

School of Engineering and Computer Science
Te Kura Mātai Pūkaha, Pūrorohiko

Retaining answer conditions in Retrieval Augmented Generation

Julian Smith

Supervisor: Andrew Lensen

AIML 589 Research Project

Master of Artificial Intelligence

October 2026

Disclosure: ChatGPT was used to assist with language polishing and paraphrasing to improve the clarity and readability of this report. All AI-assisted content was reviewed and verified by the author, who takes full responsibility for the final content.

Contents

Abstract 5
1. Introduction 6
1.1 Public sector motivation 6
1.2 The research problem 6
2. Related work and study positioning 7
3. Research question and objectives 8
4. Methodology 9
4.1 Data selection and preparation 10
4.2 Data splits and scored questions 11
4.3 Preparing passages (Component 1) 12
4.4 Searching and ranking passages (Component 2) 13
4.4.1 Supplied document search 14
4.4.2 Whole collection search 14
4.4.3 BM25, dense and hybrid ranking 15
4.5 Reranking search results (Component 3) 15
4.6 Expanding passages (Component 4) 16
4.7 Packing selected text (Component 5) 17
4.8 Evaluating condition retention (Component 6) 18
4.8.1 Deciding when a condition was retained 18
4.8.2 Complete-question counts and secondary measures 19
4.9 How settings were chosen 19
4.9.1 Passage and BM25 settings selected on the training set 20
4.9.2 Word allowances 21
4.9.3 Fusion, diversity and expansion settings 21
4.10 How the results were tested 22
4.10.1 Pre-specified and exploratory analyses 22
4.10.2 Allowing for chance across documents 22
4.10.3 Training-set replication 23
4.11 Implementation and resource constraints 24

4.12 Matched comparisons and component summary 24
5. Results 27
5.1 Reading the results 27
5.2 Configuration comparisons at 1,002 words 27
5.2.1 Passage preparation 27
5.2.2 Search and ranking 28
5.2.3 Reranking 28
5.2.4 Expansion 29
5.3 Effect of word allowance 30
5.4 Explaining observed omissions 31
5.4.1 Reference document present, conditions missing 31
5.4.2 Saved cases 32
5.5 Replication on the training set 33
6. Analysis and discussion 35
6.1 Packs with reference-document text can still miss conditions 35
6.2 How passages are prepared decides what can be retrieved 35
6.3 Dense and hybrid retrieval help most when documents compete 36
6.4 Reranking did not improve condition retention 37
6.5 What a larger allowance adds and what it costs 38
6.6 Implications for practice 38
6.7 Limitations and how they affect the conclusions 40
6.8 Why human review is still needed 41
6.9 Contribution 42
7. Conclusion 43
References 45
Appendix A Supporting results 49
A.1 Training selection 49
A.2 Supplied document results at initial allowances 50
A.3 Supplied document results at larger allowances 51
A.4 Whole collection results at initial allowances 52

A.5 Whole collection results at larger allowances 53
A.6 Changes from 15,000 to 20,000 words 54
A.7 Changes from cross-encoder reranking 55
A.8 Changes in the heading follow-up 56
A.9 Training-set replication with the supplied document 56
A.10 Training-set replication across the whole collection 57
A.11 Training-set replication across the whole collection at the other allowances 58
A.12 Source elements under dense and hybrid ranking (exploratory) 59
Appendix B Implementation details 61
B.1 Data, query and passages 61
B.2 Ranking, reranking and expansion 61
B.3 Packing and scoring 62
B.4 Settings, statistics and replication 63
B.5 Software and reproducibility 63
B.6 Supporting results 64
B.7 Exploratory checks used in Section 6 64

Abstract

A widow asks whether she qualifies for Bereavement Support Payment. The answer depends on several conditions, and one is that she was “under State Pension age” when her husband died. The four words share nothing with her question, so a retrieval system could leave them out while its answer still cites the right page. New Zealand public service guidance advises staff to check AI-generated outputs for missing context [1], and a dropped answer condition (a rule deciding whether an answer applies) is one form of it. This study tested which ways of selecting text keep such conditions in the evidence pack, the text a language model receives within a word allowance. It used 63 development questions from ConditionalQA, a dataset built on UK government guidance, and compared searching all 652 documents with searching only the document that holds each answer.

Finding the right document was not enough. When all 652 documents were searched with a 1,002-word allowance, hybrid retrieval, which combines keyword and meaning-based ranking, put text from the right document in the pack for 61 of 63 questions but kept every condition for only 50. Of the six comparisons re-tested on the training set, how documents were split into passages mattered most, as 121-word passages kept more conditions than single paragraphs or list items, especially when other documents competed. Ranking by meaning helped, but ranking for variety could push the right document’s conditions out, and larger packs helped only partly because much of the extra text came from other documents. All six comparisons met the replication criterion on 550 training-set questions, generally with smaller effects. This re-test had to reuse the training set because the test set has no public condition annotations, and the training set had already been used to choose settings, so it was not fully independent. So a citation to the right document does not show that the evidence was complete. Agencies should test whether their systems keep the conditions that matter in their own search setting, and reviewers should check answers against the full applicable source document.

1. Introduction

1.1 Public sector motivation

New Zealand's public service guidance on generative AI and accountability advises staff to check AI-generated outputs for missing context before publication [1]. One form of missing context is a condition the answer depends on, such as an age limit, that never reaches the language model. This study looks at retrieval, the step that picks which source text the model sees, and asks how the way text is selected and how much text is allowed affect whether such conditions get through.

New Zealand agencies are now trialling and adopting AI systems [2], [3], [4], and some use Retrieval Augmented Generation (RAG), in which the language model answers from passages retrieved from the agency's own documents [5] and cites the documents it used. Verifiability studies check whether a cited page supports each claim [6], but even a correct citation does not show whether the model received all the context it needed to interpret an answer. This is one reason human oversight is required [1], and it means a reviewer has to open each citation and read the text around it. Can the retrieval step itself keep more of the conditions?

1.2 The research problem

This study concerns answer conditions, the conditions in a document that determine whether an answer applies [7]. The answer conditions marked by human annotators (Section 4.1) are called "conditions" from here on. Conditional answers are common in complex documents such as UK government guidance [7]. For example in question dev-129 (Figure 4.2) a widow's claim for Bereavement Support Payment depends partly on her being "under State Pension age" when her husband died. This condition is a short list item that shares no word with her question so retrieval could drop it. In RAG, documents are first split into passages, which are ranked by how well they match the search request, called the query, before selected text is assembled into the final evidence pack (hereafter the pack) [8] and passed to the language model for generation. In this study a word allowance limits the pack's length and cost. A condition may be lost while the pack is assembled if passage preparation separates it from the text that matches the query or if its passage is not retrieved. A retrieved passage may also be left out of the pack because of its rank and the word allowance. The study contributes a word-level measure of condition retention that needs no language-model judge. It shows that a citation to the right page can hide a missing condition and that testing with the document supplied can understate effects across a collection (Section 6.9).

2. Related work and study positioning

ConditionalQA by Sun et al. is a dataset of questions whose answers may apply only under certain conditions. Its baseline models were given the source document, the question and a user scenario, and were asked to find the answers and their conditions [7]. This study instead looks at the earlier evidence selection stage and asks whether the conditions reach the evidence pack. Open-domain question answering studies score whether the answer text appears in the top passages [9], and chunking studies score retrieval and answer quality, but neither asks whether answer conditions reached the model in full [8], [10]. RAG evaluation comes closest. RAGChecker claim recall [11] and RAGAS [12] context recall [13] use a language model to check whether retrieved context covers each claim in a reference answer. Joren et al. use one to judge whether context is sufficient [14]. This study checks without a judge whether each annotated condition reached the pack word for word in two search settings.

Yuan et al. study language model reasoning over *conflicting* web references that had already been retrieved [15]. Work on each retrieval component is discussed where that component is introduced (Sections 4.3–4.6).

3. Research question and objectives

The study addressed one research question.

> *How do retrieval configurations and word allowances affect the retention of annotated answer conditions in the final evidence pack?*

The aim was to explain why conditions were retained or lost. The study measured the final evidence pack because only the selected source text would reach the language model. Stopping before answer generation separates retrieval omissions from errors the model might make when using the evidence. A condition was retained only if its source text was in the pack in full. A retrieval configuration is one combination of the methods and settings used at each step from passage preparation to packing. The question led to two objectives. One was to compare configurations with matched controls (Section 4.12) at the same word allowance to find which retained the most conditions. The other was to test how retention changed as the word allowance increased. The configuration and packing rule were held fixed so the larger pack was the only change.

The comparisons were run in two search settings. Supplied document search covered only the reference document to see if conditions were still missed when the document was already known. Whole collection search was closer to typical RAG use and searched all 652 source documents together so passages from different documents competed for the word allowance. In both settings, retention was measured with condition recall (the average share of conditions kept) and the number of questions with all conditions retained.

4. Methodology

The methodology has six components. The first five broadly follow the steps of a typical RAG workflow [8] with reranking and expansion tested as optional steps, and the sixth measures condition retention. Passage preparation determines which words are ranked together and ranking and reranking determine their priority. Expansion adds nearby text and packing decides which words fit the allowance. Evaluation then checks whether every annotated word of each condition survived. Separating the stages lets a change be compared with the same system without it, although some changes have more than one effect (Section 4.12).

Sections 4.1–4.2 describe the data, Sections 4.3–4.8 explain the six components, and Sections 4.9–4.10 justify the settings and tests, with Figure 4.1 showing the dataflow and Table 4.3 summarising the components. A choice is pre-specified when it was fixed before any development result was seen, and Section 4.10.1 lists which analyses were pre-specified and which were exploratory.

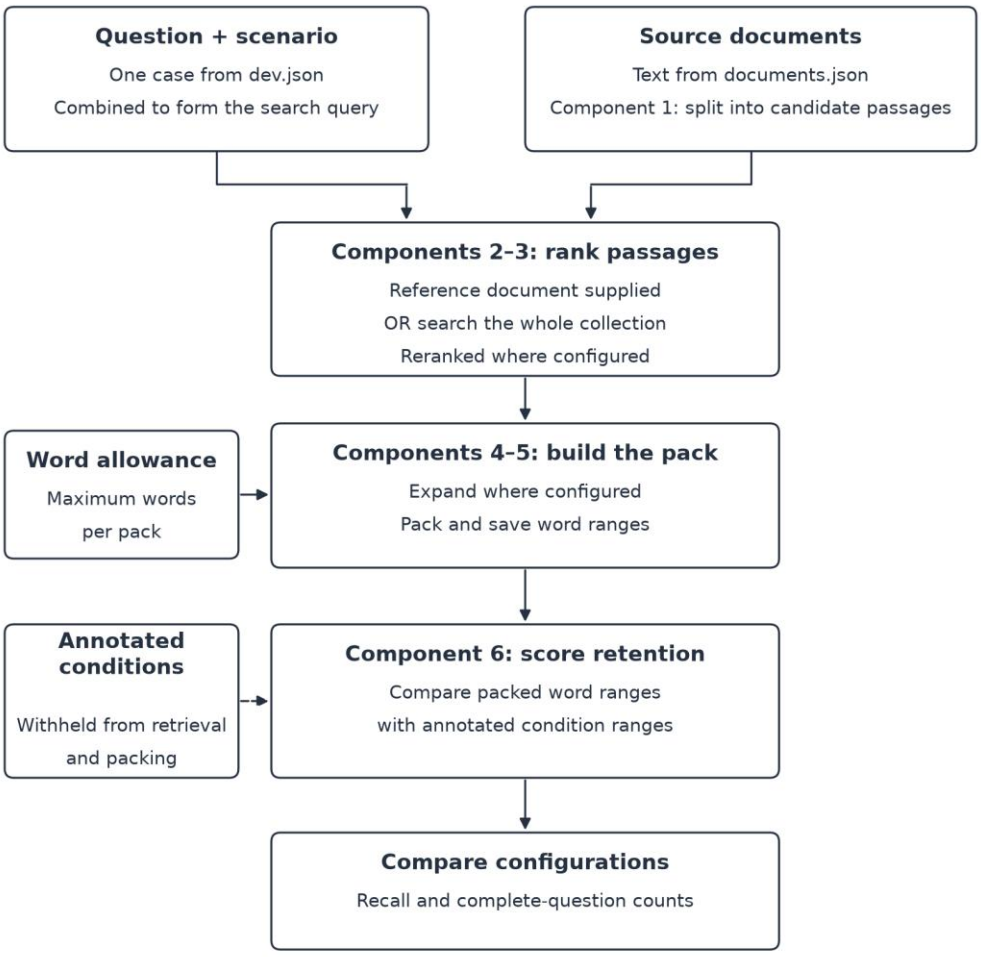

Figure 4.1. Dataflow for a development question through the six components (Sections 4.3–4.8).

4.1 Data selection and preparation

ConditionalQA was chosen because each question comes with a user scenario and a reference document. Annotators marked each answer condition as an exact HTML element (mostly a paragraph or list item) in that document so a pack can be checked word for word without a model judging it. The documents are UK government guidance where answers often depend on eligibility rules. In this study a source document is one of the 652 archived webpages in the dataset's fixed collection. The split by source document let settings be chosen on the training set and tested on the development set without training a model (Section 4.2).

Each question record also holds a URL identifying its reference document, the source document from which its reference answer is derived, and annotations comprising reference answers and answer conditions. ConditionalQA defines a conditional answer as an answer plus a list of conditions, and the answer is only true if its conditions apply [7]. Annotators selected only the requirements the scenario did not already settle [7], so all listed conditions had to reach the pack. Figure 4.2 shows the dev-129 question record from Section 1.2 and its reference document.

ConditionalQA also marks supporting excerpts that back the answer more broadly and stores them in the "evidences" field [7]. This study scored only the answer conditions because a missing condition is what can change whether an answer applies.

The reference answers, conditions and supporting excerpts were withheld from ranking and packing. The question and scenario were combined to form the query because ConditionalQA annotates conditions relative to the scenario, which can already meet some of the requirements [7]. The scenario can also supply the matching words (Section 6.6). Condition wording remained searchable in the source documents because masking it would have meant using the annotations to change what was searched.

A Question record: query input

dev.json · dev-129 · lines 2647–2648

```
2647  "scenario": "My husband died after an accident at work. He was only
      52. I am heartbroken and now strapped for money.",
2648  "question": "Do I qualify for Bereavement Support payment?",
```

The question and scenario formed the query, with the question placed first.

B Reference document: searchable text

documents.json · Bereavement Support Payment · contents excerpt

```
6434  "contents": [
6435    "<h1>Eligibility</h1>",
6436    "<p>You may be able to get Bereavement Support Payment (BSP) if
      your husband, wife or civil partner died in the last 21
      months.</p>",
      ... [five intervening source elements omitted] ...
6442    "<p>When they died you must have been:</p>",
6443    "<li>under State Pension age</li>",
6444    "<li>living in the UK or a country that pays bereavement
      benefits</li>",
```

HTML tags were removed before the source text was divided into passages.

C Question record: evaluation annotations

dev.json · dev-129 · lines 2650–2659

```
2650  "answers": [
2651    [
2652      "yes",
2653      [
2654        "<p>You may be able to get Bereavement Support Payment (BSP)
      if your husband, wife or civil partner died in the last 21
      months.</p>",
2655        "<li>under State Pension age</li>",
2656        "<li>living in the UK or a country that pays bereavement
      benefits</li>"
2657      ]
2658    ]
2659  ],
```

Annotated answer conditions were withheld from ranking and packing.

Figure 4.2. Example dataset text for dev-129, a development-set question. Panel A shows the question record (question and scenario). Panel B shows the reference document's searchable text with the paragraph "When they died you must have been:" followed by the separate four-word list item "under State Pension age", which has no word in the query. Panel C shows the question record annotations, which were used only to score the pack.

4.2 Data splits and scored questions

The training set was used to select passage and BM25 (Best Matching 25, a keyword matching ranking method [16], Section 4.4.3) settings (Section 4.9.1) and the development set was used to

compare configurations. The training set was later used again to re-test six comparisons (Section 4.10.3).

Reference documents did not overlap between the training and development sets (436 training, 59 development and 136 test plus 21 source documents with no linked question). The 804 test questions were not scored because they do not have publicly available annotated answer conditions. Table 4.1 summarises the training and development sets.

Table 4.1. Dataset composition by split, showing question records and reference documents including questions with annotated answer conditions.

What was counted	Training set	Development set
All question records	2,338	285
Questions with annotated answer conditions	550	63
Answerable questions without annotated answer conditions	1,697	208
Unanswerable questions	91	14
Reference documents	436	59
Reference documents with annotated answer conditions	307	34

Only questions with at least one annotated answer condition were scored because a question with no conditions would count as full retention without testing anything. That left 63 development questions from 34 reference documents and 550 training questions from 307 (Table 4.1).

4.3 Preparing passages (Component 1)

The experiment limited how much source text could enter each evidence pack, so the 652 documents were split into smaller passages that could be ranked and selected within that limit. In RAG this step is called chunking and the resulting units are chunks [8], [10]. This study uses the terms passage preparation and passages instead, with “passage” covering both types, because one of the two passage types tested keeps the document’s own paragraphs and list items whatever their length.

Passage length sets a trade-off. Longer passages can keep a condition with the text that matched the query but use more of the word allowance. Wang et al. found that changing passage size affected faithfulness and relevancy scores in a RAG evaluation using a Lyft report [8]. To test the trade-off for conditions, this study compared two passage types with opposite boundary rules.

Fixed passages, the common baseline [10] and the control here, cut the text by word count regardless of the document’s structure. They were windows over the continuous text including

heading text. Successive windows overlapped so text near a passage edge had another chance to appear with useful context at the cost of more indexed text. The overlap was built into the passages before ranking whereas neighbour expansion added nearby source elements after ranking (Section 4.6). Length and overlap were selected using condition recall on the training set (Section 4.9.1). The resulting setting of windows up to 121 words long with 50% overlap was pre-specified. Later ranking methods used the same passages to avoid changing passage length and ranking method together.

Source elements follow the document's structure and use each paragraph or list item as a passage, which preserves the source boundaries without a separate boundary detection model. This also matches how ConditionalQA annotates conditions (as whole HTML elements, Section 4.1) so a passage edge cannot cut a condition. But a source element can separate a short condition from the text that matches the query. For example in dev-129 "When they died you must have been:" was a paragraph whereas "under State Pension age" was a separate four-word list item (Figure 4.2, Panel B). The BM25 comparison of the two types was pre-specified and Section 5.2.1 reports its results.

To track which source words were in the pack, the system recorded the document identity and assigned each word a position (Appendix B.1). The positions let packing skip words that overlapping passages repeated and let scoring check whether a condition was retained in full (Section 4.8.1).

The original source elements left out headings so a condition present only in a heading could not be retained, which affected 3 of the 181 development conditions (23 of 1,391 on training). Fixed passages included heading text and the original design left this difference untested, so an exploratory follow-up let headings be ranked and used as neighbours (Section 4.6) with nothing else changed (Appendix B.1 and Section 5.2.1).

Semantic chunking, another passage preparation method that uses semantic similarity to determine chunk boundaries, was not tested within the local-CPU scope (Section 4.11). Qu et al. had also found that it did not consistently outperform fixed size passages in retrieval or answer generation despite the additional computational cost [10].

4.4 Searching and ranking passages (Component 2)

The collection contained more text than the allowance could hold so passages were scored for relevance to the query and ranked. An annotated condition can be missed in two ways. The pack

may contain no passages from the reference document or it may contain some of that document's passages but not the passage with the annotated condition.

Supplied document search was compared with whole collection search to distinguish reaching the reference document's passages from retaining its conditions.

4.4.1 Supplied document search

Supplied document search restricted ranking to passages from the reference document. It could represent a user who names the document they want searched.

If the whole document were packed, every condition would be retained but half the training documents were longer than the 1,002-word main allowance (Section 4.9.2). Because no other documents competed, any omission would be due to passage preparation, ranking, expansion or packing within the word allowance.

4.4.2 Whole collection search

Whole collection search tests whether conditions survive competition from other documents. All 652 documents (including those without a scored question) were ranked together so the 8 fixed passages from dev-129's reference document competed with 13,681 from the other 651 documents. Because packing followed rank order, passages from other documents could fill the allowance before a condition-bearing passage was reached. That text may still be useful. But the measure credits only reference-document text because the conditions were annotated only there and similar wording elsewhere may apply to a different policy. For example the dev-129 condition "under State Pension age" is also a list item in the Carer's Credit guidance where it limits eligibility for a National Insurance credit.

No step chose a document first so a pack could hold text from several documents. A document-first step could reduce this competition but would add another retrieval stage and a wrong choice would lose all the conditions. Supplied document search is close to the best case of such a step (Section 6.6).

The search settings differed in one way besides competition. Dense retrieval compared each passage with the query and used no statistics from other passages, so its scores were the same in both settings. But BM25 gives more weight to query words that are rare among the searched passages and its word statistics came only from the passages searched in each setting (Section 4.4.3). So BM25 could order the reference document's passages differently in the two settings, although the stage trace suggests this mattered little (Section 6.3).

4.4.3 BM25, dense and hybrid ranking

A condition may not use the same words as the query. In dev-129 (Figure 4.2) the query says the husband “died after an accident at work” and “was only 52” yet the condition “under State Pension age” shares no word with the query, so the link is one of meaning.

BM25 can rank a condition with no query word only when its passage has a query word elsewhere. BM25 ranking by shared words (keyword matching) was tested against dense retrieval ranking by meaning [9] and hybrid retrieval combining the two [17].

All ranking methods used the query as given (Section 4.1) since rewriting would add generated text and another source of variation.

BM25 scored a passage by summing a weight for each distinct query word it contained (Appendix B.2). It was used as the keyword matching baseline to show how many conditions can be retained by matching query words with no meaning-based ranking. It suits that role because the BEIR benchmark, a collection of varied retrieval tasks, uses BM25 as a strong zero-shot baseline (one used without training on the target task) [18].

Dense retrieval represents the query and each passage separately as lists of numbers called vectors and compares the vectors to estimate similarity in meaning without relying only on shared words. This study used BAAI/bge-small-en-v1.5 [19], a general-purpose English model small enough to run on a local CPU (Section 4.11). It was used as released and the passages were ranked by cosine similarity to the query (Appendix B.1).

Hybrid retrieval combined the BM25 and dense rankings to determine whether they were complementary. It used reciprocal rank fusion [20], which works on rank positions so the two scores do not need to be scaled. Each ranking adds 1/(60 + rank) to a passage’s score so top ranks gain little advantage. A passage ranked 5th by both methods then outscores one ranked 1st by BM25 and 20th by dense retrieval. Because the dense ranking was fused, passages missed by BM25 remained eligible. Bruch et al. found that a tuned convex combination of scores outperformed reciprocal rank fusion [17] so other fusion rules may retain more conditions.

4.5 Reranking search results (Component 3)

The initial ranking may not place the most useful passages first so two rerankers each reordered the top of a ranking before packing. Each reranker kept the remainder of its ranking in order and was compared with that ranking so any difference came from the reordering itself.

A cross-encoder [21] reads the query and a passage together to score their relevance whereas dense retrieval encodes them separately. BEIR found that such rerankers were among the best zero-shot methods but slow because each of the top 100 passages was rescored together with the query [18]. The cross-encoder tested whether joint query-passage scoring would place condition-bearing passages early enough in hybrid rankings to fit the word allowance. Hybrid rankings were used because they narrowly led across the collection in the development results (Table 5.2). The six-layer model, cross-encoder/ms-marco-MiniLM-L6-v2, was chosen for its compactness, which suited the local CPU setup (Section 4.11). Its reported scores were almost identical to the twelve-layer model's [22]. The first 100 hybrid candidates were sorted by cross-encoder score, a depth that limited the scoring cost and still held most of the conditions that hybrid packs missed across the collection at 1,002 words (Appendix B.2).

But the model was trained for Microsoft Machine Reading Comprehension (MS MARCO) [23] passage relevance [22], in which a passage is relevant if it answers a web search query. Since a condition may not look like an answer, the model may rank answer passages above condition passages. The test asked whether a relevance reranker used as released would still improve retention.

Overlapping fixed passages share up to 60 words so the top of a dense ranking can hold near-copies of the same text. The near-copies could push passages from elsewhere in the reference document out of the pack. Maximal marginal relevance (MMR) picked passages one at a time by balancing relevance to the query against similarity to passages already selected [24]. The relevance weight λ sets this balance and larger values favour relevance [24] (full score in Appendix B.2). MMR needs only the dense vectors already computed so it reranked the dense rankings and tested whether repeated content crowded conditions out. Reducing repetition might reveal conditions elsewhere in the reference document. MMR might instead favour other source documents or demote a passage that is similar to one already selected but has a different condition.

4.6 Expanding passages (Component 4)

A passage may match the query while the condition limiting the answer is adjacent to that passage. In dev-129 the age condition comes straight after a paragraph that shares the query word "died". So expansion added nearby source text after ranking.

For each BM25-ranked source element, one or two-neighbour expansion added up to one or two neighbours (adjacent source elements in the same section) on each side. A section is the text under

one heading up to the next heading. Neighbours needed no query word and were packed straight after their ranked element, closest first. Expansion stayed within the section because a new heading usually starts another topic. But expansion reached only conditions within two elements of a ranked element so a condition farther away was retained only if ranked directly. Each neighbour also used allowance that could otherwise hold a lower-ranked condition, as in dev-217 (Section 5.4.2).

Neighbour expansion resembles the small-to-big method Wang et al. tested, which retrieves small passages and then gives the model the larger text around them [8].

Boundary completion tested a related failure in which a fixed passage cuts off a condition at its edge. As with the cross-encoder, it used hybrid-ranked fixed passages and extended each to the boundaries of the source elements it intersected.

Boundary completion also adds text so its control was an equal-length centred span. It matched the completed passage's length and was centred on the fixed passage within the document ignoring element boundaries. If this control matched or beat boundary completion any gain would more likely come from extra text than from alignment.

4.7 Packing selected text (Component 5)

The ranked passages often held more text than the word allowance could accommodate so packing had to decide what fitted. Summarising with a language model would fit more text but was not used because it would change the words the model receives and could smooth away the conditions being measured.

Packing took the ranked passages in order and copied their words unchanged but skipped any word position already in the pack. Skipping mattered because overlapping fixed passages share up to 60 words and copying them twice would have wasted allowance. Each configuration used one untuned rule and no separate selection model.

For traceability each fragment (a run of consecutive words newly copied from one passage) had its own source label identifying its document and word range. Because the labels would also reach the language model, each label used two words of the allowance. That penalised source elements with their many short fragments but repacking without labels changed the gap between passage types by at most 1.5 percentage points at 1,002 words (Appendix B.3).

If a passage did not fit, the code kept the first words that fitted and stopped. Skipping ahead to a shorter passage would have added a length-based selection rule, and a larger allowance could then drop text that a smaller one kept. The cost of this rule is that a condition can be cut off at the end.

The allowance was counted in words so all configurations had the same maximum. With BM25 ranking, a top-10 passage limit would have given development packs across the collection about 170 words of source elements against about 980 of fixed passages.

4.8 Evaluating condition retention (Component 6)

4.8.1 Deciding when a condition was retained

A pack could hold most of a condition but miss the words that limit who qualifies. In dev-129 “State Pension age” without “under” would no longer say whether the widow qualifies. So a condition counted as retained only when the pack held all of its annotated words. The check used the condition’s word positions (Section 4.3). The words could arrive in several fragments and in any order (Appendix B.3). When the same wording appeared more than once in a document any one complete occurrence was enough because the annotations do not say which was meant. The widow’s saved pack (BM25 source elements at 1,002 words with the document supplied) held her other 2 conditions but missed all four age words (Section 5.4.2).

Some questions (12 of the 63 development questions) have more than one correct answer, each with its own conditions. In dev-38 (“Am I eligible for Gift Aid?”) the answer is yes under 3 conditions and no under 5 others. A pack that kept only the yes conditions would support half an answer. Pooling all conditions into one set would have let an answer with more conditions outweigh one with fewer (5 against 3 in dev-38). Scoring only the best-covered answer would have hidden a missing answer. So each reference answer and its conditions formed an answer group and all groups were scored. Answers without conditions had nothing to retain and were not scored.

A group could also hold the same condition in different words. The pre-specified rule removed only exact duplicate HTML strings within a group because merging differently worded versions would need subjective judgements and would change how many conditions each group had. One condition can then count twice (dev-224, septic tank, Section 5.4.2).

Each answer group scored the share of its conditions retained. The scores were averaged within each question, then within each reference document and then equally across documents so a

document with many questions or a question with many answer groups could not outweigh the rest.

$$R = \frac{1}{|D|}\sum_{d\in D}\frac{1}{|Q_d|}\sum_{q\in Q_d}\frac{1}{|A_q|}\sum_{a\in A_q}\frac{\sum_{c\in C_{qa}} I(c\ retained)}{|C_{qa}|}.$$

D is a set of scored reference documents, Q_d their scored questions, A_q the question's answer groups with at least one condition and C_qa each answer group's conditions. R is condition recall.

Condition recall is literal, so the model could receive a rule that it does not credit. The guidance can state one rule twice, as the septic tank page does (both versions were annotated in dev-224, Section 5.4.2), or state it again on another page. A pack can also lose one word of a condition where two fragments meet. So literal recall may understate retention, while a complete source element could still need nearby text to make sense.

4.8.2 Complete-question counts and secondary measures

Condition recall can overstate practical success because it gives partial credit when a pack retains only some of the conditions required for a question. So a question was classified as complete only when the pack retained every annotated condition in full. In the widow example, the pack retained two of three conditions and achieved 66.7% condition recall, but the question was still incomplete.

Three secondary measures were used to explain why complete-question performance changed. Source presence recorded whether the final pack contained any text from the question's reference document. Source presence combined with an incomplete question meant that the pack included the document but omitted at least one required condition. The widow example was such a case, so its citation could appear correct even though the pack was incomplete. Mean pack length, including source labels, indicated how fully the word allowance was used. Questions gained and lost were also reported separately because the same net change can conceal both improvements and regressions.

4.9 How settings were chosen

The settings were chosen in three ways because they served different purposes. Passage length, overlap and BM25 parameters directly affect retrieval, so they were tuned on the training set using condition recall. Word allowances represent resource limits, so the initial values came from quartiles of training-document length and later allowances were labelled exploratory. Fusion, diversity, reranking and expansion settings defined the mechanisms being compared, so tuning

them on the development questions would have biased those comparisons. So they were fixed using published values, neutral baselines, matched candidate depths or the smallest non-zero expansion.

4.9.1 Passage and BM25 settings selected on the training set

Passage length and overlap decide whether a condition shares a passage with text that matches the query, as with the dev-129 age condition and the query word "died" (Sections 4.3 and 4.6). BM25 settings then determine which passages rank high enough to be packed.

Candidate passage lengths were the quartiles of the training section lengths (36, 58 and 87 words) so they matched the scale of the source documents, and each length was tested with no overlap (the control), 25% overlap and 50% overlap. BM25 has two settings, where k1 limits the credit for a query word repeated within one passage and b adjusts scores for passage length (Appendix B.4). The grid crossed the values used by two popular libraries, Pyserini [25] (k1 = 0.9 and b = 0.4) and rank_bm25 [26] (1.5 and 0.75), which gave four BM25 pairs.

Each candidate was scored in the supplied document search, which isolates passage preparation and ranking. Its training-set condition recall was averaged across the 687-, 1,002- and 1,551-word allowances to avoid tuning to one. For each BM25 pair the best fixed-passage score and the source-element score were averaged with equal weight so the winning pair had to suit both passage types used later (Section 5.2.1). In the first grid the longest length tested (87 words) scored best so the best length might lie beyond it. A one time pre-specified extension added 23- and 121-word lengths (the 10th and 90th percentiles). Allowing one extension stopped the grid being widened repeatedly to find a preferred result. The winning configuration (121-word fixed passages with 50% overlap, k1 = 1.5 and b = 0.75) was fixed before any development-set comparison and reused unchanged in the whole collection search.

Passage length mattered most. Condition recall rose from 75% at 23 words to 90% at 121 words so keeping a condition with its nearby text was worth fitting fewer passages into the allowance. The winning 121-word length was again the longest tested, which suggests longer passages may be better (stability check in Appendix B.4). At 121 words, 50% overlap added a negligible 0.6 percentage points (hereafter "points") for roughly double the indexed text, but it was kept because the pre-specified rule considered cost only for exact ties. The BM25 pair made a negligible difference (0.2 points).

4.9.2 Word allowances

Longer packs cost more to process and to review (Section 6.5) so each pack had a word allowance. The allowance covered the copied source words and their source labels but excluded the question and scenario, which all configurations shared. An allowance is a maximum and packing stopped when no eligible text remained, while a larger allowance only extended the smaller pack so condition recall could not fall.

The allowances were counted in words because no answers were being generated and counting a generator's tokens would have tied the allowance to one answer-generation model. The initial allowances of 687, 1,002 and 1,551 words were the quartiles of training document length and were pre-specified to fit documents of different sizes, with the median of 1,002 words as the main allowance for Sections 5 and 6. A pre-specified 7,733-word BM25 check, sized to the longest training document, tested whether retention stopped increasing once whole documents fitted.

The larger allowances of 4,000, 10,000, 15,000 and 20,000 words were added in turn after earlier development results were inspected so they are exploratory (Section 4.10.1). The 4,000-word run, about three times the average source document (about 1,310 words), tested whether extra room helped expansion configurations more than their matched controls. Later values tested whether retention kept increasing or levelled off (Section 5.3). The 20,000-word allowance was declared final before its results were seen. No larger allowance was needed because fixed passages with the document supplied retained all development conditions at 4,000 words, and across the collection most added text came from other documents.

4.9.3 Fusion, diversity and expansion settings

No settings were tuned on development results so each comparison tested a mechanism at fixed settings. Equal fusion weights were used because no dense development result existed when they were set. The fusion constant was set to 60, which Cormack et al. fixed in a pilot where 60 and 80 performed almost equally [20].

MMR used two values of the relevance weight λ, both from Goldstein and Carbonell's example [24]. The main setting $\lambda = 0.7$ prioritised relevance while $\lambda = 0.3$ tested stronger diversity. A depth of 100 (the number of top candidates reordered) matched the cross-encoder so both rerankers worked on the same number of candidates. Depths of 50 and 200 at $\lambda = 0.7$ tested whether halving or doubling it mattered. The alternatives only checked sensitivity. Whole collection recall at 1,002 words increased as λ approached 1, where MMR reproduced dense ranking (Table 5.3), so tuning λ would have removed the diversity contrast being tested.

One and two-neighbour expansion, the two smallest non-zero amounts, were pre-specified for comparison with no expansion (Section 5.2.4).

4.10 How the results were tested

4.10.1 Pre-specified and exploratory analyses

An analysis is exploratory when results already seen could have influenced its design. Only the original BM25 passage-preparation and neighbour-expansion comparisons were fixed before the development results were seen. Dense, hybrid, MMR, cross-encoder and boundary-completion comparisons, and the four larger allowances (4,000–20,000 words), were added later. The heading follow-up, pack-composition measures (Appendix B.3), query-word overlap check and saved cases were also later additions. The later analyses reused the same development questions, so their patterns need confirmation and cannot count as independent tests. Further exploratory checks of section grouping, dense and hybrid source elements, the MMR penalty between overlapping passages, the supplied document ceiling, passage length, query wording, collection size, document choice and document length are reported in Section 6 where they inform the discussion (Appendix B.7).

The 18 MMR and boundary-completion comparisons were declared before their runs, and the cross-encoder settings were fixed before reranker scoring. This limited changes within the experiments but did not remove the influence of earlier development results. Six predictions were also recorded in a local replication protocol, whose limits Section 4.10.3 explains.

The four saved development cases in Section 5.4.2 were inspected by hand to see how the packs retained or lost particular conditions. The cases were chosen after the results were known and illustrate possible explanations.

A score cannot show at which step a condition was lost, so a retrospective stage trace of the saved rankings recorded where each condition's passage ranked, how far packing reached and how many other-document passages sat above it. The trace reproduced the saved scores exactly and checked the explanations in Sections 6.2–6.4.

4.10.2 Allowing for chance across documents

The development set contained only 34 reference documents, so estimated differences could be sensitive to which documents were included. A 95% confidence interval summarises this sampling uncertainty. An interval that includes zero does not establish a difference or prove that the methods are equivalent. Questions from the same document tend to succeed or fail together so

uncertainty was estimated by resampling documents. Each of 5,000 resamples drew 34 documents with replacement (one document could appear more than once and another not at all) and recalculated the paired mean difference. The middle 95% of the resampled differences (Appendix B.4) formed the paired document-bootstrap interval [27]. “Paired” means that both configurations were scored on the same resampled documents, so the comparison used each document as its own control.

The 18 declared development contrasts repeated three comparisons in both search settings at the three initial allowances. They compared MMR with dense fixed passages. Boundary completion was compared with hybrid fixed passages and equal-length centred spans. Dense and hybrid gains over BM25 had already been observed so they were tested in the replication instead (Section 4.10.3). Each contrast used a sign-flip test [28], which asks whether the difference could have arisen by chance and does not assume a normal distribution. Under the test’s null assumption, either sign of each document’s paired difference is equally likely. The procedure randomly flipped the signs in 10,000 draws. The p-value was approximately the share of draws whose mean was at least as far from zero as the observed mean (Appendix B.4). In 15 of the 18 comparisons, at most 5 of the 34 documents scored differently. With so few non-zero differences, even consistent gains can provide too little evidence to reach the .05 threshold (Section 5.2.4).

Testing 18 comparisons increases the chance of a false positive, so the p-values were Holm-adjusted [29]. Holm’s method checks the smallest p-value against the strictest threshold, then relaxes the threshold step by step, stopping at the first result that fails. It controls the risk of at least one false positive across the declared comparisons without requiring every test to pass the same strict threshold.

A single document could also drive an effect so each of the 34 documents was omitted in turn (MMR result in Section 5.2.3). Holm adjustment cannot correct for results seen before the comparisons were declared. All intervals are pointwise, covering one comparison each, so none is adjusted for multiple comparisons and some may exclude zero by chance.

4.10.3 Training-set replication

Most development results were exploratory, so six comparisons were re-tested on 550 training-set questions from 307 reference documents, none shared with the development set (Section 5.5). The training set was the only other annotated data because the test questions have no public condition annotations (Section 4.2). A local protocol fixed the statistical test, pass rule and six predictions at 1,002 words before the replication run. The predictions covered dense and hybrid gains over

BM25, passage type, MMR, expansion with the document supplied and whether hybrid retrieval gained more over BM25 across the collection than with the document supplied (the setting gap).

The replication checked different documents but was not a fully independent held-out test, because the training set had already been used to choose passage and BM25 settings. Its supplied-document expansion result had also been calculated before the protocol was fixed, so that prediction only repeated an earlier check, although no other configuration comparison had used the training questions. The protocol was stored locally (Appendix B.4).

With settings unchanged, the code first reproduced the saved development results exactly, and intervals and sign-flip tests were then calculated over the 307 training documents. A prediction met the replication criterion if its estimate was in the predicted direction and its Holm-adjusted p across the six comparisons was below .05. The criterion checks only direction, so Table 5.6 also compares effect sizes. All other comparisons were descriptive, including the cross-encoder and boundary completion.

4.11 Implementation and resource constraints

All experiments were run offline on a single desktop computer with no GPU, and both models were small enough to run on its CPU (Sections 4.4.3 and 4.5). Pinned local copies of both models meant each run used the same model versions. Each saved pack could be rebuilt from its word ranges and separate verifier scripts recalculated each score (Appendix B.5).

4.12 Matched comparisons and component summary

Each comparison changed one component and kept the rest constant so any difference pointed to that component. Configuration comparisons kept the queries, search setting, word allowance, packing rule and scoring the same, and allowance comparisons changed only the allowance. Table 4.2 lists the purpose of each comparison and its matched control. Few pairs were compared because testing all possible pairs would have required many more comparisons and raised the chance of a false positive with 34 development documents. A matched pair may still have more than one effect so each pair was checked for side effects.

For ranking methods the possible side effect was eligibility. BM25 ranks only passages that share a query word whereas dense retrieval ranks all passages. Common words such as “the” (stopwords) were kept so each fixed passage shared a query word in the supplied document search. Across the collection the few passages BM25 left out would have ranked far deeper than any pack (Appendix B.2). So on fixed passages BM25–dense differences come from ranking.

Passage types could not be matched as closely. Even with headings in both (Section 4.3) the types differed in passage length, number, overlap and BM25 statistics so their results are read more cautiously. Expansion also had two effects, because it added nearby text but used allowance that other passages could fill. So the passage-type and expansion results show which configuration retained more conditions but cannot isolate a single mechanism.

Table 4.2. Purpose and matched control for each retrieval configuration.

Configuration	Purpose or question	Matched control
BM25 fixed passages	Establish a reproducible keyword-matching baseline.	Matched control for the source-element and dense retrieval comparisons.
BM25 source elements	Do source elements retain conditions better than fixed passages?	BM25 fixed passages (same BM25 parameters).
Source elements + one or two neighbours	Does nearby text recover conditions that do not independently match the query?	BM25 source elements (same ranking, unexpanded).
Dense fixed passages	Does dense retrieval recover differently worded conditions better than BM25?	BM25 fixed passages (same passages).
Hybrid fixed passages	Does combining BM25 and dense rankings outperform either ranking by itself?	BM25 fixed passages and dense fixed passages, evaluated separately.
Hybrid + cross-encoder	Does joint query–passage scoring improve the order of hybrid candidates?	Hybrid fixed passages (same ranking, not reranked).
Dense + MMR (maximal marginal relevance)	Does diversity reranking expose additional conditions?	Dense fixed passages (same embedding vectors and candidate passages).
Hybrid + boundary completion	Does extending passages to source-element boundaries prevent conditions from being cut off? Is any gain from alignment or from extra text?	Hybrid fixed passages (unexpanded) and equal-length centred spans (same added length).

Table 4.3. Why each of the six experimental components was tested, with its methods and decision timing.

Component	Why it was tested	Methods or settings	When selected or varied
1. Passage preparation	A short passage can separate a condition from the text that matches the query.	Fixed passages (up to 121 words, about 50% overlap) or source elements (paragraphs and list items).	Length and overlap selected on the training set, BM25 passage-type comparison pre-specified and heading follow-up exploratory (Section 5.2.1).
2. Search and ranking	A condition may share no query word and other documents' passages compete with it.	BM25, dense and hybrid retrieval, with supplied document or whole collection search.	BM25 settings selected on the training set, with dense and hybrid exploratory.

Component	Why it was tested	Methods or settings	When selected or varied
3. Reranking	Reordering can move a condition passage into or out of the pack.	Cross-encoder on hybrid results and MMR on dense results.	Exploratory, with cross-encoder settings fixed before reranker scoring and MMR settings not tuned.
4. Expansion	Nearby text can recover an unmatched or cut-off condition but uses allowance that other passages could fill.	Neighbour expansion of source elements, and boundary completion of hybrid fixed passages (with equal-length centred spans as the control).	Neighbour expansion pre-specified, but boundary completion exploratory.
5. Packing	A condition ranked too low to fit or cut off at the end of the pack is lost.	One fixed rule copying ranked text and labels in order and skipping words already packed.	687, 1,002 and 1,551 words and the 7,733-word BM25 check pre-specified, but 4,000–20,000 words exploratory.
6. Evaluation	One missing condition can change whether an answer applies.	Condition recall and complete-question counts (63 development questions and 550 training).	Development results compared configurations, with main findings re-tested on the training set (local protocol, with limitations in Section 4.10.3 and results in Section 5.5).

5. Results

5.1 Reading the results

Section 5.2 compares configurations at 1,002 words, Section 5.3 changes the word allowance, Section 5.4 examines the missed conditions and Section 5.5 repeats the main tests on the training set. Section 5.4.1 reports the central finding that in whole collection search most incomplete questions still had reference-document text in the pack (Section 6.1). Unless labelled otherwise, condition recall and complete-question counts in Sections 5.2–5.4 are based on the 63 development questions with annotated answer conditions from 34 reference documents.

Tables pair supplied document results (Supplied) with whole collection results (Collection), and results are from the original runs unless otherwise indicated. Differences use unrounded scores and square brackets indicate 95% paired document-bootstrap intervals (Section 4.10.2).

5.2 Configuration comparisons at 1,002 words

5.2.1 Passage preparation

With both ranked by BM25, source elements (the document's own paragraphs and list items) retained fewer conditions than 121-word fixed passages and far fewer across the collection (Table 5.1). The exploratory heading follow-up (Section 4.3) changed recall by less than 2 points at 1,002 words (Table A.8) and source elements stayed below BM25 fixed passages at all allowances across the collection, so excluding headings does not explain the gap.

Table 5.1. Passage preparation at 1,002 words. Recall is condition recall (each reference document weighted equally). Complete is the number of complete questions (every annotated condition retained) out of 63.

Configuration	Supplied recall	Supplied complete	Collection recall	Collection complete
BM25 fixed passages	88.6%	52/63	64.2%	35/63
BM25 source elements	85.9%	42/63	23.7%	9/63

With the document supplied, the source elements were close to fixed passages in recall but 10 fewer questions were complete, and across the collection they were far behind. So Section 5.2.2 keeps fixed passages and changes only the ranking method.

5.2.2 Search and ranking

Dense retrieval (ranking by meaning) and hybrid retrieval (dense and BM25 combined) both retained more conditions than BM25 on the same fixed passages, mostly when other documents competed (Table 5.2).

Table 5.2. Ranking fixed passages at 1,002 words. Columns as in Table 5.1.

Configuration	Supplied recall	Supplied complete	Collection recall	Collection complete
BM25 fixed passages	88.6%	52/63	64.2%	35/63
Dense fixed passages	94.7%	57/63	81.1%	47/63
Hybrid fixed passages	93.3%	56/63	82.5%	50/63

In whole collection search, dense and hybrid retrieval improved recall over BM25 by 17.0 points [5.1, 29.9] and 18.4 points [8.8, 29.5] respectively. Hybrid led dense by 1.4 points [−6.8, 8.1] on development and 2.1 [−0.8, 5.2] on training, and both intervals included zero. With hybrid retrieval, 15 questions moved from incomplete to complete and none moved back.

When the document was supplied, the gains were smaller and uncertain at 6.1 points [−0.5, 14.1] for dense retrieval and 4.7 points [−1.9, 12.8] for hybrid retrieval, and both intervals included zero. Table 5.6 tests the hybrid setting gap directly.

5.2.3 Reranking

Reordering the top of a ranking for relevance (cross-encoder) or for diversity (MMR) did not clearly improve retention, and MMR sharply lowered it across the collection (Table 5.3). MMR labels give the relevance weight λ and depth, and Tables A.2 and A.4 add depths 50 and 200.

Table 5.3. Reranking at 1,002 words with dense fixed passages and hybrid fixed passages as controls. Columns as in Table 5.1.

Configuration	Supplied recall	Supplied complete	Collection recall	Collection complete
Dense fixed passages	94.7%	57/63	81.1%	47/63
Dense + MMR 0.7 / 100	94.7%	56/63	56.5%	30/63
Dense + MMR 0.3 / 100	89.4%	51/63	31.4%	14/63
Hybrid fixed passages	93.3%	56/63	82.5%	50/63
Hybrid + cross-encoder	94.4%	56/63	74.8%	41/63

For MMR (λ = 0.7, depth = 100) whole collection recall fell by 24.6 points [−34.7, −15.5] and 17 questions moved from complete to incomplete while none moved back (Holm-adjusted sign-flip p = .0018). The loss persisted with each reference document omitted in turn (−25.4 to −22.4 points)

and with stronger diversity (λ = 0.3) and depths of 50 and 200. The 1,551-word MMR loss was also significant after Holm adjustment but no other declared contrast was (Appendix B.4). For whole collection search, MMR packs used more distinct source documents than dense packs (6.49 against 4.41) and held a smaller share of words from the reference document (31.1% against 46.8%). With the document supplied, MMR at λ = 0.7 changed recall little (Table 5.3).

The cross-encoder did not consistently improve on hybrid fixed passages. Across the collection at 1,002 words, recall decreased by 7.8 points [−19.0, 2.9] and complete questions fell from 50 to 41 with 1 gained and 10 lost. At 1,551 words, recall increased by 2.2 points [−5.0, 10.8] but complete questions fell from 51 to 49 (Table A.7).

5.2.4 Expansion

On the development set at 1,002 words, neighbour expansion recovered conditions beside ranked passages mainly with the document supplied, whereas boundary completion gave at most a small gain (Table 5.4). On the training set two-neighbour expansion gained in both search settings (Section 5.5).

Table 5.4. Expansion at 1,002 words. Columns as in Table 5.1.

Configuration	Supplied recall	Supplied complete	Collection recall	Collection complete
BM25 source elements	85.9%	42/63	23.7%	9/63
Source elements + one neighbour	92.4%	51/63	24.9%	15/63
Source elements + two neighbours	93.1%	53/63	25.0%	15/63
Hybrid fixed passages	93.3%	56/63	82.5%	50/63
Hybrid + boundary completion	93.9%	57/63	83.4%	50/63
Hybrid + equal-length centred spans	92.6%	56/63	81.5%	48/63

With the document supplied, two-neighbour expansion increased recall by 7.1 points [0.8, 13.4] and completed 11 more questions. Across the collection, recall increased from 23.7% to 25.0% with 6 more questions completed, which left neighbour expansion still below BM25 fixed passages at 64.2% (Section 6.2).

Across the collection, boundary completion and hybrid fixed passages had almost the same recall (0.8 points [−0.5, 2.7]) and the complete-question count remained unchanged at 50 with 1 question gained and 1 lost. Boundary completion improved recall over equal-length centred spans by 1.9 points [0.2, 4.4]. Only 4 of the 34 documents changed, all upward, and with so few changes the sign-flip test could not reach significance (p = .128, Holm-adjusted p = 1).

5.3 Effect of word allowance

With the document supplied, retention levelled off once all eligible text fitted, whereas across the collection it generally kept rising (Figure 5.1). Extra room widened neighbour expansion's lead over unexpanded source elements but boundary completion stayed close to its controls.

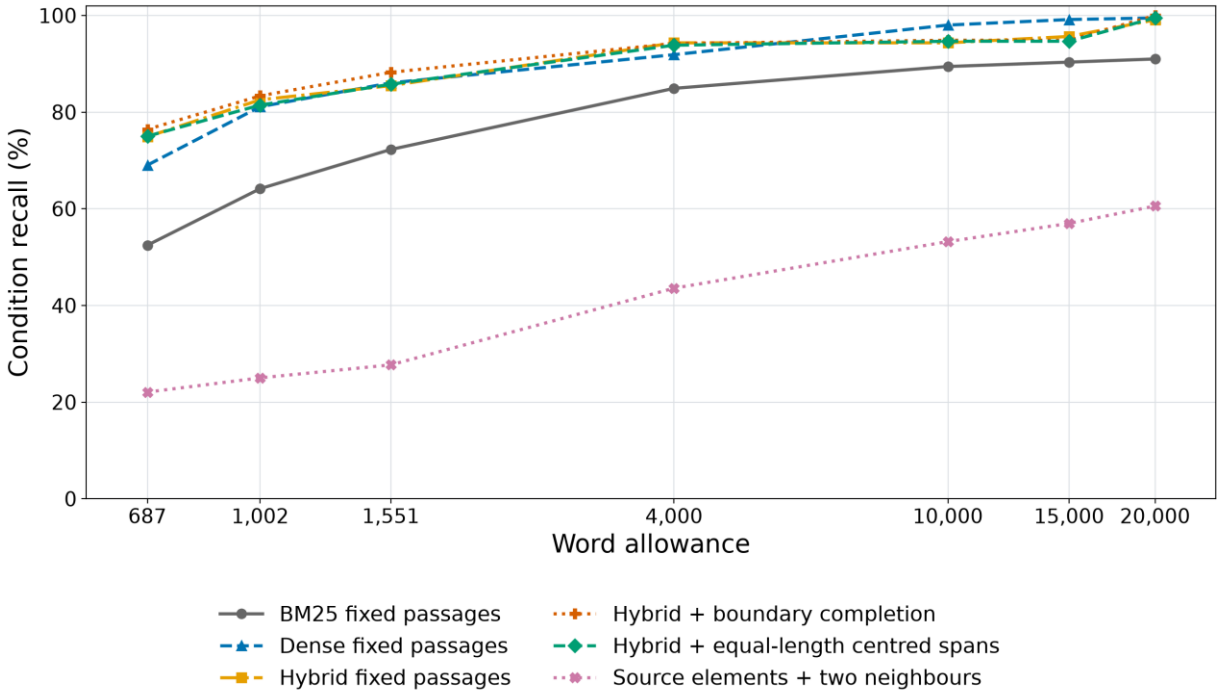

Figure 5.1. Condition recall for six configurations in whole collection search at each of the seven allowances for the 63 development questions (training results are in Tables A.10–A.11). The lines only join the measured points and the horizontal axis is logarithmic.

With the document supplied, fixed passages levelled off at full coverage. From 687 to 1,551 words, each configuration had every condition for at most 61 of 63 questions (Table A.2). From 4,000 to 20,000 words all nine fixed-passage configurations had 100% and 63 of 63 questions (Table A.3). On the training set only MMR ($\lambda = 0.3$) and the cross-encoder were one to three questions short at 4,000 words (Table A.9). Development packs stopped at about 1,340 words because all eligible text had fitted (Table A.3).

Source elements levelled off below 100% at 4,000 words, where unexpanded they reached 93.0% and 49 of 63 questions against 61 with two neighbours (Table A.3). An exploratory query-word overlap check examined why. In 12 questions including dev-129, at least one condition-bearing source element had no word in the query so BM25 could not rank it. In 2 more, dev-217 and dev-227, a condition was a heading, which the original runs excluded. The 14 were exactly the incomplete questions in the 7,733-word BM25 check and on the training set the flagged questions were exactly the 97 incomplete ones at 20,000 words, so no larger allowance could complete

them. On the development set two-neighbour expansion completed the 12 questions whose condition had no query word and left only the 2 heading questions incomplete.

Across the collection the gaps between ranking methods narrowed without closing and the best development configuration changed with the allowance (Table A.5 and Appendix B.6). From 10,000 words, MMR at depths of 50 and 100 matched or came within half a point of dense fixed passages because the packs then held nearly all the passages MMR had reordered. Boundary completion reached 100% and 63 of 63 at 20,000 words whereas BM25 fixed passages were still behind at 91.0% and 55 of 63.

Most of the extra text in larger boundary-completion packs came from other documents (Section 6.5). At 20,000 words the packs of all 12 configurations tested were essentially full (Appendix B.6). On the development set boundary-completion packs held text from about 64 source documents on average against about 4 at 1,002 words. Under 600 of their 19,000 extra words came from the reference document.

Neighbour expansion gained more from extra room than boundary completion did (Tables A.4–A.5). On the development set two-neighbour expansion added 1.3 points over unexpanded source elements at 1,002 words but about 12 at 4,000. At 20,000 words source elements were still far behind BM25 fixed passages, at 45.2% unexpanded and 60.6% with two neighbours (Table A.5). Boundary completion at that allowance was less than 1 point above hybrid fixed passages and equal-length centred spans on both sets (Appendix B.6).

Across the collection, retention was still rising at 20,000 words. The last step from 15,000 to 20,000 words added 5 complete questions for boundary completion and 1 each for dense and hybrid fixed passages (Table A.6).

5.4 Explaining observed omissions

5.4.1 Reference document present, conditions missing

In whole collection search at 1,002 words, most incomplete questions had reference-document text in the pack so finding the reference document did not guarantee retention of its conditions (Table 5.5, Section 6.1).

Table 5.5. Source presence and complete questions in whole collection packs at 1,002 words.

Configuration	Reference document absent	Reference document present, conditions incomplete	Complete questions
BM25 fixed passages	6	22	35
Dense fixed passages	2	14	47
Hybrid fixed passages	2	11	50
Hybrid + boundary completion	2	11	50

With hybrid fixed passages, the packs held reference-document text for 61 of 63 questions but retained every condition for only 50 of 63. Among the 11 packs with reference-document text but missing conditions, 6 had none of the conditions and 5 had some (103 and 84 of 187 on training). On both sets more than half of such packs held none of the question's conditions although an answer from them could still cite the right document. The gap narrowed at 20,000 words. Each development pack then had reference-document text yet only boundary completion retained every condition for all 63 questions (Table A.5). On the training set even boundary completion missed at least one condition in 33 of 550 questions (Table A.10) and 31 of the 33 packs still held reference-document text.

The stage trace showed that 41 of the 45 hybrid misses at 1,002 words were ranked too low to be packed (a median rank of 29 while packing reached about rank 11) and 4 were cut off at the end of the pack.

5.4.2 Saved cases

The saved cases were chosen after the results were known so they are illustrative only. They show that expansion can recover conditions that BM25 cannot rank. In dev-129 (Bereavement Support Payment) with the document supplied at 1,002 words, unexpanded source elements retained 2 of 3 conditions. The missed four-word list item "under State Pension age" had no query word and so no BM25 score. Because the pack used 384 words, the allowance was not the limit. One neighbour expansion added the item beside a matching element and retained all 3 conditions.

Expansion can also crowd out lower-ranked conditions. In dev-217 (VAT) with the document supplied, a condition-bearing source element ranked 32 among BM25 source elements and began "There's no threshold if neither you nor your business is based in the UK". Without expansion it was packed within the 1,002-word allowance and 3 of 4 conditions were retained. With one neighbour expansion and the same ranking, the neighbour preceding the rank-24 element exhausted the allowance. The rank-32 element was not reached so 2 of 4 conditions were retained.

The saved cases also show the limits of literal scoring. In dev-224 (septic tank) one answer group contains two differently worded versions of the same permit condition. One is a paragraph requiring a permit when the general binding rules are not met and the other reads “You’ll need a permit if you do not meet the general binding rules.” Exact-string deduplication keeps both so the condition is counted twice. A pack that retains only one version receives credit for only one.

In dev-46 (coastguard) the annotated condition is “Your local station may have further requirements, for example you may need to live closer to the station”. Because overlapping passages were packed without repeating words, the 1,002-word dense pack holds the condition in three fragments. One ends “for example you may”, another is the word “need” and the third, packed first, begins “to live closer”. Because scoring checks whether all word positions are covered, this condition and the question’s other three are retained in full. But the fragments may make the sentence harder to follow.

5.5 Replication on the training set

All six predictions met the replication criterion on the training questions and documents (Table 5.6). The expansion result was already available before the local protocol was fixed, and the training set had informed settings (Section 4.10.3).

Table 5.6. Development and training results for the predictions in the local protocol at 1,002 words. Values are differences in condition recall in points with 95% paired document-bootstrap intervals. Replicated means Holm-adjusted p < .05 on the training set in the predicted direction. The expansion result was already available (Section 4.10.3).

Prediction (1,002 words)	Development (63 questions, 34 documents)	Training (550 questions, 307 documents)	Replicated
Hybrid fixed passages − BM25 fixed passages (collection)	+18.4 [8.8, 29.5]	+16.0 [12.4, 19.7]	Yes (p < .001)
Dense fixed passages − BM25 fixed passages (collection)	+17.0 [5.1, 29.9]	+13.9 [9.3, 18.2]	Yes (p < .001)
Dense + MMR 0.7 / 100 − dense fixed passages (collection)	−24.6 [−34.7, −15.5]	−12.3 [−15.8, −8.7]	Yes (p < .001)
BM25 fixed passages − BM25 source elements (collection)	+40.5 [26.4, 54.8]	+35.1 [30.6, 39.4]	Yes (p < .001)
Source elements + two neighbours − BM25 source elements (supplied)	+7.1 [0.8, 13.4]	+10.5 [7.4, 13.6]	Yes (p < .001)
Hybrid fixed passages − BM25 fixed passages (collection − supplied)	+13.7 [0.8, 26.6]	+12.8 [8.7, 16.9]	Yes (p < .001)

Four of the six effects were smaller on training (Section 6.7), whereas the expansion gain was larger and the setting gap was about the same. The MMR loss was the only training estimate outside its development interval. It halved because dense fixed passages scored lower on training (67.3% against 81.1%) while MMR was not much different (55.0% against 56.5%).

On the development set neither the expansion gain with the document supplied nor the setting gap was significant after Holm-adjusting across the six ($p = .084$). Both still replicated in the training set partly because 307 documents gave narrower intervals than 34. For the setting gap, hybrid retrieval gained 16.0 points across the collection against 3.2 with the document supplied. The central finding (Section 5.4.1) also held on the training set. Hybrid whole collection packs held reference-document text for 518 of 550 questions but retained every condition for only 331, so 36% of the packs with that text missed a condition against 18% on development.

In the descriptive training comparisons, boundary completion and two-neighbour expansion gave clear gains across the collection at 1,002 words whereas the cross-encoder made no clear difference (Table A.10). The cross-encoder changed recall by 1.2 points [−2.5, 4.8] and complete questions from 331 to 346 (60 gained, 45 lost). Boundary completion gained 3.4 points [1.7, 5.2] over hybrid fixed passages and 1.7 points [0.8, 2.7] over equal-length centred spans. The intervals for the centred-span comparison were not adjusted for the 168 training comparisons but excluded zero at all whole collection allowances, supporting a small advantage associated with alignment. Above 1,551 words the advantage would not be significant after Holm adjustment (Appendix B.6). Across the collection two-neighbour expansion gained 8.6 points [5.3, 11.9] over unexpanded source elements but remained far below BM25 fixed passages (26.9% against 53.4%).

6. Analysis and discussion

Section 5 leaves three findings to explain. Finding the right document did not mean the pack held every condition, and the way documents were split into passages and ranked mattered most when many documents competed, yet adding nearby text or reordering the results did not reliably fill the gaps. Sections 6.1–6.5 look at why and weigh other explanations, using the larger training set for more precise estimates although its earlier use in choosing settings limits its independence (Section 4.10.3). Sections 6.6–6.9 then turn to what this means for practice, the limitations, human review and the contribution.

6.1 Packs with reference-document text can still miss conditions

Finding the correct document did not guarantee that the pack contained all the evidence needed to answer the question. At 1,002 words, 11 of the 61 hybrid packs containing reference-document text still omitted at least one required condition (Table 5.5), and the training share was twice as high (Section 5.5). The problem remained when the reference document was supplied. Each configuration completed at most 57 of the 63 development questions at 1,002 words (Table A.2). This occurs because passages are ranked separately, so one passage can establish source presence while another passage containing a required condition falls below the packing limit.

A correct source or citation is not evidence of a complete pack. In dev-129, for example, the whole collection BM25 source-element pack contained reference-document text at every allowance but missed two of the three required conditions, including the age condition.

6.2 How passages are prepared decides what can be retrieved

The main finding is that the unit used for retrieval affected whether required conditions could be found. Fixed 121-word passages retained far more conditions than source elements, especially when searching the whole collection. On the training set across the collection at 1,002 words, BM25 recall was 53.4% with fixed passages and 18.3% with source elements (Table A.10). Dense retrieval showed the same pattern, 67.3% versus 24.2% (Table A.12), so the difference was not caused only by BM25's reliance on shared words.

The most likely reason is that source elements were ranked in isolation. A short condition may share few or no words with the query and so rank below stronger-looking passages from other documents. A fixed passage keeps the condition with nearby explanatory text that is more likely to match the query. This problem was less severe when the correct document was supplied (85.9% against 23.7% recall, Table 5.1) because the condition competed only with passages from that

document. For BM25 source elements across the collection at 1,002 words, 83.9% of development misses and 86.3% of training misses were rankable but fell below the packing limit. Only 16.1% and 13.5% respectively could not be ranked at all. So document competition was the main source of loss.

An exploratory check supports this explanation. Merging each section's source elements into passages of up to 121 words raised whole collection BM25 recall on the training set at 1,002 words from 18.3% to 46.2%, compared with 53.4% for fixed passages. But the check changed passage length, grouping and heading inclusion together, so it shows that isolated elements were a problem but does not identify which change produced the gain.

Adding neighbouring elements was only a partial remedy. Expansion can add context only after an element has already been selected, so it cannot recover a condition whose surrounding elements also rank too low. It can also use space that would otherwise hold another condition. Extending selected fixed passages to whole source elements gave a small advantage over equal-length centred spans across the collection, clear on the training set but not significant on development (Sections 5.2.4 and 5.5). Passages that kept a condition with its nearby qualifying or explanatory text retained more conditions, whereas ranking each small source element separately lost them.

6.3 Dense and hybrid retrieval help most when documents compete

Dense and hybrid retrieval helped mainly when passages from many documents competed for the limited pack. On the training set, dense retrieval improved condition recall over BM25 by about 14 points in whole collection search but only about 3 points when the reference document was supplied (Tables 5.6 and A.9). This suggests that the main benefit came from filtering competing documents, while reordering passages within a known document added little.

The stage trace supports this explanation. BM25 can rank passages from other documents highly when they share words with the query. Conditions missed by BM25 had a median rank of 37 on development and 56 on training, while packing reached only about rank 11 and most passages above the missed conditions came from other documents. Hybrid retrieval placed fewer other-document passages above 70 of the 80 missed development conditions and 637 of the 713 missed training conditions. BM25 packs also already contained reference-document text for at least 9 of the 15 questions that hybrid retrieval made complete (Table 5.5). So the problem was often that competing passages displaced text containing required conditions although the pack already held reference-document text.

Other explanations fit the evidence less well. BM25's use of word statistics from only the supplied document made little difference when tested in the stage trace. Better ordering within the reference document also cannot explain most of the gain because the whole document already fitted within the 1,002-word pack for 34 of 63 development questions and 225 of 550 training questions, leaving no opportunity for ranking to change retention. After excluding them, hybrid retrieval's extra advantage in whole collection search over supplied document search fell from 13.7 to 1.3 points on development but only from 12.8 to 10.2 points on training (Appendix B.7). So the training results support document competition as the main explanation, while the development evidence is less conclusive.

6.4 Reranking did not improve condition retention

MMR reduced condition retention because its diversity penalty conflicted with the task. In whole collection search, it often treated useful passages from the same reference document as redundant and promoted passages from other documents instead. Overlap was not the main cause. Removing the penalty only between overlapping passages left most of the loss, whereas removing it within each document left no clear loss (Appendix B.7). In the stage trace a median of 18 other-document passages on development and 13 on training ranked above conditions displaced by MMR, compared with 1 under dense retrieval.

When the reference document was supplied, MMR could not promote other documents, so $\lambda = 0.7$ changed recall little. But the stronger diversity setting, $\lambda = 0.3$, still reduced recall by pushing some condition-bearing passages below less relevant passages from the same document (Tables 5.3 and A.9). MMR's aim of reducing repetition [24] is poorly matched to a task that requires retaining every relevant condition, including similar conditions from one document.

The cross-encoder also produced no reliable improvement. Across the collection at 1,002 words it retained more development conditions in total than hybrid retrieval (141 versus 136 of 181), but condition recall fell by 7.8 points because misses in questions with few conditions had greater effect. This loss did not repeat on training, where the 1,002-word interval across the collection included zero (Section 5.5), so it may be due to chance.

The reason for the cross-encoder's lack of gain remains uncertain. Training on short MS MARCO relevance queries may limit transfer to the longer study queries, which averaged about 48 words against about 6 [18].

So neither reranker improved condition retention in this study. Both rank for general relevance or diversity and neither checks whether the final pack contains all required conditions. A

completeness-aware reranker may be more suitable, but it would need evaluation on new annotated questions.

6.5 What a larger allowance adds and what it costs

When the correct document was supplied, increasing the allowance helped only until all relevant passages could fit. Every fixed-passage configuration retained all development conditions by 4,000 words and all training conditions by 10,000 words (Section 5.3 and Table A.9). In this setting smaller packs lost conditions because there was too little space for lower-ranked passages. Beyond the point where the document fitted, extra words provided no benefit.

Across the whole collection, a larger allowance partly compensated for weaker ranking, but it was not a substitute for better retrieval. Recall increased because larger packs reached further down the ranked list, yet on the development set fewer than 600 of the 19,000 added words in boundary-completion packs came from the reference document. The usefulness of the remaining text was not evaluated (Section 5.3). The training gap between hybrid retrieval and BM25 narrowed from 16 points at 1,002 words to 7 points at 20,000 words, but BM25 at 20,000 words still performed worse than dense retrieval at 10,000 words on both sets (Tables A.5 and A.10).

The training results also showed diminishing returns for hybrid retrieval and boundary completion. Each additional 1,000 words produced a smaller recall gain as the allowance increased (Tables A.10–A.11). Development results were less stable. The gains rose again at the largest allowances and the best configuration changed with the allowance (Tables A.5–A.6 and Section 5.3). The fluctuations may reflect the smaller set of 63 development questions and the all-or-nothing scoring of each condition. So the training pattern gives stronger evidence that continually enlarging the pack becomes progressively less useful.

The results favour better ranking at a moderate allowance over simply creating very large packs. A 20,000-word pack is about as long as 15 average source documents, requires more processing, and may make information in the middle harder for a model to use [30]. It also increases the amount a reviewer must inspect. Better ranking retained conditions with less text, but processing cost and reviewer workload were not measured.

6.6 Implications for practice

The first practical decision is whether the right document is already known. If the user or reliable metadata identifies it and it fits within the allowance, the system should pass the whole document in its original order, which removes competition from other documents and stops lower-ranked

conditions being dropped. A 4,000-word allowance would have held 642 of the 652 documents in this dataset. The study measured whether evidence was retained but did not test whether a model would then use it correctly.

If the document is not known, the system should search the whole collection, because the document holding the top-ranked passage is not a safe guess. That passage came from the wrong document for about one-third of training questions, and with hybrid ranking at 4,000 words, packing only that document achieved 68.0% training recall against 85.3% for searching the whole collection. Asking the user to choose from the three highest-ranked documents may help, because the list held the right document for 89% of training questions, but this safeguard was not tested.

Passages should keep each condition with the statement it qualifies. Among training conditions that shared no query word apart from stopwords, 121-word passages ranked by meaning (dense retrieval) packed 60% across the collection at 1,002 words, whereas single paragraphs and list items ranked the same way packed 1.5% (Appendix B.7). This suggests that passage windows are the safer choice when a condition may not repeat the question's wording.

For whole collection search, ranking by meaning (dense or hybrid retrieval) is a stronger candidate than the tested keyword ranking (BM25). Hybrid's advantage over BM25 grew from about 7 points with 33 competing documents to 16 points with all 651, which suggests that the ranking method matters more as competition between documents grows, although larger collections were not tested.

A system should not assume that reranking will help, because ranking for variety (MMR) reduced recall on both sets. It should also put the user's circumstances in the query, because leaving the scenario out reduced whole collection recall on the training set by 9–18 points (Appendix B.7). So rerankers and query formats should be tested directly on condition retention before deployment. Evaluations should also check whether every required condition was retained as well as whether the right document appeared, because a correct citation can hide a missing condition (Section 6.1).

Because the evidence comes from one UK dataset, the recommendations are starting points. Agencies should check them locally by annotating the required conditions in their own cases, recording which conditions reach the pack and matching each type of failure to one of the responses below.

- change how documents are split into passages when a condition matches the query weakly

- improve ranking or metadata filters, or increase the allowance, when other documents crowd a condition out
- increase the allowance when a condition is cut off at the end of the pack.

The chosen change should then be tested on held-back questions so that the same cases are not used both to design and evaluate it.

6.7 Limitations and how they affect the conclusions

The first limitation is possible tuning bias. Passage length, overlap and BM25 settings were tuned using training questions with the document supplied, which may have favoured the 121-word fixed passages and made BM25 easier to outperform. But the main findings survived alternative passage lengths. On the training set across the collection at 1,002 words, dense and hybrid retrieval still led BM25 by about 8 and 10 points with 181-word passages, and fixed passages led source elements by at least 21 points at every tested length from 61 to 1,001 words (Appendix B.7). So the direction of the findings holds up reasonably well, but their exact size may depend on the chosen settings.

The heading and pack-composition follow-ups were not repeated on the training set, so they remain possible explanations.

The development results probably overstate expected recall. Whole collection fixed-passage recall was lower on the larger training set at every allowance (Tables A.4–A.5 and A.10–A.11), and shorter development documents explained only about 3 of the 11–14 point difference at 1,002 words. The remaining gap may reflect easier development cases or chance, so training-set recall is the safer estimate of what a similar system might retain.

The passage comparison has a narrow interpretation. Source elements were ranked individually, so the results show that ranking each paragraph or list item as a separate passage performed worse than using fixed passages (Section 6.2). The study did not test retrieving the whole surrounding section, so it cannot claim that 121-word passages are the best possible unit.

Condition recall measures whether annotated words were retained. It does not measure whether their meaning was preserved, so it can understate or overstate semantic retention, as dev-224 and dev-46 show (Section 5.4.2). It also assumes that the annotations are complete, so even 100% recall would not prove that every substantive requirement was present if a requirement had not been annotated.

No answers were generated, so the results do not establish answer accuracy. The retrospective stage trace showed where conditions were lost but did not experimentally isolate why, so the mechanisms proposed in Sections 6.2–6.4 are evidence-based explanations that remain unproven.

Generalisability is limited because both sets came from one dataset of historical English-language UK documents [7], with one reference document per question and partly constructed queries. The findings have not been established for real New Zealand queries, te reo Māori, near-duplicate document versions, longer or linked policies, larger collections or questions requiring several documents. Performance and the value of including a scenario may differ in such settings.

Only a limited set of models and pipeline choices was tested. Results could differ with other embedding or reranking models, tuned hybrid fusion, stemming, packing rules or combinations of ranking and expansion. A later check found that hybrid source elements with two neighbours still trailed hybrid fixed passages across the collection on the training set at every allowance (Appendix A.12), but this does not cover all alternatives. Benchmark exposure was also unchecked, so pretrained models may have seen some documents and the dense and hybrid gains may be overstated. Among the tested configurations, fixed passages and dense or hybrid retrieval retained more conditions, but the study does not identify a universally best retrieval system.

6.8 Why human review is still needed

Guidance relies on human review to catch what AI systems miss [1]. But retrieval is hidden from the user and the reviewer, who see only the answer, its citations and at most the retrieved text. A citation to the right page could give the impression that the answer drew on the whole page, but packs holding text from that page often missed a condition (Section 6.1). Every configuration missed at least one annotated training-set condition across the collection, even at the largest allowance. So a reviewer needs to confirm that the cited document applies (Section 6.6) and check the answer against the full source as well as the retrieved excerpts. Responsibility for retention also falls on whoever configures the system, who has to test it before deployment in the setting where it will be used. This matters most for a public-facing assistant, which usually has no staff reviewer.

The risk was highest in whole collection search or with a guessed document and for short list-item conditions or tight allowances. A retained condition can also be left out of the answer. The models tested on ConditionalQA struggled most with selecting answer conditions even with the document given [7]. A reviewer should list the conditions in the source document, use saved word ranges to confirm that each is in the pack and then check that each was applied to the right person. In dev-

129 the age condition applies to the widow yet the scenario gives only her husband's age. When the whole known document is packed the word-range check is already met so review can focus on how the conditions were applied.

A system that saves the word ranges it packed, as this study did, can highlight them in the source document. A reviewer then sees what the model received and what was left out. This helps staff follow the accountability guidance to check summaries for missing context before publication and the transparency guidance to explain what information is given to AI systems [1], [31].

6.9 Contribution

ConditionalQA [7] and Yuan et al. [15] test models after the evidence has been chosen, whereas this study measured whether the conditions reached the final evidence pack. Unlike the RAG evaluation measures in Section 2, the word-range check needs no language-model judge. It showed that a pack can hold text from the right document yet miss conditions (Section 6.1), and that ranking each paragraph or list item without its surrounding text lost conditions when documents competed (Section 6.2). It also showed that testing with the document supplied understated the gain from hybrid ranking across the whole collection (Section 6.3).

7. Conclusion

This study asked how the way text is selected, and how much text is allowed, affects whether the conditions an answer depends on reach the final evidence pack. The clearest answer is that finding the right document was not enough. When the whole collection was searched with a 1,002-word allowance, hybrid packs that held text from the right document still missed a condition for 11 of 61 development questions (18%) and 187 of 518 training questions (36%).

The way text was selected mattered most when documents competed and the allowance was small. Across the collection at 1,002 words, ranking by meaning (dense and hybrid retrieval) kept far more conditions than keyword ranking (BM25) on both sets, with 69.4% for hybrid against 53.4% for BM25 on the training set. The likely reason is that fewer passages from other documents crowded the conditions out. Hybrid fixed passages with boundary completion narrowly kept the most conditions on the training set and about the same as hybrid fixed passages on development (Sections 5.2.4 and 5.5 and Table A.10).

Of the six re-tested comparisons, how documents were split into passages had the largest effect across the collection. Ranking single paragraphs and list items left short conditions open to competition from other documents, and keeping conditions with their nearby text helped, although the comparison could not separate length, grouping and overlap. Ranking for variety (MMR) lowered retention at smaller allowances and the cross-encoder made no clear difference. All six predictions met the replication criterion, but four effects were smaller and whole collection recall was lower on training (Section 5.5). Earlier use of the training data, including the already available expansion result, limits this confirmation (Section 4.10.3).

A larger allowance fixed the problem only when the document was known. In exploratory runs with the document supplied, BM25, dense and hybrid fixed passages and boundary completion kept every condition for all 63 development and all 550 training questions at 4,000 words, which was enough to hold nearly every document. Across the whole collection the same four configurations kept every condition for only 51–58 of 63 development questions and 368–455 of 550 training questions at 4,000 words. Larger allowances generally helped but most of the added text came from other documents, so a larger pack only partly makes up for poor ranking (Section 6.5).

Future work could test whether answers from the saved packs apply the retained conditions to the right person, and whether returning whole sections or a completeness-aware reranker (Sections 6.4 and 6.7) helps on annotated New Zealand guidance. Until then, agencies that configure a RAG

system should test condition retention in their own search setting and word allowance, because retrieval is hidden from users and reviewers. Human review needs to check each answer against the whole applicable source document, or a short list-item condition like the widow's age could be missed by both the system and the reviewer.

References

[1] Department of Internal Affairs, “Accountability, responsibility and GenAI,” Responsible AI Guidance for the Public Service: GenAI. Accessed: Oct. 09, 2026. [Online]. Available: https://www.digital.govt.nz/standards-and-guidance/technology-and-architecture/artificial-intelligence/responsible-ai-guidance-for-the-public-service-genai/genai-foundations/accountability-responsibility

[2] Parliamentary Counsel Office, “Chat-Based Access to Legislation: Research & Development Findings,” Parliamentary Counsel Office, Wellington, New Zealand, May 2025. Accessed: Apr. 25, 2026. [Online]. Available: https://assets.ctfassets.net/49ei4dhxj2cb/6rzMwkt0JSHe0cApvLJZa9/f06c73de475c180b4528173436d6c86e/Chat-Based-Access-to-Legislation-Report-Boost__1_.pdf

[3] Department of Conservation, “Response to Official Information Act request OIAD-5943 on AI systems,” Department of Conservation, Official Information Act response, Jan. 2026. Accessed: Apr. 25, 2026. [Online]. Available: https://www.doc.govt.nz/globalassets/documents/about-doc/oia/2026/january/oiad-5943-response.pdf

[4] Ministry of Housing and Urban Development, “Response to Official Information Act request HUD2025-008421 on AI tools,” Ministry of Housing and Urban Development, Official Information Act response, Nov. 2025. Accessed: Apr. 25, 2026. [Online]. Available: https://www.hud.govt.nz/assets/Uploads/Documents/Proactive-Releases/January-2026/HUD2025-008421-response_Redacted.pdf

[5] P. Lewis *et al.*, “Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks,” in *Advances in Neural Information Processing Systems 33*, Curran Associates, Inc., 2020, pp. 9459–9474. Accessed: Oct. 17, 2026. [Online]. Available: https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html

[6] N. F. Liu, T. Zhang, and P. Liang, “Evaluating Verifiability in Generative Search Engines,” in *Findings of the Association for Computational Linguistics: EMNLP 2023*, Singapore: Association for Computational Linguistics, Dec. 2023, pp. 7001–7025. doi: 10.18653/v1/2023.findings-emnlp.467.

[7] H. Sun, W. Cohen, and R. Salakhutdinov, “ConditionalQA: A Complex Reading Comprehension Dataset with Conditional Answers,” in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 3627–3637. doi: 10.18653/v1/2022.acl-long.253.

[8] X. Wang *et al.*, “Searching for Best Practices in Retrieval-Augmented Generation,” in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language*

Processing, Miami, FL, USA: Association for Computational Linguistics, Nov. 2024, pp. 17716–17736. doi: 10.18653/v1/2024.emnlp-main.981.

[9] V. Karpukhin *et al.*, "Dense Passage Retrieval for Open-Domain Question Answering," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online: Association for Computational Linguistics, Nov. 2020, pp. 6769–6781. doi: 10.18653/v1/2020.emnlp-main.550.

[10] R. Qu, R. Tu, and F. S. Bao, "Is Semantic Chunking Worth the Computational Cost?," in *Findings of the Association for Computational Linguistics: NAACL 2025*, Albuquerque, NM, USA: Association for Computational Linguistics, Apr. 2025, pp. 2155–2177. doi: 10.18653/v1/2025.findings-naacl.114.

[11] D. Ru *et al.*, "RAGChecker: A Fine-grained Framework for Diagnosing Retrieval-Augmented Generation," in *Advances in Neural Information Processing Systems 37*, Vancouver, BC, Canada: Curran Associates, Inc., 2024, pp. 21999–22027. doi: 10.52202/079017-0692.

[12] S. Es, J. James, L. Espinosa Anke, and S. Schockaert, "RAGAs: Automated Evaluation of Retrieval Augmented Generation," in *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, St. Julian's, Malta: Association for Computational Linguistics, Mar. 2024, pp. 150–158. doi: 10.18653/v1/2024.eacl-demo.16.

[13] Ragas, "Context Recall," Ragas documentation. Accessed: Oct. 17, 2026. [Online]. Available: https://docs.ragas.io/en/stable/concepts/metrics/available_metrics/context_recall/

[14] H. Joren, J. Zhang, C.-S. Ferng, D.-C. Juan, A. Taly, and C. Rashtchian, "Sufficient Context: A New Lens on Retrieval Augmented Generation Systems," in *International Conference on Learning Representations*, Singapore, 2025, pp. 20310–20334. Accessed: Oct. 17, 2026. [Online]. Available: https://proceedings.iclr.cc/paper_files/paper/2025/hash/33dffa2e3d2ab74a783d1a8c292f66d9-Abstract-Conference.html

[15] Y. Yuan, R. Kong, D. Li, Y. Li, and Y. Liu, "Benchmarking LLM's Capability in Reasoning over Conflicting Web References," in *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, San Diego, CA, USA: Association for Computational Linguistics, Jul. 2026, pp. 303–322. doi: 10.18653/v1/2026.acl-long.11.

[16] S. Robertson and H. Zaragoza, "The Probabilistic Relevance Framework: BM25 and Beyond," *Foundations and Trends in Information Retrieval*, vol. 3, no. 4, pp. 333–389, Dec. 2009, doi: 10.1561/1500000019.

[17] S. Bruch, S. Gai, and A. Ingber, "An Analysis of Fusion Functions for Hybrid Retrieval," *ACM Transactions on Information Systems*, vol. 42, no. 1, pp. 1–35, Jan. 2024, doi: 10.1145/3596512.

[18] N. Thakur, N. Reimers, A. Rücklé, A. Srivastava, and I. Gurevych, “BEIR: A Heterogeneous Benchmark for Zero-shot Evaluation of Information Retrieval Models,” in *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, 2021. Accessed: Oct. 17, 2026. [Online]. Available: https://datasets-benchmarks-proceedings.neurips.cc/paper_files/paper/2021/hash/65b9eea6e1cc6bb9f0cd2a47751a186f-Abstract-round2.html

[19] BAAI, “bge-small-en-v1.5,” Hugging Face. Accessed: Sep. 20, 2026. [Online]. Available: https://huggingface.co/BAAI/bge-small-en-v1.5

[20] G. V. Cormack, C. L. A. Clarke, and S. Büttcher, “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*, Boston, MA, USA: Association for Computing Machinery, Jul. 2009, pp. 758–759. doi: 10.1145/1571941.1572114.

[21] R. Nogueira and K. Cho, “Passage Re-ranking with BERT,” 2019, arXiv:1901.04085. doi: 10.48550/arXiv.1901.04085.

[22] Sentence Transformers, “cross-encoder/ms-marco-MiniLM-L6-v2,” Hugging Face. Accessed: Sep. 20, 2026. [Online]. Available: https://huggingface.co/cross-encoder/ms-marco-MiniLM-L6-v2

[23] P. Bajaj *et al.*, “MS MARCO: A Human Generated MAchine Reading COmprehension Dataset,” 2016, arXiv:1611.09268. doi: 10.48550/arXiv.1611.09268.

[24] J. Goldstein and J. Carbonell, “Summarization: (1) Using MMR for Diversity-Based Reranking and (2) Evaluating Summaries,” in *TIPSTER Text Program Phase III: Proceedings of a Workshop Held at Baltimore, Maryland, October 13–15, 1998*, Baltimore, MD, USA: Association for Computational Linguistics, Oct. 1998, pp. 181–195. doi: 10.3115/1119089.1119120.

[25] Pyserini contributors, *LuceneSearcher.set_bm25*. (Sep. 06, 2026). Python. Accessed: Sep. 07, 2026. [Online]. Available: https://github.com/castorini/pyserini/blob/139adc808413158c09510a7341cd6dcd3f122acc/pyserini/search/lucene/_searcher.py

[26] rank_bm25 contributors, *BM25Okapi*. (Oct. 08, 2024). Python. Accessed: Sep. 07, 2026. [Online]. Available: https://github.com/dorianbrown/rank_bm25/blob/47aa3ddf8dc1ebeb7ef4e65f2b4536af44594099/rank_bm25.py

[27] A. C. Davison and D. V. Hinkley, *Bootstrap Methods and their Application*. Cambridge, U.K.: Cambridge University Press, 1997. doi: 10.1017/CBO9780511802843.

[28] M. D. Smucker, J. Allan, and B. Carterette, “A Comparison of Statistical Significance Tests for Information Retrieval Evaluation,” in *Proceedings of the 16th ACM Conference on Information and Knowledge Management*, Lisbon, Portugal: Association for Computing Machinery, Nov. 2007, pp. 623–632. doi: 10.1145/1321440.1321528.

[29] S. Holm, “A Simple Sequentially Rejective Multiple Test Procedure,” *Scandinavian Journal of Statistics*, vol. 6, no. 2, pp. 65–70, 1979, Accessed: Oct. 17, 2026. [Online]. Available: https://www.jstor.org/stable/4615733

[30] N. F. Liu *et al.*, “Lost in the Middle: How Language Models Use Long Contexts,” *Transactions of the Association for Computational Linguistics*, vol. 12, pp. 157–173, 2024, doi: 10.1162/tacl_a_00638.

[31] Department of Internal Affairs, “Transparency and GenAI,” Responsible AI Guidance for the Public Service: GenAI. Accessed: Oct. 09, 2026. [Online]. Available: https://www.digital.govt.nz/standards-and-guidance/technology-and-architecture/artificial-intelligence/responsible-ai-guidance-for-the-public-service-genai/customer-experience/transparency

Appendix A Supporting results

A.1 Training selection

Table A.1 contains the four pairs of BM25 parameters used on the training set. For each pair the average fixed-passage recall was computed over the 687-, 1,002- and 1,551-word allowances using the passage setting with the highest average. All four pairs selected 121 words with 50% overlap.

The joint score is the mean of the fixed-passage recall and the unexpanded source-element recall averaged over the same three allowances. Group ranks compare the four pairs across five fixed document groups. The training documents were sorted by a hash of their URL and assigned to the groups in turn so all questions on a document stayed in one group. The settings were not refitted within each group. Training selection used source elements with headings excluded as in the original runs.

Table A.1. Training selection scores and ranks across document groups.

BM25 k1 / b	Fixed-passage recall (%)	Source-element recall (%)	Joint score (%)	Ranks in groups 0, 1, 2, 3, 4
1.5 / 0.75 (selected)	**90.285**	**75.953**	**83.119**	**2, 1, 1, 1, 2**
1.5 / 0.4	90.278	75.562	82.920	3, 3, 2, 4, 1
0.9 / 0.4	90.254	75.570	82.912	1, 2, 4, 3, 3
0.9 / 0.75	90.104	75.688	82.896	4, 4, 3, 2, 4

The fixed-passage recall at the chosen pair (k1 = 1.5, b = 0.75) was 90.3% with 50% overlap, 89.7% with no overlap and 89.2% with 25% overlap, all at 121 words. The 87-word no-overlap winner of the first search was 88.3%.

With the chosen pair, 121-word passages had the highest recall in all 5 groups. Among them 50% overlap was best in 2 groups, no overlap in 2 and 25% overlap in 1.

A.2 Supplied document results at initial allowances

Condition recall (%) on the first line and complete questions out of 63 on the second. Arrows from original to headings included in shaded source-element rows. Bold right-hand values indicate that configuration was tested with headings included. A dash indicates that configuration was not tested at that allowance.

The 7,733-word column is the separate 7,733-word BM25 check. Larger-allowance results are in Table A.3. Paired cross-encoder comparisons are in Table A.7.

Table A.2. Supplied document retention at the three initial allowances and the 7,733-word BM25 check.

Configuration	687 words	1,002 words	1,551 words	7,733 words (BM25 check)
BM25 fixed passages	83.2% 49/63	88.6% 52/63	96.2% 58/63	100.0% 63/63
BM25 source elements	80.1→80.0% 38→38/63	85.9→86.4% 42→44/63	88.8→89.1% 43→45/63	93.0→94.0% 49→51/63
Source elements + one neighbour	85.1→84.6% 42→42/63	92.4→92.2% 51→50/63	96.3→96.7% 56→58/63	98.8→99.7% 59→61/63
Source elements + two neighbours	88.3→87.0% 47→47/63	93.1→93.9% 53→54/63	95.8→97.0% 57→58/63	99.0→100.0% 61→63/63
Dense fixed passages	92.5% 55/63	94.7% 57/63	98.0% 60/63	—
Dense + MMR 0.7 / 100	91.9% 53/63	94.7% 56/63	98.4% 60/63	—
Dense + MMR 0.3 / 100	78.7% 39/63	89.4% 51/63	94.7% 55/63	—
Dense + MMR 0.7 / 50	91.9% 53/63	94.7% 56/63	98.4% 60/63	—
Dense + MMR 0.7 / 200	91.9% 53/63	94.7% 56/63	98.4% 60/63	—
Hybrid fixed passages	90.2% 53/63	93.3% 56/63	97.8% 59/63	—
Hybrid + cross-encoder	91.6% 52/63	94.4% 56/63	97.7% 59/63	—
Hybrid + boundary completion	91.8% 54/63	93.9% 57/63	98.4% 61/63	—
Hybrid + equal-length centred spans	90.4% 53/63	92.6% 56/63	98.4% 61/63	—

A.3 Supplied document results at larger allowances

Each configuration had the same scores and mean pack length at 4,000, 10,000, 15,000 and 20,000 words so one set of results covers all four. Configurations are listed in the order of Table A.2 even if the results are identical.

The arrows indicate the original → headings. Mean pack length (words) is the average of the 63 questions with annotated answer conditions. The original runs did not test cross-encoder reranking at these allowances (Appendix A.7).

Table A.3. Supplied document retention and mean pack length at all four larger allowances.

Configuration	Condition recall (%)	Complete questions / 63	Mean pack length (words)
BM25 fixed passages	100.0%	63/63	1,343.6
BM25 source elements	93.0→94.0%	49→51/63	1,286.8→1,372.3
Source elements + one neighbour	98.8→99.7%	59→61/63	1,360.5→1,486.0
Source elements + two neighbours	99.0→100.0%	61→63/63	1,369.6→1,498.7
Dense fixed passages	100.0%	63/63	1,343.6
Dense + MMR 0.7 / 100	100.0%	63/63	1,343.6
Dense + MMR 0.3 / 100	100.0%	63/63	1,343.6
Dense + MMR 0.7 / 50	100.0%	63/63	1,343.6
Dense + MMR 0.7 / 200	100.0%	63/63	1,343.6
Hybrid fixed passages	100.0%	63/63	1,343.6
Hybrid + cross-encoder	—	—	—
Hybrid + boundary completion	100.0%	63/63	1,334.1
Hybrid + equal-length centred spans	100.0%	63/63	1,334.3

A.4 Whole collection results at initial allowances

All 652 source documents were searched. Condition recall (%) and complete questions out of 63 are shown in each cell. Source-element arrows show original → headings included and the right-hand values are bold. A dash means not tested.

The 7,733-word column is the separate 7,733-word BM25 check. Table A.5 continues the comparison at larger allowances and Table A.8 gives paired changes in the heading follow-up.

Table A.4. Whole collection retention at the three initial allowances and the 7,733-word BM25 check.

Configuration	687 words	1,002 words	1,551 words	7,733 words (BM25 check)
BM25 fixed passages	52.5% 27/63	64.2% 35/63	72.3% 39/63	89.4% 54/63
BM25 source elements	16.4→16.8% 8→8/63	23.7→23.5% 9→9/63	25.2→26.5% 10→11/63	34.7→35.4% 16→17/63
Source elements + one neighbour	22.9→20.0% 14→12/63	24.9→22.9% 15→14/63	27.4→26.3% 16→15/63	46.2→47.1% 22→24/63
Source elements + two neighbours	22.1→20.8% 13→13/63	25.0→23.8% 15→14/63	27.7→30.6% 16→17/63	51.5→54.2% 27→28/63
Dense fixed passages	69.0% 37/63	81.1% 47/63	86.0% 51/63	—
Dense + MMR 0.7 / 100	54.3% 28/63	56.5% 30/63	67.1% 33/63	—
Dense + MMR 0.3 / 100	31.4% 14/63	31.4% 14/63	36.4% 15/63	—
Dense + MMR 0.7 / 50	54.3% 28/63	58.4% 31/63	70.0% 35/63	—
Dense + MMR 0.7 / 200	54.3% 28/63	56.5% 30/63	62.9% 32/63	—
Hybrid fixed passages	74.9% 42/63	82.5% 50/63	85.5% 51/63	—
Hybrid + cross-encoder	68.7% 33/63	74.8% 41/63	87.7% 49/63	—
Hybrid + boundary completion	76.4% 43/63	83.4% 50/63	88.2% 52/63	—
Hybrid + equal-length centred spans	75.0% 42/63	81.5% 48/63	85.8% 52/63	—

A.5 Whole collection results at larger allowances

The condition recall percentage and the number of complete questions out of 63 are in each cell. The source element arrows indicate the original → headings included. The right-hand values are bold. The original runs did not test cross-encoder reranking at these allowances (Appendix A.7).

All configurations tested at 20,000 words had reference-document text for all 63 questions, including source-element configurations with headings. Table A.6 shows the changes from 15,000 to 20,000 words.

Table A.5. Whole collection retention at the four larger allowances.

Configuration	4,000 words	10,000 words	15,000 words	20,000 words
BM25 fixed passages	84.9% 51/63	89.4% 54/63	90.3% 55/63	91.0% 55/63
BM25 source elements	31.4→32.4% 14→15/63	35.2→36.2% 17→18/63	40.9→41.5% 21→22/63	45.2→44.7% 23→25/63
Source elements + one neighbour	41.0→40.9% 20→21/63	49.2→48.6% 24→24/63	52.9→53.8% 25→26/63	52.9→55.8% 25→27/63
Source elements + two neighbours	43.6→42.6% 23→22/63	53.2→55.5% 28→30/63	57.0→59.1% 30→31/63	60.6→61.8% 31→32/63
Dense fixed passages	91.9% 56/63	98.0% 60/63	99.1% 61/63	99.5% 62/63
Dense + MMR 0.7 / 100	85.7% 48/63	98.0% 60/63	99.1% 61/63	99.5% 62/63
Dense + MMR 0.3 / 100	56.7% 26/63	97.5% 60/63	99.1% 61/63	99.5% 62/63
Dense + MMR 0.7 / 50	89.7% 52/63	98.0% 60/63	99.1% 61/63	99.5% 62/63
Dense + MMR 0.7 / 200	81.2% 45/63	94.4% 56/63	97.5% 60/63	99.0% 61/63
Hybrid fixed passages	94.3% 58/63	94.3% 58/63	95.6% 59/63	99.1% 60/63
Hybrid + cross-encoder	—	—	—	—
Hybrid + boundary completion	94.2% 57/63	94.9% 58/63	94.9% 58/63	100.0% 63/63
Hybrid + equal-length centred spans	93.8% 57/63	94.7% 58/63	94.7% 58/63	99.5% 62/63

A.6 Changes from 15,000 to 20,000 words

This table compares the final two allowances in whole collection search. Changes are 20,000 - 15,000 words and positive values indicate increased retention. Source-element rows include headings at both allowances.

Unrounded scores and pointwise 95% document-bootstrap intervals are used for changes. The larger-allowance comparisons are exploratory and were not adjusted for multiple comparisons. A dash indicates that the comparison was not tested.

Table A.6. Paired retention changes from 15,000 to 20,000 words.

Configuration	Recall change (pp) [95% interval]	Change in complete questions
BM25 fixed passages	0.65 [0.00, 1.96]	0
BM25 source elements	3.23 [0.41, 6.62]	3
Source elements + one neighbour	1.96 [0.00, 5.39]	1
Source elements + two neighbours	2.72 [0.22, 6.47]	1
Dense fixed passages	0.33 [0.00, 0.98]	1
Dense + MMR 0.7 / 100	0.33 [0.00, 0.98]	1
Dense + MMR 0.3 / 100	0.33 [0.00, 0.98]	1
Dense + MMR 0.7 / 50	0.33 [0.00, 0.98]	1
Dense + MMR 0.7 / 200	1.47 [0.00, 4.41]	1
Hybrid fixed passages	3.43 [0.00, 9.80]	1
Hybrid + cross-encoder	—	—
Hybrid + boundary completion	5.14 [0.25, 12.42]	5
Hybrid + equal-length centred spans	4.86 [0.25, 11.85]	4

A.7 Changes from cross-encoder reranking

The results are the hybrid + cross-encoder minus the results for hybrid fixed passages at the same allowance. Positive values favour reranking while negative values favour hybrid fixed passages. Full scores are shown in Tables A.2 and A.4.

Changes are calculated using unrounded scores and pointwise 95% paired document-bootstrap intervals. Net changes in complete-question counts are the net differences out of 63. With the document supplied, 1 question was gained and 2 lost at 687 words and 1 gained and 1 lost at 1,002 and at 1,551 words. Across the collection 2 were gained and 11 lost at 687 words, 1 gained and 10 lost at 1,002 words and 2 gained and 4 lost at 1,551 words. The original development runs did not test cross-encoder reranking above 1,551 words so Tables A.3 and A.5 show dashes. The replication code later ran it on both sets at all seven allowances (Section 5.5). On development it matched hybrid fixed passages with the document supplied from 4,000 words. Across the collection it trailed them by 1.4 points [−3.7, 0.0] at 4,000 words (0 gained, 3 lost) and matched them from 10,000 words. Training results are in Tables A.9–A.11.

Table A.7. Paired changes from cross-encoder reranking at the three initial allowances.

Search setting	Allowance (words)	Recall change (pp) [95% interval]	Change in complete questions
Supplied	687	1.4 [−2.2, 5.6]	−1
Supplied	1,002	1.1 [−2.3, 5.0]	0
Supplied	1,551	−0.2 [−0.6, 0.0]	0
Collection	687	−6.2 [−17.5, 3.6]	−9
Collection	1,002	−7.8 [−19.0, 2.9]	−9
Collection	1,551	2.2 [−5.0, 10.8]	−2

A.8 Changes in the heading follow-up

This table compares headings included with the original source-element runs (headings excluded) at 1,002 words. Change is headings included minus original. Complete-question counts are original → headings included.

The gains and losses of counts from incomplete to complete and from complete to incomplete are recorded. Otherwise the complete status is unchanged. Recall intervals are pointwise 95% paired document-bootstrap intervals (5,000 resamples, fixed seed). The six comparisons are exploratory and not adjusted for multiple comparisons.

Table A.8. Paired changes in the heading follow-up at 1,002 words.

Search setting	Source-element configuration	Recall change (pp) [95% interval]	Complete questions (original → headings included)	Gained / lost / unchanged
Supplied	No expansion	0.49 [−1.23, 2.45]	42→44/63	2 / 0 / 61
Supplied	One neighbour	−0.20 [−2.89, 2.21]	51→50/63	1 / 2 / 60
Supplied	Two neighbours	0.82 [−0.33, 2.53]	53→54/63	1 / 0 / 62
Collection	No expansion	−0.17 [−1.69, 1.23]	9→9/63	0 / 0 / 63
Collection	One neighbour	−1.92 [−4.04, −0.33]	15→14/63	0 / 1 / 62
Collection	Two neighbours	−1.23 [−5.88, 3.43]	15→14/63	1 / 2 / 60

All eight allowances are listed in Tables A.2–A.5.

A.9 Training-set replication with the supplied document

Only each question's reference document was searched. Each cell shows condition recall (%) and complete questions out of 550. Source-element rows exclude headings as in the original runs. From 10,000 to 20,000 words every configuration retained every condition for all 550 questions except the source-element rows, which reached 90.2% (453/550) unexpanded, 96.8% (525/550) with one neighbour and 97.5% (532/550) with two neighbours.

Table A.9. Training-set supplied document retention at the three initial allowances and 4,000 words.

Configuration	687 words	1,002 words	1,551 words	4,000 words
BM25 fixed passages	82.0% 409/550	91.2% 473/550	97.7% 522/550	100.0% 550/550
BM25 source elements	66.0% 282/550	76.9% 354/550	84.9% 408/550	90.0% 452/550
Source elements + one neighbour	75.0% 354/550	83.9% 417/550	92.5% 489/550	96.5% 522/550
Source elements + two neighbours	78.5% 385/550	87.4% 450/550	93.4% 496/550	97.1% 529/550
Dense fixed passages	87.3% 441/550	94.0% 494/550	98.0% 527/550	100.0% 550/550

Configuration	687 words	1,002 words	1,551 words	4,000 words
Dense + MMR 0.7 / 100	84.1% 419/550	93.3% 483/550	97.7% 524/550	100.0% 550/550
Dense + MMR 0.3 / 100	64.8% 308/550	80.1% 392/550	91.6% 477/550	99.8% 547/550
Dense + MMR 0.7 / 50	84.1% 419/550	93.3% 483/550	97.7% 524/550	100.0% 550/550
Dense + MMR 0.7 / 200	84.1% 419/550	93.3% 483/550	97.7% 524/550	100.0% 550/550
Hybrid fixed passages	86.9% 438/550	94.5% 498/550	98.4% 530/550	100.0% 550/550
Hybrid + cross-encoder	88.2% 457/550	93.8% 496/550	97.5% 527/550	99.8% 549/550
Hybrid + boundary completion	89.2% 455/550	95.0% 505/550	98.9% 533/550	100.0% 550/550
Hybrid + equal-length centred spans	88.2% 446/550	94.3% 495/550	98.7% 531/550	100.0% 550/550

A.10 Training-set replication across the whole collection

The search was performed on all 652 source documents and the cells are as in Table A.9. Source-element rows exclude headings as in the original runs. Table A.11 gives the other three allowances.

Table A.10. Training-set whole collection retention at 1,002, 4,000, 10,000 and 20,000 words.

Configuration	1,002 words	4,000 words	10,000 words	20,000 words
BM25 fixed passages	53.4% 241/550	74.4% 368/550	83.6% 423/550	88.1% 450/550
BM25 source elements	18.3% 65/550	27.4% 98/550	33.7% 126/550	39.7% 152/550
Source elements + one neighbour	25.8% 96/550	41.4% 163/550	50.0% 205/550	55.8% 234/550
Source elements + two neighbours	26.9% 112/550	44.7% 191/550	55.4% 240/550	62.1% 281/550
Dense fixed passages	67.3% 326/550	83.8% 439/550	90.7% 482/550	95.2% 512/550
Dense + MMR 0.7 / 100	55.0% 250/550	75.7% 377/550	90.6% 482/550	95.2% 512/550
Dense + MMR 0.3 / 100	24.8% 115/550	43.6% 196/550	90.5% 480/550	95.2% 512/550
Dense + MMR 0.7 / 50	55.3% 251/550	81.6% 412/550	90.7% 482/550	95.2% 512/550

Configuration	1,002 words	4,000 words	10,000 words	20,000 words
Dense + MMR 0.7 / 200	55.0% 250/550	73.9% 363/550	86.2% 443/550	95.3% 513/550
Hybrid fixed passages	69.4% 331/550	85.3% 443/550	91.9% 486/550	95.0% 512/550
Hybrid + cross-encoder	70.6% 346/550	85.3% 444/550	91.9% 487/550	95.0% 512/550
Hybrid + boundary completion	72.8% 358/550	86.6% 455/550	92.5% 494/550	95.2% 517/550
Hybrid + equal-length centred spans	71.1% 346/550	86.0% 450/550	92.1% 488/550	94.8% 513/550

A.11 Training-set replication across the whole collection at the other allowances

The cells are as in Table A.10. Source-element rows exclude headings as in the original runs.

Table A.11. Training-set whole collection retention at 687, 1,551 and 15,000 words.

Configuration	687 words	1,551 words	15,000 words
BM25 fixed passages	47.4% 213/550	61.1% 280/550	85.9% 440/550
BM25 source elements	16.4% 57/550	21.2% 75/550	36.3% 139/550
Source elements + one neighbour	21.2% 81/550	30.9% 117/550	53.6% 226/550
Source elements + two neighbours	24.5% 105/550	34.3% 143/550	59.8% 266/550
Dense fixed passages	60.3% 288/550	73.5% 364/550	93.1% 499/550
Dense + MMR 0.7 / 100	46.3% 207/550	61.2% 282/550	93.1% 499/550
Dense + MMR 0.3 / 100	23.3% 109/550	27.0% 125/550	93.1% 499/550
Dense + MMR 0.7 / 50	47.4% 212/550	63.0% 288/550	93.1% 499/550
Dense + MMR 0.7 / 200	46.2% 206/550	61.1% 281/550	92.4% 490/550
Hybrid fixed passages	61.6% 290/550	76.3% 374/550	93.6% 505/550
Hybrid + cross-encoder	63.9% 306/550	77.0% 384/550	93.6% 505/550

Configuration	687 words	1,551 words	15,000 words
Hybrid + boundary completion	64.2% 315/550	79.1% 395/550	94.3% 509/550
Hybrid + equal-length centred spans	62.6% 300/550	77.9% 384/550	93.9% 504/550

A.12 Source elements under dense and hybrid ranking (exploratory)

The checks were run after the replication and changed no reported result. Dense source elements were ranked by the cosine similarity between each element's own vector and the query vector, and hybrid source elements combined that ranking with the BM25 source-element ranking by reciprocal rank fusion with the constant 60 (Section 4.4.3). Expansion, packing and scoring were unchanged and source elements excluded headings as in the original runs. Each cell shows condition recall (%) and complete questions at 1,002 words, and the BM25 and fixed-passage cells repeat Tables 5.1, 5.2, 5.4, A.9 and A.10. Across the collection on the training set hybrid source elements with two neighbours stayed 19.0 to 28.0 points below hybrid fixed passages at every allowance, and every interval excluded zero.

Table A.12. Passage type by ranking method at 1,002 words.

Set and search setting	Ranking	Source elements	Source elements + one neighbour	Source elements + two neighbours	Fixed passages
Training, collection	BM25	18.3% 65/550	25.8% 96/550	26.9% 112/550	53.4% 241/550
Training, collection	Dense	24.2% 86/550	35.4% 131/550	40.4% 164/550	67.3% 326/550
Training, collection	Hybrid	27.9% 107/550	37.9% 149/550	41.5% 171/550	69.4% 331/550
Training, supplied	BM25	76.9% 354/550	83.9% 417/550	87.4% 450/550	91.2% 473/550
Training, supplied	Dense	82.3% 405/550	84.3% 416/550	86.9% 443/550	94.0% 494/550
Training, supplied	Hybrid	81.0% 394/550	86.5% 434/550	89.4% 457/550	94.5% 498/550
Development, collection	BM25	23.7% 9/63	24.9% 15/63	25.0% 15/63	64.2% 35/63
Development, collection	Dense	27.1% 14/63	32.8% 18/63	35.2% 18/63	81.1% 47/63
Development, collection	Hybrid	32.0% 16/63	36.2% 18/63	35.7% 16/63	82.5% 50/63
Development, supplied	BM25	85.9% 42/63	92.4% 51/63	93.1% 53/63	88.6% 52/63
Development, supplied	Dense	92.2% 53/63	93.6% 54/63	93.8% 55/63	94.7% 57/63

Set and search setting	Ranking	Source elements	Source elements + one neighbour	Source elements + two neighbours	Fixed passages
Development, supplied	Hybrid	92.3% 50/63	95.0% 55/63	92.7% 54/63	93.3% 56/63

Appendix B Implementation details

This appendix gives the implementation details and exploratory checks behind Sections 4 to 6. The items follow the order of the report and each one names the section it supports.

B.1 Data, query and passages

Query text (Sections 4.1 and 4.4.3). The query was the question, a space and then the scenario. For dense retrieval it began with the model's retrieval instruction, "Represent this sentence for searching relevant passages: ", and passages had no instruction. BM25 and the cross-encoder used the query without it. Dense retrieval ranked passages by cosine similarity of normalised vectors. Record identifiers linked each query to its saved packs and scores but were not part of the query.

Text preparation (Section 4.3). HTML tags were removed from each heading, paragraph and list item separately so that retrieval used only the source text. Preserving word order and the boundaries between elements allowed both source elements and fixed passages to be built from that text.

Word positions (Sections 4.3 and 4.8.1). Each whitespace-separated word in a document was given a position starting at zero. Word ranges recorded a start position and an exclusive end position so 0:121 includes positions 0–120. In dev-129 the annotated age condition occupied word positions 116–119 of the reference document (range 116:120) so retention required all four positions.

Fixed-passage overlap (Section 4.3). Successive fixed passages started 61 words apart and shared 60 words when both were full length so the first two word ranges were 0:121 and 61:182.

Document boundary (Section 4.3). The final passage of each document stopped at the document boundary because crossing into the next document would have joined unrelated pages with no single source label.

Headings (Section 4.3). In the original source-element runs headings only marked the section boundaries that limited neighbour expansion (Section 4.6) so they were not candidates or neighbours and their text was not attached to source elements. The exploratory follow-up let all 13,072 non-empty headings be ranked and used as neighbours in their sections. Heading text was not added to other passages because that would also have changed their length and the words they were ranked on. So heading eligibility was the only change.

B.2 Ranking, reranking and expansion

BM25 tokenisation and scoring (Section 4.4.3). The BM25 query and passage texts were tokenised into lowercase Unicode words without stemming or stopword removal. Each distinct query word was counted once so passage scores reflected only the exact words shared with the query. Inverse document frequency used the Lucene form, $\log(1 + (N - n + 0.5)/(n + 0.5))$, where N was the number of passages searched in that setting and n the number containing the word. So no word scored below zero and only passages with a positive score were ranked. In a later exploratory training check at 1,002 words Porter stemming changed whole collection BM25 recall by +2.5 points [−0.6, 5.8] leaving hybrid retrieval 13.5 points ahead.

Counting each repeat of a query word, as rank_bm25 does, lowered whole collection BM25 recall by 12.4 points [−15.8, −9.2].

BM25 eligibility (Section 4.12). With stopwords kept each fixed passage had a query word in the supplied document search. Across the collection BM25 excluded 15 of 13,689 passages on average (at most 80) and the excluded passages would have ranked below 13,600 others.

Reciprocal rank fusion (Section 4.4.3). Ranks started at 1 and both full rankings were fused with no cut-off. A passage that BM25 did not rank, because it shared no query word, received nothing from the BM25 term.

Cross-encoder depth (Section 4.5). All packs at the three initial allowances remained within the 100 reranked hybrid candidates so every packed passage had been scored by the cross-encoder. Every query–passage pair also fitted within the model's input limit (next item) so truncation cannot explain the missed conditions. Hybrid packs across the collection at 1,002 words used at most the first 15 candidates whereas 33 of the 45 development conditions they missed lay within the first 100.

Encoder input lengths (Sections 4.4.3 and 4.5). Every 121-word fixed passage fitted within the dense encoder's 512-token input limit (at most 277 tokens). The longest query–passage pair was 344 tokens, within the cross-encoder's 512-token input limit, so every pair was scored in full.

MMR score (Section 4.5). MMR started from the most relevant candidate and at each later step selected the candidate with the highest score, λ times its cosine similarity to the query minus $(1 - \lambda)$ times its highest cosine similarity to any passage already chosen, following Goldstein and Carbonell [24]. Similarities used the normalised dense retrieval vectors.

Centred-span control (Section 4.6). When an equal-length span would have crossed a document edge it was shifted inside the document and kept its full length. This happened for 1,214 of the 13,689 fixed passages and for 1,211 of them the shifted span was the same as the boundary-completed span. This makes the comparison conservative because those passages added the same text in both configurations.

Expansion reach (Section 4.6). Across the collection at 1,002 words development packs reached about rank 51 of the BM25 source-element ranking without expansion but only 24 with one neighbour and 18 with two.

B.3 Packing and scoring

Source labels and overlap (Sections 4.7 and 4.12). Each run of new words from one passage was a separate fragment with its own label, even when it continued a fragment already packed, and fragments stayed in rank order. At 1,002 words labels took about 22 words in fixed-passage packs and 83 to 103 words in source-element packs. Repacking without labels changed the fixed-passage lead over source elements by at most 1.5 points. With the document supplied, the 10-question gap in complete questions on the development set stayed at 10 and the 119-question gap on the training set fell to 109. Without overlap BM25 121-word fixed passages scored 50.0% across the collection on the training set at 1,002 words against 53.4% with overlap and 18.3% for source elements.

Fragment order (Section 4.8.1). Packing followed rank order so a later passage could add words that came before an earlier one as in dev-46. At 1,002 words, 87% to 93% of the conditions retained by BM25, dense and hybrid fixed passages sat in one fragment and 3% to 7% were covered out of document order. The scores describe which words were packed so they would be the same if the packed text were put back in document order. Requiring the fragments to be in document order would have lowered fixed-passage

recall on the training set by 1.5 to 3.4 points across the collection and by 6 to 7 points with the document supplied.

Pack-composition measures (Section 4.10.1). These were the number of distinct source documents per pack and the amount and share of reference-document body text excluding source labels.

B.4 Settings, statistics and replication

BM25 pair (Section 4.9.1). The k1 parameter limits the extra credit when a query word appears many times in one passage. The b parameter sets how strongly a score is adjusted for passage length against the average so a long passage does not rank higher just because it holds more words. The choice of BM25 pair made a negligible difference mostly on source elements. One possible but untested reason is that b adjusts for length against the average, and fixed passages are almost all the same length whereas source elements vary widely.

Stability check (Section 4.9.1). Each candidate was also scored within 5 groups of training documents fixed before tuning to see whether a winner held across different documents. It played no part in the choice. The 121-word length, the longest tested, ranked first in all 5 groups so longer passages may be better. The chosen BM25 pair ranked first in 3 (Table A.1), in line with its negligible effect.

Sign-flip p-value (Sections 4.10.2 and 5.2.3). The two-sided p-value was (1 + the number of draws with an absolute mean difference at least as large as observed) / 10,001. So with 18 contrasts the smallest possible Holm-adjusted p was 18/10,001 (.0018). Of the 18 declared development contrasts only the whole collection MMR losses at 1,002 and 1,551 words had a Holm-adjusted p below .05 (both .0018). The 687-word MMR loss did not (unadjusted p = .021, Holm-adjusted p = .33) and the other 15 had a Holm-adjusted p of 1.

Bootstrap interval (Section 4.10.2). The interval ends were the linearly interpolated 2.5th and 97.5th percentiles of the 5,000 resampled mean differences, drawn with random seed 589 (seed check in the next item). The sign-flip draws used the same seed.

Seed check (Section 4.10.2). To test whether the intervals and sign-flip tests depended on the random seed all 274 configuration comparisons were rerun on both sets with 20 other seeds. The 274 were 245 of the 336 comparisons of 12 configuration pairs in both search settings at seven allowances on the two sets, the six predictions on each set and 17 of the 18 declared development contrasts. Comparisons whose two configurations gave identical results were left out. The declared contrasts and five predictions on each set were also among the 245 but were rerun separately so their Holm adjustments could be rechecked. The interval ends moved by at most 1.3 points (0.4 on the training set) and no Holm-adjusted result changed.

Replication protocol (Section 4.10.3). The protocol's SHA-256 fingerprint was recorded before the replication ran so later changes to it would be detectable. The replication scripts checked this fingerprint at the start of each stage. The protocol was kept locally with no external registry, so it shows the order of steps but is not independent proof of when they happened. Setting selection had already scored two-neighbour expansion with the document supplied (87.4% against 76.9% for source elements at 1,002 words) although the choice of settings did not use it. So the training result for the expansion prediction existed before the protocol was fixed.

B.5 Software and reproducibility

Software (Section 4.11). Experiments used Python 3.12. Dense encoding and cross-encoder scoring used sentence-transformers 5.7.0 with transformers 5.15.0 and PyTorch 2.13.0. The models were BAAI/bge-small-en-v1.5 and cross-encoder/ms-marco-MiniLM-L6-v2. Each experiment was run once as a Python script from the command line except the MMR, boundary-completion and larger-allowance experiments,

which were run in full from a new kernel in Jupyter notebooks so that results could not depend on hidden notebook state. The experiments ran on a Windows 11 desktop computer with an Intel Core i5-13400F processor. Each experiment loaded the documents, built the passages and indexes and encoded the passages once, and later experiments reused the verified vectors.

Saved packs and score checks (Section 4.11). Each experiment ran automatically over every question, ranking and packing at each word allowance, and saved each pack with its word ranges (Figure 4.1) so any pack could be rebuilt and checked. All 285 development questions, including questions without annotated conditions, were processed to verify that each saved pack could be rebuilt from its recorded word ranges. Annotations were used only to score completed packs, and separate verifier scripts recalculated every score. The code, saved packs, verifier scripts and replication protocol are available from the author on request.

B.6 Supporting results

Larger allowances across the collection (Section 5.3). On the development set hybrid fixed passages and boundary completion were effectively tied for best at 4,000 words (94.3% and 94.2%). At 10,000 and 15,000 words dense fixed passages tied for best with several MMR settings. The check that all 12 configurations filled their 20,000-word packs used the source-element runs with headings included. On the development set boundary-completion packs averaged 64.24 distinct source documents at 20,000 words against 3.86 at 1,002 words. Their reference-document body text (excluding source labels) averaged 1,086.4 words against 507.1. At 20,000 words boundary completion led hybrid fixed passages by 0.92 points [0.00, 2.66] and equal-length centred spans by 0.49 points [0.00, 1.47] on the development set. On the training set the leads were 0.2 points [−0.6, 1.0] and 0.4 points [0.1, 0.8] (Table A.10).

Training paired comparisons (Section 5.5). The intervals and gained and lost counts quoted for the cross-encoder, boundary completion and two-neighbour expansion on the training set came from the saved replication runs and were not adjusted for multiple comparisons. At 1,002 words across the collection the three gains (boundary completion over hybrid fixed passages and over equal-length centred spans, and two-neighbour expansion over unexpanded source elements) each had the smallest possible sign-flip p (1/10,001) so each would stay below .05 after a Holm adjustment across all 168 training comparisons (12 configuration pairs in both search settings at seven allowances). Above 1,551 words the boundary-completion gains over equal-length centred spans would not stay below .05.

B.7 Exploratory checks used in Section 6

Section grouping (Section 6.2). This exploratory BM25 check merged each section's consecutive elements, heading first, while the merged text stayed within 121 words, and an element over 121 words stayed as its own passage. That gave 14,425 passages with a mean of 59 words, ranked with the study's BM25 settings and packed and scored with its rules. On the training set at 1,002 words recall was 46.2% (217/550) across the collection against 18.3% (65/550) for source elements and 53.4% (241/550) for fixed passages, and 91.8% (483/550) with the document supplied against 91.2% (473/550) for fixed passages. On development across the collection it was 42.5% against 23.7% and 64.2%. Adding the heading path and list lead-in to each element for ranking only gave 36.9% across the collection on training and 49.6% with the document title as well.

Dense source-element check (Sections 6.2 and 6.6). This exploratory check used the study's dense model to encode each of the 56,123 non-empty source elements that were not headings, ranked them by cosine similarity to the query and packed and scored them with the study's rules. The dev-129 age condition ranked 5,219th of the elements across the collection. A content word was a query word not on the scikit-learn English stop-word list. On the training set 539 of the 1,368 non-heading conditions shared no content word with the query, and at 1,002 words across the collection dense 121-word passages packed 59.6% of

them against 1.5% for dense source elements (65.8% against 0% for the 73 such conditions on development). These shares count conditions whereas the 24.2% and 67.3% in Section 6.2 are condition recall (Table A.12).

Supplied document ceiling (Section 6.3). At 1,002 words the BM25, dense and hybrid fixed-passage packs with the document supplied stopped short of the allowance for the same 34 development and 225 training questions, and all three retained each of their conditions. Without those questions the hybrid and BM25 recalls in both search settings were recalculated from the saved question results with each document weighted equally. This left 29 development questions from 14 documents and 325 training questions from 169 documents.

MMR and overlap (Section 6.4). In the dense top 100 across the collection the median cosine similarity was 0.908 between overlapping passages of the same document, 0.814 between other passages of the same document and 0.726 between passages of different documents on development (0.909, 0.818 and 0.741 on training). Ignoring the MMR penalty between overlapping passages changed the loss at 1,002 words from 12.3 to 9.7 points on training and from 24.6 to 14.9 on development. Ignoring it for all passages of the same document left no clear loss (differences from dense fixed passages of −0.2 points [−4.3, 4.0] on training and −2.9 [−9.3, 3.7] on development).

Allowance gains (Section 6.5). Because recall could not decrease as the allowance grew (Section 4.9.2) the analysis focused on the size and location of gains. Gains were compared per additional 1,000 words because the steps between allowances were unequal such as 1,002 to 4,000 and 4,000 to 10,000 words. Each gain was point change in recall was divided by the increase in allowance in thousands of words. Table A.6 reports intervals for the increase from 15,000 to 20,000 words. Under the full-coverage rule a condition counts only when its last word arrives so gains across the 63 development questions occur in steps. The development gains per 1,000 words for hybrid fixed passages and boundary completion both increased at the largest allowances (Tables A.5 and A.6) probably as a single step. Across the collection boundary completion narrowly led at all training allowances below 20,000 words and nearly tied at 20,000 (Tables A.10–A.11) whereas the best development configuration changed with the allowance (Section 5.3).

Automatic document choice (Section 6.6). This exploratory check reused the saved whole collection rankings. Documents were ordered by their highest-ranked passage and the first was taken as the system's choice. If it was the reference document the question kept its saved supplied document result at the same allowance and otherwise its recall was zero because conditions were scored only in the reference document. With hybrid ranking the choice was right for 68.4% of training questions (82.5% on development), or 68.0% with each document weighted equally, and the reference document was among the three highest-ranked documents for 89.1% (93.7%). At 4,000 words recall was 68.0% against 85.3% for whole collection search on training (−17.3 points [−21.7, −12.8]). Because hybrid packs of the known document kept every condition at 4,000 words (Table A.9) a choice rule would need a document-weighted accuracy above 85.3% to beat whole collection search.

Collection size (Section 6.6). This exploratory check used the training set at 1,002 words with 121-word fixed passages. Each draw kept a random 5%, 10%, 25% or 50% of the 652 documents (33, 65, 163 or 326), with five draws for each share, and each question's reference document was added when it was not drawn. With the full collection, that made 21 collections. BM25 word statistics were recalculated for each collection. The loss per doubling was the least-squares slope of condition recall on the base-2 logarithm of the number of documents, with a 95% interval from resampling documents. It was 5.7 points [4.8, 6.5] for BM25, 4.2 [3.5, 4.9] for dense and 3.5 [2.9, 4.2] for hybrid retrieval. The slope is an average and the BM25 loss grew from 3.9 to 8.2 points per doubling between the smallest and largest steps.

Query wording (Section 6.6). This exploratory check reran BM25, dense, hybrid and MMR ($\lambda = 0.7$, depth = 100) fixed passages in both search settings on both sets at all seven allowances with the question without the scenario as the query. The scenario was dropped and nothing else changed. On the training set

whole collection recall fell by between 8.6 points (hybrid retrieval at 20,000 words) and 18.1 points [−23.1, −13.2] (hybrid retrieval at 1,002 words). With the document supplied training recall changed by at most 3.4 points. On the development set whole collection recall fell by 4.0 to 22.5 points.

Passage length (Section 6.7). This exploratory check reran fixed passages of 61, 121, 181, 241, 361, 601 and 1,001 words with 50% overlap in both search settings on both sets at all seven allowances. BM25 used all lengths but dense and hybrid retrieval stopped at 241 words because 1.4% of 361-word passages exceeded the dense encoder's 512-token input limit. On the training set at 1,002 words across the collection, 181-word passages raised condition recall over 121-word passages by 8.7 points [5.4, 12.0] for BM25 (62.1% against 53.4%) but by only 2.8 points for dense retrieval [−0.3, 5.9] and hybrid retrieval [−0.1, 5.6] (70.0% and 72.2% against 67.3% and 69.4%). That cut the lead of hybrid retrieval over BM25 from 16.0 to about 10 points. With the document supplied BM25 gained 1.3 points [−0.6, 3.3]. Even 61-word BM25 passages led source elements across the collection at 1,002 words (39.6% against 18.3% on training and 45.8% against 23.7% on development).

Document length (Section 6.7). Training recall at 1,002 words was averaged within thirds of the 307 training reference documents by length and reweighted to the development mix of 17 short, 10 medium and 7 long documents. Across the collection this closed about 2.9 points of each of the 10.7-point (BM25), 13.8-point (dense) and 13.1-point (hybrid) fixed-passage gaps between the sets. A linear fit on the logarithm of document length gave about 2 points.