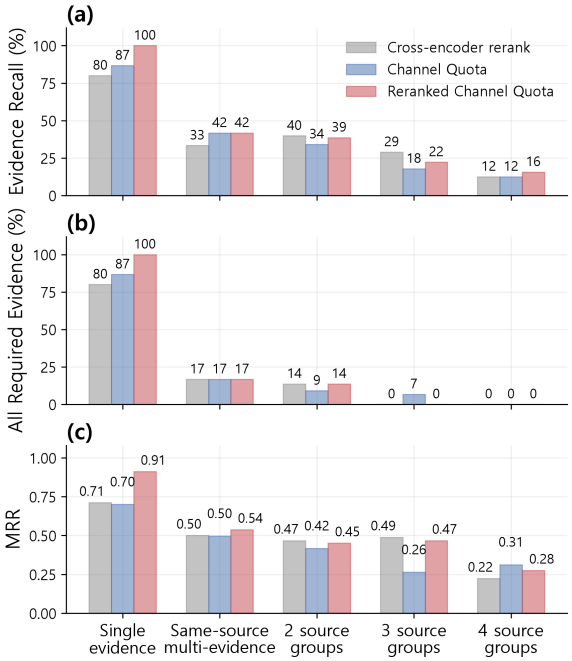


Figure 2: Comparison by question type for cross-encoder rerank, source-aware channel quota, and source-aware reranked channel quota. Panels show (a) evidence recall, (b) all required evidence, and (c) MRR.

Cross-encoder rerank provides a useful reference because it uses the same reranking model on an unconditioned candidate pool. Reranked channel quota performs better on the single-evidence questions and has higher evidence recall and MRR for same-source multi-evidence questions. Cross-encoder rerank has higher evidence recall and MRR for the two-source and three-source groups, whereas reranked channel quota is higher on these two measures for the four-source group. For all required evidence, the two methods give the same

Method	Evidence Recall	Any Evidence	All Required Evidence	Source Coverage	MRR
Dense Retrieval	32.1	52.8	15.3	61.5	0.374
Hybrid RRF	32.4	55.6	15.3	52.0	0.345
Cross-encoder rerank	41.8	65.3	23.6	78.6	0.501
Channel Quota	40.5	61.1	25.0	100.0	0.445
Reranked Channel Quota	**46.0**	**68.1**	**27.8**	**100.0**	**0.545**

Table 1: Retrieval performance at $k = 5$. Evidence recall, any evidence, all required evidence, and source coverage are percentages; MRR is on a 0–1 scale.

result for every multi-evidence question type in Figure 2. The difference seen in Table 1 therefore arises from the single-evidence questions. Across the full evaluation, the clearer distinction is coverage of required sources. Both source-aware methods cover every required source group, whereas cross-encoder rerank covers 78.6% of them.

5 Discussion

5.1 Required-Source Coverage as a Separate Retrieval Objective

Required-source coverage measures a part of retrieval that relevance ranking alone does not fully reflect. Within the source-aware retrieval layer, both methods include every required source in the final retrieved results. Relevance-based baselines still miss some required sources, even when cross-encoder rerank improves evidence-level ranking. Reranking changes candidate order, but it does not by itself ensure that every required source is included. This distinction becomes important when a review draws requirements from several standards. The source-aware retrieval layer therefore supports selective retrieval from external memory by covering the required sources before selecting evidence within those sources for the limited context window. In this workflow, the practical concern is whether a required standard is absent from the retrieved results, not only whether relevant evidence appears in a different order. Source-group information attached to each retrieval unit acts as a minimal structural layer over the corpus, and retrieval is conditioned on that structure rather than on relevance alone. Richer structure over standards, such as knowledge-graph relations (Luo et al., 2025) or citation links across documents (Ju and Lee, 2026), could extend the same separation beyond source groups.

5.2 Evidence Selection after Source Coverage

After the required source groups are covered, the quality of the retrieved evidence becomes more important. Within-channel reranking improves this part of the retrieval process. Source-aware reranked channel quota shows higher evidence recall and MRR than source-aware channel quota, and the improvement is especially noticeable for single-evidence questions. A cross-encoder scores each query-candidate pair independently, so better-ranked candidates do not necessarily produce a retrieved set containing all required evidence. This difference becomes more apparent when several evidence locations are required for the same question. Multi-evidence retrieval should therefore consider how the retrieved results work together rather than scoring each candidate separately. With five retrieved results per question, retrieval of all required evidence remains limited even when all required source groups are covered, indicating that source coverage and evidence completeness remain distinct challenges. Future work should examine how the number of retrieved results interacts with the number of prioritized candidates per channel, particularly for questions requiring multiple evidence locations. Another direction is to improve the selection of remaining results while retaining the priority given to required-source representation.

6 Conclusion

Engineering review across technical standards from multiple organizations requires both coverage of the required source groups and selection of appropriate evidence within those sources. The source-aware retrieval layer separates these two parts of retrieval. Both source-aware methods improve source coverage, and within-channel reranking further improves evidence recall and MRR. The same design supports selective retrieval from external memory by using source-group structure to ensure required-source coverage before refining evidence selection within each source. This retrieval layer can serve as an external-memory component in future agent architectures for technical standards.

Limitations

The comparison focuses on retrieval behavior rather than equalizing computation across methods, and the source-aware variants process candidates separately within source-conditioned channels. The two source-aware methods remain directly comparable, but the evaluation is limited to 15 standards and 72 questions. The underlying standards cannot be released because of copyright restrictions. The full evaluation set is not released, but selected evaluation questions and answers are provided in Appendix A.

Declaration on Generative AI

The authors used a generative AI assistant for language editing only. It was not used to generate ideas, design experiments, or produce any reported results.

References

Leif Azzopardi and Vishwa Vinay. 2008. Retrievability: an evaluation measure for higher order information access tasks. In *Proceedings of the 17th ACM conference on Information and knowledge management*, CIKM '08, pages 561–570, New York, NY, USA. Association for Computing Machinery.

Jaime Carbonell and Jade Goldstein. 1998. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '98, pages 335–336, New York, NY, USA. Association for Computing Machinery.

Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. M3-Embedding: Multi-Linguality, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2318–2335, Bangkok, Thailand. Association for Computational Linguistics.

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 '09, pages 758–759, New York, NY, USA. Association for Computing Machinery.

Tuba Gokhan and Ted Briscoe. 2025. Grounded Answers from Multi-Passage Regulations: Learning-to-Rank for Regulatory RAG. In *Proceedings of the Natural Legal Language Processing Workshop 2025*, pages 135–146, Suzhou, China. Association for Computational Linguistics.

Yeong-Joon Ju and Seong-Whan Lee. 2026. Citation-Closure Retrieval and Per-Rule Attribution for Real-World Regulatory Compliance Question Answering. *arXiv preprint*. ArXiv:2605.29742 [cs.AI].

Vladimir Karpukhin, Barlas Oguz, 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, Online. Association for Computational Linguistics.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, pages 9459–9474, Red Hook, NY, USA. Curran Associates Inc.

Zhongze Luo, Weixuan Wan, Tianya Zhang, Dan Wang, and Xiaoying Tang. 2025. KG2QA: Knowledge Graph-enhanced Retrieval-augmented Generation for Communication Standards Question Answering. *arXiv preprint*. ArXiv:2506.07037 [cs.CL].

Janosch Luttmer, Vitalijs Prihodko, Dominik Ehring, and Arun Nagarajah. 2023. Requirements extraction from engineering standards – systematic evaluation of extraction techniques. *Procedia CIRP*, 119:794–799.

Ali Modarressi, Ayyoob Imani, Mohsen Fayyaz, and Hinrich Schütze. 2024. RET-LLM: Towards a General Read-Write Memory for Large Language Models. *arXiv preprint*. ArXiv:2305.14322 [cs.CL].

Rodrigo Nogueira and Kyunghyun Cho. 2020. Passage Re-ranking with BERT. *arXiv preprint*. ArXiv:1901.04085 [cs.IR].

Charles Packer, Sarah Wooders, Kevin Lin, Vivian Fang, Shishir G. Patil, Ion Stoica, and Joseph E. Gonzalez. 2024. MemGPT: Towards LLMs as Operating Systems. *arXiv preprint*. ArXiv:2310.08560 [cs.AI].

Stephen Robertson and Hugo Zaragoza. 2009. The Probabilistic Relevance Framework: BM25 and Beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389.

Vitali Schuk, María Elena Pombo Jiménez, and Ullrich Martin. 2022. Technical specifications to meet the requirements of an Automatic Code Compliance Checking tool and current developments in infrastructure construction. *Results in Engineering*, 16:100650.

Ellen M. Voorhees and Dawn M. Tice. 2000. The TREC-8 Question Answering Track. In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'00)*, Athens, Greece. European Language Resources Association (ELRA).

Colin Wilkie and Leif Azzopardi. 2017. Algorithmic Bias: Do Good Systems Make Relevant Documents More Retrievable? In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, CIKM '17, pages 2375–2378, New York, NY, USA. Association for Computing Machinery.

Zeyu Zhang, Quanyu Dai, Xiaohe Bo, Chen Ma, Rui Li, Xu Chen, Jieming Zhu, Zhenhua Dong, and Ji-Rong Wen. 2025. A Survey on the Memory Mechanism of Large Language Model-based Agents. *ACM Transactions on Information Systems*, 43(6):155:1–155:47.

Appendix

A Evaluation Question Examples

The evaluation questions were written from clauses and tables in the corpus and linked to annotated evidence locations. Two examples from the evaluation set are shown below.

Method	k	Evidence Recall	Any Evidence	All Required Evidence	Source Coverage	MRR
Dense Retrieval	10	44.1	65.3	25.0	71.8	0.392
	15	49.6	69.4	31.9	80.7	0.395
	20	52.9	75.0	34.7	83.3	0.398
Hybrid RRF	10	39.1	65.3	18.1	57.8	0.357
	15	43.4	68.1	20.8	69.1	0.359
	20	49.4	76.4	23.6	73.2	0.363
Cross-encoder rerank	10	53.0	76.4	33.3	85.5	0.515
	15	55.6	79.2	37.5	87.4	0.517
	20	58.5	81.9	40.3	92.1	0.519
Channel Quota	10	53.5	75.0	37.5	100.0	0.465
	15	58.9	84.7	38.9	100.0	0.472
	20	64.6	88.9	43.1	100.0	0.475
Reranked Channel Quota	10	55.1	80.6	36.1	100.0	0.561
	15	60.3	86.1	40.3	100.0	0.565
	20	65.3	90.3	44.4	100.0	0.568

Table 2: Retrieval performance for $k \in \{10, 15, 20\}$. Evidence recall, any evidence, all required evidence, and source coverage are percentages; MRR is reported on a 0–1 scale. Both source-aware methods maintain 100% source coverage across all tested values of k.

Question 1. How are the minimum hold times for an industrial valve body pressure test divided by DN range?

Answer. The minimum hold times are 15 s for DN 50 and below, 60 s for DN 65 through DN 200, and 180 s for DN 250 and above.

This is a single-evidence question linked to one annotated evidence location in a table.

Question 2. For pressure testing of a steel gate valve for oil and gas service, what is the basis for calculating the body pressure test pressure and which test types are required?

Answer. The body pressure test pressure is at least 1.5 times the pressure corresponding to the 38°C pressure rating. The required tests are the shell pressure test, closure leakage test, and stem backseat test.

This is a same-source multi-evidence question linked to two annotated evidence locations.

B Sensitivity to k

Table 2 shows the results for k values of 10, 15, and 20. The percentage of questions with all required evidence increases with k for every method, while both source-aware methods continue to cover all required source groups. Source-aware reranked channel quota gives the highest evidence recall, any evidence, and MRR at each tested k. It also gives the highest all-required-evidence rate at $k = 15$ and $k = 20$. At $k = 10$, source-aware channel quota is slightly higher, reaching 37.5% compared with 36.1%.