大規模言語モデルを基盤とした法務契約文書リスク評価手法

A risk assessment method for legal contract documents based on large language models

楓川　滉人　Hiroto Kaedegawa
立川　雄一　Yuichi Tachikawa

キーワード
大規模言語モデル（LLM），自然言語処理（NLP），生成AI，検索強化生成（RAG），AIによる情報検索，AI良否判定

本稿は，大規模言語モデル（LLM）と検索強化生成（RAG）を基盤とした法務契約文書のリスク評価手法を提案する。従来の手法では，自社基準への適応や複雑な文脈を持つ長文条項への対応に課題があり，リスク評価の精度に限界があった。本手法は，RAGを活用した参考情報の統合に加え，偽陰性削減と動的ウィンドウサイズ調整を組み合わせることで，リスク評価と修正案生成の精度を大幅に向上させた。これにより，法務契約文書のみならず，他分野の専門的文書への応用可能性も期待できる。

This paper proposes a risk assessment method for legal contract documents based on large language models (LLMs) and retrieval-augmented generation (RAG). Traditional methods have struggled in adapting to company-specific standards and addressing lengthy clauses with complex contexts, which limited the accuracy of risk assessments. The proposed method significantly improves the accuracy of both risk assessment and the generation of revision proposals by integrating reference information using RAG and combining false negative reduction with dynamic window size adjustment. This approach is expected to be applicable not only to legal contract documents but also to specialized documents in other fields.

1. はじめに

法務部門では，コンプライアンス強化に伴い契約書のリスク評価やチェック業務が増加し，業務負担の軽減と効率化が急務となっている。

一方，人工知能（AI）技術の進展により，自然言語処理（NLP）分野では大規模言語モデル（LLM：Large Language Model）が注目を集めている。LLMは膨大なテキストデータを用いて訓練され，文章の生成，要約，質問応答など，人間の言語処理を模倣する高度な能力を持つ[(1)]。代表的なモデルにはGPTやBERTなどがあり[(2)]，それぞれ異なる特性を活かして様々なタスクに適用されている。また，生成AI（Generative AI）はデータを学習して新しいコンテンツを生成する技術である。その応用の一つである検索強化生成（RAG:Retrieval-Augmented Generation）は，外部データベースから関連情報を検索し，生成タスクに組み込むことで，従来のAIモデルよりも精度の高い回答生成を可能にする。

本稿では，LLMとRAGを組み合わせた法務契約文書リスク評価手法を提案する。RAGを活用した参考情報の統合に加え，偽陰性削減や動的なウィンドウサイズ調整を通じて，リスク評価と修正案生成の精度向上を目指す。

2. 既存AIリスク評価サービスの課題

既存AIリスク評価サービスを自社に導入する際には，以下の課題があると考えられる。

- 自社基準に対応できない：汎用的な評価が多く，自社基準や独自の運用要件に適合しない。
- 評価結果の説明が不十分：評価の理由が明確でなく，活用が難しい。

3. システムの仕組み

本システムは，法務契約文書を解析し，最新の法律および自社の法務実務経験・ガイドラインに基づき，契約文書全体の文脈を考慮した各条項のリスク評価（4段階）と修正提案を生成する（図1）。また，利用者からのフィードバックを反映し，継続的な精度向上と進化を実現する仕組みを提供する。

条項名	条項内容	リスク度	コメント	修正案
第２条（請負業務の内容及び範囲）	第２条（請負業務の内容及び範囲）本契約に基づく請負業務の内容及び	1	この条項は、請負業務の内容と範囲を明確に定義しており、特別なリスク要素は含まれていないと判断されます。したがって、リスクレベルは1と評価します。	修正不要
第３条（基本契約と個別契約）	第３条（基本契約と個別契約） 本契約は個別契約に関する事項について適用する。ただし、別途甲が定めて	2	この条項は、基本契約が個別契約に適用されることを定めていますが、甲が別途定めた事項については乙が了承する必要があるとしています。これは、甲が一方的に契約内容を変更する可能性を示しており、乙にとってはリスクがあると考えられます。また、乙が了承しなければならない事項の具体的な範囲や内容が明示されていないため、予見性が低くなっています。	「ただし、別途甲が定めて乙が了承した事項についてはこの限りでない。」という文言を削除し、個別契約に関する事項はすべて本契約に基づくことを明確にするか、乙が了承するべき事項の具体的な範囲や内容を明示することを提案します。
第４条（個別契約の成立）	第４条（個別契約の成立） 個別契約は、注文書その他注文内容を明示した文書又は電磁的方法による甲からの申込みに対し、次の各号のいずれ	3	この条項は、甲が乙に対して注文を行った場合、乙が7日以内に諾否の通知を発しなかった場合には、個別契約が成立すると規定してい	「甲が申込みを発した日から７日以内に乙が諾否の通知を発しなかったとき」の部分を

図1　評価結果例–条項ごとにリスク度評価と修正提案

本章では，図2に示すシステム構成を基に，各処理や技術の役割について説明する。

・3.1節：システム全体の構成と各要素の相互関係を示す。
・3.2節：RAGによる動的情報検索と統合，Few-shot学習を活用した評価手法を説明する。
・3.3節：各LLMの特性を活かし，情報検索と評価プロセスに適用する方法を説明する。
・3.4節：検索精度を高めるための参考情報の前処理手法を説明する。
・3.5節：ベクトル検索とプロンプト設計手法を説明する。

3.1　システム構成図

システム構成図を図2に示す。白色部分は一般タスク処理，灰色部分は参考情報関連処理，黒色部分はリスク評価処理である。

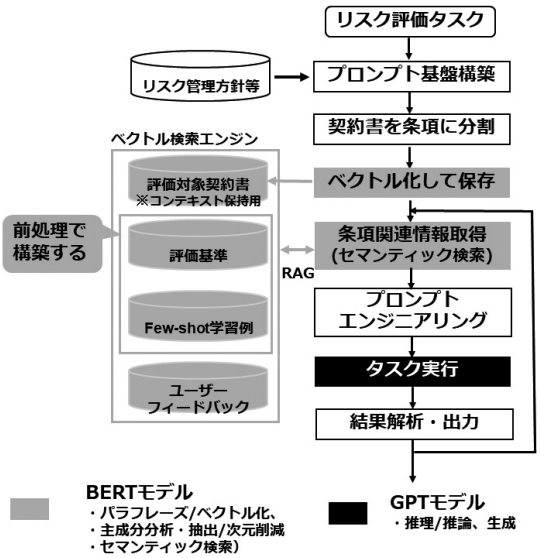

図2　システム構成図

3.2　RAGによる動的情報統合とFew-shot学習

自社基準や運用要件に特化したリスク評価を実現する一般的な手法として，モデルのFine-tuningがある。しかし，大量のデータや学習時間を要し，運用コストが高く，最新情報を迅速に反映するのが難しい課題がある。本システムでは，RAGを活用した動的情報統合とFew-shot学習による柔軟な適応を組み合わせ，これらの課題を解決している。

RAGは，LLMの生成能力と動的検索を統合した技術である[3]。これにより，外部データベースや社内ナレッジベースから最新の法令や基準，ユーザーフィードバックを検索して評価タスクに反映する。この仕組みにより，トレーニングで未学習の情報も動的に利用可能となり，常に最新の状況に基づいたリスク評価を提供できる。また，RAGで一般的に用いられるベクトル検索は，意味的類似性を精度高く捉える柔軟性を持ち，法務契約文書のような文脈解釈が重要なタスクで特に効果を発揮する。

Few-shot学習では，プロンプト（指示・質問・例示）に少量の自社事例や運用要件を入力するだけで，モデルの再学習を行わずに高精度なリスク評価やコメント生成が可能となる。本システムは，RAGによる動的情報統合とFew-shot学習による柔軟なカスタマイズを組み合わせ，迅速かつ精度の高いリスク評価を実現し，業務効率を向上させている。

3.3　言語モデルの選定

近年，LLMの開発が進み，高度な自然言語処理能力を持つモデルが数多く登場している。代表的なモデルには，OpenAI社のGPTモデルやGoogle社のBERTモデルがある。

表1　GPTモデルとBERTモデルの性能比較[2]

Method	CoLA	SST-2	MRPC		STS-B		QQP		MNLI		QNLI	RTE	GLUE
	Mcc.	*Acc.*	*Acc.*	*F1*	*Pear.*	*Spea.*	*Acc.*	*F1*	*m.*	*mm.*	*Acc.*	*Acc.*	avg.
BERT-base	56.4	88.0	90.0	89.8	83.0	81.9	80.0	80.0	82.7	82.7	84.0	70.0	79.2
BERT-large	62.4	96.0	92.0	91.7	88.3	86.8	88.0	88.5	82.7	88.0	90.0	82.0	85.4
RoBERTa-base	61.8	96.0	90.0	90.6	90.2	89.1	84.0	84.0	84.0	88.0	92.0	78.0	84.7
RoBERTa-large	65.3	96.0	92.0	92.0	92.9	91.1	90.0	89.4	88.0	90.7	94.0	84.0	87.8
ChatGPT	56.0	**92.0**	**66.0***	**72.1***	**80.9**	**72.4***	**78.0**	79.3	**89.3***	81.3	84.0	**88.0***	78.7

Table 2: Overall comparison between ChatGPT and fine-tuned BERT-style models on GLUE benchmark. The results in green denote that ChatGPT surpasses the BERT-base model by a clear margin (> 2% (↑) score), while the red results denote ChatGPT under-performs BERT-base (> 2% (↓) score)). More specifically, "*" means that the performance difference between ChatGPT and BERT-base is larger than 10%.

本システムでは，法務契約文書の複雑なリスク評価タスクに対応するため，BERTモデルとGPTモデルの特性を活かして組み合わせて活用している。各モデルの役割は以下の通りである。

・BERTモデル：表1に示すように，パラフレーズ（paraphrase）や類似性（similarity）の評価に優れるので，本システムではベクトル化やセマンティック検索で活用する。BERTモデルには法務分野に特化したLegal-BERTも存在するが[4]，主に英語で学習されており，日本語契約文書への適用には言語的な制約が生じる可能性がある。本システムでは，この制約を考慮し，通常のBERTモデルを採用している。
・GPTモデル：表1に示すように，推論（inference）や推理（reasoning）の能力に優れている。また，CoT（思考の連鎖）やFew-shot学習をサポートする特性を持つため，本システムではリスク評価，コメント，修正提案の生成を担当する。

このように，BERTモデルの検索能力とGPTモデルの生成能力を補完的に活用し，高精度かつ柔軟なリスク評価を実現している。

3.4　法務参考情報の前処理

3.4.1　参考情報の構造化

RAGの参考情報には，PDF，Excel，画像など多様な形式のデータが含まれる。これらのデータをそのままベクトル化したり，LLMに与えたりすると，文脈の理解や検索精度が低下する可能性がある。そのため，本システムでは，データをJSONやXML形式に変換して統一的に構造化している(図3)。

JSON形式を利用することで，階層的な構造を通じて情報間の関係を明確化でき，また，メタデータ（例:作成日時，出所など）を付加してLLMに与える場合，応答生成に信頼性や背景情報を加えることができ，回答のトレーサビリティを確保できる。また，この構造化により，システム処理が効率化され，モデルが情報を正確に理解しやすくなる。これにより，検索結果の精度が向上し，LLMを活用した高度なリスク評価や修正提案の生成が可能となる。

```
},
"所有権移転": {
    "移転時期": [
        {
            "情報種類": "参考事例",
            "契約内容": "代金完済よりも早いタイミング（例：検査合格時）で所有権が移転する。",
            "リスク度": 2,
            "コメント": "*****",
            "修正ポイント": "******",
            "関連情報": ["***", "***", "***"],
            "タグ": ["所有権", "移転", "時期", "*****"],
            "メタデータ": {
                "作成日時": "2021-06-01T00:00:00",
                "出所": "******"
            }
        }
    ],
    "所有権移転の条件": [
```

図3　データ構造化(JSON形式)例

3.4.2　参考情報のベクトル化・次元削減

(1) 参考情報のベクトル化

参考情報に付与されたタグ（例：["所有権", "移転", "時期", …]）をスペースで連結し，テキストシーケンスとしてBERTモデルの入力データとする。このシーケンスをBertJapaneseTokenizerでトークナイズし，生成されたトークン列をBERTモデルに入力してベクトル化を実施する。BERTモデルの隠れ層は768次元で構成されているため，各参考情報は768次元のベクトルで表現される。このベクトルは検索や類似度計算に使用される。

(2) 主成分分析(PCA)による次元削減

PCAは，情報の重要な要素を保持しつつベクトルの次元を削減する手法であり，計算効率の向上やメモリ使用量の削減が期待される。図4に示す本システムの参考情報では，タグのベクトルを32次元に削減した場合，累積分散比が95%に達することが確認された。これにより，検索精度をほぼ維持しながら処理効率を向上させる可能性が示唆される。一方，評価対象である条項部分は未知の内容を含むため，次元削減が累積分散比や検索精度に与える影響は評価対象ごとに異なる可能性がある。そのため，次元削減の適用には精度と効率のトレードオフを考慮し，適切な運用が求められる。

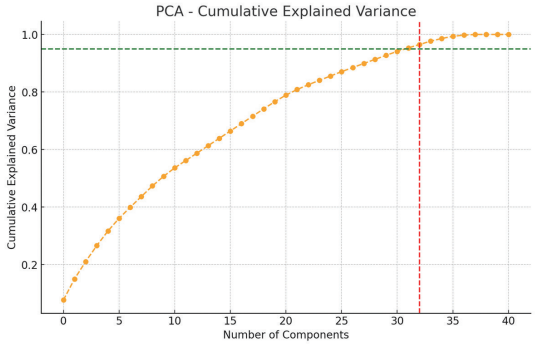

図4　PCAによる次元削減と累積分散比の関係

3.4.3　参考情報の前処理結果格納

前処理で生成された参考情報のベクトルや，構造化された参考情報内容をベクトル検索エンジン（例：Azure AI Search)に格納する。図5は，本システムにおけるAzure AI Searchインデックスの構造例を示しており，RAG参考情報の検索と効率的な取得を支える基盤である。

フィールド名	型	説明
id	String	参考情報id
refer_vector	SingleCollection	参考情報ベクトル
reduced_vector	SingleCollection	参考情報ベクトル（次元削減）
refer_info	String	参考情報内容（構造化済）
risk_level	Int32	リスク度（リスク事例：1〜4、その他参考情報：0）
metadata	String	メタデータ

図5　Azure AI Searchインデックス例

3.5　LLMによるリスク評価

3.5.1　コサイン類似度によるベクトル検索

法務契約条項をベクトル化し，それを使用して，参考情報のタグのベクトルをコサイン類似度で検索する(図5のインデックスを使用)。この検索プロセスでは，評価対象と参考情報の間の類似性を評価し，関連情報を抽出する。

コサイン類似度は，ベクトルの長さに影響されず，ベクトル間の方向性を基に高次元ベクトル間の意味的な類似性を効率的に評価できる手法である[5]。これにより，契約条項やタグのようにデータの長さが異なる場合でも，検索精度を維持しつつ契約文書の文脈的関連性を的確に評価可能である。

本システムでは類似度が0.8を超えるFew-shot学習例を含む関連情報を抽出し，リスク評価に活用する。

3.5.2　プロンプト設計による業務適応

本システムでは，systemロールとuserロールを活用して情報を役割ごとに整理している。図2に示すように，プロンプト基盤構築では，systemロールに自社のリスク管理方針などの全体ルールや指示を，プロンプトエンジニアリングでは，userロールにFew-shot用事例などの具体的な参考情報を設定する。これにより，プロンプトが自社の判断基準や運用要件へ柔軟に適応する設計が可能となる。

また，評価結果を構造化形式(例：JSON)で生成する指示

を加えることで，後続プロセスの効率化を図っている。

4. 偽陰性削減によるリスク検出強化

リスク評価においては，偽陰性（リスクの見逃し）の削減が重要な課題である。本章では，リスク検出度を高める手法を説明する。

4.1　リスク関連参考情報の拡充

過去の法務実務や類似契約事例を基に，リスク要素を軸として事例を体系的に分類し，参考情報のタグを具体化・階層化してベクトル検索の精度を向上させる（図2の前処理構築部分）。また，ユーザーフィードバックを組み込む動的更新の仕組みを導入し，参考情報を迅速に補強することで，検出網羅性を高める。

4.2　プロンプト設計によるトレードオフ

(1) リスク検出優先指示と根拠提示

図2のプロンプト基盤構築では，systemロールへリスク検出を最優先とする方針を明示し，誤検出を許容するよう指示する。また，検出結果には参考情報と類似度スコアを提示させることで，誤検出（偽陽性）時の人間対応の負担を軽減する。

(2) コサイン類似度の動的しきい値調整

参考情報（リスク事例，図5）のリスク度R_c(1〜4)に基づき，類似度しきい値を以下の式で動的に調整する。

$$T = T_b - \beta (R_c - 1) \qquad \text{式(1)}$$

ここで，T_bは基準しきい値（例：0.8），βは調整係数（例：0.05）である。事例のリスク度が高い場合にはしきい値を低く設定し，リスク事例を広く抽出することでリスク検出の網羅性を高める。

5. リスク重要度と文脈分散度に基づく動的ウィンドウサイズ調整

ウィンドウサイズはLLMが一度に処理するテキストの範囲（評価対象）を指す。本章では，動的にウィンドウサイズを調整する手法により，リスク評価の精度向上を目指す方法について説明する。

5.1　動的ウィンドウサイズ調整の必要性と仕組み

5.1.1　ウィンドウサイズと検索・評価精度の関係

主流のLLMはTransformer構造を基盤とし[7]，自己注意機構（Self-Attention）でテキストを処理するが，不適切なウィンドウサイズにより次のような問題が発生する。

(1) ウィンドウサイズが大きい場合

長文をベクトル化すると文脈が平均化され，特定の用語やリスク要素に関連する情報が埋もれる（情報の希薄化）。これにより検索精度が低下し，評価精度にも影響を与える。また，ウィンドウサイズの大きさによりモデルが重要情報に十分な注意を割けず，評価精度が低下する（注意分散）。さらに，トークン数が増加すると自己注意機構内の誤差が累積し，特に文脈が複雑な場合には評価精度がさらに低下する（累積誤差の増加）。

(2) ウィンドウサイズが小さい場合

文を細かく分割すると文脈が分断され，文章の前後関係が失われる（文脈損失）。これにより検索精度が低下するだけでなく，評価の一貫性も損なわれる。

5.1.2　法務契約文書の特性とウィンドウサイズ設定

法務契約文書は，以下の特性を持つため，一つの条項をそのまま処理する場合や固定的なウィンドウサイズで処理する場合には，検索および評価精度が低下する。

(1) 条項ごとのリスク重要度の多様性

損害賠償や情報提供など，条項ごとにリスクの重要度が異なる。リスク重要度の高い条項では詳細な分析が必要なため，小さいウィンドウサイズが適する。一方，リスク重要度の低い条項では，大きなウィンドウサイズを適用して効率を優先できる。この特性を考慮しない場合，リスク要素を見逃したり，過小評価したりする可能性がある。

(2) 長文条項

一つの条項をそのまま処理したり，大きいウィンドウサイズで処理したりすると，情報の希薄化，注意分散，累積誤差の増加により，検索および評価精度が低下する。

(3) 複雑な文脈構造

文脈が絡み合う条項では，小さい固定的なウィンドウサイズで処理すると文脈が分断される（文脈損失）。

5.1.3　本手法の仕組み

(1) リスク重要度に基づく調整

条項別のリスク重要度の違いに対応するため，以下のようにウィンドウサイズを調整する。

- リスク重要度が高い条項：小さいウィンドウサイズを適用し，詳細な分析を可能にすることで重要なリスクを見逃さない。
- リスク重要度が低い条項：大きなウィンドウサイズを適用し，処理効率を向上させる。

(2) 文脈分散度に基づく調整

文脈分散度は，文同士の関連性の強さを示す指標であり，分散度が高い場合は文脈が広範囲に分散していること，低い場合は文脈が一貫していることを意味する。

- 文脈分散度が高い場合：小さいウィンドウサイズを適用することで，分散した文脈を個別に扱い，関連性の低い内容の平均化（情報の希薄化）を防ぐ。また，局所的な文脈や条件間の依存関係を正確に捉えることが可能になる。
- 文脈分散度が低い場合：大きなウィンドウサイズを適用して処理効率を優先する。この場合，文脈の一貫性が強いため，広範囲な内容を同時に処理しても評価精度を損なうリスクが少ない。

5.2　ウィンドウサイズ動的調整プロセス

本手法は，評価対象のリスク要素（リスク特性に基づく重

要度）と文脈分散度に応じてウィンドウサイズを動的に調整する。そのプロセスは次のようになる（図6）。

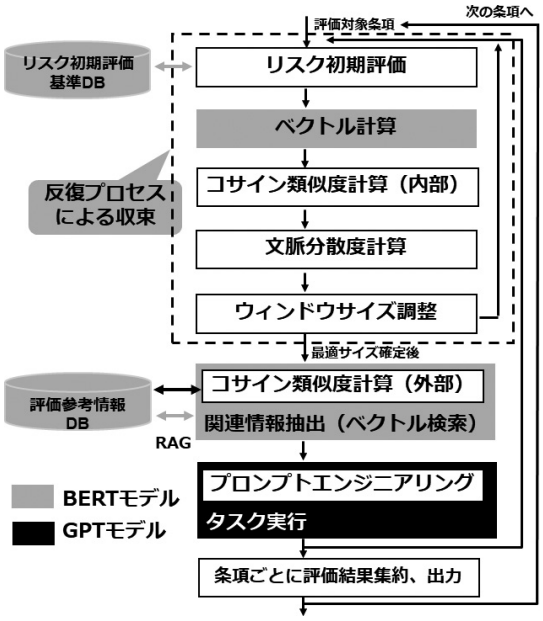

図6　動的ウィンドウサイズ調整手法

5.2.1　評価対象のリスク初期評価

（1）初期評価リスク度Rの設定

評価対象に含まれるリスク要素に応じて，リスク特性に基づく重要度を示すリスク度Rを設定する。重要項目（例：損害賠償，責任範囲）は高く，低リスク項目（例：情報提供）は低く設定する。

・リスク事前評価基準：条項全体リスク$R_{article}$と個別リスクR_{indi}を，リスク要素名とリスク度（1〜4）でリスク事前評価基準DBに登録する。

・リスク度計算：

a.総合リスクR'を計算：

$$R'=R_{article}\times\prod_{i=1}^{n}R_{indi} \qquad 式(2)$$

ただし，$R'>R_{max}$の場合，$R'=R_{max}$とする。

b.スケーリング：R'を0.5〜1の範囲にスケーリング：

$$R=R_{scaled}=0.5+0.5\times\left(\frac{R'-R_{min}}{R_{max}-R_{min}}\right) \qquad 式(3)$$

ここで，R_{min}=1とする。

（2）ウィンドウサイズの初期設定と初期評価

評価対象の条項に対し，トークン数を初期ウィンドウサイズ$W_{initial}$に設定する。以下の制約を設ける。

$$W_{min}\leq W_{initial}\leq W_{max} \qquad 式(4)$$

・最低トークン数：文書の適切なコンテキスト保持のためW_{min}=50を設定。

・上限トークン数：情報希薄化を防ぐためW_{max}=300を設定。

・条項のトークン数がW_{min}未満の場合，$W_{initial}=W_{min}$。W_{max}を超える場合は$W_{initial}=W_{max}$とする。

5.2.2　文脈分散度計算

本手法では，文脈分散度は評価対象各文同士のコサイン類似度を基に計算し，0〜1の値を取る。

a.類似度の計算：各文をベクトル化し，全ての文ペア間でコサイン類似度を計算する。

b.平均類似度の計算：

$$avg\_cos\_sim=\left(\frac{\sum_{i<j}cos\_sim\ (v_i,v_j)}{n(n-1)/2}\right) \qquad 式(5)$$

ここで，$cos_sim(v_i,v_j)$は文iと文jのコサイン類似度。

c.文脈分散度の定義：

$$cd\_score=1-avg\_cos\_sim \qquad 式(6)$$

5.2.3　動的ウィンドウサイズ調整

ウィンドウサイズは初期評価リスク度と文脈分散度に基づき，次の式で動的に調整される。

$$W_{new}=W_{current}\times(1+\alpha\cdot(1-R)\cdot(1-cd\_score)-\beta\cdot R\cdot cd\_score) \qquad 式(7)$$

W_{new}：調整後のウィンドウサイズ

$W_{current}$：現在のウィンドウサイズ

α：拡大係数（ウィンドウサイズ拡大時に使用する係数。例：0.3〜0.5）

β：縮小係数（ウィンドウサイズ縮小時に使用する係数。例：0.3〜0.5）

R：初期評価リスク度（0.5〜1）

cd_score：文脈分散度（0〜1）

この式は，Transformerの自己注意機構の原理を基に構築されている[7]。これにより，リスク重要度と文脈分散度に応じた動的なウィンドウサイズ調整が可能となる。リスク重要度や文脈分散度が高い場合には小さいウィンドウサイズを適用して詳細な分析を行い，低い場合にはウィンドウサイズを拡大して処理効率を向上させる。

5.2.4　反復プロセスの収束

（1）収束条件

反復プロセスは以下の条件を満たした場合に終了する。

$$|W'_{new}-W_{current}|<W_{current}\times\varepsilon \qquad 式(8)$$

ε：収束しきい値（2〜3%が妥当）

（2）文単位でのトークン数調整

最適トークン数W'_{new}前後で最も近い文の終わりを特定し，W_{new}を決定する。必要に応じて文脈変化位置の特定手法も併用可能である。

6. 評価・考察

6.1　評価結果

評価結果を表2～表4に示す。RAG構成には以下の3つのパターンがある。

- パターンA：偽陰性削減対策および動的ウィンドウサイズ調整を未実施
- パターンB: 偽陰性削減対策を実施，動的ウィンドウサイズ調整は未実施
- パターンC: 偽陰性削減対策と動的ウィンドウサイズ調整を両方実施

表2　リスク度に基づく各手法の精度比較

リスク度（4段階評価）	GPT	RAG構成（BERT+GPT）		
		パターンA	パターンB	パターンC
正解率（%） $R_s = R_e$	34	70	71	80
過大評価率（%） $R_s > R_e$	51	18	22	17
偽陽性率（%） $R_s > 1 \wedge R_e = 1$	48	18	21	16
過小評価率（%） $R_s < R_e$	15	12	7	3
偽陰性率（%） $R_s = 1 \wedge R_e > 1$	7	10	4	2

R_s：システムのリスク評価値
R_e：専門家のリスク評価値

表3　リスク検出有無（2値分類）に基づく各手法の精度比較

リスク検出有無※ Binary Classification	GPT	RAG構成（BERT+GPT）		
		パターンA	パターンB	パターンC
正解率（%） $(TP + TN) \div N$	44	72	75	82
適合率P（%） $TP \div (TP + FP)$	20	37	44	54
再現率R（%） $TP \div (TP + FN)$	59	50	82	91
F値（%） $(2 \times P \times R) \div (P + R)$	30	42	57	68

※リスク度=1：リスクなし　リスク度=2～4：リスクあり
TP：真陽性　TN：真陰性　FP：偽陽性　FN：偽陰性　N：全データ数

表4　修正案に基づく各手法の精度比較

修正案の正確性※	GPT	RAG構成（BERT+GPT）		
		パターンA	パターンB	パターンC
完全一致率（%）	9	36	45	73
部分一致率（%）	55	14	36	18
不一致率（%）	36	50	18	9

※リスクありの条項について、修正案を法務専門家が提示する修正案と比較し、一致率を計算

6.2　分析・考察

表2～表4の評価結果から，RAG構成（cl-tohoku/bert-base-japanese（Hugging Face Transformersライブラリv4.41.2）+GPT-4o）を用いることで，GPT-4o単独と比較して法務契約文書のリスク評価と修正案生成の精度が大幅に向上することが確認された。特に，パターンCではリスク度の正解率80%（表2），リスク検出有無の再現率91%（表3），修正案の完全一致率73%，部分一致率18%（表4）に達し，リスク検出と修正案生成の両面で最も優れたパフォーマンスを示した。一方，GPT-4o単独ではリスク度の正解率34%（表2），リスク検出有無の再現率59%（表3），修正案の完全一致率9%，部分一致率55%（表4）と低く，RAG構成の重要性が明確となった。適合率の一部低下は偽陰性削減を優先した結果であり[8]，業務要件に応じた調整が必要である。また，RAG構成に動的ウィンドウサイズ調整を組み合わせることで，さらなる精度向上が実現された。

7. 今後の展望

提案手法は法務契約文書のリスク評価と修正案生成に有効であることが確認されたものの，さらなる改善が必要である。本手法を補助的ツールとして活用しつつ，法務部署では判定基準や事例の精査・詳細化，リスク事例の追加が求められるほか，特定の契約について法務を介さずに契約チェックを行う運用方法の検討も重要である。一方，開発部署では，現場での適用性を評価し，法務以外の文書への応用可能性を探ることで，本手法の汎用性を高めることが期待される。

8. おわりに

本システムは，RAG構成を採用し，外部情報を動的に統合することで，大規模なトレーニングを必要とせず，最新情報を迅速かつ低コストで反映可能な法務契約文書リスク評価手法を実現した。さらに，Few-shotを活用し，偽陰性削減と動的ウィンドウサイズ調整を組み合わせることで，評価精度の向上を達成した。これらの成果は，参考情報の拡張や精査を通じ，他分野の専門文書への応用基盤となると考えられる。

＜参考文献＞

(1) Qiu, X., et al.: A Survey on Pre-trained Language Models, 2020, Retrieved from https://arxiv.org/abs/2003.08271.
(2) Zhong, Q., et al.: Can ChatGPT Understand Too? A Comparative Study on ChatGPT and Fine-tuned BERT, 2023, Retrieved from https://ar5iv.labs.arxiv.org/html/2302.10198.
(3) Lewis, P. et al.: Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks, 2020, Retrieved from https://arxiv.org/abs/2005.11401.
(4) Chalkidis, I., et al.: LEGAL-BERT: The Muppets straight out of Law School, arXiv preprint arXiv:2010.02559, 2020.
(5) Mikolov, T., Chen, K., Corrado, G., & Dean, J.: Efficient Estimation of Word Representations in Vector Space, 2013, Retrieved from https://arxiv.org/abs/1301.3781.
(6) Holtzman, A., et al.: The Curious Case of Neural Text Degeneration, 2020, Retrieved from https://arxiv.org/abs/1904.09751.
(7) Vaswani, A., et al.: Attention is All You Need, 2017, Retrieved from https://arxiv.org/abs/1706.03762.
(8) Manning, C., Raghavan, P., & Schütze, H.: Introduction to Information Retrieval, 2008, Cambridge University Press.

＜著者所属＞
楓川　滉人　アズビル株式会社
デジタルそうぞう部
立川　雄一　アズビル株式会社
法務・リスク管理本部