AQUA RESEARCH · TECHNICAL REPORT 007

Contextual Retrieval は日本語でも効くか

Anthropic の手法と、見出しで切る反論を、本番の日本語 RAG で確かめる

Does Contextual Retrieval Work in Japanese? Testing Anthropic's Method Against Structure-Aligned Chunking in a Production Hybrid RAG

湯村 信司（Shinji Yumura）　AQUA合同会社　2026年10月17日

この報告が確かめた論文

Anthropic「Introducing Contextual Retrieval」2024年9月

A. Kuehlkamp, P. C. S. Moreira, S. Rund「Does Document Structure Help Dense Retrieval?」arXiv:2610.10170（2026年10月）

切れ端ごとに LLM の説明文を添える Anthropic の手法と、「説明文より、見出しで切るだけの方が良い」という反論を、反論が試していなかった BM25 の側と日本語を含めて、同じ文書・同じ問題で比べた。

67 → 87 %

本番の YUA の答えの正しさ（100問）。切れ端単位で探し、LLM の並べ替え役を足したとき。AI に渡す文脈は約3万字から約4千字に

1勝1敗

Anthropic の説明文は、埋め込みでは見出しで切るだけの形に負け（反論の論文と同じ）、BM25 では効いた（反論の論文が試していない側）。両方を合わせると差は消えた

18.0 → 9.8 %

上位20件に答えが入らない割合。並べ替え役を足したとき。並べ替え後は、説明文や見出しの差はなくなった

ABSTRACT

Anthropic's Contextual Retrieval prepends an LLM-written context to every chunk before embedding and BM25 indexing, and reported a 49% reduction in top-20 retrieval failures (67% with reranking). A placebo-controlled study by Kuehlkamp et al. (arXiv:2610.10170, October 2026) found instead that, for dense retrieval in English, 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

compared seven indexing variants (fixed 500-character chunks as in production, a length-matched control, Anthropic's 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, Holm $p = 0.011$) and real headings beat the placebo (+0.024, Holm $p = 0.019$), supporting Anthropic; after hybrid fusion no difference remained. An LLM listwise reranker over the top 50 halved top-20 failures (18.0% → 9.8%, $p < 0.001$) and erased the remaining differences between variants. The best fusion weight depended on how users phrase questions. Deployed in the production assistant, chunk-level hybrid retrieval with reranking raised the share of questions whose evidence reached the model 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 a cost of about 1.2 s added latency.

要旨

Anthropic の Contextual Retrieval は、文書を切り分けた各切れ端に、LLM が書いた短い説明文を添えてから、埋め込みと BM25 の両方に入れる手法である。上位20件の探し漏れが49%減り、並べ替え役を足すと67%減ったと報告された。これに対し、Kuehlkamp ほか（arXiv:2610.10170、2026年10月）は、偽の見出しを対照に置いた実験で、英語の埋め込み検索では、この説明文は「見出しで切るだけ」に負けると示した。どちらも日本語では試されておらず、反論の論文は、Anthropic が最も効いたとした BM25 の側を試していない。
本稿は、AI アシスタント YUA の本番の索引（AQUA合同会社のブログ150記事）で、7つの索引の作り方を比べた。本番と同じ約500文字の切り方、長さをそろえた対照、Anthropic の説明文、見出しで切る形、本物の見出しの道筋、入れ替えた偽の見出し（偽の対照）、見出しと説明文の両方である。問題は、本文と違う言葉で聞く295問（人の確認で96%が妥当）と、本文の言葉に近い150問で、埋め込み・BM25+・両方を合わせる形（RRF）で探した。
埋め込みでは、見出しで切る形が Anthropic の説明文に勝った（nDCG@10 +0.070、Holm 補正後 $p < 0.001$）。反論の論文と同じ結果である。BM25 では、説明文が効き（+0.030、$p = 0.011$）、本物の見出しが偽の見出しに勝った（+0.024、$p = 0.019$）。こちらは Anthropic を支持する。両方を合わせると、どの差も消えた。上位50件を LLM で並べ直すと、上位20件の探し漏れは 18.0% から 9.8% に半減し（$p < 0.001$）、作り方どうしの差もなくなった。合わせ方の一番良い重みは、利用者の聞き方で変わった。この形を本番の YUA に入れると、答えの根拠が AI に届く割合は59%から81%に、答えの正しさは67%から87%に上がり（片方だけ正解

21対1)、AI に渡す文脈は約3万字から約4千字になった。代わりに、検索は約1.2秒長くなった。

1 はじめに

RAG（検索拡張生成）は、文書を数百文字ずつの切れ端に分け、質問に近い切れ端を探して AI に渡す。切れ端は短いので、それだけでは「何の話か」がわからなくなる。「料金は月5万円です」という切れ端からは、どのサービスの料金かがわからない。

Anthropic は2024年に、この弱点への手当てとして Contextual Retrieval を発表した [1]。各切れ端に、文書全体を読んだ LLM が「これは何の文書の、どの部分か」を短く書き、それを切れ端の先頭に添えてから索引に入れる。手軽で効果が大きいとされ、広く使われている。

2026年10月7日、Kuehlkamp ほか [2] は、これに疑問を投げた。説明文の効き目の一部は、中身ではなく「切れ端に文字を足したこと」そのものかもしれない。そこで、文書の構造（見出し）を使う手法と並べ、偽の見出しを付けた対照まで置いて比べたところ、LLM の説明文は、見出しの区切りで切るだけの形に負けた。

ただし、この反論は英語の文書と、埋め込みによる検索だけで確かめられている。Anthropic が最も効いたとしたのは、説明文を埋め込みと BM25（言葉で探す検索）の両方に入れたときであり、反論の論文は BM25 を試していない。日本語の会社のブログで、どちらの主張が成り立つのかはわかっていない。

本稿は、AI アシスタント YUA が本番で使う日本語の索引で、この2つの主張を同じ文書・同じ問題で比べる。さらに、実験で一番効いたものを本番に入れ、本番の答えがどう変わったかまで測る。

2 2つの主張

2.1 Anthropic：Contextual Retrieval

- **やり方**：文書全体と切れ端を LLM（Claude 3 Haiku）に渡し、「この切れ端を文書の中に位置づける短い説明を書いて」と頼む（50〜100トークン）。説明文を切れ端の先頭に足し、埋め込みと BM25 の両方の索引に入れる
- **報告**：上位20件の探し漏れが、説明文を埋め込みに入れると 5.7% → 3.7%（35%減）、BM25 にも入れると 2.9%（49%減）、上位150件を並べ替え役（Cohere の reranker）で並べ直すと 1.9%（67%減）。文書はプログラムのコード・小説・論文など

2.2 反論：見出しで切るだけで十分

- **やり方**：Wikipedia の秀逸な記事200本と、論文1,585本で、決まった長さで切って説明文を添える形（Anthropic の手順そのまま）と、見出しの区切りで切る形、見出しの道筋を添える形、2段階で探す形を比べた。偽の対照として、見出しの道筋を入れ替えた形を置いた。埋め込みは BGE（小・大）
- **報告**：説明文を添えた形は、見出しで切るだけ（何も添えない）形に負けた（cov-nDCG@10 で −0.012〜−0.022）。効いていたのは説明文の中身ではなく、切る場所だった。切れ端を短くすると、説明文はかえって悪くした。著者のすすめは「説明文に使う LLM のお金を、構造を読み取ることに使え」

同じ週に出た PRIME [3] も、決まった手順で作る見出しなどの情報が、LLM で切り直す形と同じくらい効き、計算は45分の1で済むと報告した。ただし、強い並べ替え役を入れると、索引の工夫の差は消えるとも述べている。

---

3 実験のしかた

3.1 文書と問題

- **文書**：YUA が答えに使う AQUA合同会社のブログの全記事150本（公開の WordPress の API から、見出しを残したまま取得。合わせて約214万字）。記事は見出しが多く、100本で数えると1記事あたり中央値32個の見出し（h2〜h4）があった
- **問題（言い換え）**：各記事の前半と後半から、どの切り方の境目にもそろわない3〜5文を乱数で選び、そこだけで答えられる質問を Gemini 3.8 Flash に作らせた。答えの根拠になる一文（20〜80字）を本文からそのまま抜き出させ、本文にあることを確かめた。質問は「一節と同じ言葉をなるべく使わず言い換える」よう指示した。300問のうち295問ができた
- **問題（本文の言葉に近い）**：同じ手順で、本文の言葉をそのまま使ってよい質問を、各記事1問、150問作った（6節）
- **人の確認**：言い換えの問題から乱数で50問を選び、著者が一節と見比べた。48問（96%）は妥当、1問は根拠と答えがずれ、1問は一節では答えられなかった。この2問を除いても結論は変わらなかった（8節）

3.2 7つの索引の作り方

形	切り方	添えるもの	切れ端の数	平均の長さ
A	本番と同じ（約500字で、段落と文で区切る）	なし	4,979	426字
As	本番と同じ切り方で、平均の長さを C にそろえる	なし（長さの対照）	6,104	347字
B	A と同じ	LLM の説明文（Anthropic）	4,979	426字＋説明文
C	見出しの区切りで切る（長い節は本番と同じやり方で分け、短い節は同じ大見出しの中でつなぐ）	なし	5,729	350字
D	C と同じ	本物の見出しの道筋（題名 > 大見出し > 小見出し）	5,729	350字＋道筋
E	C と同じ	偽の見出しの道筋（同じ記事の別の切れ端の道筋と入れ替え）＝偽の対照	5,729	350字＋道筋
F	C と同じ	本物の見出しの道筋＋LLM の説明文	5,729	350字＋両方

表1　7つの索引の作り方。B–A が Anthropic の手法の効き目、C–As が「切る場所」の効き目（長さはそろえてある）、D–E が「本物の見出しの中身」の効き目、B–C が2つの主張の直接の比べに当たる。

- **説明文**：Anthropic の公開した指示文をそのまま使い、最後に「文書と同じ言葉で書く」とだけ足した（英語の説明文が付くと、日本語の BM25 が働かないため）。作るモデルは Gemini 3.5 Flash-Lite（Anthropic は小さなモデルの Claude 3 Haiku）。10,707個の説明文で、1個の失敗を除きすべて作れた。費用は合わせて $13.5（入力の約7割は自動のキャッシュが効いた）
- **埋め込み**：本番と同じ gemini-embedding-001（3,072次元。索引は RETRIEVAL_DOCUMENT、質問は RETRIEVAL_QUERY）。添えた文も含めて埋め込む

3.3 探し方

- **埋め込み**：全切れ端との近さ（コサイン類似度）の順
- **BM25+**：本番と同じ式と係数（k1 = 1.2、b = 0.75、δ = 1）。言葉の切り方も本番と同じ（登録した辞書の言葉・英単語・2文字ずつ）。日本語の品詞で切る形（kuromoji、名詞・動詞・形容詞・副詞の元の形）も

別に測った

- **両方（RRF）**：2つの順位を Reciprocal Rank Fusion [4]（k = 60）で合わせる。重みは等しく（Anthropic と同じ）と、本番の重み（埋め込み 0.4、BM25 1.6）の2通り
- **並べ替え役**：上位50件の切れ端を、質問とともに Gemini 3.5 Flash-Lite に見せ、役に立つ順に番号を並べさせる（LLM で一度に並べる形 [5]）。並べ直した上位20件を使う

3.4 物差しと統計

- **当たり**：切れ端の本文に、根拠の一文の70%以上が続けて入っていれば当たりとした。切り方によらず同じ物差しで比べるためで、添えた文は判定に使わない
- **探し漏れ@20**：上位20件に当たりが1つもない問題の割合（Anthropic と同じ物差し）。ほかに、上位5件の当たり、nDCG@10、MRR
- **統計**：同じ問題で形どうしを対応させて比べた。探し漏れは McNemar の正確検定、nDCG@10 は Wilcoxon の符号付き順位検定と、記事ごとにまとめて4,000回取り直した95%区間。探し方ごとに、主な6つの比べ（B−A、C−As、D−C、D−E、F−D、B−C）に Holm の補正をかけた

---

4 説明文と見出しは効いたか

形	埋め込み 漏れ@20	埋め込み nDCG@10	BM25 漏れ@20	BM25 nDCG@10	両方 漏れ@20	両方 nDCG@10
A 本番の切り方	14.9%	0.569	34.2%	0.385	18.0%	0.535
As 長さの対照	15.6%	0.584	37.3%	0.380	17.6%	0.538
B 説明文（Anthropic）	16.6%	0.549	31.5%	**0.415**	17.3%	0.551
C 見出しで切る	16.6%	**0.619**	36.9%	0.371	15.6%	0.538
D 本物の見出し	17.3%	0.595	33.6%	0.377	18.0%	0.536
E 偽の見出し	18.6%	0.576	35.9%	0.352	18.6%	0.520
F 見出し＋説明文	17.6%	0.586	32.2%	0.406	17.3%	0.551

表2　言い換えの295問での成績。漏れ@20 は低いほど、nDCG@10 は高いほど良い。両方は等しい重みの RRF。

比べ	意味	埋め込み	BM25	両方
B − A	説明文の効き目	−0.020	**+0.030**（0.011）	+0.016
C − As	切る場所の効き目	+0.034（0.27）	−0.009	+0.000
D − C	見出しの道筋を添える	−0.024	+0.005	−0.003
D − E	本物 対 偽の見出し	+0.019（0.16）	**+0.024**（0.019）	+0.016
F − D	見出しの上に説明文	−0.008	**+0.030**（0.048）	+0.015
B − C	Anthropic 対 見出しで切る	**−0.070**（<0.001）	+0.043（0.077）	+0.013

表3　nDCG@10 の差（同じ問題で対応をとった平均）。かっこ内は Holm の補正後の p（0.3 未満のものだけ）。太字は補正後に p < 0.05。探し漏れ@20 の差は、どれも補正後に有意でなかった。

4.1 埋め込みでは、反論の論文が正しかった

Anthropic の説明文は、埋め込みの検索を良くしなかった（B − A：−0.020、95%区間 −0.046〜+0.006）。そして、見出しの区切りで切るだけの形が、説明文を添えた形に、はっきり勝った（B − C：−0.070、95%区間 −0.102〜−0.037、補正後 p < 0.001）。英語と BGE で示された反論 [2] が、日本語と Gemini の大きな埋め込みでも再現された。

ただし、切る場所の効き目そのもの（C − As：+0.034、p = 0.067）と、見出しの道筋を添える効き目（D − C：−0.024）は、どちらも偶然の範囲だった。見出しの道筋を添えると、むしろ下がる向きにあった。本物の見出しは偽の見出しより上の向き（+0.019、補正前 p = 0.031）だったが、補正後は有意でなかった。

4.2 BM25 では、Anthropic が正しかった

BM25 では、説明文がはっきり効いた（B − A：+0.030、95%区間 +0.014〜+0.047、補正後 p = 0.011）。見出しの上に説明文を足しても効いた（F − D：+0.030、p = 0.048）。本物の見出しは偽の見出しより良かった（D − E：+0.024、p = 0.019）。言葉で探す検索では、説明文や見出しが足す「言葉」そのものが役に立つ。反論の論文が試していなかった側で、Anthropic の主張が成り立った。

4.3 両方を合わせると、差は消えた

本番と同じく埋め込みと BM25 を合わせると、どの比べも偶然の範囲だった（表3の右の列）。埋め込みで勝つ形と BM25 で勝つ形が違うため、合わせると打ち消し合ったと考えられる。

なお、日本語の品詞で切る BM25 は、本番の言葉の切り方（辞書＋2文字）と差がなかった（A で探し漏れ 32.5% 対 34.2%、p = 0.40）。

5 並べ替え役

形と探し方	並べ替え前 漏れ@20	並べ替え後 漏れ@20	並べ替え後 上位5の当たり	並べ替え後 MRR
A 本番の切り方・両方（等しい重み）	18.0%	**9.8%**	80.0%	0.657
B 説明文・両方	17.3%	12.9%	74.6%	0.613
C 見出しで切る・両方	15.6%	12.9%	77.6%	0.663
F 見出し＋説明文・両方	17.3%	11.5%	79.0%	0.660
A 本番の切り方・埋め込みだけ	14.9%	12.2%	77.6%	0.664
A 本番の切り方・本番の重み	29.8%	22.7%	69.5%	0.589

表4　上位50件を Gemini 3.5 Flash-Lite で並べ直した結果（295問）。A・両方では、並べ替えで探し漏れが半分になった（片方だけ見つけた問題 30対6、p = 0.00007）。上位5件の当たりは 63.1% → 80.0%。

一番大きく効いたのは、索引の工夫ではなく、並べ替え役だった。本番と同じ切り方で、両方を等しく合わせた上位50件を並べ直すと、探し漏れは 18.0% から 9.8% に半減し、上位5件に答えが入る割合は 63% から 80% に上がった。

並べ替えのあとは、説明文や見出しの工夫の差はなくなった。本番の切り方（A）が最も低い探し漏れで、B・C・F との差はどれも偶然の範囲だった（p = 0.09〜0.44）。PRIME [3] の「強い並べ替え役があると、索引の工夫の差は消える」と同じである。費用は、並べ替え1回あたり約 $0.004 で、説明文を作る費用（$13.5）を使う必要はなかった。

本番の重み（BM25 を4倍重く）で合わせた候補を並べ直しても、22.7% までしか下がらなかった。並べ替え役は、候補の50件に入っていない答えは拾えない。候補をどう集めるかが、並べ替えの効き目を決める。

6 質問の言い方で、一番良い合わせ方が変わる

言い換えの問題では、埋め込みだけ（14.9%）が、本番の重みの両方（29.8%）よりはるかに良かった。しかし、この問題集は「本文と違う言葉で聞く」ように作ったので、言葉で探す BM25 に不利である。そこで、本文の言葉に近い150問でも測った。

探し方（形 A）	言い換えの295問 漏れ@20	本文の言葉に近い150問 漏れ@20
埋め込みだけ	**14.9%**	11.3%
BM25 だけ	34.2%	3.3%
両方（等しい重み）	18.0%	2.7%
両方（本番の重み）	29.8%	**2.0%**
両方（等しい重み）＋並べ替え役	**9.8%**	**2.0%**

表5　質問の言い方による違い。本番の重みは、本文と同じ言葉で聞く人には最良で、言い換えて聞く人には最悪だった。等しい重み＋並べ替え役は、どちらでも最良だった。

結果は逆になった。本文の言葉に近い質問では BM25 が強く（3.3%）、本番の重みが最良（2.0%）だった。一番良い重みは、利用者がどんな言葉で聞くかで決まる。本番の YUA の利用者の聞き方はわからないので、どちらにも強い形を選ぶのがよい。等しい重みで合わせて並べ直す形は、どちらの問題でも最良（9.8%・2.0%）だった。

7 本番に入れた

7.1 本番の検索の弱点

ここまでは切れ端単位の実験である。本番の YUA の検索は、これと違い、記事単位で動いていた。埋め込みと BM25 の結果を記事ごとにまとめ、記事の種類や言葉による加点をして上位3記事を選び、各記事の**先頭から25個の切れ端**をまとめて AI に渡す。本番の検索のプログラムをそのまま呼んで測ると、正しい記事を上位3記事に選べたのは76.9%だったが、答えの根拠が AI に渡る文脈に入ったのは59.3%だった。長い記事では、答えが後半にあると先頭25切れ端に入らない。しかも、AI に渡す文脈は毎回約3万字あった。

7.2 新しい形

実験の結果をもとに、本番の検索に新しい形を足した（切り替えのスイッチ付き）。全切れ端から、埋め込みの上位50件と BM25+ の上位50件を、等しい重みの RRF で合わせ、上位50件を Gemini 3.5 Flash-Lite で並べ直し、上位8切れ端を記事ごとにまとめて渡す。並べ替えが失敗したり6秒を超えたりしたときは、並べ替えの前の順で続ける。返事の読み取りを丈夫にした後の測定（295問と150問）では、並べ替えの失敗は0回だった。

本番の検索のプログラムで測定	今までの形	新しい形
根拠が文脈に入った（言い換えの295問）	59.3%	**81.0%**
根拠が文脈に入った（本文の言葉に近い150問）	82.0%	**96.7%**
正しい記事が上位3記事に入った（295問）	76.9%	85.1%
答えが正しかった（100問）	67	**87**
AI に渡す文脈の長さ	約29,800字	約4,100字
答える AI への入力（1問あたり）	15,357トークン	2,148トークン
検索の時間（中央値・90%）	0.52秒・0.57秒	1.68秒・2.16秒

表6　本番の検索のプログラム（同じ索引）で、今までの形と新しい形を比べた。根拠が文脈に入った問題は、新しい形だけ81問・今までの形だけ17問（$p < 0.000001$）。本文の言葉に近い問題では25問対3問（$p = 0.00003$）。答えは、新しい形だけ正解21問・今までの形だけ正解1問（$p = 0.00001$）。

答えの正しさは、言い換えの問題から乱数で選んだ100問で、本番の検索が実際に渡す文脈だけを Gemini 3.8 Flash（本番のチャットと同じモデル）に見せて答えさせ、Gemini 3.8 Flash が正解と比べて採点した。どちらの形の答えかは、採点役に知らせていない。答えの正しさは 67問 から 87問 に上がり、正解が入れ替わった問題は21対1だった。

AI に渡す文脈は約7分の1になった。答える AI の入力の料金（3.8 Flash）と並べ替えの料金を足すと、1問あたりの費用は約 $0.012 から約 $0.005 に下がった。代わりに、検索は中央値で約1.2秒長くなった。この形は2026年10月17日に本番の YUA に入れた。

答える指示は「資料だけを使って短く答える」という簡単なもので、本番の YUA の指示文ではない。本番の YUA は、ネット検索の結果も合わせて答えることがある。

8 結果の確かさ

- **問題の質**：人が見た50問のうち、妥当でなかった2問を除いても、主な結論（B − C の埋め込み −0.070、B − A の BM25 +0.030 など）は小数第3位まで同じだった
- **長さ**：見出しで切る形は切れ端が短くなるので、本番の切り方のまま平均の長さをそろえた対照（As）と比べた。As は A とほぼ同じ成績で、長さの違いは結果をほとんど動かさなかった
- **偽の対照**：偽の見出し（E）は、どの探し方でも最も低いか、それに近い成績だった。見出しの効き目は、文字を足したことではなく、中身による部分がある（BM25 で補正後に有意）

- **正解の判定**：切り方によって切れ端の境目が違うので、根拠の一文の70%以上が続けて入っていれば当たりとした。根拠が2つの切れ端にまたがるときも、どちらかに7割あれば当たりになる
- **検出力**：295問の対応のある比べで、nDCG@10 の差の95%区間の幅は ±0.02〜0.04 だった。反論の論文の主な差（0.012〜0.022）より小さい差は、この問題の数では見分けられない

---

9 考察

9.1 2つの主張は、どちらも半分正しい

埋め込みでは、LLM の説明文は効かず、見出しで切るだけの形に負けた。埋め込みは文の意味をまとめて1つの数の並びにするので、切れ端の先頭に足した説明文は、本文の意味をぼかす方向にも働く。一方、BM25 は言葉の一致を数えるので、説明文や見出しが足す「記事の題名・話題の言葉」が、そのまま当たりを増やす。Anthropic の報告で、BM25 にも入れたときに効き目が大きかったのは、このためと考えられる。反論の論文が BM25 を試していれば、結論は「埋め込みには効かないが、BM25 には効く」になった可能性がある。

9.2 実務では、索引より並べ替え

両方を合わせて並べ替え役を足すと、索引の工夫の差は消え、探し漏れは半分になった。説明文の作成は、文書全体を切れ端の数だけ読ませるので、索引を作り直すたびに費用がかかる。並べ替えは1問あたり約0.6円で、索引を変えずに足せる。少なくともこの規模（150記事）の日本語の会社のブログでは、索引の工夫より、並べ替え役を先に入れるのがよい。

9.3 本番の弱点は、手法より作りにあった

本番の YUA で一番効いたのは、手法の選び方ではなく、「記事単位で選び、先頭25切れ端を渡す」という作りを直したことだった。正しい記事を選んでも、根拠の4分の1近くが文脈に届いていなかった。RAG の論文の多くは、切れ端単位の探し漏れで比べる。しかし本番では、選んだものを AI にどう渡すかで、答えは大きく変わる。

---

10 限界

- **文書は1種類**：1つの会社のブログ150記事だけである。見出しの多い解説記事が中心で、マニュアルや法律の文書、もっと大きな文書の集まりでは、結果が違う可能性がある
- **問題は AI が作った**：問題と答えの根拠は Gemini 3.8 Flash が作り、人が確かめたのは50問だけである。実際の利用者の質問ではない。質問の言い方で結果が変わることは、6節で示した
- **モデルは Gemini だけ**：説明文・並べ替え・埋め込み・答え・採点がすべて Gemini の系統である。Anthropic は説明文を Claude 3 Haiku で作り、並べ替えに専用のリランカーを使った

- **答えの採点は AI**：答えの正しさは Gemini 3.8 Flash が正解と比べて採点した。人による答えの採点はしていない
- **本番の指示文ではない**：答えは簡単な決まった指示で作らせた。本番の YUA の答えは、指示文やネット検索の結果によっても変わる
- **時間**：並べ替えで検索が約1.2秒長くなる。利用者の満足への影響は測っていない

---

11 まとめ

Anthropic の Contextual Retrieval と、「見出しで切るだけで十分」という反論を、日本語の本番の索引で、同じ問題で比べた。埋め込みでは、見出しで切る形が説明文に勝ち、反論が再現された。BM25 では説明文が効き、Anthropic が支持された。両方を合わせると、差は消えた。

一番効いたのは、上位50件を小さな LLM で並べ直すことで、探し漏れは半分になり、索引の工夫の差もなくなった。本番の YUA では、記事単位で選んで先頭だけを渡す作りを、切れ端単位で探して並べ直す形に変えたことで、答えの正しさは67%から87%に上がり、AI に渡す文脈は7分の1になった。

索引をどう作るかの論争より、候補をどう集め、どう並べ、何を AI に渡すかの方が、本番の答えを大きく変える。

参考文献

1. Anthropic. “Introducing Contextual Retrieval.” anthropic.com/news/contextual-retrieval, 2024年9月19日.
2. A. Kuehlkamp, P. C. S. Moreira, S. Rund. “Does Document Structure Help Dense Retrieval? A Placebo-Controlled Ablation of Four Mechanisms Across Two Corpora.” arXiv:2610.10170, 2026.
3. C. H. Chung, L. Liu. “PRIME: Pre-Retrieval Index Modeling Enhancement for Structured Document Retrieval.” arXiv:2610.13296, 2026.
4. G. V. Cormack, C. L. A. Clarke, S. Büttcher. “Reciprocal Rank Fusion Outperforms Condorcet and Individual Rank Learning Methods.” *SIGIR*, 2009.
5. W. Sun ほか. “Is ChatGPT Good at Search? Investigating Large Language Models as Re-Ranking Agents.” *EMNLP*, 2023. arXiv:2304.09542
6. Y. Lv, C. Zhai. “Lower-Bounding Term Frequency Normalization.” *CIKM*, 2011.

データ：問題集（言い換えの295問・本文の言葉に近い150問）・人の確認・切れ端の一覧・問題ごとの検索の結果・並べ替えの結果・本番の測定と答え・集計のプログラムを、GitHub で公開し、Zenodo（doi:10.5281/zenodo.23426281）に保管した。公開した結果から集計をやり直すと、本稿の比べの数字がすべて同じになる。記事の本文は含めていない（公開の API から取り直すプログラムを付けた）。