不動産情報抽出業務の効率化に向けた大規模言語モデルを用いたアンサンブル手法

齊藤 佑太郎 [1]　叶内 晨 [1,2]　松本 健太郎 [1]　岩成 達哉 [1]
[1] 株式会社 estie　[2]NLPeanuts Inc.
{yutaro.saito, kentaro.matsumoto, tatsuya.iwanari}@estie.co.jp
shin.kanouchi@nlpeanuts.com

概要

本研究では，不動産業界におけるプレスリリースからの情報収集業務を効率化するため，大規模言語モデル（LLM）の出力を統合したアンサンブル手法を提案する．LLM による情報抽出は精度が向上しているものの，複雑なタスクでは抽出漏れや誤抽出が依然として課題であり，結果の信頼性を担保するためには人手による確認が必要とされている．提案手法では，複数の LLM の出力をルールベースで統合し，モデル間の出力の一致を活用して人手チェックの優先順位を付けることで，作業を効率化する．実験の結果，提案手法は LLM を単体で利用した場合と比較して抽出精度が向上し，確認作業の負担を大幅に削減できることが示された．

1　はじめに

不動産業界では，物件情報や取引データ，プレスリリースなどが PDF 形式で流通しており，物件の売買や投資判断に活用されている．その中でも，J-REIT（不動産投資法人）に関連する取引や鑑定情報を含むプレスリリース資料は，数少ない公開取引事例として注目を集めている．これらの資料を効率的に収集・構造化・管理することは，不動産業務の効率化の観点から大きなニーズとなっている．

しかし，J-REIT に関連するプレスリリース資料は内容が多岐にわたり，資料ごとに記載項目やレイアウトが異なる．また，物件数や記載内容の詳細度が資料によって異なるため，事前に想定した固定的な処理フローでは対応が困難である．そのため，従来の OCR 技術を活用した情報抽出手法では，書式の多様性に対応しきれず，精度に限界が生じている．

近年注目されている大規模言語モデル（LLM）は，非構造化ドキュメントからの情報抽出精度を大幅に向上させている．しかし，LLM を用いた抽出にはいくつかの問題が残っており，抽出漏れや誤抽出，さらにはハルシネーションと呼ばれる不正確な生成などがある．これらの誤りを完全に排除することは難しく，現場では抽出された情報を結局人手でチェックする必要があり，作業負担が未だ大きいという課題がある．

これらの課題を解決するため，本研究では PDF データからの情報抽出精度の向上と人手チェック作業の効率化を目的として，複数の LLM の出力を突合させるアンサンブル手法を提案する．本手法では，複数の LLM の出力結果を比較し，モデル間で一致した項目を高信頼度の情報として自動的に採用する．一方，不一致が見られる箇所については優先順位を付けて重点的に確認する仕組みを構築した．これにより，抽出精度の向上とともに，人的確認作業の大幅な削減を実現することを目指している．

実験では，不動産業界で実際に使用されているプレスリリース資料を対象に抽出を検証し，評価を行った．その結果，単体の LLM を用いた場合と比較し，提案するアンサンブル手法は情報抽出の精度が向上した．さらに，モデル間の一致・不一致情報を利用することで，人手によるチェック箇所を大幅に削減できることが示された．

2　先行研究

不動産業界では物件情報や取引データ，プレスリリースといった資料のデジタル化が進む中 [1]，これらの情報を管理，収集，構造化する研究が進められている [2]．

大規模言語モデル（LLM）は自然言語処理の分野で顕著な進展を遂げており [3, 4]，非構造化テキストの情報抽出や構造化タスクにおいて高い性能を発揮している [5]．さらに，近年では画像や PDF 形式

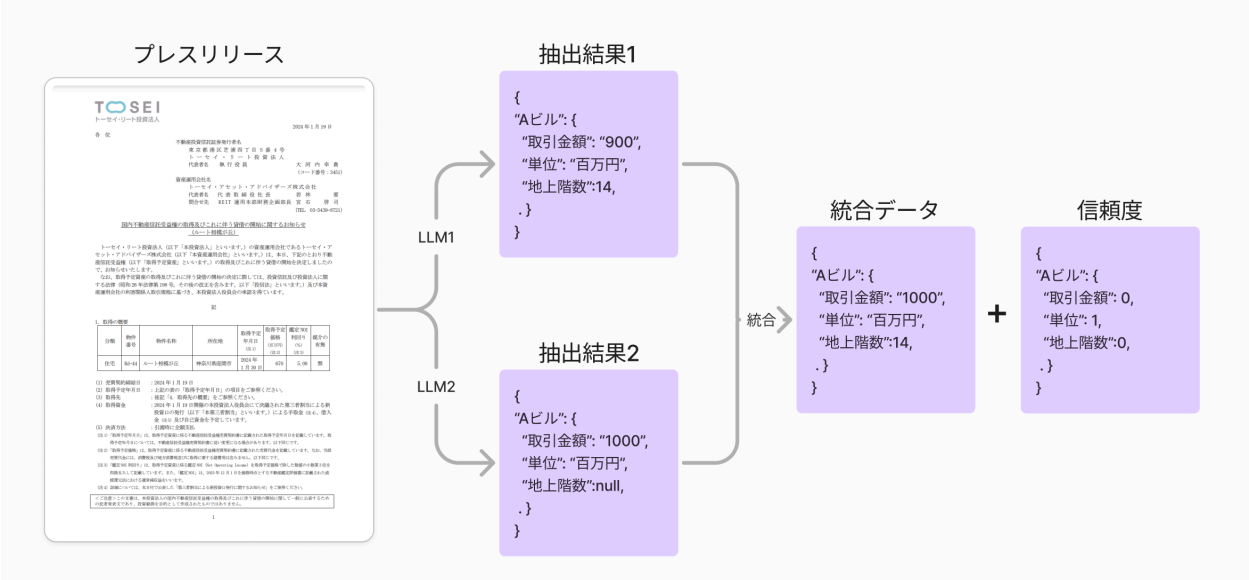

図 1　提案手法の流れ

のデータを直接入力として受け付けるモデルも登場しており [6]，従来のテキスト処理に加え，より柔軟なデータ入力が可能となっている．これにより，情報抽出の適用範囲が大きく広がっている．

一方で，既存の LLM には抽出漏れや誤抽出，ハルシネーションが発生する問題がある [7, 8]．その対策として，複数 LLM の結果を統合して性能を向上させるアンサンブル手法が提案されており [9, 10]，異なるモデルの出力を組み合わせることで，単一モデルでは達成できない高精度な推論や情報抽出が可能となっている．いくつかの実応用タスクにおいても研究が進んでおり，医療分野ではアンサンブル手法により診断精度が向上する報告や [11, 12]，プログラミングのコード生成を複数モデルでアンサンブルさせることで精度が向上する報告がある [13]．不動産情報領域の情報抽出においても，複数 LLM を用いることにより精度改善が期待される．

3　タスク概要

本研究では，J-REIT 関連のプレスリリース（PDF）から物件情報を抽出するタスクに取り組む．プレスリリースに記載される情報は様々だが，今回のタスクでは鑑定評価額，還元利回りなど，投資判断に必要と考えられる 34 種類の項目の抽出を目指す．抽出したいデータには以下の特徴がある．

- **多様な書式・レイアウト**　同一の発行主体による資料であっても，公表時期や作成者により見出し，文字装飾，段組みなどが異なり，統一的な処理が困難．
- **専門用語・略語の混在**　不動産関連の専門用語や鑑定手法の略語が多用されており，抽出タスクの難易度を高める要因となっている．
- **数値情報の表記ゆれ**　物件取得価格，想定賃料，稼働率などの定量情報が複数掲載されている一方で，「百万円」や「百万円（税抜）」といった表記の揺れがあり，文脈を理解した処理が求められる．
- **各項目情報の有無が不明**　収集したい情報が PDF 内に記載されていない場合があり，項目の記載有無も同時に判定する必要がある．

また，今回の抽出タスクの特徴として，与えられた PDF データが何件の物件を含むのかが自明でない点がある．例えば PDF に物件が 4 件含まれる場合，34 項目 x 4 物件 で 136 項目の抽出が必要となる．対象とするプレスリリースは 1 物件あたり 1〜4 ページ程度で，PDF 全体では 3〜40 ページ程度となり，物件数や記載される投資指標に応じてページ数が大きく変動する．

4　提案手法

本研究では，複数 LLM の出力を統合するアンサンブル手法を提案し，情報抽出における精度向上を目指す．図 1 に提案手法の流れを示す．まずそれぞれの LLM を用いて情報抽出を行い，その後モデルの抽出結果を比較・統合することで，単一モデルでは対応が難しい誤抽出や抽出漏れを解決する．

4.1 複数 LLM による情報抽出

各 LLM に対して，同一のプロンプトと JSON スキーマを与えて情報抽出を行い，それぞれのモデルから全項目を含む JSON 形式の出力を得る．

LLM に入力するプロンプトには「与えられた不動産プレスリリースのテキストから，以下の JSON スキーマに従って各項目を抜き出してください．値が存在しない場合は null としてください」という指示をした．利用する JSON スキーマは抽出対象となる 34 項目を JSON 形式で整理したもので {"物件名称": {"type": "string"},"鑑定評価額": {"type": "number"},...} という構造になっている．GPT-4o は PDF から直接テキストを抽出して入力データとし，Gemini は PDF を直接モデルに入力した．

本研究では，GPT-4o[14] と Gemini-2.0-flash-exp[15] の 2 つの LLM を使用する．また，今回対象とする PDF データにはテキストが埋め込まれているため，GPT-4o へは直接テキスト入力データとし，Gemini へは PDF を直接モデルに入力した．

4.2 抽出結果の統合

抽出結果に各項目に対して，以下のようにルールベースで出力を統合する．

A.一致項目の自動採用： 両モデルが同一の値を返した項目は，信頼できる回答としてその値を採用する．

B.不一致項目の処理： 一方が null で他方が有効値の場合は有効値を採用し，両方が異なる有効値の場合はより信頼性が高いと判断される GPT-4o の結果を優先する．

さらに，各項目に対して一致・不一致をそれぞれ「1」「0」で示す信頼度フラグを出力する．

信頼度フラグを参照し，両モデルの出力が一致している場合に人間がレビューを省略することができれば，業務の大幅な効率化につながる．また，このルールベースのフローは両モデルの出力を単純に比較するだけでよいため，実装が簡易で運用フローの理解もしやすいという利点を持つ．

表 1 AI モデルの精度比較

モデル	正解数	総数	正解率 (%)
Gemini	966	986	97.97
GPT-4	979	986	99.29
提案手法	979	986	99.29

表 2 モデル間の予測一致・不一致数

モデル間の予測	正解	不正解
一致 (n=960)	960	0
不一致 (n=26)	19	7

5 実験

5.1 実験データおよび設定

本研究では，3 節で述べた多様なフォーマットを持つ不動産プレスリリースを対象に情報抽出を行う．実験対象として，オフィスだけでなく住宅や商業施設など幅広い物件タイプを扱う総合型の投資法人であるサンケイリアルエステート投資法人 [16] が公表した不動産取引プレスリリースを用いた．対象のプレスリリースは 9 件の PDF データで，合計 29 物件に関する取引情報や鑑定評価情報が含まれている．

正解データとして，本研究で抽出対象とした物件名称・所在地・鑑定評価額・稼働率など計 34 項目に対して，事前に人手による正解ラベルの付与を行い，実験に用いた．評価時は正解ラベルとモデル出力が完全一致した場合のみを正解とし，29 物件 × 34 項目の全 986 項目に関する正解率をチェックした．

5.2 実験結果

表 1 に，GPT-4o，Gemini，および両者をルールベースでアンサンブルする提案手法の抽出精度を示す．全ての手法において正解率は 97%以上であり，非常に高い結果となった．

モデル間の出力項目の一致・不一致について

表 2 は，2 つのモデルの予測が一致したか否かによって回答を分類し，実際の正解率を分析した結果である．

両モデルの回答が一致した 960 件のうち，正解は 960 件，不正解は 0 件であった．また，不一致が発生した 26 件中，19 件は null でない結果を取り GPT-4o を優先する提案手法では正解であり，7 件は誤りであった．

key	Gemini	GPT-4o	正解
DCF 法による価格	NaN	2780	2780
PM フィー	1	3	3
その他費用	3	0	NaN
収益価格	NaN	2830	2830
直接還元法による価格	NaN	2870	2870

表 3　宮崎台ガーデンオフィスの不一致例

両モデルの抽出結果が不一致だった場合の抽出例を表 3 に示す．不一致は主に以下の 3 パターンで起きていることがわかった．

- **どちらかのモデルが抽出漏れをしている場合**
  該当の key 情報を抽出できていないケースであり，多くの不一致がこのパターンであった．なお，提案手法では一方が null で他方が有効値の場合は有効値を採用している．そのため，提案手法ではこのケースによる抽出漏れは起きていない．
- **値を取り違えている場合**　他の key の値を誤って参照している事例を確認した．例えば Gemini は「その他費用」をは 3 と答えているが，実際には NaN が正解であり，誤抽出となっている．本研究で扱うデータには似た数値情報が多く記述されているため，数値の取り間違えが起きることを確認した．
- **0 と NaN を取り違えている場合**　対応する値が対象データに存在しない場合にモデルは NaN を出力しなければならないが，Nan の代わりに誤って 0 を出力するケースを確認した．例えば「その他費用」の GPT-4o の結果は 0 と出力されているが，正解は NaN である．0 と NaN は大きな違いであり，実運用におけるタスクではこのエラーが後続処理に伝播するため，改善することが求められる．

5.3　人手チェック工数の削減

本研究の実験では，両モデルの結果が一致した箇所については，実験上はすべて正解となった．今回使用していない他のデータセットにおいても常に正解を保証できるわけではないが，人手によるチェックでも一定のミスが生じる可能性を考慮すれば，モデル出力の一致箇所に関してレビューを省略しても一定の品質を維持できると考えられる．

実務においては「すべての項目を確認しないと安心できない」という声もあるが，現場での大量処理を想定すると，不一致箇所だけを優先的にチェックすることの恩恵は大きい．実際に，今回の実験ではチェック対象を「不一致となった 26 件のみ」に絞り込むことで，1 件あたりのレビュー時間を従来比で約 1/40 に削減できた．特に物件数や抽出項目数が増加すればするほど，この効果はより顕在化すると考えられる．また，人間のレビュー担当者からは「モデルが同じ回答を出しているのを確認すると，安心感がある」という定性的なフィードバックも得られた．

6　考察と今後の展望

本研究では，複数の LLM の出力を統合するアンサンブル手法を提案し，不動産関連の情報抽出タスクに適用した．実験結果から，単体で高精度を発揮する LLM であっても，複数モデルの出力を照合することでさらなるメリットが得られることが示唆された．特に，下記の点が顕著な利点として確認された．

1. **不一致部分の可視化による誤り検知** モデルごとの誤りパターンが異なる場合，予測一致・不一致をフラグとして誤りを発見しやすくなる．
2. **高精度で安心感のある結果** モデル間の合意が得られた項目は実際に正解率が極めて高く，人間のレビュー工数を削減しつつ，ビジネス上のリスクも低減できる．
3. **モジュール化された拡張性** 今回は 2 種類の LLM のアンサンブルを想定したが，3 種類以上を活用した多数決や重み付け投票制，さらにはモデルの自己評価プロンプトを組み合わせることで，今後さらなる精度向上や汎用性の拡張が期待できる．

また，実務においては追加の照会プロセスや人間の判断を組み合わせることで，最小限のステップで高精度を実現するワークフロー設計が求められる．今回提案したアンサンブル手法は，不動産業界に限らず，幅広いドメインにおける非構造化ドキュメント処理の課題に適用可能な汎用的アプローチとしての可能性を示している．今後は対象文書の多様化やさらなるモデルバリエーションの適用を進めるとともに，不一致が生じる原因の定量分析や，自己評価プロンプトを用いた自動的な再照会フローの開発などを行い，実運用レベルでの安定性や拡張性をさらに高めていく予定である．

参考文献

[1] Daniel Piazolo and Utku Cem Dogan. Impacts of digitization on real estate sector jobs. **Journal of Property Investment & Finance**, Vol. 39, No. 2, pp. 47–83, 2021.

[2] 齊藤 佑太郎本郷 槙一, 岩成達哉. 大規模言語モデルを用いたマイソク pdf からの情報抽出. 言語処理学会第 30 回年次大会 (NLP2024), 3 2024.

[3] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. GPT-4 technical report. **arXiv:2303.08774**, 2023.

[4] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. **arXiv preprint arXiv:2302.13971**, 2023.

[5] Alexander Dunn, John Dagdelen, Nicholas Walker, Sanghoon Lee, Andrew S. Rosen, Gerbrand Ceder, Kristin Persson, and Anubhav Jain. Structured information extraction from complex scientific text with fine-tuned large language models. **arXiv:2212.05238**, 2022.

[6] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. **arXiv:2312.11805**, 2023.

[7] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. **ACM Transactions on Information Systems**, 2023.

[8] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. **ACM Computing Surveys**, Vol. 55, No. 12, pp. 1–38, 2023.

[9] Chenhao Fang, Xiaohan Li, Zezhong Fan, Jianpeng Xu, Kaushiki Nag, Evren Korpeoglu, Sushant Kumar, and Kannan Achan. Llm-ensemble: Optimal large language model ensemble method for e-commerce product attribute value extraction. In **Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval**, pp. 2910–2914, 2024.

[10] Yichong Huang, Xiaocheng Feng, Baohang Li, Yang Xiang, Hui Wang, Bing Qin, and Ting Liu. Ensemble learning for heterogeneous large language models with deep parallel collaboration, 2024.

[11] Han Yang, Mingchen Li, Huixue Zhou, Yongkang Xiao, Qian Fang, and Rui Zhang. One llm is not enough: Harnessing the power of ensemble learning for medical question answering. **medRxiv**, 2023.

[12] Gioele Barabucci, Victor Shia, Eugene Chu, Benjamin Harack, and Nathan Fu. Combining insights from multiple large language models improves diagnostic accuracy. **arXiv:2402.08806**, 2024.

[13] Tengfei Xue, Xuefeng Li, Tahir Azim, Roman Smirnov, Jianhui Yu, Arash Sadrieh, and Babak Pahlavan. Multi-programming language ensemble for code generation in large language model. **arXiv:2409.04114**, 2024.

[14] OpenAI et al. Gpt-4o system card, 2024.

[15] Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. **arXiv:2403.05530**, 2024.

[16] サンケイリアルエステート投資法人. Home. https://www.s-reit.co.jp/. Accessed: 2025-01-09.