TU Delft

Delft University of Technology

Document Version
Final published version

Licence
CC BY

Citation (APA)
Jumle, V., Ağaoğlu, O., Mol, L., Salas Gironés, E., & Bai, N. (2026). Retrieving Climate Change Adaptation Concepts from World Heritage Management Plans in the Netherlands. In T. Schimanski, A. Usmanova, G. Morio, J. Ni, K. Dutia, S. Saeid Vaghefi, D. Stammbach, & M. Leippold (Eds.), *Proceedings of the 3rd Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2026)* (pp. 170-182). Association for Computational Linguistics (ACL). https://doi.org/10.18653/2026.climatenlp-1.13

Important note
To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Copyright
In case the licence states “Dutch Copyright Act (Article 25fa)”, this publication was made available Green Open Access via the TU Delft Institutional Repository pursuant to Dutch Copyright Act (Article 25fa, the Taverne amendment). This provision does not affect copyright ownership.
Unless copyright is transferred by contract or statute, it remains with the copyright holder.

Sharing and reuse
Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

Takedown policy
Please contact us and provide details if you believe this document breaches copyrights.
We will remove access to the work immediately and investigate your claim.

This work is downloaded from Delft University of Technology.

Retrieving Climate Change Adaptation Concepts from World Heritage Management Plans in the Netherlands

Vedant Jumle[1], Orhan Ağaoğlu[1], Laurens Mol[1], Edgar Salas Gironés[2,3], Nan Bai[1]
[1]Delft University of Technology, Delft, the Netherlands
{vvjumle, oagaoglu1, L.P.J.Mol}@student.tudelft.nl, n.bai@tudelft.nl
[2]Technopolis Group Amsterdam, Amsterdam, the Netherlands
edgar.girones@technopolis-group.com
[3]University of Amsterdam, Amsterdam, the Netherlands
e.salasgirones@uva.nl

Abstract

Monitoring climate change adaptation (CCA) in the heritage sector, especially in the context of World Heritage (WH) management, requires systematic analyses of policy documents for CCA strategies, a process that is difficult to scale. We ask whether automated retrieval systems can *prioritise* expert annotation effort: given a CCA concept, can a system rank passages of a management plan so that those an expert would annotate appear near the top, without any in-domain supervision? This objective is formulated as an information retrieval problem with a new benchmark dataset on climate-adaptive heritage management strategies, created from expert-labelled text chunks in official Management Plans of 11 Dutch WH properties. Four retrieval approaches (BM25, a bi-encoder, a dense retriever with reranker, and a cross-encoder reranker) are evaluated and compared. The cross-encoder reranker achieves the strongest performance across labels. An LLM-as-judge evaluation further shows that top-ranked unlabelled chunks could be relevant, suggesting that the existing annotations are precise but not exhaustive. Per-label analysis identifies which concepts are automatically retrievable and which remain beyond the reach of general-purpose retrieval. Our benchmark establishes a foundation for scalable mapping and monitoring of climate change adaptation in the context of heritage management.

1 Introduction

Monitoring climate change adaptation (CCA) at scale requires structured analysis of policy documents in which adaptation strategies are recorded. In the context of UNESCO World Heritage (WH), Management Plans (MPs) are one such instrument: World Heritage managers use them to translate risks into conservation responses. Heritage properties face compounding and accelerating climate pressures (Sesana et al., 2021). Chen et al. (2025b) find empirically that 80% of UNESCO sites already experience harmful heat- and moisture-related climate risks. Climate risks and adaptation strategies are increasingly being addressed in MPs as heritage properties are especially vulnerable amidst global climate change. Properties worldwide are also expected to develop tailored climate action plans by 2030 (UNESCO, 2023; Chatziandreou and Haupt, 2026). Yet policy documents systematically fail to translate this awareness into operational actions. Analysing 1,868 World Heritage documents from 535 locations, Chen et al. (2025a) find a persistent awareness-action gap that sites with greater vulnerability recognition implement fewer concrete local management interventions. Fatorić and Biesbroek (2020) document that in the Netherlands, the largest perceived barrier for heritage sites to adapt to climate change impacts is lack of adaptation policy for heritage, and Guzman (2025) identifies similar challenges of cross-sector synergy in Norwegian national climate planning. Cheang et al. (2025a) developed a domain-expert taxonomy of 27 climate-adaptive heritage management strategy codes and manually annotated hundreds of sentences from 11 World Heritage MPs in the Netherlands, thereby providing a reusable and systematic evidence base for understanding and assessing CCA in heritage management at scale. This annotation is accurate but expert-intensive and cannot readily extend to the growing corpus of heritage policy documents worldwide, a shortfall made more acute by the scale and urgency of the underlying climate problem.

Parallel progress in Natural Language Processing (NLP) and Information Retrieval (IR) for climate policy text provides a methodological potential for addressing this challenge. Webersinke et al. (2021) show that domain-adapted language models, continued-pretrained on climate-related text, substantially outperform general-purpose models on climate text classification. Bonenkamp et al. (2025) demonstrate that fine-tuned encoder-only models achieve the strongest results on adaptation-

Proceedings of the 3rd Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2026), pages 170–182
October 28, 2026 ©2026 Association for Computational Linguistics

relevance classification in climate policy text, with text granularity as a critical design factor. Schimanski et al. (2024) release ClimRetrieve as a retrieval benchmark, covering over 8,500 question-answer pairs from corporate sustainability reports, and show that generic embeddings require expert augmentation to handle domain-specific language. The scale of the challenge motivates automation. The IPCC AR6 WGII alone drew on more than 34,000 papers (De Rijke et al., 2025), and Sietsma et al. (2024) show that ML-based analysis of 8,691 adaptation policy documents surfaces geographic patterns that manual review cannot detect.

While these methods have been developed and evaluated on climate corpora, applying them to practitioner-authored heritage policy documents remains largely unexplored. Hall and Ram (2016) demonstrate through lexical analysis of all five IPCC assessment reports (1990–2014) that heritage terminology is minimally present in IPCC discourse, reflecting a conceptual separation between the two domains that extends to the vocabulary level. Chen et al. (2025a) use keyword search and LLM classification to respectively evaluate climate awareness and IPCC local action in WH properties, but the analyses on both domains are separate. Bai et al. (2025) apply sentence embeddings to identify shared CCA topics across Dutch WH properties, but do not systematically validate the obtained similarity scores. We address this cross-domain evaluation gap directly. The manually labelled sentence-level dataset of climate-adaptive heritage management strategies in Cheang et al. (2025a) is treated as the ground-truth retrieval reference. We formalise the evaluation process using comparable metrics and test four common IR approaches, with the following question:

Given a CCA concept and its expert definition, can an automated retrieval system rank the text chunks of a World Heritage MP so that those expert-annotated passages with that concept appear near the top, without any in-domain supervision?

This effectively functions as a benchmark[1] for ClimateNLP researchers to optimise retrieval methods and for heritage scholars to potentially scale up the analysis on the intersection of CCA and heritage management.

[1]Code, the extracted chunks, the concept definitions, the alignment output and the LLM-as-judge prompts and judgements are available at https://github.com/vedant-jumle/Heritage-Doc-Retrival.

2 Methodology

2.1 Datasets and Labels

Using all the textual data in MPs from 11 Dutch WH properties $\boldsymbol{\mathcal{H}} := \{\mathcal{H}_l | l = 1, 2, ..., n\}$ (see Table A2), two separate datasets were constructed:

Definition 1 (Sentence-level Labelled Dataset $\mathcal{D}_1$)**.** Let $\mathcal{D}_1 = (\mathcal{T}_1, \mathcal{C}, \mathcal{L})$, where:

- $\mathcal{T}_1 = \{t_1^{(i)(l)} | i = 1, 2, ..., M, t_1^{(i)(l)} \in \mathcal{H}_l\}$ is the set of $M = 763$ sentences manually extracted from the MP source texts.
- $\mathcal{C} = \{c^{(k)} | k = 1, 2, ..., P\}$ is the set of $P = 27$ *climate-adaptive heritage management strategy labels* assigned by experts.
- $\mathcal{L} : \mathcal{T}_1 \rightarrow \mathcal{C}$ is the manual labelling function that maps each sentence to one label.

$\mathcal{D}_1$ originates from Cheang et al. (2025b), who manually reviewed MPs from 11 Dutch WH properties and came up with an inductive coding system of 27 climate-adaptive heritage management strategies in the Netherlands (see Table A1). These 27 labels encompass both strategies and operational considerations related to CCA. The strategies are categorised into past, present, and future approaches, including '*Adaptation*', '*Conservation*', and '*Knowledge Transfer*'. The operational considerations comprise challenges, such as '*Climate Change*' and '*Soil Subsidence*', as well as managerial considerations, such as '*Management Variation*' and '*Environmental Variation*'.

Definition 2 (Chunk-level Unlabelled Dataset $\mathcal{D}_2$)**.** Let $\mathcal{D}_2 = (\mathcal{T}_2, \mathcal{C}_{\text{def}})$, where:

- $\mathcal{T}_2 = \{t_2^{(j)(l)} | j = 1, 2, ..., N, t_2^{(j)(l)} \in \mathcal{H}_l\}$ is the set of N text chunks extracted from all MPs and indexed by a unique Chunk ID j.
- $\mathcal{C}_{\text{def}} = \{[c^{(k)}, d^{(k)}] | k = 1, 2, ..., P'\}$ is a smaller set of $P' = 24$ labels, where $d^{(k)}$ is an expert-authored definition for concept $c^{(k)}$.

$\mathcal{D}_2$ originates effectively from the same sources of MPs as $\mathcal{D}_1$, but with a different format to organize information. The MPs are digitally authored PDFs with no scanned pages. We extract text $\mathcal{T}_2$ using Docling (Auer et al., 2024) with GPU-accelerated layout analysis while OCR (Optical Character Recognition) is disabled, producing clean markdown output free of the character-level artefacts (e.g., vertical logo text, page number bleed-through) common in naïve PDF parsing. Each extracted document is split into text chunks

using a sliding window over characters with a default setting of `chunk_size`$= 1000$ and an overlap of `chunk_overlap`$= 200$, snapping boundaries to the nearest sentence end to avoid mid-sentence splits. This yields $N = 2{,}926$ chunks across 10 MPs ($\mathcal{H}_9$ from *Beemster*[2] is excluded throughout, since this is the only Dutch-language MP). Chunk size is a hyperparameter with non-trivial effects on retrieval, as larger chunks reduce the candidate pool and provide more context per match, while smaller chunks offer finer granularity. We report a chunk-size ablation over six log-spaced configurations (200–3,000 characters) in Section 3. Three labels (*Environmental variation*, *Preservation*, and *Restricted greenery*) are excluded from $\mathcal{C}_{\text{def}}$ since they lack clear expert-authored definitions or representativeness in our dataset (see Table A1).

Because $\mathcal{T}_2$ was drawn from the same source documents as $\mathcal{T}_1$, every sentence $t_1^{(i)(l)} \in \mathcal{T}_1$ is fully or partially contained within one chunk $t_2^{(j)(l)} \in \mathcal{T}_2$ or two near-duplicate chunks from the same MP $\mathcal{H}_l$. The two datasets are linked at the text level through an alignment function ψ described below.

Definition 3 (Text Alignment ψ)**.** A *text alignment* maps each labelled sentence in $\mathcal{T}_1$ to the set of one or more chunks in $\mathcal{T}_2$ that contain it.

For each labelled sentence $t_1^{(i)(l)}$, we locate the chunk(s) in $\mathcal{T}_2$ that contain(s) it using a five-pass matching procedure: (i) exact substring match, (ii) split-pair match (first or second half of the sentence), (iii) split-triplet match, (iv) fuzzy match (token-sort ratio ≥ 85), and (v) fuzzy split match. Sentences with no match under any pass are recorded as unmatched. Every labelled sentence from the 10 evaluated MPs is matched, while 32 unmatched sentences all belong to *Beemster* ($\mathcal{H}_9$), which is excluded throughout. In our experiment, the alignment is dominated by exact substring matches (95.3% of matched sentences). The remaining 4.7% rely on the split and fuzzy passes, which are the only passes that can in principle produce a false-positive alignment. Exact matches cannot misalign by construction, so alignment noise is limited to the 4.7% of matches produced by the split and fuzzy procedures, and we release these 35 approximate matches alongside the code for inspection. Since the labelled dataset $\mathcal{D}_1$ has been reviewed and annotated using the same corpora $\boldsymbol{\mathcal{H}}$ as the retrieval dataset $\mathcal{D}_2$, there could naturally be an implicit assumption suggesting that those blocks not associated with any labelled sentences would be negative examples, not relevant to any concepts in $\mathcal{C}_{\text{def}}$, which we will challenge with our experiments. Using the mapping function ψ, the labelled dataset $\mathcal{D}_1$ can be aligned with $\mathcal{D}_2$, such that:

Definition 4 (Aligned Chunk-level Labelled Dataset $\mathcal{D}_2'$)**.** Let $\mathcal{D}_2' = (\mathcal{T}_2, \mathcal{C}_{\text{def}}, \mathcal{L}')$, where:

- $\mathcal{T}_2$ is the same set of $N = 2{,}926$ text chunks.
- $\mathcal{C}_{\text{def}}$ is the same set of $P' = 24$ concepts enhanced with their definitions.
- $\mathcal{L}' : \mathcal{T}_2 \to 2^{\mathcal{C}_{\text{def}}}$ is the function that maps each text block to none, one or more labels.

Formally, $\mathcal{D}_2'$ is potentially multi-label (71 chunks, 2.4%), since one chunk could contain sentences related to different labels, meaning that a labelled chunk can result in several valid label-chunk pairs.

2.2 Problem Definition

We address the following problem of retrieval consistency: given a heritage strategy label $c^{(k)} \in \mathcal{C}_{\text{def}}$ and its definition $d^{(k)}$, our goal is to identify the subset of text chunks in each MP that are relevant to that label. This requires a method to assign a relevance score to every chunk, so that chunks can be ranked by their relevance to a given concept. We evaluate how well such rankings agree with expert annotation. Converting these scores into a variable-size relevant set per concept-MP pair is left to future work. We formalise this as a retrieval function f for the aligned dataset $\mathcal{D}_2'$, where: $f : \mathcal{C}_{\text{def}} \times \mathcal{T}_2 \to \mathbb{R}$ assigns a score to every label-chunk pair. Figure 1 gives an overview of the full pipeline described in this section.

2.2.1 Query Formation

Each climate-adaptive heritage management strategy label $c^{(k)}$ is represented as a single query string by concatenating the label name and its definition, formally as $q^{(k)} =$ "$c^{(k)}$: $d^{(k)}$". This gives the retrieval system access to both the label name (useful for lexical matching) and its semantic definition (useful for embedding-based matching), without requiring any fine-tuning or query expansion.

2.2.2 Ground Truth

The ground truth for retrieval systems is derived from the aligned dataset $\mathcal{D}_2'$. Multiple labelled sentences may align to the same chunk under

[2]See Table A2. *Droogmakerij de Beemster (Beemster Polder)* as a property inscribed in 1999, available at https://whc.unesco.org/document/218353, accessed 20 September 2026.

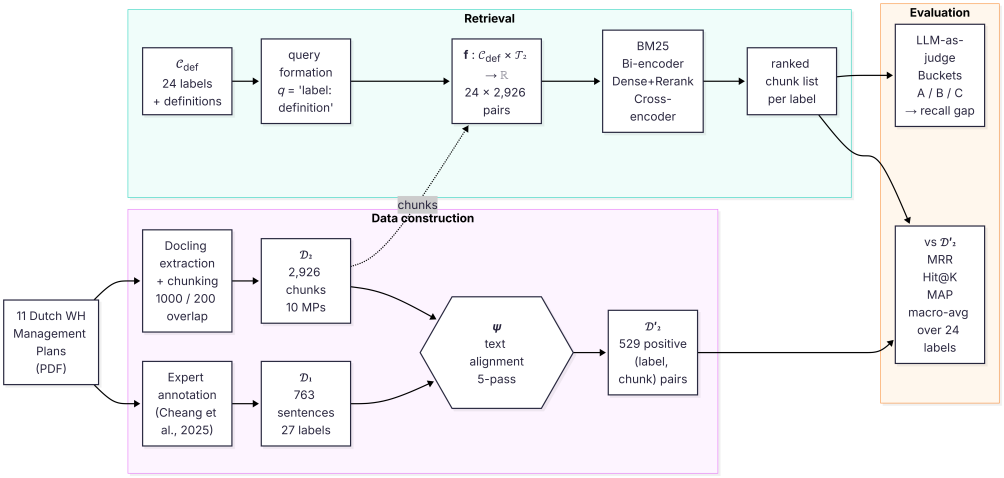


Figure 1: Overview of the pipeline. The same 11 Management Plans give rise to two datasets: the expert-annotated sentences $\mathcal{D}_1$ and the Docling-extracted chunks $\mathcal{D}_2$. The alignment ψ links them into the chunk-level ground truth $\mathcal{D}'_2$. Each concept in $\mathcal{C}_{\text{def}}$ is turned into a query, the retrieval function f scores every label-chunk pair, and the resulting ranking is evaluated both against $\mathcal{D}'_2$ and by an LLM judge.

the same label. After deduplication, 529 unique query-chunk pairs remain across 24 labels. A detailed explanation of how the original 763 labelled sentences become 529 query-chunk pairs is provided in Appendix A. The retrieval function f is evaluated against the human-annotated labels in $\mathcal{D}'_2$, measuring how consistently automated ranking agrees with expert judgement. For each label $c^{(k)} \in \mathcal{C}_{\text{def}}$, the retrieval function f scores all $N = 2{,}926$ chunks in $\mathcal{T}_2$ and produces a ranked list of chunks ordered by their retrieval scores. The ground truth for label $c^{(k)}$ is the set of positive chunks $\mathcal{G}^{(k)} = \{t_2^{(j)(l)} \mid \mathcal{L}'(t_2^{(j)(l)}) \ni c^{(k)}\}$ derived from the alignment ψ. Evaluation metrics are then computed over this ranked list against $\mathcal{G}^{(k)}$, and macro-averaged across all P' evaluable labels.

2.2.3 Retrieval Systems

Four common retrieval systems of increasing sophistication are evaluated, spanning the lexical–semantic spectrum.

BM25. A bag-of-words lexical baseline using the *Okapi BM25* algorithm (Robertson and Zaragoza, 2009). Chunks are tokenised by whitespace. BM25 scores are computed for each query–chunk pair and used to rank all N chunks. BM25 requires no training and serves as a strong lexical baseline that directly rewards surface-form overlap between query terms and chunk text.

Bi-encoder. A dense retrieval system based on `all-mpnet-base-v2` (Song et al., 2020), a general-purpose sentence encoder pre-trained on over one billion sentence pairs. Query and chunk texts are encoded independently into 768-dimensional vectors and ℓ_2-normalised, while scores are computed as cosine similarities. The bi-encoder captures semantic similarity beyond surface overlap, at the cost of treating query and chunk in isolation.

Dense + Rerank. A two-stage pipeline: the bi-encoder first retrieves the top-50 chunks by cosine similarity, which are then reranked by a cross-encoder `ms-marco-MiniLM-L-6-v2` (Reimers and Gurevych, 2019). The cross-encoder scores each query-chunk pair jointly, enabling fine-grained relevance judgement at the cost of higher inference time. Only the top-50 candidates are passed to the cross-encoder to keep inference tractable.

Cross-encoder. The cross-encoder is also evaluated as a standalone system, scoring all $P' \times N = 24 \times 2{,}926 = 70{,}224$ query-chunk pairs jointly without any bi-encoder pre-filter. This approach is well-suited to our goal, as it assigns a fine-grained relevance score to every chunk in the corpus for a given query. Since it produces a comparable score for every pair rather than only an ordering within a shortlist, it is also the system best suited to a future thresholding step. The trade-off is computational cost, which remains feasible at the scale of this corpus, but would suffer when further scaled up.

2.3 Evaluation

2.3.1 Evaluation Metrics

We evaluate with rank-based metrics, which measure if the systems assign high scores to known-

relevant chunks so that they surface near the top of the ranking. To quantify the evaluation, we adopt three standard ranking metrics in the IR literature.

First, *mean reciprocal rank* (MRR) captures the rank of the first positive chunk for each label, averaged across all evaluated labels:

$$\text{MRR} = \frac{1}{P'} \sum_{k=1}^{P'} \frac{1}{\text{rank}_k^*},$$

in which rank_k^* is the rank of the highest-scoring positive chunk for label $c^{(k)}$ in the retrieval output. MRR penalises systems that rank a known-relevant chunk lower, directly reflecting the goal of surfacing relevant content near the top. We complement MRR with Hit@K, which measures if at least one positive chunk appears within the top-K results:

$$\text{Hit@}K = \frac{1}{P'} \sum_{k=1}^{P'} 1[\text{rank}_k^* \leq K],$$

where $1[\cdot]$ is the indicator function and we report $K \in \{1, 3\}$, while a larger range of K is left for future studies. Hit@K is used since multiple chunks may be positive for a single label (due to conceptual overlap across passages), and surfacing any one of them within the top K is practically useful.

Lastly, *mean average precision* (MAP) rewards systems that rank *all* known positive chunks highly:

$$\text{MAP} = \frac{1}{P'} \sum_{k=1}^{P'} \text{AP}_k,$$

$$\text{AP}_k = \frac{1}{|\mathcal{G}^{(k)}|} \sum_{r=1}^{N} P_k(r)\, \text{rel}_k(r),$$

where $\mathcal{G}^{(k)}$ is the set of positive chunks for label $c^{(k)}$, N is the total number of ranked chunks, $P_k(r)$ is the precision at rank r, and $\text{rel}_k(r) = 1$ if the chunk at rank r is in $\mathcal{G}^{(k)}$ and 0 otherwise. MAP is particularly suitable when multiple chunks per label are relevant, as it measures retrieval completeness rather than just the top hit.

2.3.2 LLM-as-Judge Evaluation

The original dataset $\mathcal{D}_1$ may contain several expert-labelled sentences per label per MP. However, because the labels were provided through manual read-through, experts may have unintentionally missed other relevant sentences. This means that the ground truth has high precision but inherently low recall, where genuinely relevant chunks that were not selected during annotation are treated as negatives. To quantify this recall gap, we employ an LLM-as-judge protocol (Zheng et al., 2023). For each of the 24 labels, we sample three buckets of label-chunk pairs: Bucket A (ground-truth positive chunks), Bucket B (top-ranked chunks *not* in the ground truth), and Bucket C (low-ranked chunks not in the ground truth). Each pair is scored 1–5 for relevance by an LLM[3] using a structured prompt that provides the label name, its definition, and the chunk text (see Appendix A.3). The intended design is 2 from Bucket A, 3 from Bucket B and 1 from Bucket C per label, giving 144 pairs in total. Because of a shortage of samples for two labels as described in Appendix A.3, we obtained 140 pairs (47 Bucket A, 70 Bucket B and 23 Bucket C pairs), scored using 50 concurrent API calls. Bucket B chunks with scores of at least 4 indicate those that the system ranked highly that ground truth does not credit, thus potentially false-negative candidates.

3 Results

3.1 Overall Retrieval Performance

Table 1 reports macro-averaged metrics across the $P' = 24$ evaluated labels for each retrieval system. The Cross-encoder obtains the highest value on every metric by ranking all $N = 2{,}926$ chunks jointly against each query without any pre-filtering. With only 24 labels, however, the margins are not all equally reliable. Bootstrapping over labels (10,000 replicates) further separates the Cross-encoder from BM25 and the Bi-encoder on MRR ($+0.13$, 95% CI $[+0.03, +0.26]$ and $+0.22$, CI $[+0.07, +0.39]$), but the Hit@1 and MAP differences are indistinguishable from zero for every pair of systems, as are all differences between the Cross-encoder and Dense + Rerank. We therefore rest the comparison on MRR and Hit@3 and do not read the small MAP differences as evidence of ranking quality. Dense + Rerank scores lower than the Cross-encoder alone, indicating that the Bi-encoder pre-filter discards relevant chunks before the Cross-encoder can evaluate them, while BM25 outperforms the Bi-encoder on most metrics. Surface-form overlap between label definitions and chunk text is a strong signal here, and independent query and chunk encoding with a generic embed-

[3] `google/gemma-4-31B-it-qat-w4a16-ct`, a 31B instruction-tuned model quantised to 4-bit weights, served on-premises through the TULIP API (an OpenAI-compatible Chat endpoint) at temperature 0. Running the judge locally avoids sending heritage policy text to third-party providers.

ding model is shown as insufficient for this inter-domain cross-vocabulary task.

System	MRR	Hit@1	Hit@3	MAP
BM25	0.32	0.21	0.33	0.13
Bi-encoder	0.23	0.17	0.21	0.13
Dense + Rerank	0.36	0.21	0.50	0.12
Cross-encoder	**0.45**	0.29	**0.54**	0.14

Table 1: Macro-averaged retrieval metrics across 24 predefined labels. Bold marks the Cross-encoder's MRR and Hit@3 as the only margins a label-level bootstrap separates from BM25 and the Bi-encoder. No metric separates the Cross-encoder from Dense + Rerank, and Hit@1 and MAP do not separate any pair of systems.

3.2 Per-Label and Per-Document Analysis

Performance varies substantially across labels (Table 2). Seven climate-adaptive heritage management strategy labels achieve Cross-encoder MRR of 1 with varying MAP scores, indicating that for these concepts the annotated chunks consistently rank highly. For *Maritime management*, the Cross-encoder misses Hit@1 (with MRR of 0.50, i.e. the first positive chunk appears at rank 2), but its sufficiently high MAP of 0.50 indicates that both of its positive chunks are still ranked highly. At the other extreme, five labels score Cross-encoder MRR lower than 0.06, suggesting either ambiguous concept definitions or insufficient label-specific content in the MPs. Combined with Table A1, it is evident that the frequency of concepts is not the main constraint for retrieval systems to perform.

Notably, system rankings sometimes invert for individual labels. BM25 achieves MRR of 1 on *Restoration*, where the ground-truth chunk contains the keyword directly. Dense + Rerank, however, scores 0.02, and the Cross-encoder recovers only to 0.50 by de-prioritising it in favour of semantically adjacent but lexically distinct chunks. Conversely, *Maritime management* yields Bi-encoder MRR of 1 against BM25's 0.10, a case where semantic similarity succeeds and keyword overlap fails.

Performance also varies substantially across the 10 evaluated MPs (see Table A2 for basic statistics and Table A3 for detailed MAP scores per metric), as shown in Figure 2. *Schokland* ($\mathcal{H}_7$) is the easiest property for the three systems other than the Cross-encoder (BM25 MAP of 0.39, Bi-encoder and Dense + Rerank MAP of 0.37). It is the smallest MP (28 Docling chunks) and its content closely matches the vocabulary of the relevant strategy labels. *Rietveld Schröder House* ($\mathcal{H}_{10}$) achieves the highest Cross-encoder MAP (0.35) despite moderate BM25 performance (0.25), suggesting its content is semantically well-aligned with label definitions but lexically distinct. *Van Nellefabriek* ($\mathcal{H}_6$) is the most challenging property for BM25 and the Cross-encoder (MAP of 0.10 and 0.09 respectively), as it covers only 4 evaluable labels and its industrial heritage vocabulary diverges from the climate-adaptive management strategy definitions. The Bi-encoder scores lowest on *Colonies* (0.06) and Dense + Rerank on *Amsterdam* (0.07). *Amsterdam Canal Ring* ($\mathcal{H}_5$) and *Dutch Water Defence Lines* ($\mathcal{H}_3$) show below-average performance across all systems, indicating that large, topic-diverse MPs dilute the label-specific signal.

Label	MRR	Hit@3	MAP
Top 7 (Cross-encoder)			
Soil subsidence	1.00	1	0.54
Limited heritage inclusion	1.00	1	0.26
Climate change	1.00	1	0.25
Management variation	1.00	1	0.22
Human activities	1.00	1	0.17
Fisheries management	1.00	1	0.17
Stakeholder collaboration	1.00	1	0.07
Maritime management	0.50	1	0.50
Bottom 5 (Cross-encoder)			
Legislation	0.06	0	0.05
Research	0.05	0	0.02
Prohibition	0.01	0	0.00
Coherence	0.01	0	0.01
Research gap	0.01	0	0.01

Table 2: Best and worst performing labels under the generally best-performing Cross-encoder system.

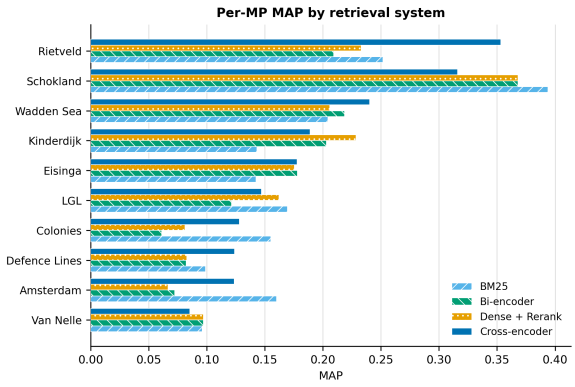


Figure 2: Per-MP MAP for all systems. MPs are sorted by the scores on Cross-encoder MAP.

3.3 Ablation Study on Chunk Sizes

We ablate chunk size while holding all other parameters fixed. As shown in Figure 3, all four systems

are evaluated at each configuration. BM25 and Dense + Rerank trend upward with chunk size. The Cross-encoder peaks at chunk size of 1,384 (MRR of 0.54) and then declines slightly at 3,000, while the bi-encoder improves gradually. At chunk size of 3,000, BM25 reaches MRR of 0.46 and Dense + Rerank reaches 0.42, substantially above the default configuration. However, larger chunks mechanically reduce the candidate pool (904 chunks at 3,000 vs. 2,881 at 940) and pack more text per chunk, making it easier to contain ground-truth sentences, conflating retrieval quality with corpus granularity. This confound also affects any *relative* quantity computed over chunks, including alignment match rates, since larger chunks absorb more labelled sentences by construction. Such ratios are thus not comparable across chunk sizes. We treat the ablation as a sensitivity analysis over independently re-chunked corpora, and retain the 1,000-character configuration for all primary results.

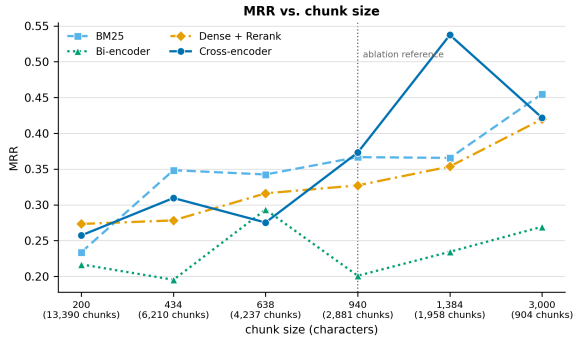


Figure 3: MRR computed over changing chunk sizes for all systems. The dotted line marks the 940-character grid point, the ablation's closest configuration to the 1,000-character setting used for all primary results. Chunk counts are given under each grid point, since chunk size and corpus granularity vary together.

3.4 Relevance of Unlabelled Chunks

The Cross-encoder assigns a real-valued score to every query-chunk pair. Figure 4 shows the distribution of these scores separately for ground-truth positive chunks and all other chunks, aggregated across all labels. Positive chunks cluster at higher scores while negatives span the full range, but the distributions overlap substantially, reflecting the recall gap established in Section 2.3.2. Many high-scoring, unlabelled chunks can be genuinely relevant. The weak but significant Spearman correlation ($r = -0.259$, $p = 0.030$) between Cross-encoder rank and LLM relevance score for Bucket B chunks indicates that score order carries some signal in the unlabelled region while being far from a restatement of the retrieval score itself.

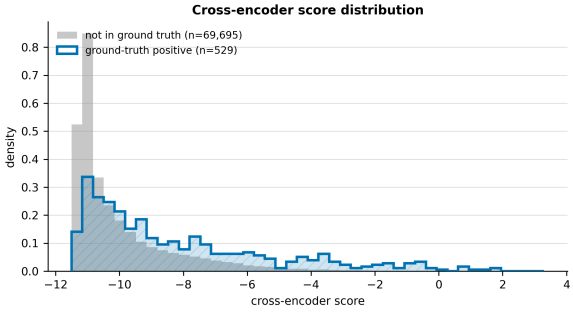


Figure 4: Cross-encoder score distribution for ground-truth positive chunks (GT) and all other chunks, aggregated across 24 labels. Distributions overlap because genuinely relevant but unannotated chunks score similarly to annotated positives.

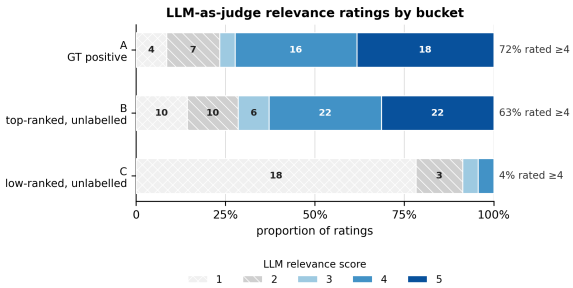


Figure 5: LLM relevance ratings by sampling bucket, shown as the proportion of ratings at each of the five discrete score levels (counts printed inside the bars, $n = 140$). Bucket B (top-ranked, unlabelled) is rated at least 4 almost as often as Bucket A (ground-truth positives), 62.9% vs. 72.3%, while Bucket C (low-ranked) is rated at least 4 only 4.3% of the time.

Figure 5 shows the full distribution of relevance scores by bucket, while the statistical effects are recorded in Table A4. High scores for Bucket A (ground-truth positives, $n = 47$) confirm that the labelled chunks are relevant, while significantly lower scores for Bucket C (low-ranked, unlabelled, $n = 23$) confirm that low-ranked chunks are genuinely irrelevant. Crucially, the scores for Bucket B (top-ranked, unlabelled, $n = 70$) are nearly as high as Bucket A, with substantial overlap in distribution and no statistically significant difference. Notably, 62.9% of top-ranked unlabelled chunks score at least 4 (relevant or highly relevant) by LLM judgment. These results demonstrate that a retrieval strategy based on a fixed-K cutoff (e.g., Hit@K) would fail to identify a large amount of relevant

content that the system correctly surfaces. The high relevance scores of top-ranked, unlabelled chunks (Bucket B) suggest that the Cross-encoder provides a strong signal for relevance that extends beyond the items in the original ground truth. This supports a future thresholding-based approach to identify the complete set of relevant chunks per document, as the gap between reported rank-based metrics (Cross-encoder MRR of 0.45) and true retrieval quality is structural, arising from the sparse labelling in $\mathcal{D}_1$, not from system failure. Table A5 in the Appendix further demonstrates a few qualitative examples of sentences from all three buckets.

4 Discussion

4.1 Reflection on the Retrieval Systems

Label names for climate-adaptive heritage management strategies are domain-specific but often literal, e.g., *Soil subsidence*, *Fisheries management*, *Stakeholder collaboration*. These terms and their definitions appear verbatim in the MPs, making surface-form overlap a genuine retrieval signal rather than a trivial one. BM25's overall MRR (0.32) exceeds the Bi-encoder (0.23), demonstrating that the vocabulary gap between heritage practice and climate science is partially bridgeable without complex semantic reasoning. This finding aligns with the observation in Schimanski et al. (2024) that lexical baselines remain competitive in specialised climate domains where terminology is consistent.

The Bi-encoder treats the query and each chunk independently, and then ranks them by cosine similarity. For broad labels such as *Coherence* or *Conservation*, the embeddings of the query would capture general semantic neighbourhoods that are not discriminative enough to separate on-topic MP passages from the background information. Without in-domain fine-tuning, the embedding space conflates semantically adjacent but task-irrelevant content. The Bi-encoder's independent encoding also misses the interaction between a label definition and specific chunk content, which Cross-encoder handles by scoring query-chunk pairs jointly.

Dense + Rerank passes only the top-50 Bi-encoder candidates, so a positive outside that window can never be scored. This is decisive for the six concepts whose positives never survive the pre-filter (Table A1). Elsewhere, the pre-filter can help by discarding distractors, and Dense + Rerank matches or exceeds the Cross-encoder on 18 of the 24 labels: the macro-averaged difference (+0.09 MRR) is driven by four labels and is not distinguishable from zero by the label-level bootstrap. The trade-off for using Cross-encoder is inference cost: with a fixed set of P' concept queries, exhaustive scoring is $O(P'N)$, i.e. linear in the number of chunks but with a large constant, since every chunk is passed through the cross-encoder for every query. At a small $N = 2{,}926$, full joint scoring is still feasible, but larger collections would require pre-filtering, which re-introduces the candidate truncation problem identified here. Increasing the number of pre-filter candidates may help balance the two, and could be further ablated.

4.2 Failure Cases based on Problem Definition

The worst-performing labels split into two distinct failure modes that require different remedies. *Research gap*, *Prohibition*, and *Maritime Management* are concepts that are rare or absent in the MPs, suggesting **corpus sparsity**. Improvement requires either more documents or augmentation with synthetic samples. Meanwhile, *Coherence* is a broad term with **concept ambiguity** that appears throughout the MPs in many contexts, making it difficult to distinguish relevant from irrelevant chunks. Improvement requires more precise query specification, definition enrichment, or query expansion, thus not necessarily a better retrieval model. Furthermore, the **recall gap** mentioned in Section 2.3.2 is a structural constraint in our problem definition. The LLM-as-judge analysis reveals that 62.9% of top-ranked, unlabelled chunks are independently judged relevant. This finding demonstrates that the Cross-encoder's relevance scores are meaningful and can be used to identify a set of relevant chunks much larger than the original ground truth. Relying on rank-based metrics like MRR or a fixed top-K retrieval would arbitrarily exclude this content. The ground truth $\mathcal{D}_1$ was constructed as a labelling instrument rather than an exhaustive relevance annotation. Therefore, instead of viewing this as a limitation for retrieval evaluation, we see it as motivation for our future task formulation: using the system's scores to determine for each concept-MP pair how many relevant chunks exist without forcing a uniform top-K retrieval. The high quality of the top-ranked unlabelled chunks suggests that a thresholding approach on the cross-encoder scores could be a viable method for achieving this (Zheng et al., 2023). Iterative human-in-the-loop reviews and updates on the labelled set of ground truth can benefit further optimisation of retrieval systems.

4.3 Limitations and Future Directions

The main objective of this report was to build a pipeline that can correctly analyse at scale World Heritage MPs and rank relevant chunks to a set of pre-defined climate-adaptive heritage management labels. Hall and Ram (2016) have emphasised the importance of identifying more effective ways to communicate the value of heritage in relation to climate adaptation, arguing that clearer articulation of heritage value would increase its prominence in climate change documentation such as IPCC reports. The findings of this research represent an initial step towards this ultimate objective.

Several limitations bound the generalisability of these results. The ground-truth labels have high precision but inherently low recall by design, and only $P' = 24$ labels are currently evaluated. Yet even the original $P = 27$ labels from Cheang et al. (2025a) would also only be a subset of the 915-concept full taxonomy provided by IPCC. Connecting the two jargon-compact domains of heritage management and CCA remain a challenge. In the current paper, only a small corpus (10 MPs, 2,926 chunks) in the Netherlands is evaluated. All queries are built with English label definitions applied to MPs written in English, ignoring the Dutch-language MP from *Beemster*. Cross-lingual mismatch may suppress recall for passages that are more fluently expressed in Dutch. Results may also not generalise well to other heritage corpora, other languages, or larger document collections of MPs for WH properties across the globe. Furthermore, all four retrieval systems operate in a zero-shot setting. Fine-tuning on in-domain label-chunk pairs may improve all metrics and may change the relative ordering of systems. No inter-annotator agreement nor cross-system consistency for chunk-level relevance was evaluated.

The retrieval quality could be improved by taking actions in a few directions. The Bi-encoder could be fine-tuned with in-domain knowledge on (label definition, positive chunk) pairs derived from the aligned dataset $\mathcal{D}_2'$, using contrastive losses with hard negatives sampled from Bucket B pairs that the LLM judges as irrelevant. This would address the vocabulary gap between label definitions and MP text that limits zero-shot embedding performance. Replacing the general-purpose encoder models in this study with a climate-domain encoder, such as ClimateBERT (Webersinke et al., 2021) continued-pretrained on climate and heritage policy text, may improve semantic alignment for retrieval at the nexus of heritage management and CCA. Label definitions could be augmented with synonyms, related terms, or example sentences drawn from $\mathcal{D}_1$, providing richer query signal particularly for labels with few matched sentences and reducing chances for false positives. As stated earlier, the retrieval system should optimally incorporate an adaptive threshold for each label-MP pair. However, operationalising a threshold requires deciding which score separates the relevant from the irrelevant. Practical approaches can include: (i) a global percentile cutoff (e.g., top-5% of scores per label), which adapts to per-label score range differences; (ii) a fixed score threshold calibrated against the distribution of known-positive scores; or (iii) using LLM-as-judge scores to calibrate the threshold on a small sample, then applying it to the full corpus. Alternatively, (ii) and (iii) can be combined by selecting a threshold that maximises agreement with LLM judge scores on a calibration sample, then applying it corpus-wide to produce a variable-size relevant set per label-MP pair rather than a fixed top-K. Once a reliable score threshold is calibrated using LLM-as-judge feedback, the pipeline can be applied beyond the 10 Dutch MPs analysed here to the broader corpus of WH management plans worldwide, enabling scalable monitoring of CCA uptake in heritage governance.

5 Conclusions

Being at the nexus of heritage management and climate governance, we formulate an information retrieval problem with a benchmark dataset to map climate change adaptation concepts in World Heritage management documents. Four retrieval systems are evaluated against expert-annotated labels, offering insights on where to further improve both the retrieval methods and the ground-truth annotation. With a ranking approach, human experts can be potentially aided to efficiently prioritise what to annotate. The study prepares for future studies in retrieving and evaluating climate adaptation strategies developed in heritage sector at scale.

Acknowledgments

This study is supported by the Climate Action Seed Fund at TU Delft. We thank the TU Delft TULIP team for providing access to the on-premises LLM infrastructure used in the evaluation. We thank the anonymous reviewers for their useful feedback.

References

Christoph Auer, Maksym Lysak, Ahmed Nassar, Michele Dolfi, Nikola Stanisljevic, Kateryna Abramova, and Peter Staar. 2024. Docling technical report. *arXiv preprint arXiv:2408.09869*.

Nan Bai, Kai Ut Cheang, and Ana Pereira Roders. 2025. Bridging world heritage management with climate change adaptation in the netherlands using artificial intelligence. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 48:79–86.

Jetske Bonenkamp, Robbert Biesbroek, and Ioannis N. Athanasiadis. 2025. Transforming adaptation tracking: benchmarking transformer-based NLP approaches to retrieve adaptation-relevant information from climate policy text. In *Proceedings of the 2nd Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2025)*, pages 266–277, Vienna, Austria. Association for Computational Linguistics.

Konstantinos Chatziandreou and Wolfgang Haupt. 2026. Integrating cultural heritage and climate policy? insights from unesco world heritage cities. *European Planning Studies*, pages 1–23.

Kai Cheang, Nan Bai, and Ana Pereira Roders. 2025a. Unveiling climate-adaptive world heritage management strategies: The netherlands as a case study. *Sustainability*, 17(12):5555.

Kai Cheang, Nan Bai, and Ana Pereira Roders. 2025b. Data from: 'unveiling climate-adaptive world heritage management strategies: the netherlands as a case study'.

Yang Chen, Dayang Wang, Luchen Zhang, Chongxiao Wang, Cheng Sun, and Qi Dong. 2025a. World heritage documents reveal persistent gaps between climate awareness and local action. *Nature Climate Change*, pages 1–7.

Zihua Chen, Qian Gao, Yajing Wu, Jiaxin Li, Xiaowei Li, Xiao Li, Zhenbo Wang, and Haiyang Cui. 2025b. World cultural heritage sites are under climate stress and no emissions mitigation pathways can uniformly protect them. *Communications Earth & Environment*, 6(1):628.

Maarten De Rijke, Bart Van Den Hurk, Flora Salim, Alaa Al Khourdajie, Nan Bai, Renato Calzone, Declan Curran, Getnet Demil, Lesley Frew, Noah Gießing, and 1 others. 2025. Report on the 1st workshop on information retrieval for climate impact (manila24) at sigir 2024. In *ACM SIGIR Forum*, volume 59, pages 1–23. ACM New York, NY, USA.

Sandra Fatorić and Robbert Biesbroek. 2020. Adapting cultural heritage to climate change impacts in the netherlands: barriers, interdependencies, and strategies for overcoming them. *Climatic Change*, 162(2):301–320.

Paloma Guzman. 2025. Cultural heritage in climate planning: An analysis of the norwegian national climate documents and guidelines. *Journal of Cultural Heritage*, 74:35–47.

C Michael Hall and Yael Ram. 2016. Heritage in the intergovernmental panel on climate change assessment reports: a lexical assessment. *Journal of Heritage Tourism*, 11(1):96–104.

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*, pages 3982–3992.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: BM25 and beyond. In *Foundations and Trends in Information Retrieval*, volume 3, pages 333–389. Now Publishers Inc.

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*, pages 17509–17524.

Elena Sesana, Alexandre S Gagnon, Chiara Ciantelli, JoAnn Cassar, and John J Hughes. 2021. Climate change impacts on cultural heritage: A literature review. *Wiley Interdisciplinary Reviews: Climate Change*, 12(4):e710.

Anne J Sietsma, Emily Theokritoff, Robbert Biesbroek, Iván Villaverde Canosa, Adelle Thomas, Max Callaghan, Jan C Minx, and James D Ford. 2024. Machine learning evidence map reveals global differences in adaptation action. *One Earth*, 7(2):280–292.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. MPNet: Masked and permuted pre-training for language understanding. In *Advances in Neural Information Processing Systems*, volume 33, pages 16857–16867.

UNESCO. 2023. Policy document on climate action for world heritage. Technical report, UNESCO.

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

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P Xing, and 1 others. 2023. Judging LLM-as-a-Judge with MT-Bench and chatbot arena. In *Advances in Neural Information Processing Systems*, volume 36.

A Appendix

A.1 Pre-processing of Datasets

Table A1 shows the number of labelled sentences per label in $\mathcal{D}_1$, excluding the MP of *Beemster* ($\mathcal{H}_9$). Table A2 gives an overview of basic statistics for all the MPs from the Netherlands being used as the dataset in this paper.

Concept	Count	#Pairs	R@50
Adaptation	16	14	0.36
Assessment	26 (27)	22	0.14
Climate change	65 (66)	42	0.33
Coherence	28 (29)	25	0.00
Conservation	63 (64)	53	0.00
*Environmental variation**	5	—	—
Fisheries management	10	7	1.00
Framework	55 (56)	44	0.14
Greenery	19	8	0.25
Heritage inclusion	33 (34)	27	0.11
Human activities	28	12	0.17
Knowledge transfer	96 (97)	63	0.06
Legislation	52 (54)	35	0.17
Limited heritage inclusion	4	4	0.00
Management variation	5	5	0.00
Maritime management	2	2	1.00
Monitoring	45 (47)	32	0.06
Multi-level management	19	18	0.28
*Preservation**	1	—	—
Prohibition	6	6	0.00
Research	32 (33)	23	0.13
Research gap	1	1	0.00
Restoration	5	5	0.20
*Restricted greenery**	1	—	—
Soil subsidence	4	2	1.00
Stakeholder collaboration	52 (53)	39	0.13
Water management	58	40	0.17
Total / macro-avg.	731 (744)	529	0.24

*Excluded from evaluation

Table A1: Labelled sentences per concept in $\mathcal{D}_1$ (excl. Beemster), the resulting positive label-chunk pairs, and the share of those positives that survive the Bi-encoder's top-50 pre-filter. Count gives unique annotated sentences, with the number of sentence-chunk rows in parentheses where a sentence spans two overlapping chunks. #Pairs is the number of unique positive pairs after deduplication, which is the quantity retrieval is evaluated against. R@50 is recall of those positives within the Bi-encoder's top-50 candidates. The definition used for each concept to formulate the queries can be found in our GitHub repository.

To clarify the steps taken from 763 labelled sentences to 529 query-chunk pairs, we describe the procedure as follows. The source file derived from Cheang et al. (2025b) records corresponds to 763 distinct annotated sentences, of which 731 are matched onto the Docling chunks (Table A1) and 724 remain after excluding the three labels without expert definitions. These 724 annotations across $P' = 24$ labels that have at least one matched sentence yield 737 sentence-chunk rows, because 13 sentences each span two overlapping chunks. Deduplicating by (label, chunk) then gives the 529 unique positive pairs used as "ground truth".

A.2 Extended Results

The R@50 reported in Table A1 shows what share of the known positives for each concept appears within the Bi-encoder's top-50 candidates, i.e. the fraction that the Dense + Rerank pre-filter passes on to the cross-encoder. For six concepts, no positive survives the pre-filter, so the reranker in Dense + Rerank can never score one. This illustrates our hypothesis on why Dense + Rerank performs worse in Section 4.1. The exact effect of this is still open to further ablation experiments

Table A3 gives the detailed results of retrieval systems performed on each of the MPs.

A one-way analysis of variance (ANOVA) was conducted to examine differences in LLM-as-judge scores across the three buckets. The results revealed a statistically significant effect of group, $F(2, 137) = 29.44$, $p < .001$, $\eta^2 = .30$, indicating a large effect. Post-hoc comparisons are presented in Table A4. Bucket C ($M = 1.35$, $SD = 0.78$) scored significantly lower than both Bucket A ($M = 3.79$, $SD = 1.33$) and Bucket B ($M = 3.51$, $SD = 1.43$), both $p < .001$. The difference between Bucket A and Bucket B was not statistically significant ($p = .295$), supporting the results from Section 3.4. We note that the 1–5 judgements are ordinal and nested within labels, whereas this analysis treats them as independent continuous observations; the bucket separation is large enough that this simplification is unlikely to change the qualitative conclusion, but a mixed-effects ordinal model would be more appropriate.

ID	Name of the World Heritage MP	Short Name	#Chunks	#Labels	#Lab. Chunks	#Pairs	Year
$\mathcal{H}_0$	*Colonies of Benevolence*	*Colonies*	847	11	30	43	2020
$\mathcal{H}_1$	*Frontiers of the Roman Empire – The Lower German Limes*	*LGL*	313	12	32	45	2021
$\mathcal{H}_2$	*Eise Eisinga Planetarium in Franeker*	*Eisinga*	334	10	32	43	2021
$\mathcal{H}_3$	*Dutch Water Defence Lines*	*Defence Lines*	525	13	14	32	2018
$\mathcal{H}_4$	*The Wadden Sea*	*Wadden Sea*	203	21	125	233	2023
$\mathcal{H}_5$	*The seventeenth-century canal ring area of Amsterdam inside the Singel-gracht*	*Amsterdam*	338	13	27	41	2023
$\mathcal{H}_6$	*Van Nellefabriek*	*Van Nelle*	48	4	5	6	2021
$\mathcal{H}_7$	*Schokland and surroundings*	*Schokland*	28	8	17	34	2019
$\mathcal{H}_8$	*Mill Network at Kinderdijk-Elshout*	*Kinderdijk*	143	13	19	28	2022
$\mathcal{H}_9$	*Droogmakerij de Beemster*	*Beemster*	—	—	—	—	2022
$\mathcal{H}_{10}$	*Rietveld Schröder House*	*Rietveld*	147	10	16	24	2024
Total (excl. $\mathcal{H}_9$)			**2,926**	**24**	**317**	**529**	

Table A2: The 11 Dutch World Heritage MPs used in this study. #Chunks: Docling text chunks (with the default setting of chunk_size= 1,000, overlap= 200); #Labels: distinct strategy labels annotated in that MP (from $\mathcal{D}_1$) being evaluated; #Lab. Chunks: number of chunks that get assigned at least one label; #Pairs: unique positive label-chunk pairs via alignment ψ. Total #Pairs= 529 corresponds to the ground truth used in retrieval evaluation. $\mathcal{H}_9$ (Beemster) is excluded because it is the only Dutch-language MP. The links to all the PDFs of the MPs for the Dutch WH properties are provided as hyperlinks, accessed on 22 September, 2026. The years when the current collection of the MPs is published are also provided, as a reference to potential research on temporal development of policy corpora across timeframes.

MP	BM25	Bi-enc	D+Rnk	CE
Rietveld	0.25	0.21	0.23	**0.35**
Schokland	**0.39**	**0.37**	**0.37**	0.32
Wadden Sea	0.20	0.22	0.21	0.24
Kinderdijk	0.14	0.20	0.23	0.19
Eisinga	0.14	0.18	0.18	0.18
LGL	0.17	0.12	0.16	0.15
Colonies	0.16	0.06	0.08	0.13
Defence Lines	0.10	0.08	0.08	0.12
Amsterdam	0.16	0.07	0.07	0.12
Van Nelle	0.10	0.10	0.10	0.09

Table A3: Per-MP MAP score by system (sorted by Cross-encoder MAP descending). Bi-enc means Bi-encoder; D+Rnk means Dense + Rerank; and CE means Cross-encoder.

Comparison	ΔM	t	p	g
A vs. B	0.27	1.05	.295	0.19
A vs. C	2.44	9.64	$< .001$	2.04
B vs. C	2.17	9.20	$< .001$	1.65

Table A4: Pairwise post-hoc comparisons (Welch's t-test) for the LLM-as-judge scores across the buckets. g shows the Hedges' g effect size.

A.3 LLM-as-Judge Prompt and Results

When making the buckets, 4 samples are missing from the intended 2 from Bucket A, 3 from Bucket B, and 1 from Bucket C. Two labels fall short of this intention: *Research gap* has only one ground-truth pair in total and therefore contributes a single Bucket A item, and *Water management* is missing two Bucket B items and one Bucket C item due to dropped TULIP API calls.

The LLM judge is run at temperature 0 with the following template, instantiated per label-chunk pair. Chunk text is truncated to 1,500 characters. Responses are parsed as JSON; the `reason` field is retained for qualitative inspection but not used in any quantitative result. Table A5 shows qualitative examples of the LLM-as-judge evaluation across three sampling buckets for two climate-adaptive management strategy labels together with the LLM reasoning.

```
You are an expert in heritage management and
climate adaptation.

Heritage concept label: {label}
Definition: {definition}

Text chunk:
"""{chunk_text}"""

Task: Rate how relevant the text chunk is to
the heritage concept label above, on a scale of
1-5:
1 = not relevant
2 = slightly relevant
3 = moderately relevant
4 = relevant
5 = highly relevant

Respond with ONLY JSON, no other text: {"score":
<1-5>, "reason": "<one short sentence>"}
```

Bucket	Label	Score	Chunk excerpt	LLM reasoning
A (labelled with ground truth)	*Adaptation*	5	*"Integrate and mainstream the Trilateral Climate Change Adaptation Strategy into all activity fields... Continue monitoring the Trilateral Climate Change Adaptation Strategy and embed the results in long-term trilateral climate change policies."*	Explicitly implements and monitors a Climate Change Adaptation Strategy.
B (unlabelled top-ranked)	*Adaptation*	5	*"Climate change adaptation requires the ability to anticipate sharp fluctuations in water supply and deal with heat stress. To this end, dykes are strengthened and water retention sites built, including in the World Heritage Site."*	Describes specific adaptation measures; not in ground truth, yet clearly relevant.
B (unlabelled top-ranked)	*Assessment*	5	*"assessed in terms of recognisability and authenticity: 1. Attribute possesses authenticity and integrity, and gives a good and perceptible impression of how the attribute looked in the historical context in the period concerned. 2. Attribute illustrates what it was historically, but the historical context has not been well preserved. 3. At[tribute...]"*	The text explicitly outlines a structured assessment framework to evaluate the authenticity and recognisability of heritage attributes.
B (unlabelled top-ranked)	*Assessment*	4	*"Spacial quality advisory team (as a complement to the ICOMOS guidelines), have drawn up recommendations regarding situations in which a Heritage Impact Assessment is desirable, and on formulating the terms of reference of future Heritage Impact Assessments and how to implement them. Over recent years, an extensive network of participating part[ners...]"*	The text explicitly mentions the use and implementation of Heritage Impact Assessments to evaluate impacts on the site.
C (unlabelled low-ranked)	*Adaptation*	1	*"The roads are damaged by speeding cars as well as by heavy (agricultural) traffic. Walkers, cyclists and horse riders sometimes come into conflict with the motorised traffic."*	Traffic management only; no climate content.

Table A5: Qualitative examples of LLM-as-judge evaluation for two labels. Bucket A shows a confirmed ground-truth positive (score 5). Bucket B shows a few high-ranked chunks not in the ground truth that the LLM judges equally relevant (score 5 or 4) as potential candidates of false negatives. Bucket C shows a low-ranked chunk the LLM correctly judges irrelevant (score 1). The column of Chunk excerpt is subsets of the complete 1000-character chunks selected by the evaluating LLM, and the column of LLM reasoning records its judgement criteria.