日本語金融ドメインにおける大規模言語モデルの原因・結果表現抽出能力強化

辻拓真 [1]　平野正徳 [2]　今城健太郎 [2]
坂地泰紀 [1]　野田五十樹 [1]
[1] 北海道大学　[2] 株式会社 Preferred Networks
tsujipon.0718@gmail.com　research@mhirano.jp　sakaji@ist.hokudai.ac.jp

概要

金融文書からの原因・結果表現抽出は投資判断などに有用だが，大規模言語モデル（LLM）を用いた手法では過抽出による精度の低下，ルールベース手法は再現率の低さが課題となっている．本論文では，既存のルールベース手法のアルゴリズムを自然言語で記述し，LLM を実行器として用いる手法を提案する．決算短信を用いた評価実験の結果，適切な粒度のアルゴリズムの提示により，意味的一致評価において，提案手法は主要モデル（GPT-5，Gemini 2.5 Pro，gpt-oss-120b）を含む全 8 モデル中 5 モデルで，ゼロショット抽出および既存ルールベース手法を同時に上回る F1 スコアを達成した．

1　はじめに

企業分析・投資判断では，財務指標に加え，決算短信等の金融文書に含まれる記述の活用が重要である．金融文書には因果的説明（例：「○○の増加により売上が伸長した」）が多く，これらを抽出できれば業績変動の要因理解や検索・要約などに資する．

因果関係抽出には，大規模言語モデル (LLM) の応用が期待される．しかし，LLM のみを用いた抽出では，因果関係のない記述を誤って抽出しやすく（過抽出），結果の信頼性を損なうおそれがある [1]．また，ルールベース手法は抽出根拠が明確で制御しやすいが，定義されたパターン以外の表現を取り逃がす課題がある．そこで，ルールベースの手順による制約で過抽出を抑制しつつ，LLM の言語理解能力で多様な表現への対応力を補うことができれば，高い精度と再現率を両立できると考えられる．本研究では，ルールベース因果抽出手法のアルゴリズムを自然言語で記述し，LLM をその解釈・実行器として用いることで，原因と結果を表す表現の抽出を行う．

本論文の主な貢献は以下の通りである．

1. 既存のルールベース原因・結果表現抽出手法のアルゴリズムを自然言語で記述し，LLM をその解釈・実行器として用いる原因・結果表現抽出手法（提案手法 1）を提示する．
2. 提案手法 1 に対し，アルゴリズムの一部を決算短信特有の表現に合わせて調整した手法（提案手法 2）を実装し，汎用枠組みに対するドメイン知識注入の有効性を検証する．
3. 複数 LLM および複数のプロンプト条件で体系的に比較評価し，失敗様式と有効な指示/調整を整理する．

2　関連研究

2.1　因果関係抽出

因果関係抽出の研究は，因果の有無を判定するテキスト分類，エンティティ間の関係を判定する関係抽出，および原因と結果を原文中のスパンとして特定する系列ラベリングの 3 つの設定に整理されている [2]．金融ドメインにおいても，国際共有タスク FinCausal[3] 等を通じて，因果関係抽出の枠組みが整備されている．

因果関係抽出手法は，ルールベース，教師あり学習ベース，および LLM ベースに大別できる．ルールベースは説明可能性に優れるが表現の多様性への対応に課題があり，教師あり学習ベースは汎化性能が高いが，教師データの整備が必要である．一方，LLM ベースは In-context Learning により少量データで適用可能であり [4]，ゼロショット抽出も現実的になりつつある [5]．しかし，抽出範囲の揺れや形式逸脱といった精緻な出力制御に課題がある [6] ほか，過抽出の問題も指摘されている [1]．

2.2 プロンプトによるアルゴリズムの提示と実行

プロンプトを用いてアルゴリズムを提示する試みとして，Algorithm of Thoughts (AoT) [7] がある．これは，探索アルゴリズムの実行例を提示することで，LLM に探索的推論を行わせる手法である．

これに対し本手法は，探索的推論ではなく，文書からの情報抽出を対象タスクとする．また提示内容も，AoT のような実行例ではなくアルゴリズムの実行手順そのものである点で異なる．LLM を推論器としてではなく，与えられた手順の解釈・実行器として用いる点が特徴である．

3 原因・結果表現抽出手法

本章では，問題設定を説明した後に，提案手法の土台となる既存手法と，提案手法を説明する．

問題設定 本研究では，決算短信に含まれる記述から原因・結果表現を抽出するタスクを対象とする．坂地ら [8] に倣い，文献 [9] に準拠し，原因・結果表現は「出来事（結果）」と「その理由（原因）」の組から構成されるとし，1 文中または隣接 2 文中に直接表現された明示的な記述に限定する．

入力は決算短信文書の断片とし，出力は原因表現（*basis*），結果表現（*result*），および両者を結び付ける手がかり表現（*clue*）からなる三つ組のリストと定義する．手がかり表現とは，「により」「を背景に」「その結果」等のように因果関係を示唆する表現を指す [8]．

3.1 既存手法: 坂地らのルールベース手法

坂地らのルールベース手法 [8] は，まず手がかり表現を検出する．次に，検出された手がかり表現と周囲の文節との係り受け関係（CaboCha[10] による解析結果）を分析し，原因・結果表現がどのような位置関係にあるかを構文情報に基づく 5 つのパターン（A–E）のいずれかに分類する．この分類過程では，「手がかり表現の係り先文節は動詞句か？」といった言語特徴に基づく Yes/No の判定手順（Step 1–5）が適用される（図 1）．最後に，特定されたパターンごとに定義された規則に従い，原因表現と結果表現の範囲を特定・抽出する．パターン（A–E）の概要を表 1 に示す．詳細な定義および識別手順（Step 1–5）については，付録 A を参照されたい．

決算短信向けの拡張 決算短信では，「主な要因は〜によるものです。」のように，文頭の定型句

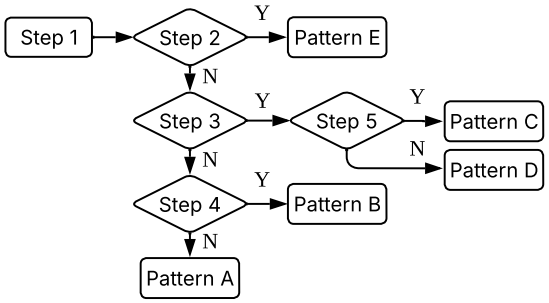

図 1 坂地らのパターン識別手順のフローチャート

表 1 坂地らによる構文情報に基づくパターン概要

Pattern	構造概要
Pattern A	原因 → 手がかり → 結果
Pattern B	結果 (主) → 原因 → 手がかり → 結果 (述)
Pattern C	結果 → 原因 → 手がかり （倒置）
Pattern D	文 1(結果) → 文 2(原因 → 手がかり)
Pattern E	文 1(原因) → 文 2(手がかり → 結果)

（「主な要因は」）と文末の手がかり表現（「によるものです。」）の組み合わせで原因表現が記述されることが多い．坂地らは，この文頭定型句（Prefix Pattern）を用いて Step 5 の識別条件を拡張し，Prefix Pattern を持つ文を原因表現（その前文が結果表現，Pattern D）と判定することで，課題であった Pattern C / D の誤分類を改善している [8]．

3.2 提案手法: ルールベース手法の自然言語化による原因・結果表現抽出

図 2 に提案手法の全体像を示す．まず，坂地らの手法に基づく 5 つのパターン（Pattern A–E）と，それらを識別する手順（Step 1–5）を自然言語の手順として記述し，プロンプトテンプレートを作成する．次に，抽出対象文をテンプレートに挿入する．係り受け解析を使う条件の場合は，係り受け解析器 CaboCha[10] による抽出対象文の解析結果を補助情報としてテンプレートに挿入し，LLM に入力する．LLM は，手がかり表現を同定して適用パターンを決定し，原因表現（*basis*），結果表現（*result*），手がかり表現（*clue*），および適用パターン（*pattern*）を抽出する．出力はこれらの要素を持つオブジェクトのリストを格納した JSON とし，JSON Schema で規定された構造化出力により生成させる．

プロンプト設計 プロンプトは (1) アルゴリズム部，(2) 出力制約部から成る．アルゴリズム部には，3.1 節のパターン定義（Pattern A–E）および識別手順（Step 1–5）を記述する（提案手法 1）．ただし，Step 5 を Prefix Pattern による判定に置換したものを提案

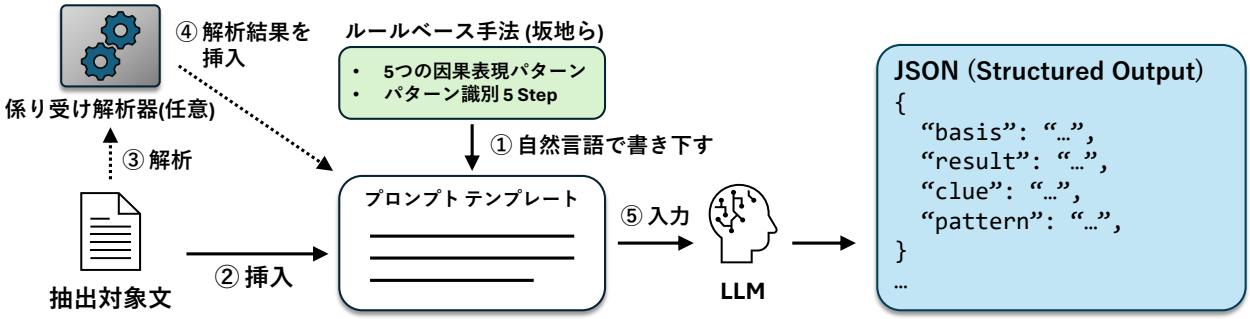

図 2　提案手法の全体パイプライン

手法 2 とする．出力制約部では，アノテーション規則に準拠した制約を明示する．

補助情報（係り受け解析結果）の利用　係り受け解析結果は識別手順の分岐判断に利用できる一方，係り受け解析器の誤解析が入力ノイズになる可能性もある．係り受け解析結果を挿入する/しないの両設定を用意し，性能への寄与を比較する．

4　評価実験

データセット　坂地ら [11] による，決算短信 PDF から抽出した 1–3 文からなるファイル 370 件を用いる．各ファイルには，1 件以上の原因・結果表現のペアがアノテーションされている（総ペア数 452）．

比較条件　比較対象は，(i) 既存手法（坂地らのルールベース手法; 3.1 節），(ii) アルゴリズム部を与えず LLM に抽出をさせるゼロショット抽出，(iii) ルールベース手順を自然言語化した提案手法 1（3.2 節），および (iv) Step 5 の条件文を調整した提案手法 2（同）の 4 条件とする．

提案手法では，パターン定義から識別ステップ（Step 1–5）まで情報を段階的に追加するアブレーション分析も行う．さらに，係り受け解析器 CaboCha[10] の解析結果をプロンプトに付与する/しない条件を設け，補助情報の影響を比較する．

対象モデルと推論設定　本実験では，商用 API 経由で利用可能なクローズドモデル（GPT-4o, GPT-5, Gemini 2.5 Pro, Claude Sonnet 4.5）および，オープンウェイトモデル（gpt-oss-20b, gpt-oss-120b, Llama 3.1 Swallow Instruct 8B v0.5, Gemma-2-Llama Swallow 9B IT v0.1）の計 8 モデルを対象とする．推論の深さが調整可能なモデルについては，API コストを考慮しクローズドモデルでは低〜中負荷，オープンウェイトモデルでは性能を優先し高負荷の設定を採用する．各実験は 3 回試行し（seed 指定時は 0, 42, 777），平均値を報告する．詳細な設定は付録 B（表 5）を参照されたい．

評価指標　評価指標には，予測および正解の原因・結果表現ペアを最大二部マッチングで対応付けて算出された，マイクロ平均 F1 スコアを用いる．正解判定の基準は，(i) 完全一致，(ii) 部分一致，(iii) 意味的一致（LLM-as-a-Judge[12] として GPT-5 による意味的等価判定）の 3 通りとする．各基準の詳細な定義は付録 C を参照されたい．

4.1　実験結果

各一致基準における実験結果（F1）を表 2 – 4 に示す．表の行はモデル[1)]，列は条件を表す（Zero: ゼロショット抽出，提案手法（Pat.: パターンのみ，+Sk: Pat. + Step 1–k（例: +S3 は Step 1–3 までを含む提案手法），+S5/+S5*: Pat. + Step 1–5；Step 5 はそれぞれ提案手法 1/2））．最高値を太字，次点を下線で示す．

表 2　実験結果：完全一致（係り受け解析結果なし，F1）

Model	Zero	Pat.	+S1	+S2	+S3	+S4	+S5	+S5*
GPT-4o	**0.28**	0.26	0.25	0.25	0.25	0.21	0.24	0.22
GPT-5	0.23	0.23	0.28	**0.30**	0.29	0.20	0.23	0.25
Gemini-2.5	0.36	0.41	0.44	**0.46**	0.45	0.34	0.38	0.39
Claude-4.5	0.10	0.14	0.20	0.17	0.19	0.31	0.30	**0.32**
gpt-oss-20b	0.18	0.22	0.28	0.29	0.29	0.28	0.27	**0.30**
gpt-oss-120b	0.25	0.28	0.31	**0.38**	0.37	0.28	0.25	0.26
Llama-Sw-8b	0.07	0.05	0.06	**0.08**	0.07	0.04	0.05	0.05
Gemma-Sw-9b	0.02	0.01	0.01	0.01	0.01	0.02	0.02	**0.02**
Rule-based				0.45				

表 3　実験結果：部分一致（係り受け解析結果なし，F1）

Model	Zero	Pat.	+S1	+S2	+S3	+S4	+S5	+S5*
GPT-4o	0.74	0.73	0.75	0.73	0.75	0.77	0.77	**0.78**
GPT-5	0.76	0.76	0.74	0.76	0.75	0.76	0.74	**0.77**
Gemini-2.5	0.84	0.88	0.89	0.88	**0.89**	0.87	0.88	0.87
Claude-4.5	0.80	0.81	**0.81**	0.81	0.80	0.79	0.79	0.79
gpt-oss-20b	0.65	0.69	0.72	**0.73**	0.72	0.70	0.69	0.72
gpt-oss-120b	0.73	0.75	0.74	0.74	**0.75**	0.72	0.72	0.75
Llama-Sw-8b	0.51	0.51	0.51	0.52	0.53	0.56	0.56	**0.57**
Gemma-Sw-9b	0.49	0.38	0.47	0.47	0.47	**0.50**	0.48	0.47
Rule-based				0.72				

1)　モデル名は略称．Sw は Swallow を表す．

表 4　実験結果：意味的一致（係り受け解析結果なし，F1）

Model	Zero	Pat.	+S1	+S2	+S3	+S4	+S5	+S5*
GPT-4o	0.73	0.73	0.77	0.75	0.76	0.77	**0.78**	0.77
GPT-5	0.74	0.73	0.73	0.75	0.73	**0.77**	0.76	0.76
Gemini-2.5	0.79	0.83	**0.86**	0.85	0.86	0.85	0.85	0.85
Claude-4.5	**0.80**	0.80	0.80	0.79	0.78	0.80	0.79	0.78
gpt-oss-20b	0.72	0.74	0.77	0.77	0.75	0.74	0.72	**0.77**
gpt-oss-120b	0.76	**0.79**	0.76	0.76	0.77	0.73	0.74	0.75
Llama-Sw-8b	0.52	0.51	0.58	0.60	**0.61**	0.60	0.59	0.60
Gemma-Sw-9b	0.53	0.37	0.47	0.50	0.49	0.55	**0.56**	0.54
Rule-based					0.65			

5　考察

ルールの詳細度と知識注入の効果　図 3 に 8 モデル平均のアブレーション分析結果を示す．最も大きな性能向上は Step 1 の導入時（Pat→+S1）に確認された．Step 1 では抽出対象の手がかり表現リストをプロンプト内で明示しており，手順の提示に加えてこの知識注入の効果が大きく寄与したと考えられる．その後，手順を詳細化（Step 2–4）するにつれて精度は向上する傾向にあり，自然言語による指示が LLM の探索空間を適切に制約していることが推測される．F1 スコアは Step 4 で最高値（平均 0.727）に達し，Step 5（+S5）では飽和した．この結果は，詳細なルール記述が精度向上に有効であることを示す一方で，LLM への指示には F1 スコアを最大化する最適な粒度が存在することを示唆している．また，ドメイン固有の表現規則を用いた提案手法 2 の Step5（+S5*）の効果は，一部のモデルで改善が見られたものの大局的には限定的であった．これは，明示的なルールを与えずとも，LLM がその文脈理解能力によって当該パターンを十分にカバーできているためと考えられる．

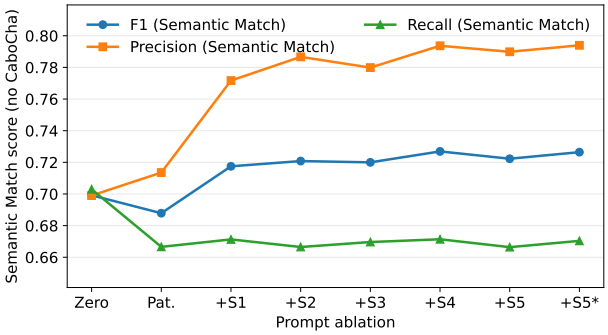

図 3　アブレーション分析における精度，再現率，F1 スコアの変化 (8 モデル間平均，意味的一致，係り受け解析結果なし)

意味的一致評価では，主要モデル（GPT-5，Gemini 2.5 Pro，gpt-oss-120b）に適切な粒度の手順を提示することにより，ゼロショット以上の精度とルールベース以上の再現率を両立し，両手法を上回る F1 スコアを達成した．これは手順制約による過抽出の抑制と，LLM の言語理解能力が多様な表現の捕捉を後押ししたためだと考えられる．

係り受け解析情報の有効性とモデル差　図 4 に係り受け解析情報の付与による意味的一致スコアの変化を示す．付与効果はモデルにより異なり，Gemini，Claude のみで向上した．これは解析誤りがノイズとなる可能性に加え，構造化情報の解釈能力やプロンプト長への耐性の差異が影響したと推察される．

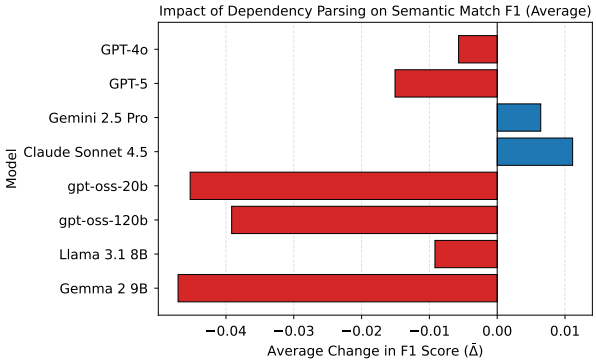

図 4　係り受け解析情報の付与による意味的一致スコアの平均的な変化（全プロンプト条件の平均差分）

エラー分析と課題　誤りの要因は大きく 2 点に分けられる．第一に，手がかり表現の見落としによる再現率低下であり，これが F1 スコアを下げる主要因であった．第二に，結果先行型（Pattern C, D）における精度低下である．事前学習データにおいては原因・結果の出現順序が「原因 → 結果」となる記述が比較的多いと考えられるため，逆の順序（結果 → 原因）の記述に対しては予測が不安定になり，関係のない記述を過剰に抽出する傾向が見られた．

6　おわりに

本研究では，決算短信からの原因・結果表現抽出において，ルールベース手法の手順を LLM に実行させる手法を提案した．実験の結果，適切な粒度の手順提示により，LLM によるゼロショット抽出とルールベース手法の欠点であった，精度と再現率のトレードオフの改善が確認された．また，一部のモデルでは係り受け解析情報の有効性も確認された．今後は，外部ツールを用いた手がかり表現候補の提示による再現率向上や，結果先行型パターンの抽出精度改善に向けた検証ステップ導入に取り組む．

謝辞

本研究は JST さきがけ JPMJPR2267 の支援を受けたものです.

参考文献

[1] Takehiro Takayanagi, Masahiro Suzuki, Ryotaro Kobayashi, Hiroki Sakaji, and Kiyoshi Izumi. Is chatgpt the future of causal text mining? a comprehensive evaluation and analysis, 2024.

[2] Jinghang Xu, Wanli Zuo, Shining Liang, and Xianglin Zuo. A review of dataset and labeling methods for causality extraction. In Donia Scott, Nuria Bel, and Chengqing Zong, editors, **Proceedings of the 28th International Conference on Computational Linguistics**, pages 1519–1531, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics.

[3] Dominique Mariko, Hanna Abi-Akl, Estelle Labidurie, Stephane Durfort, Hugues De Mazancourt, and Mahmoud El-Haj. The financial document causality detection shared task (fincausal 2020). In **Proceedings of the 1st Joint Workshop on Financial Narrative Processing and MultiLing Financial Summarisation**, pages 23–32, 2020.

[4] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020.

[5] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. In **Advances in Neural Information Processing Systems**, volume 35, pages 27730–27744, 2022.

[6] Darren Yow-Bang Wang, Zhengyuan Shen, Soumya Smruti Mishra, Zhichao Xu, Yifei Teng, and Haibo Ding. Slot: Structuring the output of large language models, 2025.

[7] Bilgehan Sel, Ahmad Tawaha, Vanshaj Khattar, Ruoxi Jia, and Ming Jin. Algorithm of thoughts: Enhancing exploration of ideas in large language models. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, **Proceedings of the 41st International Conference on Machine Learning**, volume 235 of **Proceedings of Machine Learning Research**, pages 44136–44189. PMLR, 21–27 Jul 2024.

[8] 坂地泰紀, 酒井浩之, and 増山繁. 決算短信 pdf からの原因・結果表現の抽出. **電子情報通信学会論文誌**, 2014.

[9] 庵 功雄. **新しい日本語学入門：ことばのしくみを考える**. スリーエーネットワーク, 第 2 版 edition, 4 2012. 355p.

[10] 工藤 拓 and 松本 裕治. チャンキングの段階適用による日本語係り受け解析. 43(6):1834–1842, 2002.

[11] Hiroki Sakaji and Kiyoshi Izumi. Financial causality extraction based on Universal Dependencies and clue expressions. **New Generation Computing**, 41(4):839–857, November 2023. Issue date: 2023-11-01.

[12] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena, 2023.

A　既存手法のパターン識別手順

坂地らによるルールベース手法では，手がかり表現が含まれる最後尾の文節を核文節，核文節の係り先の文節を基点文節とし，以下の手順で適用パターンを識別する [8].

1. 手がかり表現を含む文を取得する．（提案手法ではこのとき手がかり表現リストの提示を行う）
2. 手がかり表現が文頭に出現する場合は Pattern E を適用し，処理を終了する．
3. 手がかり表現に句点が含まれる，または手がかり表現の直後に句点がある場合は Step 5 へ進む．それ以外は Step 4 へ進む．
4. 基点文節が動詞句であり，かつ基点文節が係り先である文節中に係り助詞または格助詞を含むものがあれば Pattern B，そうでなければ Pattern A を適用して終了する．
5. 核文節に係っている文節に係り助詞が含まれている場合は Pattern C，それ以外は Pattern D を適用して終了する．

【決算短信特化の Step 5（+S5*)】
末尾の手がかり表現が「によります。」等であり，かつ文頭が Prefix Pattern（「主な要因といたしましては」等）である場合は，Pattern D を適用する．

B　LLM の詳細設定

本実験で使用した各 LLM の推論設定を表 5 に示す．表中の「Temp.」は Temperature パラメータを，「Reasoning」は推論モデル等における推論の深さを指定するパラメータ（例：reasoning_effort）を表す．「–」は該当するパラメータの指定がない（デフォルト設定）ことを示す．

表 5　LLM の推論・デコード設定

Model	Temp.	Reasoning
GPT-4o	0.0	–
GPT-5	1.0*	minimal
Gemini 2.5 Pro	0.0	default
Claude Sonnet 4.5	0.0	–
gpt-oss-20b	0.0	High
gpt-oss-120b	0.0	High
Llama 3.1 Swallow Instruct 8B v0.5	0.0	–
Gemma-2-Llama Swallow 9B IT v0.1	0.0	–

共通条件：Structured output（同一 JSON Schema, strict）．
* 温度は指定不可（固定）．

C　評価指標の詳細

完全一致（Exact Match）　正解と予測の原因表現および結果表現の文字列が一文字も違わずに一致する場合．正規化は行わず，空白や句読点を含めた厳密な一致を要求する．

部分一致（Partial Match）　原因表現と結果表現の両方において，正解と予測の一方が他方を連続した部分文字列として含む場合．完全一致の場合も含む．

意味的一致（Semantic Match）　正解と予測の原因表現と結果表現が，それぞれ意味的に同義であるか LLM-as-a-Judge で判定し，一致と判定された場合．完全一致の場合は，無条件で一致と判定する．

意味的一致の判定では，GPT-5（reasoning_effort=minimal）に対し，予測されたペアと正解ペアの意味的な類似度を 6 段階で評価させた．使用した評価基準を以下に示す．

- **5: 完全一致**。表現も意味も完全に同じ。
- **4: ほぼ同義**。表現は少し異なるが、因果関係の意味内容は同じと解釈できる。
- **3: 包含関係**。どちらか一方が、もう一方を具体化・要約した関係にある。
- **2: 関連あり**。トピックは関連しているが、因果関係の捉え方が異なる、または不正確。
- **1: 因果逆転**。原因と結果が逆になっている。
- **0: 無関係**。全く異なる内容を指している。

本研究では，スコア 4 以上を正解（True Positive）として扱った．

LLM-as-a-Judge の検証　LLM-as-a-Judge を用いた評価を行った正解-予測ペアのサンプル（N = 219,367）から，LLM が一致と判定したペア（スコア 4 以上）から 200 件，不一致と判定したペア（スコア 3 以下）から 200 件をランダムサンプリングした．サンプリングしたペアに対して，LLM の判定を伏せた状態で人手（著者 1 名）で match / non-match のラベリングを行い，最終的に Accuracy，Precision，Recall，および F1 スコアを算出した．

表 6　LLM-as-a-Judge の検証結果

指標	値
Accuracy	0.9063
Precision	0.8283
Recall	0.9820
F1 Score	0.8986