# 企業法務向け日本語文書検索評価データセットの構築と分析

菅原祐太 [1] 丸山拓海 [1] 西野裕貴 [1] 稲垣有二 [1]
[1] 弁護士ドットコム株式会社
{y.sugawara,ta.maruyama,nishino,inagaki}@bengo4.com

## 概要

本研究では，日本語検索拡張生成（RAG）システムの性能評価と改善を目的とした企業法務向け評価データセットを構築する．構築した評価データセットを用いて既存モデルの性能検証とエラー分析を行う．一般ドメインの評価データセットでの性能評価結果と異なることを示し，特定のドメインに特化した評価データセットを作成することの重要性を明らかにする．さらに，より信頼性の高い評価データセットの確立のため，評価データの収集方法の違いがモデルの性能評価結果に与える影響について追加検証を実施する．

## 1 はじめに

近年，高性能な大規模言語モデル（LLM）の台頭に伴い，LLM を用いた質疑応答のユースケースが急速に拡大している．しかし，LLM は既知の知識に依存するため，専門性の高い分野や最新情報を必要とする場面では誤った情報を提供したり，適切な回答を生成できないといった課題がある．この課題を解決する有望な手法として，検索拡張生成（Retrieval-Augmented Generation, RAG）が注目を集めている [1, 2]．RAG は外部の知識ベースを参照することで，LLM の回答精度を向上させ幻覚を低減できるアプローチである [3]．

企業の法務部門では，日々の法律相談に対応するため関連する書籍や判例，ガイドラインといった文献の迅速な検索が不可欠である．法律は複雑で常に変化し続ける分野であり，最新かつ正確な法的情報の提供が求められる．適切な情報検索システムは，法務部門の業務効率を大幅に改善し，時間を有効活用できる可能性がある．さらに，法的リスクの低減や，より質の高い法的助言の提供にもつながると期待される．

近年，文書検索評価のための日本語データセットに関する研究が増加している [4, 5]．これらのデータセットは，RAG システムの性能評価や改善に不可欠なリソースとなっている．RAG は LLM に与える文書が推論結果に直接影響を及ぼすと考えられるため，文書検索の性能が重要になる．しかし，既存のデータセットは一般ドメインに焦点を当てており，法律分野特有の専門用語や複雑な概念を含む質問に対する RAG システムの性能を適切に評価するためには，法律分野に特化したデータセットが必要不可欠である．Harvey は法律分野における LLM の性能を評価するための包括的な英語の評価データセットである BigLaw Bench を構築している[1)]．BigLaw Bench の中には検索評価のためのデータセットも含まれている．しかし，BigLaw Bench は日本と英米法圏の法体系の違いにより、そのまま日本の法務に適用することは困難であるため，日本の法体系を反映した検索評価データセットが必要である．我々の知る限り日本語の法律分野に特化した書籍検索評価データセットは存在しない．

そこで本研究では，企業法務に特化した日本語文書検索評価データセットを構築する．このデータセットは，実際の企業法務に関する Q&A 解説記事から抽出した質問と，弁護士・企業法務向けの専門書籍から抽出した関連チャンクから構成されている．構築した評価データセットに対して，既存のモデルで性能評価を行い，リランキング性能の評価とエラー分析を行う．さらに，評価データの収集方法の違いがモデル性能の評価結果に与える影響について追加検証を実施する．本研究の成果は，法律分野における RAG システムの文書検索における手法の改善と，より信頼性の高い評価データセットの確立に貢献することが期待される．

## 2 データセットの構築

本章では，企業法務に特化した日本語書籍検索評価データセットの構築方法について述べる．文書検索タスクの評価には，検索クエリとそれに対応する

1） https://github.com/harveyai/biglaw-bench

正例文書が必要となる．法科大学院生による専門的なアノテーションと，ハードネガティブサンプリング技術を組み合わせることで，高品質かつモデルの性能を厳密に測定できる評価データセットの実現をめざす．

## 2.1 データ収集

Q&A のデータとして弁護士ドットコム株式会社の BUSINESS LAWYERS LIBRARY[2] が提供する企業法務の Q&A 解説記事[3] を用いる．これは企業法務に関わる問題について，弊社が独自に作成したコンテンツである．現行のリランキングモデルでは入力長制限があるため，現在ある 1430 件の記事のうち質問の文字列の長さが 160 以下である質問を 100 件ランダムに抽出した．

また，質問とペアを成す検索対象となるチャンクには書籍のチャンクを用意する．書籍は弁護士ドットコム株式会社が提供する BUSINESS LAWYERS LIBRARY および弁護士ドットコム LIBRARY[4] で閲覧できる弁護士・企業法務向けの法律書籍・雑誌のうち約 2,000 件を対象としている．

書籍チャンクは文分割した後，文を multilingual-e5-large[5] の tokenizer で 512 トークンを上限として結合したものである．また，弊社が開発している RAG 検索システムから，上位最大 10 件の推薦されたチャンクを法科大学院生の注釈付けに用いる．検索システムでは HyDE[6] を参考して構築されており，HyDE で生成された仮解答と質問と書籍チャンクに bge-reranker-v2-m3[6] を利用して，最終的な順位を決めている．

## 2.2 アノテーション

書籍チャンクへの注釈付けには法に関する専門知識が必要となるため，法科大学院生 2 名に依頼した．作成したアノテーションガイドラインに従い，質問のペアとなる書籍チャンク 10 件に対して以下の 3 つの定義にしたがって注釈付を依頼した．

- **関連なし**：質問に対して有用な情報を含んでいない．質問のトピックとは異なる法律分野や問題を扱っている．
- **部分的に関連あり**：質問に直接答えていないが，関連する法的背景や文脈を提供している，または質問の一部にのみ答えている．
- **関連あり**：質問に完全に，または大部分に答えている．

2) https://www.businesslawyers.jp/lib
3) https://www.businesslawyers.jp/practices
4) https://library.bengo4.com/about
5) https://huggingface.co/intfloat/multilingual-e5-large
6) https://www.google.com/url?q=https://huggingface.co/BAAI/bge-reranker-v2-m3

表 1: 書籍検索評価データセットの統計

| | 評価データ | JQaRA |
|---|---|---|
| 質問数 | 99 | 1667 |
| 質問の平均長 | 90.0 | 51.9 |
| 質問毎の検索データの平均長 | 721.4 | 205.4 |
| 質問毎のタイトルの平均長 | 21.6 | 6.8 |
| 関連なし | 57.4 | 90.3 |
| 部分的に関連あり | 1.3 | 0 |
| 関連あり | 8.0 | 9.7 |

表 2: 書籍チャンクの例．下線は法令・判例の引用を表している．

| | 書籍チャンク |
|---|---|
| 例 1 | ... 決議を行うとともに，会社法第 317 条による続行の決議を得て ... |
| 例 2 | ... 控訴審 (東京高判平 24.10.25 労経速 2164 号 3 頁) も原審の判断 ... |

## 2.3 ハードネガティブサンプリング

モデルの識別能力をより厳密に評価するためハードネガティブサンプリングを行う [7]．既存の研究に倣い BM25 と multilingual-e5-large のスコアで調整したものを降順に並べたとき，30 番目から 100 番目をハードネガティブサンプルとして採用する [8]．しかし，この候補中にも実際には関連性のある文書が含まれている可能性がある．他の研究では，回答である固有表現がチャンクに含まれているかどうかで判別する方法がある [9]．しかし，今回元々付属している回答は固有表現ではなく長い文章で構成されているためそのまま適応することはできない．今回は質問とその回答，およびサンプリングされた書籍チャンクを LLM（Claude 3.5 Sonnet[7]）に入力し，書籍チャンクが質問に対する回答に役に立つかどうかを判断させ，役にたつ書籍チャンクを除外した．

## 2.4 データセットの統計

表 1 に得られた評価データおよび一般ドメインで用いられている評価データの統計を示す．「関連あり」のアノテーションがされていない 1 件は除外

7) https://www.anthropic.com/news/claude-3-5-sonnet

表 3: リランキングモデルの性能

| | ndcg@10 | ndcg@30 |
|---|---|---|
| japanese-reranker-cross-encoder-large-v1 | 0.472 | 0.620 |
| japanese-bge-reranker-v2-m3-v1 | **0.535** | 0.663 |
| bge-reranker-v2-m3 | 0.528 | 0.654 |
| Ruri-Reranker-Large | 0.411 | 0.564 |
| JaColBERTv2.5 | 0.523 | **0.668** |

している．JQaRA は検索拡張評価のための日本語 Q&A データセットでテストデータのみの統計を示している．質問，質問毎の検索データおよびタイトル（本研究では書籍のタイトル）の平均長が JQaRA と比べて長いことがわかる．また表 2 に書籍チャンクの例を載せている．文中で法令の条文や判例を参照していることがわかる．

# 3 実験

作成した評価データセットに対して，複数のリランキングモデルを用い性能の検証と分析を行った．性能が低かった質問を法科大学院生に依頼してどのようなエラーがあるかを定性的に評価した．また，評価データの収集方法が結果に与える影響について追加検証した．

## 3.1 実験設定

性能検証の比較に用いるリランキングモデルには japanese-reranker-cross-encoder-large-v1[8)],japanese-bge-reranker-v2-m3-v1 [9)]，bge-reranker-v2-m3[10)]，Ruri-Reranker-Large[8]，JaColBERTv2.5[10] を使用する．入力データの前処理として，質問文と書籍チャンクの先頭に対応する書籍タイトルを付加した．さらに，入力長が 512 トークンを超過する場合，全体の入力長が 512 トークンに収まるように書籍チャンクの後半部を省略した．評価指標には情報検索やランキングシステムの評価に使用される ndcg@10 と ndcg@30 を用いる．

## 3.2 実験結果

表 3 に各リランキングモデルの評価結果を示す．japanese-bge-reranker-v2-m3-v1 が ndcg@10 で最も高いスコア，JaColBERTv2.5 が ndcg@30 で最も高いスコアを示した．一般ドメインの評価データセットの結果に反し、Ruri-Reranker-Large が最も低い結果となった．[8]．これは一般的な評価データセットで示された性能が、法律分野のタスクで同等程度の性能を達成できない可能性があることを示唆している．一般的な評価データセットでは法律分野の性能を正確に予測できないため、法律ドメイン特有の評価データセットの構築が重要であると考えられる．

表 4: 質問の長さと ndcg@10 の相関分析結果（すべての p 値が 0.05 未満）

| モデル | 相関係数 |
|---|---|
| japanese-reranker-cross-encoder-large-v1 | -0.38 |
| Ruri-Reranker-Large | -0.33 |
| japanese-bge-reranker-v2-m3-v1 | -0.30 |
| bge-reranker-v2-m3 | -0.28 |
| JaColBERTv2.5 | -0.51 |

表 5: 収集方法変更時のリランキングモデルの性能

| | ndcg@10 | ndcg@30 |
|---|---|---|
| japanese-reranker-cross-encoder-large-v1 | 0.480 | 0.637 |
| japanese-bge-reranker-v2-m3-v1 | 0.491 | 0.606 |
| bge-reranker-v2-m3 | 0.465 | 0.576 |
| Ruri-Reranker-Large | 0.437 | 0.566 |
| JaColBERTv2.5 | 0.520 | 0.653 |

表 4 に質問の長さと ndcg@10 の相関係数を示す．質問文と ndcg@10 の相関を分析した結果，すべてのモデルにおいて統計的に有意な負の相関が観察された. 特に JaColBERTv2.5 は-0.51 と相関係数が他のモデルと比較と比べて低く，質問の長さが性能に及ぼす影響が大きいことが考えられる．この結果は、質問の長さがモデルの性能に影響を与える可能性を示唆している．

## 3.3 エラー分析

ndcg@10 の値が低かった質問と書籍チャンクについて定性的な評価を行う．bge-reranker-v2-m3 と Ruri-Reranker-Large それぞれで ndcg@10 が低かった結果 5 件をランダムに抽出し，法科大学院生にエラー分析を依頼した．

表 6 にエラーの例を示す．これは「関連なし」のチャンクが，「関連あり」のチャンクより上位に挙げられていた例である．１行目は，質問が明確に海外子会社での不祥事に焦点を当てているのに対し，引用文が一般的な子会社での不祥事対応についての内容が書かれている．この乖離は，モデルが文脈や質問の特殊性（この場合，海外子会社という要素）を十分に考慮せず，キーワードの一致や一般的な関連性に基づいて評価を行っている可能性を示唆している．２行目に関しては質問の核心（異なる種類の

8） https://huggingface.co/hotchpotch/japanese-reranker-cross-encoder-large-v1
9） https://huggingface.co/hotchpotch/japanese-bge-reranker-v2-m3-v1
10） https://huggingface.co/BAAI/bge-reranker-v2-m3

表 6: エラーの例：「関連なし」のチャンクが，「関連あり」のチャンクより上位に挙げられていた例

| 質問 | 書籍チャンク | 関連度 |
|---|---|---|
| 海外子会社で不正行為・不祥事が発生した場合，どのように調査を実施すればよいでしょうか. | ... 子会社で不正が発覚した場合，親会社は有事対応として事実調査等に加え，III で述べた危機対応も併せて行う必要がある．この危機対応の場面では，中立性・客観性が強く求められる事実調査等とは異なり，事業内容等と関連した経営的な判断も踏まえた対応が求められる ... [11] | 関連なし |
| 株式会社と合同会社の間での合併，株式会社と合名会社の間での会社分割など，異なる種類の会社間で組織再編を行うことは可能でしょうか. | ... 株式会社が持分会社となること (会 744)，または，持分会社が株式会社となること (会 746) を組織変更という．合併・会社分割・株式移転・株式交換を総称する用語は会社法にはないが，本文のように組織再編と俗称される．組織変更または組織再編をするには株主総会の特別決議が必要である (会 309[2]12)．なお，持分会社が別の種類の持分会社になることもできるが，これは持分会社の定款変更と位置付けられている... [12] | 関連なし |

会社間の組織再編可能性）には直接言及していない文章だが，「組織再編」「持分会社」「株式会社」等のキーワードが一致しているため，上位にあげられてしまった可能性がある．特に法律テキストのような専門的な文脈では，単なるキーワード一致ではなく，法的な関係性や文脈の理解が重要となる．

### 3.4 追加検証

従来の収集方法では，bge-reranker-v2-m3 を用いて推薦された上位 10 件の書籍チャンクをアノテーション対象としたが，この手法にはモデルバイアスが含まれる可能性がある．モデルが得意とする特定のタイプの関連性や特徴を持つチャンクに偏る可能性があり，他のタイプの関連性を持つチャンクが除外される可能性がある．

そこで，新たに Ruri-Reranker-Large を用いて同様の手順で 40 件の質問とアノテーション済みの書籍チャンクペアを収集し，比較分析を行った．表 5 に収集方法を変更した場合のリランキングモデルの性能を示す．JaColBERTv2.5 の性能は表 3 の bge-reranker-v2-m3 での収集方法と比較して性能が変わらないことがわかる．しかし，Ruri-Reranker-Large の性能が ndcg@10 で 0.411 から 0.437 に改善した．また，japanese-bge-reranker-v2-m3-v1 と bge-reranker-v2-m3 の性能が相対的に低下した．これは，bge-reranker-v2-m3 での収集方法で bge-reranker-v2-m3 と bge-reranker-v2-m3-v1 のモデルに有利なバイアスがあった可能性を示唆している．二つの異なる方法で収集された評価データセットで結果が異なることから，評価データの収集方法が結果に影響を与えることを示唆している．

## 4 今後の課題と展望

本データセットの特徴である長文質問の構造的特徴や複雑さを詳細に分析することで、モデルの性能向上と実用的な法律文書検索システムの開発につながることが期待される．今回構築したデータセットの質問数は 100 件と他の評価データセットと比べて小さいため、より頑健な評価を行うにはデータ量の増加が不可欠である．データセット拡張の際には、複数のモデルを用いてアノテーションの対象とする書籍チャンクデータを収集することで、より公平で信頼性の高い評価データセットの作成が可能となる．

本データセットの特徴である長文の質問に対しても高い性能を発揮するリランキングモデルの開発が今後の課題として挙げられる．加えて、書籍チャンクに含まれる法令や判例の引用情報を効果的に活用することで、文脈をより深く理解したリランキングモデルの構築が期待される．具体的には、これらの情報を書籍チャンクと共に埋め込むアプローチが有効であると考えられる．

## 5 結論

本研究では，RAG システムの性能評価と改善を目的とした企業法務分野向け日本語書籍検索評価データセットを構築する．構築した評価データセットに対して，リランキングモデルの性能検証をし，専門分野に特化した評価データセットの重要性を明らかにした．さらに，評価データの収集方法の違いがモデル性能の評価結果に影響を与えることを示した．本研究の成果は，法律分野における RAG システムの性能評価手法の改善と，より信頼性の高い評価データセットの確立に貢献することが期待される．

# 参考文献


[1] Patrick S. H. 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 Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, **Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual**, 2020.

[2] Ehsan Kamalloo, Nouha Dziri, Charles L. A. Clarke, and Davood Rafiei. Evaluating open-domain question answering in the era of large language models. In Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki, editors, **Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023**, pp. 5591–5606. Association for Computational Linguistics, 2023.

[3] Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. Retrieval augmentation reduces hallucination in conversation. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, **Findings of the Association for Computational Linguistics: EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 16-20 November, 2021**, pp. 3784–3803. Association for Computational Linguistics, 2021.

[4] Xinyu Zhang, Nandan Thakur, Odunayo Ogundepo, Ehsan Kamalloo, David Alfonso-Hermelo, Xiaoguang Li, Qun Liu, Mehdi Rezagholizadeh, and Jimmy Lin. Making a MIRACL: multilingual information retrieval across a continuum of languages. **CoRR**, Vol. abs/2210.09984, , 2022.

[5] Xinyu Zhang, Xueguang Ma, Peng Shi, and Jimmy Lin. Mr. tydi: A multi-lingual benchmark for dense retrieval. **CoRR**, Vol. abs/2108.08787, , 2021.

[6] Luyu Gao, Xueguang Ma, Jimmy Lin, and Jamie Callan. Precise zero-shot dense retrieval without relevance labels. In Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki, editors, **Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023**, pp. 1762–1777. Association for Computational Linguistics, 2023.

[7] Joshua David Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. Contrastive learning with hard negative samples. In **9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021**. OpenReview.net, 2021.

[8] Hayato Tsukagoshi and Ryohei Sasano. Ruri: Japanese general text embeddings, 2024.

[9] 正敏鈴木, 潤鈴木, 耕史松田, 京介西田, 直也井之上. JAQKET: クイズを題材にした日本語 QA データセットの構築. 言語処理学会第 26 回年次大会, 2020.

[10] Benjamin Clavié. Jacolbertv2.5: Optimising multi-vector retrievers to create state-of-the-art japanese retrievers with constrained resources, 2024.

[11] レクシスネクシス・ジャパン. Business law journal 2020 年 1 月号. p. 27, 2020.

[12] 大垣尚司. 金融から学ぶ会社法入門. 勁草書房, 2017. p. 416.

[13] 青木荘太郎, 池田浩一郎, 鈴木貴泰. 会社法実務マニュアル 第 2 版 ―株式会社運営の実務と書式― 第 4 巻 組織再編・事業承継. 会社法実務研究会, 2017. p. 159.

[14] M. L. McHugh. Interrater reliability: the kappa statistic. **Biochemia medica**, Vol. 22, No. 3, pp. 276–282, 2012.

表 7: エラーの例：「関連あり」のチャンクが，「関連なし」のチャンクより下位に挙げられていた例

| 質問 | 書籍チャンク | 関連度 |
|---|---|---|
| このたび、当社（A 社）は B 社との間で、X 社を新設の完全親会社とする共同株式移転を実施します．当社は、当該株式移転の承認にかかる株主総会を開催する予定ですが、株主総会において株式移転対価の相当性に関する質問があった場合、どの程度説明すればよいのでしょうか． | ... 株式移転比率算定方法及び算定根拠 甲は D 監査法人、乙は E 監査法人に株式移転比率の算定を依頼しております．D 監査法人は、甲の評価を市場株価法、収益還元法、時価純資産法に基づき算定し、乙の評価を類似会社比準法、収益還元法、時価純資産法に基づき算定しております．E 監査法人は、甲の評価を市場株価法に基づき算定し、乙の評価を類似会社比準法、DCF 法に基づき算定しております．(2) 株式移転に際して交付する株式の数及びその割当 両社は、双方の株式移転比率算定書をもとに協議を行い ... [13] | 関連あり |

## A　アノテーションの一致度

アノテーションプロセスとして，法科大学院生には訓練用サンプルを 8 件着手した後，本番用サンプルのアノテーションを依頼するようにした．訓練用サンプルでの Cohen’s kappa 係数 [14] は 0.22，本番用サンプル 10 件での Cohen’s kappa 係数は 0.49 であった．この結果から，訓練プロセスを経ることでアノテーションの一致度が向上したことが示唆される．しかしながら，0.49 という値は「中程度の一致（Moderate agreement）」に分類され，さらなる改善の余地がある．アノテーションガイドラインの詳細化やアノテーター間の議論セッションの実施を行うことで，アノテーションの一致度をさらに向上させより信頼性の高いデータセットの構築が期待できる．

## B　他に見られたエラー

表 7 にその他のエラーの例を示す．表は「関連あり」のチャンクが，「関連なし」のチャンクより下位に挙げられていた例である．質問の核心は「株式移転対価の相当性」にあるが、引用文では代わりに「株式移転比率」という法律専門用語を使用して回答している．リランキングモデルは用語を同義として認識できておらず，質問と引用文の関連性が適切に評価されていない可能性がある．