Lost in the Files：長コンテキスト LLM による複数専門文書からの網羅的情報抽出とその限界

佐藤正太 [1]　古川慧 [2]　生駒和也 [2]　園田亜斗夢 [1]
[1] 株式会社 Lightblue [2] 清水建設株式会社
{shota.sato,atom}@lightblue-tech.com
{k.furukawa,kazuya.ikoma}@shimz.co.jp

概要

RAG は外部知識の参照において有効であるが，文書群全体からの網羅的な情報抽出においては構造的な課題を抱えている．本研究では，RAG の代替アプローチとして，長コンテキスト対応マルチモーダル LLM への PDF 直接入力に着目し，その情報抽出の網羅性を体系的に評価した．複数モデルを用いた比較実験の結果，単一文書では高い抽出精度を示す一方，複数文書の同時処理においては情報の欠落や混同が生じやすくなること，および PDF の構造的理解においてモデルごとの特性が顕著に現れることが明らかになった．

1　はじめに

実務の意思決定やリスク管理では，複数の文書を横断して関連記述を網羅的に抽出する作業が不可欠である．監査・金融・法務など多様な領域で，情報の見落としが重大な損失やコンプライアンス違反を引き起こしうる．現状では専門家による精読に依存した属人的プロセスが中心であり，対象文書の増加に伴い作業負荷が増大するという構造的課題を抱えている．

近年，大規模言語モデルと外部文書検索を組み合わせた Retrieval-Augmented Generation（RAG）が広く利用されている [1]．RAG は事前学習に含まれない知識や最新情報へのアクセスを可能にし，回答の事実性と具体性を向上させることが知られている．しかし，本研究が対象とする網羅的情報抽出という観点では，RAG のアーキテクチャは構造的な限界を持つ．第一に，検索アルゴリズムは関連性の高い情報を上位に順位付けするが，関連する可能性のある記述を全て取得することを保証しない．第二に，文書を断片に分割して処理するため，文書全体の構造や

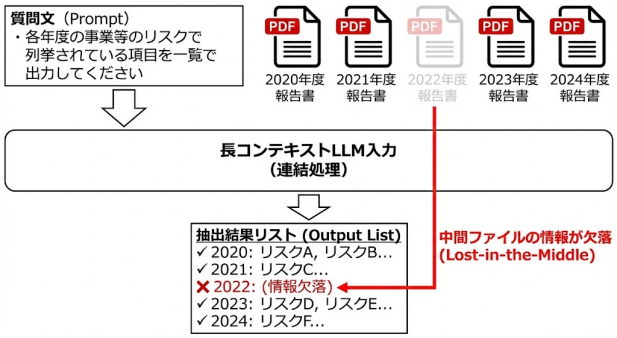

図 1　本実験の概要

複数記述にまたがる文脈情報が損失される．

こうした検索・合成に依存するアプローチとは対照的に，近年の大規模言語モデルにおける長コンテキスト化の進展 [2, 3] は，文書群を断片化せず一つの連続した情報単位としてモデルに直接入力する新たなアプローチを可能にする．このアプローチではモデルが文書全体の構造を直接把握し，記述間の関連性を自律的に探索・統合できる．これにより，検索プロセスに起因する情報の取りこぼしやチャンク化による文脈の断片化といった問題を回避可能であり，網羅的情報抽出において有望な選択肢となる．

本研究では，有価証券報告書を対象として，マルチモーダル対応大規模言語モデルへの PDF 直接入力による網羅的情報抽出能力を体系的に評価する．具体的には，総ページ数を固定してファイル数を変化させる条件と，単一ファイルでページ数を変化させる条件の二つを設定し，複数モデルを比較することで，長コンテキストの割り当て方が抽出性能に及ぼす影響を分析する．さらに，これらの知見に基づき，長大な PDF 文書群に対する網羅的情報抽出能力を評価するためのベンチマーク要件を提示する．

2　関連文献

RAG の長コンテキスト化　Jin ら [4] は，長文対

応 LLM と RAG の組み合わせが網羅的抽出において必ずしも有効でないことを示した．取得段落数を増やすと関連情報の拾い漏れは減少する一方で，誤情報（ハードネガティブ）の混入が増加し，LLM は位置バイアス（Lost-in-the-Middle）や文脈の希釈により，関連情報が入力に含まれていても正答に到達できない．この知見は，検索ベースのアプローチが抱える構造的課題を示しており，本研究が検討する直接入力方式の動機付けとなる．

長コンテキストにおけるドキュメント理解のベンチマーク　既存の長コンテキストベンチマーク [5, 6, 7, 8] は，主にテキスト形式での評価に焦点を当てている．文書理解に関するベンチマークにおいても，長コンテキストを扱うものは存在するが，画像 [9, 10] やマークダウン [11, 12] への事前変換を前提とするものが多い．生 PDF 形式を入力とする場合は比較的短いコンテキストに限定されている [13]．本研究では，長文脈の生 PDF を直接入力し，類似構造を持つ複数文書に対する網羅的情報抽出能力を評価する．この評価を通じて，実務的な文書処理タスクにおけるモデル性能を測定するためのベンチマーク要件を明らかにする．

3　実験

本研究では，マルチモーダル LLM の網羅的情報抽出能力を検証するため，有価証券報告書を対象とした比較実験を実施した．

3.1　比較モデル

実験実施時点で長コンテキスト入力および PDF ファイルの直接入力に対応している以下の 2 モデルを対象とした．

- **Gemini 2.5 Pro**: Google 開発のモデルで最大 100 万トークンに対応．PDF を視覚情報とテキスト情報の両面から処理し，1 ページあたり 258 トークンの固定コストで入力する．
- **GPT-5.1-2025-11-13**: OpenAI 開発のモデルで最大 40 万トークンに対応．PDF 入力時はテキスト抽出とページ画像化を併用する．

なお，生成時のパラメータについて，GPT-5.1 の仕様上 Temperature の設定が変更できないため，比較の公平性を期して両モデルにおいてデフォルト値を採用した．

表 1　実験条件の一覧．ファイル分割条件は複数ファイル入力の影響を，入力長変化条件はコンテキスト長の妥当性を検証する．

検証条件	ファイル数	総頁数	総文字数	正解要素数
ファイル分割				
(複数年度)	1	160	166,804	178
	2	160	172,805	365
	3	160	180,569	542
	4	160	190,314	703
	5	160	198,656	854
入力長変化				
(単一年度)	1	160	166,804	178
	1	80	84,023	178
	1	54	57,868	178
	1	40	45,644	178
	1	32	37,846	178

3.2　評価タスクと指標

評価タスクとして，複数年度の有価証券報告書を横断的に参照する必要がある質問を 15 件設定し，各モデルに回答を生成させた．プロンプトには抽出対象の観点を明示的に指定した．評価指標として，正解要素集合に対する再現率（Recall）を用いる．正解判定の基準は以下の 3 段階を設定した．要素抽出と年度対応については定量評価を行い，参照元（ページ番号）の正確性についてはモデル間で出力形式が異なるため定性的に分析する．(1) **要素存在基準**：回答に必要な要素がテキストとして含まれていれば正解とする．(2) **年度対応基準**：要素存在基準を満たし，かつその情報が対応する年度が正しく認識されていれば正解とする．(3) **参照元一致基準**：年度対応基準を満たし，かつ情報の根拠として出力されたページ番号が正しければ正解とする．評価は第一著者が実施し，あらかじめ定められた各基準を満たした要素数を全正解要素数で除した値を再現率として算出した．

3.3　データセットと実験設定

評価用データセットとして有価証券報告書を選定した．有価証券報告書は，定型的な様式で複数年度にわたり類似構造を持つため年度間の混同や欠落を検出しやすい．また，数百ページに及ぶ長文文書であることから，長コンテキスト LLM の処理能力を評価する上で適切な規模を有している．

本研究では，清水建設株式会社の有価証券報告書 2020〜2024 年度の 5 年分を使用した．この期間は，年度間で十分な差分が現れる一方で，文書構造の一貫性が保たれており，評価に適している．トークン

数とファイル構成の影響を分離するため，表 1 に示す 2 つの実験条件を設定した．各ファイルは，回答生成に必要なページを目視で抽出し，指定ページ数に達するまで前後の文脈ページを追加することで構成した．

ファイル分割条件: 総ページ数を 160 ページに固定し，入力ファイル数を 1〜5 に変化させることで，複数文書の同時処理能力を検証する．ファイル数の増加に伴い参照すべき年度および抽出対象の情報量が増加するため，固定された入力容量内での網羅的抽出能力を評価できる．

入力長変化条件: 単一ファイルのページ数を 32〜160 ページに段階的に変化させ，コンテキスト長が抽出性能に及ぼす影響を検証する．必須情報を含むページの前後に文脈ページを追加することで入力長を増大させ，長文脈下での情報検索精度の変化を評価する．

4 結果・考察

4.1 コンテキスト長と抽出精度の関係

表 2 に入力長変化条件における再現率を示す．両モデルともに，ページ数を 32 から 160（約 16 万文字）へ増加させても網羅性の低下は見られず，高い精度を維持した．この結果から，本実験規模においてはトークン数の増加自体が性能低下の直接要因にはならず，コンテキストサイズの拡大は有効に活用できることが確認された．

4.2 ファイル分割数と情報抽出精度の関係

一方で，表 3 に示すように，ファイル分割条件では異なる傾向が観察された．両モデルともファイル数の増加に伴い再現率が低下した．特に 5 ファイル条件では，要素存在基準と年度対応基準の乖離が拡大している．GPT-5.1 では両基準の差が 0.036，Gemini では 0.044 となり，正解要素は抽出できているが，抽出された情報の年度の誤りが増加していることが確認された．

誤りの詳細を分析すると，この乖離は単なる年度ラベルの付与ミスのみに起因するものではない．特定年度の情報が入力から抽出されずに欠落し，その欠落箇所を他年度の情報で誤って補完する事例が多数確認された．また，出力順序のずれにより年度の対応関係が破綻する事例も見られた．つまり，情報の欠落が一次的な要因となり，結果として出力全体の整合性が損なわれている．

表 2 入力長変化条件における再現率（要素存在基準）．

	32 頁	40 頁	54 頁	80 頁	160 頁
GPT-5.1	1.000	1.000	1.000	1.000	1.000
Gemini	0.994	0.994	0.994	0.994	0.994

表 3 ファイル分割条件における再現率（Recall）．上段：要素存在基準，下段：年度対応基準．

	1pdf	2pdf	3pdf	4pdf	5pdf
GPT-5.1（要素）	1.000	0.997	0.967	0.983	0.941
GPT-5.1（年度）	1.000	0.997	0.967	0.977	0.905
Gemini（要素）	0.994	0.992	0.982	0.881	0.857
Gemini（年度）	0.994	0.992	0.976	0.835	0.813

情報の欠落が発生する箇所には一定の傾向が見られた．両モデルにおいて，入力系列の中間付近に配置されたファイルの情報が選択的に欠落する傾向が確認された．この傾向は，長文脈 LLM において入力の中間位置にある情報の参照精度が低下する Lost-in-the-Middle 現象 [14] と一致する．本実験の結果は，単一文書内のコンテキスト位置だけでなく，複数ファイルを連結した場合においても，ファイル単位での位置バイアスが情報の検索・保持能力に影響することを示している．

ただし，この位置バイアスに対する耐性にはモデル間で差異が見られた．GPT-5.1 はファイル数が増加しても性能低下は緩やかであり，5 ファイル条件においても要素存在基準で 0.941 を維持した．中間位置での情報欠落や年度のずれは確認されたものの，その頻度は低く，文脈全体の整合性は保たれていた．対照的に Gemini は，3 ファイル条件までは高い再現率を示すが，4 ファイル以上で急激に性能が低下した．特に中間ファイルの欠落が顕著となり，情報の消失に伴って異なる年度の情報を混同して出力する事例が頻発した．この結果は，GPT-5.1 が長文脈および複数文書の処理において比較的堅牢なコンテキスト保持能力を有することを示している．一方，Gemini は入力分割数の増加に対して脆弱であり，特に中間情報の保持において課題を抱えている．

前節では，単一ファイルで 160 ページまで増加させても性能は維持された．しかし本条件では，同じ 160 ページであっても複数ファイルに分割することで性能が劣化した．この対比から，性能低下の主要因は単純なコンテキスト長の増加ではなく，複数ファイルの連結に伴う情報の分散と，抽出対象となる要素数の増加であると考えられる．特に，中間位置のファイルで情報欠落が顕著であったことから，

Lost-in-the-Middle 現象がファイル単位でも発生している．この知見は実運用上も重要である．入力を短縮する目的で文書を分割しても，複数文書を同時に入力し広範囲に分散した多数の要素を網羅的に抽出させる場合，同様の性能低下が生じうる．

4.3 参照情報の正確性と文書構造の認識

本研究では，回答の根拠となるページ番号の出力も求めたが，モデル間でページ番号の表記方法（印字番号と絶対ページ番号）が異なり，統一的な定量評価が困難であったため，定性的な分析を行った．

GPT-5.1 は，PDF のフッター等に印字されたページ番号を正確に認識し，一貫して出力した．一方で，参照先の特定においては，正解から前後 1 ページずれる誤りが頻出した．特に，ページ上部に記載された情報を直前のページの情報として扱う事例が多く確認された．この挙動は，GPT-5.1 が文書を物理的なページの集合としてではなく，ページ境界を跨ぐ連続したテキストデータとして処理していることを示唆する．そのため，文脈の連続性は保たれるものの，ページ境界の認識精度は後述の Gemini と比較して低い．

対照的に Gemini は，紙面の印字番号と PDF ファイル上の絶対ページ番号が混在して出力されるなど，表記形式の統一性に欠ける．しかし，参照箇所の特定精度という点では GPT-5.1 とは異なる特性を示した．具体的には，セクションがページをまたぐ箇所において，記述が次のページに移った時点で即座に参照番号が切り替わる挙動が確認された．これは，Gemini が各ページを独立した視覚情報として処理しているため，ページ境界を明確な区切りとして認識していることを示唆している．

これらの結果は，モデルの設計思想の違いが情報の網羅性だけでなく参照元の特定においても異なる特性として現れることを示している．GPT-5.1 はテキストの連続性を重視し，Gemini はページ単位の視覚的処理を重視しており，それぞれ異なる実務要件に適合する可能性がある．

5 おわりに

本研究では，長コンテキスト対応マルチモーダル LLM への PDF 直接入力による網羅的情報抽出能力を比較評価した．実験の結果，単一文書では両モデルとも高い再現率を維持したが，複数文書条件では挙動が異なり，設計上の優先順位の差が明らかになった．

5.1 設計思想の差異と実務上の含意

Gemini はページ単位の処理で効率性とスケーラビリティを重視するが，ファイル数が増えると情報の欠落と混同が増加し，再現率と整合性が損なわれる傾向が見られた．特に入力中間部での位置バイアスがリスクとなりうる．参照元の表記形式には課題が残るが，ページ境界の認識精度は比較的高かった．

GPT-5.1 は完全性を優先し，複数ファイルでも高い再現率と整合性を維持する．一方で，テキストの連続性を重視するため参照ページのずれが頻出した．また，トークン消費量の大きさが大量処理時の制約となる．

これらの知見から，長大 PDF 群の評価では，コンテキスト長だけでなく，複数文書処理時の欠落と混同，および参照元の認識精度を含めた多面的な評価が必要である．

5.2 網羅的抽出に特化したベンチマークへの提言

本研究の結果は，既存の評価が単一文書に偏り，実務で重要な複数文書横断の網羅的抽出を十分に測定できていないことを示している．今後構築すべきベンチマークは，少なくとも以下の要件を満たす必要がある．

1. 類似構造を持つ複数文書の同時入力を前提とし，記述が類似する条件下での情報欠落と混同を評価する
2. 単一解の検索ではなく全件抽出を対象とし，情報の有無や対応関係の正確さを段階的に測定する
3. 抽出内容の出典（どの文書のどの箇所に基づくか）を評価対象に含め，参照元の対応付け精度を検証する
4. 文書の並び順を変えた条件を含め，入力位置に依存した性能劣化（位置バイアス）への耐性を定量化する

これらの要件を満たす評価枠組みの構築により，長コンテキスト LLM が実務の精読作業をどの程度代替できるかを多角的に評価できる．本研究で明らかになった複数文書処理時の課題を踏まえ，今後はこのような評価基盤の整備が重要である．

参考文献

[1] Xinyu Gao Kangxiang Jia Jinliu Pan Yuxi Bi Yi Dai Jiawei Sun Meng Wang Haofen Wang Yunfan Gao, Yun Xiong. Retrieval-augmented generation for large language models: A survey. **arXiv preprint arXiv:2312.10997**, 2023.

[2] Mingyang Zhang Qiaozhu Mei Michael Bendersky Zhuowan Li, Cheng Li. Retrieval augmented generation or long-context llms? a comprehensive study and hybrid approach. **arXiv preprint arXiv:2407.16833v2**, 2024.

[3] Zhuyun Dai Dheeru Dua Devendra Singh Sachan Michael Boratko Yi Luan Sébastien M. R. Arnold Vincent Perot Siddharth Dalmia Hexiang Hu Xudong Lin Panupong Pasupat Aida Amini Jeremy R. Cole Sebastian Riedel Iftekhar Naim Ming-Wei Chang Kelvin Guu Jinhyuk Lee, Anthony Chen. Can long-context language models subsume retrieval, rag, sql, and more? **arXiv preprint arXiv:2406.13121v1**, 2024.

[4] Jiawei Han Sercan Ö. Arık Bowen Jin, Jinsung Yoon. Long-context llms meet rag: Overcoming challenges for long inputs in rag. **arXiv preprint arXiv:2410.05983v1**, 2024.

[5] Jiajie Zhang Hongchang Lyu Jiankai Tang Zhidian Huang Zhengxiao Du Xiao Liu Aohan Zeng Lei Hou Yuxiao Dong Jie Tang Juanzi Li Yushi Bai, Xin Lv. Longbench: A bilingual, multitask benchmark for long context understanding. **Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 3119–3137, 2024.

[6] Ming Zhong Xingjian Zhao Mukai Li Jun Zhang Lingpeng Kong Xipeng Qiu Chenxin An, Shansan Gong. L-eval: Instituting standardized evaluation for long context language models. **Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 14388–14411, 2024.

[7] Shengding Hu Zihang Xu Junhao Chen Moo Khai Hao Xu Han Zhen Leng Thai Shuo Wang Zhiyuan Liu Maosong Sun Xinrong Zhang, Yingfa Chen. ∞ bench: Extending long context evaluation beyond 100k tokens. **arXiv preprint arXiv:2410.05983v1**, pp. 15262–15277, 2024.

[8] Samuel Kriman Shantanu Acharya Dima Rekesh Fei Jia Yang Zhang Boris Ginsburg Cheng-Ping Hsieh, Simeng Sun. Ruler: What's the real context size of your long-context language models? **arXiv preprint arXiv:2404.06654v3**, 2024.

[9] C.V. Jawahar Minesh Mathew, Dimosthenis Karatzas. Docvqa: A dataset for vqa on document images. **Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)**, pp. 2200–2209, 2021.

[10] Liangyu Chen Meiqi Chen Yizhu Jiao Xinze Li Xinyuan Lu Ziyu Liu Yan Ma Xiaoyi Dong Pan Zhang Liangming Pan Yu-Gang Jiang Jiaqi Wang Yixin Cao Aixin Sun Yubo Ma, Yuhang Zang. Mmlongbench-doc: Benchmarking long-context document understanding with visualizations. **arXiv preprint arXiv:2407.01523v3**, 2024.

[11] Łukasz Borchmann Michał Pietruszka Pawel Joziak Rafal Powalski Dawid Jurkiewicz Mickael Coustaty Bertrand Anckaert Ernest Valveny Matthew Blaschko Sien Moens-Tomasz Stanislawek Jordy Van Landeghem, Rubèn Tito. Document understanding dataset and evaluation (dude). **Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)**, pp. 19528–19540, 2023.

[12] Viet Dac Lai Michael Krumdick Charles Lovering Chris Tanner Varshini Reddy, Rik Koncel-Kedziorski. Docfinqa: A long-context financial reasoning dataset. **arXiv preprint arXiv:2401.06915v3**, 2025.

[13] Hongming Zhang Kaixin Ma Deng Cai Zhuosheng Zhang Hai Zhao Dong Yu Anni Zou, Wenhao Yu. Docbench: A benchmark for evaluating llm-based document reading systems. **Proceedings of the 4th International Workshop on Knowledge-Augmented Methods for Natural Language Processing**, pp. 359–373, 2025.

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

A　実験設定の詳細

本実験において使用したシステムプロンプトおよび評価用の問題文（プロンプト）を以下に示す.

A.1　システムプロンプト

ユーザーから質問文と PDF として参考資料が入力されます．参考資料は各年度の有価証券報告書であり，例えば"2025_03.pdf"というファイルは 2024 年度の有価証券報告書です.
入力される全ての PDF の情報をもとに質問の内容に答えて下さい．ただし，全ての PDF から出来るだけ網羅的に情報を検索し，必ず情報の抜けがないようにして下さい．回答においては，提供する情報ごとに，その情報が記載されている PDF のファイル名とページ番号を直後に"2024 年度有価証券報告書 p.30"のように追記して下さい.

A.2　評価用問題文一覧

実験に使用した 15 件の質問プロンプトを以下に列挙する.

- 資料の中に記載されている，連結子会社または持分法適用関連会社の中で住所を東京都内に置く企業を各年度ごとにすべて挙げてください.
- 連結子会社の中で，管理職に占める女性労働者の割合が 5%を超えている企業をすべて各年度ごとに数値とともに情報をまとめてください.
- 各年度の事業等のリスクで列挙されている項目を一覧で出力してください.
- 「事業等のリスク」として列挙されている項目のうち，担い手不足リスク（中長期的な担い手不足リスク），建設資材価格及び労務単価の変動リスク，サイバーリスク，気候変動リスク（長期的な気候変動リスク）について，各年度の主なリスクの概要と主な対応策・取組みの内容を事細かに記載してください．事業等のリスクの表以外に記載の情報は入れないでください.
- 2020 年度から 2024 年度の自己資本比率の推移をまとめ，負債と純資産と総資産の観点からそれらの増減金額と要因を資料に記載されている範囲で明示した上で説明してください.
- 研究開発活動について，各年度における研究開発費の総額，および各年度の報告書に記載されている主な成果をすべてリストアップしてください.
- 各年度の研究開発活動について，「AI」というキーワードがタイトルまたは説明に含まれる成果をすべて列挙してください.
- 各年度の研究開発活動における「カーボンニュートラル・脱炭素」に関連する成果として記載されているものをすべて列挙してください.
- 各年度の研究開発活動の成果のうち，「Shimizu」「Shimz」「SC」「シミズ」のいずれかがプロダクト・システム・手法等の名前に含まれるものに関する成果として記載されている内容をすべて列挙してください.
- 各年度の研究開発活動の成果のうち，その開発にいずれかの大学が関わった成果として記載されているものをすべて列挙してください.
- 各年度ごとの当社グループの設備投資総額と，その主な目的（例：生産能力増強，研究開発拠点強化など）として挙げられている内容をすべて抜き出してください．実際の値だけで予定等の情報は不要です.
- 各年度ごとに，当社グループ・提出会社における主要な設備の帳簿価格の合計が大きい順に 3 つの事業所名とその合計値を教えてください.
- 各年度の執行役員の氏名をすべてリストアップしてください．2024 年度の執行役員から出力してください.
- 各年度末において，期末時点で残高のある社債のうち，償還期限が期末時点から 2 年以内または利率が 0.5%以上の銘柄を全て教えてください.
- 各年度末において，当連結会計年度末時点での負債の総額，および各社債の当期末残高が全負債のうち占める割合をリストアップして教えてください.