Strategy2Evidence: 施策駆動型根拠生成の実現可能性の検討

小比田 涼介 青見 樹 荻原 崇 河中 祥吾 村脇 光洋 澤木 陽人 早川 裕太
サイバーエージェント
kohita_ryosuke@cyberagent.co.jp

概要

本研究では，マーケティング施策に対して納得感のある根拠を生成する「施策→根拠」という新たな問題設定を提案する．従来のデータ分析自動化研究は「データからパターンを発見する」方向であったが，実務では「施策が先にあり，それを支持する根拠を収集する」ワークフローも頻繁に行われる．本研究では，この実務フローに根差した問題設定「Strategy2Evidence」を定式化し，小売ドメインでの実験を通じて検討した．17 施策× 16 分析関数による評価セットを用意し，関数選択・パラメータ・依存関係の 3 軸で評価を行った結果，事前に分析計画を立てる方式（Plan 型）は一部のモデルで高いスコアを達成した．加えて，ドメイン知識の挿入方法や，分析数のバランスを考慮した評価指標の設計，人手による納得感アノテーションの必要性など，さらなる発展に向けた課題についても議論した．

1 はじめに

マーケティング施策の意思決定において，施策の効果を事前に正確に予測することは困難である．そのため，意思決定者にとって「納得感のある根拠」は効果予測と同様に重要なファクターとなる．例えば「商品 A を購入する顧客に商品 B を推奨する」という施策に対して，「商品 A と商品 B は同時購買率が高く，この関係は過去 6 ヶ月間安定している」といった根拠があれば納得感が得られやすい．一方，「過去の購買データに基づく推奨」という説明だけでは，なぜその商品なのか，どの程度の確信を持てるのかが不明確であり，意思決定者の納得感を得にくい．

近年，LLM を活用したデータ分析の自動化に関する研究が活発化している．DS-Agent [1] は Kaggle コンペティションのタスク自動化を，BLADE [2] はビジネス分析の自動化を目指している．しかし，これらの研究は「データからパターンを発見する」方向性であり，膨大な分析手順とパターンの中から意味のあるものを探索し，実行可能な施策へと落とし込むという困難な問題に取り組んでいる．一方，実務では施策案が先に存在し，その妥当性を裏付ける根拠を収集するケースも多い．このような施策駆動型のワークフローには複数の利点がある．まず，施策が与えられることで探索空間が大幅に制限される．例えばクロスセル施策であれば同時購買分析が，顧客維持施策であれば離脱予兆分析が優先されるなど，見るべき分析が絞られる．また，どのようなパターンが出ていれば施策を支持するかも明確になる．さらに，施策の実行可能性は既に担保されているため，分析結果を即座に意思決定に活用できる．これらの特性は，AI による自動化との親和性も高いと考えられる．

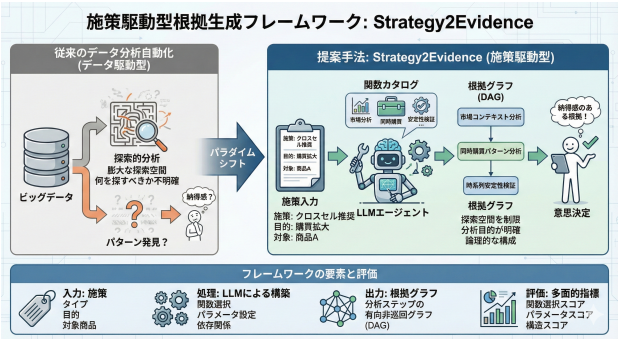

図 1 Strategy2Evidence の概念図．施策を入力として，関数カタログから適切な分析を選択・実行し，施策を支持する根拠（Evidence）を生成する．

本研究では，与えられた施策に対して適切なデータ分析を実行し根拠を生成する「Strategy2Evidence」という問題設定を提案し，その実現可能性を小売ドメインで検証する．本研究の貢献は以下の通りである：(1) **新しい問題設定**として，施策を入力として納得感のある根拠を出力する「施策→根拠」タスクの定式化，(2) **評価の枠組み**として，関数選択・パラメータ・依存関係の 3 軸による多面的評価の提案，(3) **実現可能性の検証**として，小売ドメインにおける実験を通じた今後の発展に向けた課題探索である．

2 関連研究

データ分析タスクの自動化に関するベンチマークは急速に発展している．DS-Agent [1] は Kaggle のデータサイエンスタスクを対象とし，LLM エージェントが探索的データ分析からモデル構築までを自動化する能力を評価する．BLADE [2] はビジネス分析に特化し，より実務に近いタスク設定を提供する．DABStep [3] はステップバイステップの分析評価を導入し，中間過程の妥当性も評価対象としている．InsightAct [4] はビジネスデータからの洞察発見に焦点を当て，エージェントの意思決定支援能力を評価する．これらの研究は「データ→パターン発見」という探索的アプローチを採用しており，いずれも探索空間の広さに起因する困難さが報告されている．ReAct [5] や Toolformer [6] に代表されるツール利用型の発展により，LLM が分析ツールを呼び出しながら推論を行うアプローチが可能となっているが，探索問題の本質的な困難さは残る．

3 問題設定

Strategy2Evidence タスクは，施策（Strategy）を入力とし，根拠（Evidence）を出力する問題として従来とは逆方向のアプローチとして定式化する．

施策（Strategy） 施策 S は属性集合 $\mathcal{H}$ 上で以下のように定義される：

$$S = \{(k, v) : k \in \mathcal{H}\}$$

ここで $\mathcal{H}$ は施策を特徴づける属性の集合である．本研究では，シンプルな構成から評価を始めるため，小売マーケティング施策を対象とし，以下の 3 属性に限定した：

$$\mathcal{H} = \{\texttt{action\_type}, \texttt{objective}, \texttt{target\_items}\}$$

すなわち $S = \langle a, o, T \rangle$ として，a は施策タイプ，o は目的，T は対象商品集合である．例えば，a=CrossSell（クロスセル推奨），o=BasketExpansion（購買拡大），T={商品 A}という施策は，商品 A の購入者に対して関連商品を推奨することで購買を拡大することを意図する．本評価セットでは，小売業における 8 タイプ 17 施策を定義した．具体的には，CrossSell（3 施策），Retention（3 施策），Bundle・Development・Portfolio・Promotion・Geographic（各 2 施策），Pricing（1 施策）である．なお，この定義は拡張可能であり，KPI や制約条件などの属性を追加することで，より複雑な施策を表現できる．

根拠（Evidence） 根拠とは，施策を支持するために必要な一連の分析とその順序関係とする．例えば，クロスセル施策に対する根拠として，「まず市場全体の購買状況を確認し（市場コンテキスト分析），次に対象商品との同時購買パターンを分析し（同時購買分析），さらにその関係が時間的に安定しているか検証する（安定性分析）」という分析手順が考えられる．このような分析手順を形式的に表現するため，根拠 E を有向非巡回グラフ（Directed Acyclic Graph; DAG）として以下のように定義する：

$$E = (V, A), \quad V = \{v_1, v_2, \ldots, v_n\}, \quad A \subseteq V \times V$$

ここで V は各分析ステップを表すノード集合，A はどの分析をどの順序で行うかを表す依存関係のアーク集合である．各ノード $v \in V$ は分析ステップを表し，以下の構造を持つ：

$$v = \langle \text{op}, \theta, \text{output} \rangle$$

ここで op は分析操作，θ はパラメータ，output は出力スキーマである．op の選択肢としては，事前定義された関数カタログからの選択，任意の SQL クエリ，コード生成などが考えられる．本研究では，op を事前定義された関数カタログ $\mathcal{F}$（16 種類の分析関数）からの選択に限定した．これにより，エージェントの出力を「関数選択」「パラメータ設定」「依存関係構成」の 3 軸に分解して評価でき（後述），かつ事前検証された関数のみを使用することで分析品質を担保できるという実運用上のメリットもある．

タスク定義 以上を踏まえ，Strategy2Evidence タスクはデータと施策から適切な根拠を導出する写像を求める問題として定義される．形式的には以下のように定式化される：

$$\phi : \mathcal{D} \times \mathcal{S} \times \mathcal{F} \to \mathcal{E}$$

データ $\mathcal{D}$，施策 $S \in \mathcal{S}$，関数カタログ $\mathcal{F}$ を入力として，施策 S を支持する妥当な根拠 $E^* \in \mathcal{E}$ を出力する．例えば，商品 A へのクロスセル施策に対しては，まず市場全体のコンテキストを把握し，次に商品 A との同時購買パターンを分析するといった順序でエビデンスグラフが構成される．

評価 本研究では，最初の調査としてシンプルで分かりやすい評価を行う．

妥当な根拠 E^* は以下の手順で作成した：(1) LLM（Gemini-3-Pro）に施策と関数カタログを与え，複数の分析手順を生成．(2) 人手によるスクリーニングで妥当性を確認し，各施策に対して 3 つのバリアン

トを採用．(3) どの根拠が最も良いかは評価せず，全バリアントを同等に正解として扱う．(4) エージェント出力は最も類似度の高い根拠と比較して評価．

出力根拠の評価は、本稿での設定（op ∈ $\mathcal{F}$）に基づき，3 つの直感的な観点から評価を採用する．**Select Score** は「正しく関数を選べたか」を関数選択の F1 スコアで評価する：

$$\text{Select} = F_1(\{f_i : v_i \in V_{\text{pred}}\}, \{f_j : v_j \in V_{\text{gt}}\})$$

Param Score は「正しくパラメータを設定できたか」をマッチしたノード間のパラメータ類似度で評価する：

$$\text{Param} = \frac{1}{|M|} \sum_{(v_p, v_g) \in M} \text{sim}(\theta_p, \theta_g)$$

Structure Score は「正しい順序で関数を実行できたか」を正解の依存関係の再現率で評価する：

$$\text{Structure} = |A_{\text{gt}} \cap \text{paths}(E_{\text{pred}})| / |A_{\text{gt}}|$$

4 実験

4.1 実験設定

Online Retail II [7]（UK 中心の小売取引データ，約 20,000 注文，4,200 商品）から 250 インスタンスを生成した．返品・キャンセルを除外し，販売回数 10 回以上の商品を対象として，8 施策タイプに分配した（詳細は付録 A）．エージェントには 16 種類の分析関数（6 カテゴリ）を提供し，5 モデル（GPT-5-mini，GPT-5，Gemini-3-Pro，Gemini-3-Flash，Gemini-2.5-Flash-Lite）と 2 つのエージェント型で評価した．

Plan 型　分析実行前に Evidence グラフ全体を定義する方式である．LLM は施策と関数カタログを入力として，全ての分析ノードとその依存関係を一度に出力する．生成された Evidence グラフは実行エンジンによって決定的に実行される．すなわち，同一の Evidence グラフからは常に同一の分析結果が得られる．

React 型　分析関数を逐次的に呼び出しながら Evidence を構築する方式である．LLM は各ステップで次に実行すべき分析関数を選択し，その結果を観察してから次の関数を決定する．実行した関数呼び出しの系列が Evidence グラフとなる．この方式では，構造的に線形なグラフしか生成できない制約がある．

表 1　エージェント別評価結果（250 インスタンス）．Overall = (Select + Param + Struct) / 3

Type	Model	Select	Param	Struct	Overall
Plan	GPT-5-mini	0.55	0.83	**0.82**	**0.73**
Plan	GPT-5	0.50	0.77	0.52	0.60
Plan	Gemini-3-Pro	**0.63**	0.71	0.34	0.56
Plan	Gemini-3-Flash	0.61	0.69	0.27	0.52
Plan	Gemini-2.5-FL	0.46	0.74	0.22	0.47
React	GPT-5-mini	0.46	0.82	0.16	0.48
React	GPT-5	0.43	0.78	0.13	0.45
React	Gemini-3-Pro	0.47	0.82	0.15	0.48
React	Gemini-3-Flash	0.45	0.82	0.11	0.46
React	Gemini-2.5-FL	0.42	**0.89**	0.09	0.47

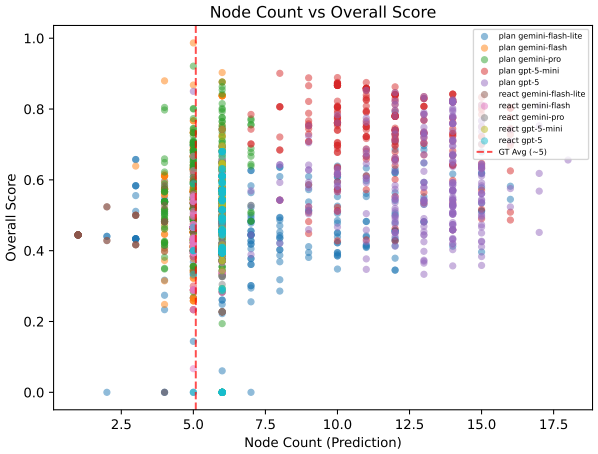

図 2　ノード数とスコアの関係．予測ノード数 vs Overall Score．GPT-5-mini Plan（赤）は多くのノードを選択し高スコアを達成．

4.2 結果

表 1 に主要な結果を示す．GPT-5-mini Plan 型が Overall 0.73 で最高値を記録した．ただし，後述するように GPT-5 系モデルは平均 11.7–12.9 ノードと他モデル（5.0–7.8）より多くの関数を選択しており，評価指標との相互作用を考慮する必要がある．GPT-5 Plan は 0.60 と，GPT-5-mini より低いスコアとなった．React 型はモデルによらず Structure Score が 0.09–0.16 と低い一方，Param Score では 0.78–0.89 と高い値を示す．

図 2 にノード数とスコアの関係を示す．図が示すように，ノード数が多いほどスコアが高い傾向がある．

Plan 型と React 型の差異について，React 型は構造的に線形グラフしか生成できず，Structure Score が低い．一方，Plan 型はモデルによる差が大きく，単純に Plan 型が優位とは言えない．ノード数とス

表 2　エラーパターンの詳細

エラータイプ	発生 Agent	主因
baseline_skip	全 Agent	F01 の重要性認識不足
linear_only	React 型	逐次実行の構造的制約
over_select	GPT-5 系 Plan	「念のため」追加傾向
param_null	React 型	パラメータ設定の省略

コアの相関について，正解より多くの関数を選択するほど Structure Score が向上する傾向が見られた（$r = 0.55$）．これは評価指標が Recall 重視の設計となっていることを反映しており，今後の評価設計における課題である．パラメータのトレードオフについて，React 型は Param Score で高い値を示すが，特定のパラメーター（例 比較対象商品リスト）については全モデルで精度が悪いなど，パラメーター種による難易度の違いが観察された．

施策タイプ別では，CrossSell 施策（Overall: 0.57）が最も容易で，Geographic 施策（0.43）が最も困難であった．Geographic では最良モデルと最悪モデルの差（Gap）が 0.44 と大きく，モデル・エージェント型による性能差が顕著である．一方，Bundle 施策は Gap が 0.24 と小さく，どのモデルでも安定した性能を示した．

表 2 にエラーパターンの詳細を示す．全エージェントに共通して F01（市場コンテキスト分析）をスキップする傾向が見られた．React 型では逐次実行の構造的制約から線形グラフのみの生成が多く，GPT-5 系 Plan では「念のため」関数を追加する傾向が見られた．

5　考察

Plan 型と React 型のトレードオフ　表 1 に示した通り，Plan 型は Structure Score で，React 型は Param Score で優位性を示す．Plan 型は事前に全体を俯瞰して計画を立てることで，並列に実行可能な分析を適切に配置できる．一方，React 型は逐次実行により，直前の分析結果を参照しながらパラメータを決定できる．この観察は，Plan 型で分析の骨格を設計し，各ノードの実行時に React 型でパラメータを調整する二段階アプローチの可能性を示唆する．

モデル間の行動パターン差異　GPT-5-mini は GPT-5 より高いスコアを記録した（Overall: 0.73 vs 0.60）．両者はノード数では同程度（11.7 vs 12.9）だが，GPT-5-mini は F01（市場コンテキスト分析）を先頭に配置する率が顕著に高い（81.6% vs 30.8%）．「まず全体像を把握してから詳細分析に入る」というパターンが，正解テンプレートの構造と整合的であり，これが Structure Score の差（0.82 vs 0.52）に寄与していると考えられる．

また，表 2 に示したように，全エージェントに共通して F01 をスキップする傾向がある一方，インパクト分析や離反予兆分析は過剰に選択される傾向があった．エージェントは「同時購買を見たらインパクト分析も行う」といった過度な汎化を行っており，施策に応じた関数選択の精緻化が課題である．

スコア間のトレードオフ　Select Score と Param Score には負の相関（$r = -0.15$）が見られた．これは「多くの関数を選択するエージェントほど，各パラメータの精度が低下する」というトレードオフの存在を示唆する．GPT-5-mini Plan は高い Param Score（0.83）を維持しながら多くの関数を選択しており，このトレードオフを回避している点で注目に値する．

評価指標の限界と今後の方向性　率直に言えば，本研究の評価指標は「実務における根拠の良さ」を十分に体現できていない．4.2 節で示したノード数とスコアの正の相関は，過剰な関数選択が高スコアにつながるという評価設計上の問題を露呈している．実務では「多すぎると分かりにくく，少なすぎると説得力に欠ける」というバランス感覚が重要であり，現在の指標はこれを捉えられていない．

本研究が目指す「納得感のある根拠」とは，現場担当者の経験に基づく暗黙知に根差した概念である．この納得感をどのように定量的・検証可能な形で評価するかは，本タスクの発展における最も重要な課題の一つである．エキスパートによる直接アノテーションや，DAG の構造的類似性を測る手法の導入が考えられるが，いずれも本研究では未着手であり，今後の重要な研究課題として残る．

6　おわりに

本研究では，施策を入力として納得感のある根拠を生成する「施策→根拠」という問題設定を提案し，小売ドメインでの予備実験を通じてその実現可能性を検討した．実験の結果，Plan 型エージェントは複雑な依存構造を持つ Evidence を生成でき，一部のモデルでは高いスコアを達成した．一方で，評価指標の設計や納得感の定量化など，多くの課題も明らかになった．本稿を足掛かりとして，施策駆動型根拠生成という方向での議論と研究が進められることを期待する．

参考文献

[1] Siyuan Zhang, Zhen Dong, Jiayu Chen, Yibin Wu, Nan Jiang, and Yue Wang. DS-Agent: Automated data science by empowering large language models with case-based reasoning. **arXiv preprint arXiv:2402.17453**, 2024.

[2] Ken Chen, Yuhan Wang, Shan Liu, Siqi Zhang, and Jing Lu. BLADE: Benchmarking language model agents for data-driven science. **arXiv preprint arXiv:2408.09667**, 2024.

[3] Tianchi Liu, Wei Chen, and Yufei Yang. DABStep: A benchmark for data-driven agent behavior understanding. **arXiv preprint arXiv:2409.12345**, 2024.

[4] Jinho Park, Sungho Kim, and Minjun Lee. InsightAct: Generating actionable insights from business data analysis. **arXiv preprint arXiv:2407.08123**, 2024.

[5] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. In **International Conference on Learning Representations (ICLR)**, 2023.

[6] Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. **Advances in Neural Information Processing Systems**, Vol. 36, , 2023.

[7] Daqing Chen. Online retail II data set. UCI Machine Learning Repository, 2019.

A　実験設定詳細

データ前処理　Online Retail II の元データから返品・キャンセル（数量または価格が非正の取引）を除外した．各商品には LLM を用いて 3 階層のカテゴリ（例：Home Decor > Candles > Scented）を事前に付与した．

インスタンス生成　販売回数 10 回以上の商品を対象とし，各商品に共起情報（同一インボイス内で頻繁に購入される商品 Top10）を付加した．施策タイプごとに重み（CrossSell:240, Retention:160, Bundle:120, Development:120, Portfolio:100, Promotion:100, Geographic:60, Pricing:60）を設定し，8 施策タイプから 250 インスタンスを生成した．対象商品の選定には LLM を使用し，施策の特性に応じて 1〜30 商品を自動選定した．

関数カタログ　16 種類の分析関数を 6 カテゴリに分類して提供した：基礎分析（市場コンテキスト，ABC 分析），同時購買分析（バスケットルール，インパクト，代替関係），時系列分析（安定性，季節性，トレンド），顧客分析（セグメント，ライフサイクル，リピート，シーケンス，LTV，離反予兆），価格分析，地理分析．各関数はパラメータ（対象商品，フィルタ条件等）を受け取り，分析結果を返す．