Does Contextual Retrieval Work in Japanese?

Testing Anthropic's Method Against Structure-Aligned Chunking in a Production Hybrid RAG

Shinji Yumura
AQUA LLC, Tokyo, Japan
official@aquallc.jp

October 2026

Abstract

Anthropic's Contextual Retrieval prepends an LLM-written context to every chunk before embedding and BM25 indexing, and was reported to cut top-20 retrieval failures by 49% (67% with reranking). A recent placebo-controlled study (Kuehlkamp et al., 2026) found instead that, for dense retrieval over English corpora, such contexts lose to simply cutting documents at their headings. Neither claim has been tested in Japanese, and the second did not test the BM25 side, where Anthropic reported its largest gain. We compare seven indexing variants—production fixed-size chunks, a length-matched control, Anthropic-style contexts, heading-aligned chunks, real heading paths, shuffled heading paths as a placebo, and headings plus contexts—on the 150-article Japanese blog behind a deployed assistant, with 295 paraphrased questions (96% judged valid by a human) and 150 lexically close questions, under dense, BM25+ and hybrid (RRF) retrieval. With dense retrieval, heading-aligned chunks beat Anthropic's contexts (nDCG@10 +0.070, Holm $p < 0.001$), replicating the counter-claim; with BM25, contexts helped (+0.030, $p_{\text{Holm}} = 0.011$) and real headings beat the placebo (+0.024, $p_{\text{Holm}} = 0.019$), supporting Anthropic; after hybrid fusion no difference remained. An LLM listwise reranker over the top 50 candidates halved top-20 failures (18.0% → 9.8%, $p < 0.001$) and erased the differences between variants. The best fusion weight depended on how questions were phrased. Deployed in the production assistant, chunk-level hybrid retrieval with reranking raised the share of questions whose evidence reached the generator from 59% to 81%, and answer accuracy from 67% to 87% (21 vs. 1 discordant), while cutting the context from about 30,000 to 4,000 characters, at the cost of about 1.2 s added latency.

1 Introduction

Retrieval-augmented generation (RAG) [5] splits documents into chunks of a few hundred characters, retrieves those closest to a question and passes them to a language model. Short chunks lose their context: a chunk stating "the fee is 50,000 yen per month" does not say which service it describes. Anthropic's Contextual Retrieval [1] addresses this by asking an LLM to write, for each chunk, a short context that situates it within the whole document, and by prepending that context to the chunk before both embedding and BM25 indexing. The method is simple and has been widely adopted.

On 7 October 2026, Kuehlkamp et al. [4] questioned it. Part of its benefit might come not from the content of the context but from perturbing the chunk at all. Comparing it with structure-based alternatives, with shuffled heading paths as a placebo, they found that LLM-written contexts on fixed windows lose to bare heading-aligned chunks. Their evidence, however, comes from English

corpora and dense retrieval only. Anthropic reported its largest gain when contexts were also added to BM25, which the counter-study did not test, and neither study used Japanese text.

We test both claims on the same documents and questions in the Japanese index used in production by the assistant YUA. We then deploy what worked best and measure the production pipeline before and after. Our contributions are:

- a controlled comparison of seven indexing variants, including a length-matched control and a heading placebo, under dense, BM25+ and hybrid retrieval in Japanese;
- evidence that both claims hold in part: structure wins for dense retrieval, contexts win for BM25, and the differences vanish after fusion;
- evidence that an LLM reranker matters more than either indexing method, and that the best fusion weight depends on question phrasing;
- a before/after measurement of a production RAG pipeline, showing that how retrieved material is assembled for the generator affected answer accuracy more than indexing.

2 The Two Claims

Contextual Retrieval. For each chunk, Claude 3 Haiku receives the whole document and the chunk and is asked to "give a short succinct context to situate this chunk within the overall document for the purposes of improving search retrieval of the chunk" (50–100 tokens). On codebases, fiction and papers, the top-20 retrieval failure rate fell from 5.7% to 3.7% with contextual embeddings (−35%), to 2.9% when BM25 was also contextualised (−49%), and to 1.9% when the top 150 were reranked (−67%) [1].

Structure-aligned chunking. Kuehlkamp et al. [4] compared, on 200 Wikipedia featured articles and 1,585 QASPER papers with BGE embeddings, contextualised fixed windows (Anthropic's recipe verbatim), structure-aligned chunks, heading-path metadata and hierarchical two-stage retrieval, with shuffled heading paths as a placebo. Contextualised windows lost to bare structural chunks (−0.012 and −0.022 coverage-aware nDCG@10), the gain came from boundary placement, and under tight token budgets contextualisation hurt. In the same week, PRIME [2] reported that deterministic metadata matches structure-aware chunking at a fraction of the cost, and that index-side effects become negligible with a strong reranker.

3 Experimental Design

3.1 Corpus and questions

The corpus is the full blog of AQUA LLC (150 articles, 2.14 million characters), fetched with headings from the public WordPress API. Articles are heavily sectioned (median 32 headings, h2–h4, in 100 articles).

Paraphrased questions. From the first and second half of each article we sampled 3–5 consecutive sentences not aligned to any chunk boundary and asked `gemini-3.8-flash` for one question answerable from that passage alone, worded differently from the passage, together with a verbatim 20–80-character evidence sentence, which we verified against the passage. 295 of 300 questions were produced.

Table 1: Indexing variants.

Variant	Chunking / prepended text	Chunks	Mean length
A	production fixed size (≈500 chars) / none	4,979	426
As	production chunker, length-matched to C / none	6,104	347
B	as A / LLM context (Anthropic)	4,979	426 + ctx
C	heading-aligned / none	5,729	350
D	as C / real heading path	5,729	350 + path
E	as C / shuffled heading path (placebo)	5,729	350 + path
F	as C / heading path + LLM context	5,729	350 + both

Lexically close questions. With the same procedure but allowing the passage's own wording, we produced one question per article (150).

Human check. The author compared 50 randomly sampled paraphrased questions with their passages: 48 (96%) were valid, one had misaligned evidence and one could not be answered from the passage. Excluding these two changes no conclusion (Section 8).

3.2 Indexing variants

Table 1 lists the seven variants. B−A estimates Anthropic's method; C−As the effect of boundary placement at matched length; D−E the content of real headings against a placebo; B−C compares the two claims directly. Contexts used Anthropic's prompt verbatim with one added sentence, "Write the context in the same language as the document" (English contexts would not match Japanese BM25 terms), generated by `gemini-3.5-flash-lite`; 10,707 of 10,708 contexts were produced, for $13.5 in total. All variants were embedded with `gemini-embedding-001` (3,072 dimensions; RETRIEVAL_DOCUMENT for chunks, RETRIEVAL_QUERY for questions), as in production, including any prepended text.

3.3 Retrieval

Dense: cosine similarity over all chunks. *BM25+* [7, 8]: the production formula and parameters ($k_1 = 1.2$, $b = 0.75$, $\delta = 1$) and the production tokenizer (dictionary terms, English words and character bigrams); we also tested a morphological tokenizer (kuromoji, base forms of content words). *Hybrid*: Reciprocal Rank Fusion [3] ($k = 60$) with equal weights, as in Anthropic's setup, and with the production weights (dense 0.4, BM25 1.6). *Reranking*: the top 50 fused candidates were shown with the question to `gemini-3.5-flash-lite`, which returned their order (listwise LLM reranking [9]); we kept the top 20.

3.4 Metrics and statistics

A chunk is relevant if its own text (not the prepended text) contains at least 70% of the evidence sentence contiguously, ignoring whitespace and punctuation; this yields the same criterion for every chunking. The primary metric is the top-20 failure rate (no relevant chunk in the top 20), as in [1]; we also report hit@5, nDCG@10 and MRR. Variants are compared on the same questions: McNemar's exact test for failures; for nDCG@10, the Wilcoxon signed-rank test and a 95% interval from 4,000 bootstrap resamples clustered by article. Holm's correction is applied over the six contrasts within each retrieval mode.

Table 2: Results on the 295 paraphrased questions. Fail@20: lower is better; nDCG@10: higher is better. Hybrid uses equal-weight RRF.

	Dense		BM25+		Hybrid	
Variant	Fail@20	nDCG	Fail@20	nDCG	Fail@20	nDCG
A production	14.9%	.569	34.2%	.385	18.0%	.535
As length control	15.6%	.584	37.3%	.380	17.6%	.538
B contexts	16.6%	.549	31.5%	**.415**	17.3%	.551
C headings cut	16.6%	**.619**	36.9%	.371	15.6%	.538
D real paths	17.3%	.595	33.6%	.377	18.0%	.536
E placebo paths	18.6%	.576	35.9%	.352	18.6%	.520
F paths + contexts	17.6%	.586	32.2%	.406	17.3%	.551

Table 3: Paired differences in nDCG@10 (Holm-adjusted p in parentheses where < 0.3; bold: $p_{\text{Holm}} < 0.05$). No difference in Fail@20 survived correction.

Contrast	Meaning	Dense	BM25+	Hybrid
B − A	contexts	−0.020	+**0.030** (.011)	+0.016
C − As	boundary placement	+0.034 (.27)	−0.009	+0.000
D − C	add heading path	−0.024	+0.005	−0.003
D − E	real vs. placebo	+0.019 (.16)	+**0.024** (.019)	+0.016
F − D	contexts on top of paths	−0.008	+**0.030** (.048)	+0.015
B − C	Anthropic vs. headings	−**0.070** (<.001)	+0.043 (.077)	+0.013

4 Do Contexts and Headings Help?

Dense retrieval: the counter-claim holds. Contexts did not improve dense retrieval (B−A −0.020, 95% CI −0.046 to +0.006), and heading-aligned chunks clearly beat them (B−C −0.070, −0.102 to −0.037, $p_{\text{Holm}} < 0.001$; Tables 2–3). The finding of [4] replicates in Japanese with a large commercial embedding model. The boundary effect itself (C−As +0.034, $p = 0.067$) and the heading-path effect (D−C −0.024) were not significant; adding paths tended to hurt. Real paths beat the placebo before but not after correction (+0.019, $p = 0.031$).

BM25: Anthropic's claim holds. Contexts improved BM25 (B−A +0.030, +0.014 to +0.047, $p_{\text{Holm}} = 0.011$), also on top of heading paths (F−D +0.030, $p_{\text{Holm}} = 0.048$), and real paths beat the placebo (D−E +0.024, $p_{\text{Holm}} = 0.019$). For lexical retrieval, the words that contexts and headings add are what helps. This is the side the counter-study did not test.

Hybrid: no difference. After equal-weight fusion no contrast was significant; the variant that wins under dense retrieval and the one that wins under BM25 offset each other. The morphological tokenizer did not change BM25 (A: 32.5% vs. 34.2% failures, $p = 0.40$).

5 Reranking

Reranking had the largest effect (Table 4). For the production chunking with equal-weight fusion, top-20 failures fell from 18.0% to 9.8% (30 vs. 6 discordant questions, $p = 0.00007$) and hit@5

Table 4: LLM reranking of the top 50 (295 paraphrased questions).

Candidates from	Fail@20 before	Fail@20 after	Hit@5 after	MRR after
A, hybrid (equal)	18.0%	**9.8%**	80.0%	.657
B, hybrid	17.3%	12.9%	74.6%	.613
C, hybrid	15.6%	12.9%	77.6%	.663
F, hybrid	17.3%	11.5%	79.0%	.660
A, dense only	14.9%	12.2%	77.6%	.664
A, hybrid (production weights)	29.8%	22.7%	69.5%	.589

Table 5: Top-20 failure rate by question phrasing (variant A).

Retrieval	Paraphrased (295)	Lexically close (150)
Dense only	**14.9%**	11.3%
BM25+ only	34.2%	3.3%
Hybrid, equal weights	18.0%	2.7%
Hybrid, production weights	29.8%	**2.0%**
Hybrid, equal weights + reranking	**9.8%**	**2.0%**

rose from 63.1% to 80.0%. After reranking, the indexing variants no longer differed: A had the lowest failure rate, and its differences from B, C and F were not significant ($p = 0.09$–0.44), in line with [2]. Reranking cost about $0.004 per question; the $13.5 spent on contexts bought nothing once a reranker was used. Reranking candidates fused with the production weights reached only 22.7%: a reranker cannot recover evidence absent from its candidate pool.

6 Question Phrasing Determines the Best Fusion

Because the paraphrased questions were deliberately worded unlike the passages, they disadvantage BM25. On the 150 lexically close questions the ranking of retrieval modes reversed (Table 5): BM25 alone failed on 3.3% and the production weights were best (2.0%), whereas on paraphrased questions they were worst. The best fusion weight therefore depends on how users phrase questions, which is rarely known. Equal-weight fusion followed by reranking was best on both sets.

7 Deployment

The production pipeline. The experiments above are chunk-level. The production pipeline was document-level: dense and BM25 results were aggregated per article, boosted by article type and keywords, the top three articles were selected, and the *first 25 chunks* of each were passed to the generator. Calling the production search code unchanged, the correct article was among the top three for 76.9% of questions, but the evidence reached the generator's context for only 59.3%: in long articles, evidence beyond the first 25 chunks was cut. The context averaged about 30,000 characters.

The new pipeline. Behind a switch, we added a chunk-level path: the top 50 chunks by dense similarity and by BM25+ are fused with equal-weight RRF, reranked by `gemini-3.5-flash-lite`,

Table 6: Production search code, same index: previous vs. new pipeline.

	Previous	New
Evidence in context, paraphrased (295)	59.3%	**81.0%**
Evidence in context, lexically close (150)	82.0%	**96.7%**
Correct article in top three (295)	76.9%	85.1%
Answers correct (100)	67	**87**
Context length (characters)	≈29,800	≈4,100
Generator input tokens per question	15,357	2,148
Search latency, median / 90th pct.	0.52 / 0.57 s	1.68 / 2.16 s

and the top eight chunks are grouped by article and passed on. If reranking fails or exceeds 6 s, the fused order is used; after hardening the response parsing, no reranking failure occurred in either measurement run.

Evidence reached the context for 81.0% instead of 59.3% of paraphrased questions (81 vs. 17 discordant, $p < 10^{-6}$) and for 96.7% instead of 82.0% of lexically close ones (25 vs. 3, $p = 0.00003$; Table 6). For answer accuracy, 100 randomly chosen paraphrased questions were answered by `gemini-3.8-flash` (the production chat model) from the context each pipeline actually returned, with a fixed instruction to answer only from the material, and scored against the reference answer by `gemini-3.8-flash`, blind to pipeline. Correct answers rose from 67 to 87 (21 vs. 1 discordant, $p = 0.00001$). The context shrank sevenfold; counting generator input and reranking, cost per question fell from about \$0.012 to \$0.005. Search latency rose by about 1.2 s. The new pipeline was deployed on 17 October 2026.

8 Robustness

- **Question validity.** Excluding the two questions flagged by the human check leaves the main contrasts unchanged to three decimals (e.g. dense B−C −0.070, BM25 B−A +0.030).
- **Length.** The length-matched control As performed like A under every mode, so the shorter heading-aligned chunks do not explain the results.
- **Placebo.** Shuffled heading paths (E) were the worst or near-worst variant under every mode, so the heading effect is partly due to content (significant for BM25).
- **Relevance criterion.** The 70%-contiguous criterion counts evidence split across a chunk boundary when one side holds most of it.
- **Precision.** Bootstrap intervals for nDCG@10 differences had half-widths of 0.02–0.04; effects as small as those reported by [4] (0.012–0.022) cannot be resolved with 295 questions.

9 Discussion

Both claims are half right. Dense embeddings compress a chunk's meaning into one vector, so prepended context can blur the chunk as well as situate it; BM25 counts term matches, so the title and topic words that contexts and headings add translate directly into hits. This explains why Anthropic saw its largest gain once BM25 was contextualised, and suggests that the counter-study's conclusion is specific to dense retrieval.

Rerank before re-indexing. A reranker halved failures and removed the differences between indexing variants, for a fraction of a cent per question and without rebuilding the index, whereas contextualisation must be repeated whenever documents change. For a corpus of this size, reranking is the first thing to add.

Assembly matters more than indexing. The largest production gain came not from choosing an indexing method but from fixing how retrieved material reached the generator: a quarter of the evidence in correctly retrieved articles never reached it. Most RAG evaluations stop at chunk-level recall; in production, what is passed to the model, and how much (cf. [6]), changed answers more.

10 Limitations

- **One corpus.** A single company blog of 150 heavily sectioned articles; manuals, legal text or larger collections may differ.
- **Synthetic questions.** Questions and evidence were generated by an LLM and only 50 were checked by a person; they are not real user queries, and Section 6 shows that phrasing matters.
- **One model family.** Contexts, reranking, embeddings, answers and scoring all used Gemini models; Anthropic used Claude 3 Haiku and a dedicated reranker.
- **LLM scoring of answers.** Answer correctness was judged by an LLM against reference answers, without human scoring.
- **Simplified generation.** Answers used a fixed instruction, not the production system prompt, and ignored the assistant's web search.
- **Latency.** The effect of the added 1.2 s on users was not measured.

11 Conclusion

On a Japanese production index, LLM-written chunk contexts lost to heading-aligned chunking under dense retrieval and won under BM25, and neither mattered after hybrid fusion. Reranking the top 50 with a small LLM halved retrieval failures and erased the differences. Replacing a document-level pipeline that passed only the beginning of each article with chunk-level retrieval and reranking raised answer accuracy from 67% to 87% while cutting the context sevenfold. How candidates are gathered, ordered and handed to the model mattered more than how the index was built.

Use of AI assistants. The author used AI assistants for code development, experiment orchestration, and drafting of this manuscript, and takes full responsibility for its content.

Data and code. Questions, the human check, chunk metadata, per-question retrieval results, reranked lists, production measurements and answers, and all scripts are available at `https://github.com/YUALAB/contextual-retrieval-japanese` and archived at `https://doi.org/10.5281/zenodo.23426281`. Re-running the analysis script on the released results reproduces every contrast reported here. Article texts are not redistributed; a script re-downloads them from the public API. A Japanese version of this report is available at `https://www.aquallc.jp/research/contextual-retrieval-japanese/`.

References

[1] Anthropic. Introducing contextual retrieval. https://www.anthropic.com/news/contextual-retrieval, September 2024.

[2] Cheuk Hei Chung and Li Liu. PRIME: Pre-retrieval index modeling enhancement for structured document retrieval, 2026.

[3] Gordon V. Cormack, Charles L. A. Clarke, and Stefan 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*, pages 758–759, 2009.

[4] Andrey Kuehlkamp, Priscila Correa Saboia Moreira, and Samuel Rund. Does document structure help dense retrieval? A placebo-controlled ablation of four mechanisms across two corpora, 2026.

[5] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474, 2020.

[6] Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024.

[7] Yuanhua Lv and ChengXiang Zhai. Lower-bounding term frequency normalization. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, pages 7–16, 2011.

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

[9] Weiwei Sun, Lingyong Yan, Xinyu Ma, Shuaiqiang Wang, Pengjie Ren, Zhumin Chen, Dawei Yin, and Zhaochun Ren. Is ChatGPT good at search? Investigating large language models as re-ranking agents. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14918–14937, 2023.