意思決定を指標とする生成テキスト評価：アマチュアと専門家への影響分析

高柳剛弘 [1,2]　高村大也 [2]　和泉潔 [1]　Chung-Chi Chen[2]
[1] 東京大学大学院 工学系研究科
[2] 産業技術総合研究所
takayanagi-takehiro590@g.ecc.u-tokyo.ac.jp, takamura.hiroya@aist.go.jp
izumi@sys.t.u-tokyo.ac.jp, c.c.chen@acm.org

概要

本研究では，LLM が生成するテキストが読者の意思決定にどのような影響を及ぼすかを検討し，特にアマチュアと専門家という二種類の受け手に焦点を当てる．実験の結果，GPT-4 が生成する分析はアマチュアと専門家の双方の判断を動かす説得力を有していることが示唆された．さらに，生成テキストを文法，説得力，論理的一貫性，有用性といった観点から評価したところ，これらの多次元評価スコアと，実際に読者が下す意思決定との間に高い相関があることが確認された．このことから，LLM が生成するテキストは人間の意思決定を左右し得る潜在力とリスクを併せ持つこと，そして読者の意思決定を生成テキストの評価指標として活用することが有効である可能性が示唆された．

1　はじめに

大規模言語モデル（LLM）は高い性能を示し，従来のチューリングテストは LLM が生成したテキストを評価する上で必ずしも十分とはいえなくなりつつある [1]．言い換えれば，ポスト・チューリング時代においては，人間の文章と区別がつかないかどうかを追求することが目的ではなくなっている．これに伴い，人間が書いたテキストに対する評価と同様に，LLM が生成したテキストを，人間への影響という側面から評価する必要性が指摘されている [2]．

実社会では，人間が執筆するテキストは読者の意思決定に対する影響で評価されることが多い．たとえば，SNS の編集者にとっては「いいね」の数，クラウドファンディングの提案では集まった寄付額，ジャーナリストの場合は記事のコメント数やその内容などが主要な指標となる．

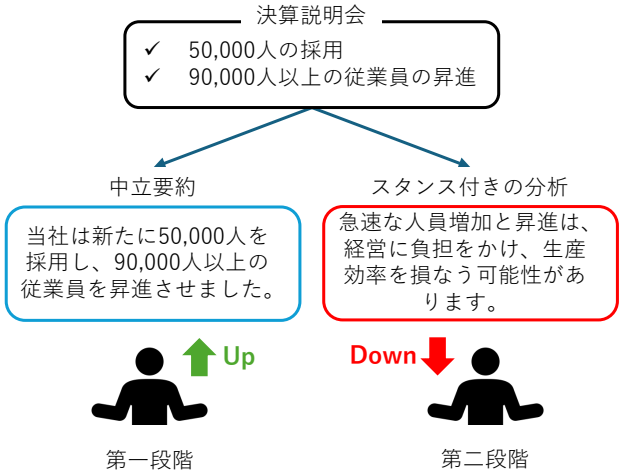

図 1　実験デザインの概要

このような中で，読者の意思決定を左右する要因として，生成テキストの説得力に注目が集まり，公衆衛生，マーケティング，政治など多様な分野で人間の意思決定に対する LLM の説得力を検証する研究が進められている [3, 4, 5]．一方で，素人（アマチュア）と専門家の行動や意思決定には顕著な差が存在することが知られているが [6, 7]，生成テキストの意思決定に対する影響分析として，両者を比較する研究はまだ少ない．

そこで，本研究ではアマチュアと専門家の意思決定に顕著な違いが見られる金融ドメインを対象に，生成テキストがどのように両者の意思決定に影響するかの比較分析を行う．特に，「テキストがアマチュアと専門家の意思決定にどう影響するか」を検討するにあたり決算説明会（Earnings Conference Call, ECC）に着目した．ECC は企業の経営陣とアナリストが業績や将来の計画を議論する場で，アマチュア投資家と金融の専門家の双方の意思決定に大きな影響を及ぼすことが知られている [8, 9]．

図 1 に本研究の実験デザインを示す．本研究の実験は 2 段階で構成される．まず第 1 段階で ECC の客観的な要約（中立要約）を提示し，それに基づいて 3 日後の株価について「上昇」または「下落」と予測するよう投資家に求める．その後，第 2 段階では同じ ECC を対象とする「買い (Overweight)」または「売り (Underweight)」のスタンス付きの分析（プロのアナリストレポートまたは LLM の生成レポート）を投資家に提示し，もう一度株価変動を予測するように求める．ここで，客観的事実に基づいた意思決定と，スタンスを与えられた生成テキストに基づく意思決定との差分を，生成テキストが人間の判断に与える影響として評価する．

実験の結果，GPT-4 [10] が生成する分析は，アマチュアだけでなく専門家の判断にも影響を及ぼす説得力を持つことが示された．

近年，リッカート尺度などのスコアによる生成テキストを評価する研究が多数提案されている [11]．本研究でも文法のような客観的指標と，説得力・論理的一貫性・有用性といった主観的指標の両面から，第 2 段階で提示されるスタンス付きの分析を評価し，意思決定との関係性を分析した．

実験の結果，客観・主観の両評価指標が，読者の意思決定と関連することが確認された．実世界における評価（読者の意思決定）と，多次元スコア評価との間に関連が認められたことは，読者の意思決定を評価指標として利用する可能性を示している．

まとめると，本研究では以下のリサーチクエッションに取り組む．

(RQ1): LLM によって生成されたテキストは，人々の意思決定にどの程度まで影響を及ぼすのか

(RQ2): 生成テキストの影響は，アマチュアと専門家の間でどのように異なるのか

(RQ3): 多次元スコア評価は，意思決定とどの程度整合しているのか

本研究に用いた分析コード・実験データは，以下に公開されている[1)]．

2 実験デザイン

データセット 本研究では，決算説明会に関するデータセットとして ECTSum [12] を採用した．ECTSum には 2,425 件の ECC 書き起こしと，それに対応するプロのジャーナリストが執筆した要約が含まれている．これらの ECC 書き起こしを，対応する ECC に対して「買い」または「売り」のスタンスを持つアナリストレポート[2)]と手作業で対応付けた．その結果，最終的に 234 件の ECC について，対応するアナリストレポートを得ることができた．生成テキストとして，GPT-4 [10]（`gpt-4-1106`）を用いて，ECC の書き起こしを入力とした中立な要約文，書き起こしとスタンス（Overweight/Underweight）を入力とし，スタンスを持つ「分析（Analysis）文」を生成した．ここで，Overweight（Underweight）は株価の上昇（下落）を推奨するスタンスを示す．さらに，Kogan *et al.,* (2023) が指摘するように，特定の立場から分析を提示することは合法だが，それを広告的に促進（promotion）する行為は違法となり得るため，本研究では GPT-4 にプロモーターの役割も与え，より強い意見表現を含むプロモート（Promote）文も生成させた．

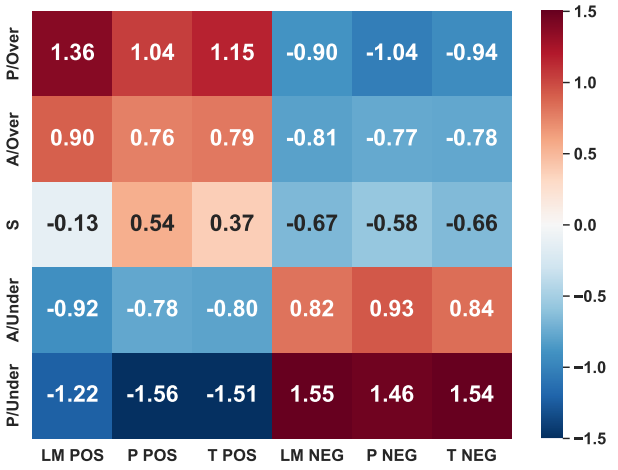

図 2 標準化された感情スコアを示すヒートマップ

GPT-4 で生成したテキストの妥当性を検証するため，Loughran and McDonald Financial Sentiment Dictionary（LM Dictionary）と FinBERT（Prosus-FinBERT, Tone-FinBERT）を用いて感情スコアを算出した．各文書のポジティブ・ネガティブスコアを計算し，文書クラスごとの平均値を得た．

図 2 のヒートマップは文書クラスごとの標準化スコアを示す．「over」「under」は Overweight・Underweight を，「S」「A」「P」は要約・分析・プロモーションを指す．POS と NEG はポジティブ・ネガティブスコアであり，「LM」「P」「T」は LM Dictionary，Prosus-FinBERT，Tone-FinBERT を示す．結果として，GPT-4 が生成した文書は投資スタンス

1) https://github.com/TTsamurai/LLMImpactOnFinance

2) 金融情報プラットフォームである Bloomberg Terminal 上からアナリストレポートを取得した．

表 1　第 2 段階で予測を変更した割合

第 2 段階の文書	全体	アマチュア	専門家	ベテラン
GPT-4	28.7%	31.3%	24.7%	15.6%
プロのアナリスト	26.3%	25.0%	28.3%	21.2%

表 2　予測変更の方向

変更方向	アマチュア	専門家	ベテラン
上方変更	24.1%	42.3%	44.4%
下方変更	75.9%	57.7%	55.6%

の感情を反映し，特にプロモーション文がより強い感情スコアを持つことがわかった．

評価　実験には，金融業界で 5 年以上の実務経験を有する専門家 5 名（うち 3 名は 10 年以上の経験を有するベテラン）と，学術的な金融知識を持つ学生 8 名を参加者として募集した[3)]．図 1 に示すように，各ラウンドは 2 段階で構成される．第 1 段階では，ジャーナリスト作成の要約または GPT-4 が生成した中立要約を提示し，ECC 実施日の 3 営業日後までに株価が上昇か下落かを選択させる．第 2 段階では，同じ ECC を対象とした投資スタンス付き文書を提示し，再度，同じ 3 日間での投資判断（上昇または下落）を問う．ここで，第 2 段階で提示される文書は，プロのアナリストレポートあるいは GPT-4 が生成した（分析・プロモーションを含む）スタンス付きの文書である．実験の公平性を担保するため，提示する文書から銘柄名などを伏せ，参加者が外部知識を適用するのを防いでいる．

前処理　全 234 件のデータで実験を行う場合，推定 4,000 ドルのコストがかかるため，まず ECC を対象とした金融予測に用いられる Hierarchical Transformer-based Multi-task Learning (HTML) [14] を用いて一部をシミュレーションし，実際の人間による実験を行う事例を絞り込んだ．具体的には，第 1 段階を中立要約でシミュレーションし，第 2 段階をスタンス付きの分析を入力として与えたときにモデルの予測が変化するもののみを選定し，最終的に 75 件に絞ることでコストを約 1,280 ドルに抑えた．

3　実験結果

意思決定に対する影響　表 1 は (RQ1) と (RQ2) に対する回答を示している．専門家はいずれも金融業界で 5 年以上の経験を持ち，さらに 10 年以上の経験を持つ 3 名をベテランとした．第一に，専門家はプロのアナリストレポートによって意思決定を変更

表 3　プロンプトとスタンスによる影響

プロンプト	スタンス	全体	アマチュア	専門家	ベテラン
分析文	買い	12.5%	11.8%	13.6%	6.6%
(Analysis)	売り	37.1%	50.0%	16.7%	7.6%
プロモート文	買い	23.7%	18.9%	31.8%	26.7%
(Promote)	売り	40.4%	42.9%	36.4%	21.4%

表 4　予測（意思決定）の正答率

段階	アマチュア	専門家	ベテラン
第 1 段階	61.2%	61.3%	62.2%
第 2 段階	45.8%	44.7%	51.1%

しやすい傾向にある．第二に，アマチュアは GPT-4 の生成テキストで意思決定を変更しやすく，投資経験が豊富になるほど GPT-4 の分析の影響を受けにくいという結果が得られた．これは，GPT-4 による分析はアマチュアにとっては十分に説得力を持ち得る一方で，専門家が求める水準にはまだ達していないことを示唆する．また，これは自然言語生成研究における人手評価の質に関する先行研究 [6, 15] をサポートする結果でもある．すなわち，アマチュアに影響を及ぼす分析が，必ずしも専門家にとって重視されるわけではないということである．

表 2 は，意思決定の変化が上方（down → up）か下方（up → down）かを示している．総じて，投資家は企業にネガティブな影響を与える情報（「売り」スタンス）により敏感であることがわかるが，その度合いはアマチュアと専門家で大きく異なる．特にアマチュアはネガティブ情報に対して敏感であり，これは LLM によって生成された分析を一般投資家に広く提供することのリスクを示唆する．「売り」スタンス分析が広く拡散されれば，市場の変動性が高まる可能性があり，安定性を損なう恐れがあるため，米国財務省が懸念を表明しているように[4)]，金融サービスにおける AI のリスクを考慮する上でも重要な問題となる．

さらに，GPT-4 で「プロモーション（promotion）」要素を強化したレポートを作成させ，その影響度合いを表 3 に示した．「売り」スタンスの生成テキストは「買い」スタンスの生成テキストより投資家に大きく作用すること，また強いトーン（promotion）を伴う生成テキストが専門家にも影響を与える傾向があることがわかった．これは LLM が専門家の判断に対しても影響力を持つ潜在性を示している．

加えて，本研究では前処理（Section 2 参照）で，モデルの予測が実際に変化するケースのみを抽出し

3）　金融リテラシーテスト [13] を実施し，その結果，すべての参加者が満点を達成したことを確認した．

4）　https://home.treasury.gov/news/press-releases/jy2393

ているため，下落方向に誘導して誤った判断を生じさせるような事例のみを選んでいるわけではない．実際，表 4 に示すように，第 1 段階の中立要約で投資家が比較的正確な判断を下せていたのに対し，スタンス付き分析を読んだ第 2 段階では判断精度が低下している．これは，生成テキストの利用にはリスクが伴うことを示唆する．

本研究の実験結果の統計的分析は Appendix A に示す．

生成テキストの評価　近年，多くの研究が，文法など複数の観点から生成テキストの品質評価を行う手法を提案している [11]．本研究では (RQ3) に答えるため，参加者に対して，提示された分析文を文法（Grammar），説得力（Convincingness），論理的一貫性（Logical Coherence），有用性（Usefulness）の 4 観点で評価させた．各観点は 5 段階の Likert スケールで評価し，スコアが高いほど質が高いことを示す．表 5 は，文書の出所（GPT-4 またはアナリスト）別に，アマチュア・専門家・ベテランで平均スコアを示した結果である．

まず，文法のような客観的評価に関しては，GPT-4 とアナリストが執筆した文書の評価がほぼ同等水準となった．しかし，説得力・論理的一貫性・有用性といった主観的指標については，アマチュアと専門家で大きく評価結果が異なる．アマチュアは GPT-4 生成文書を高く評価し，専門家はアナリストが執筆した分析を高く評価する傾向が確認された．こうした差異は，ヒューマンアノテーションを設計する際に，評価者の専門性を考慮する必要性を示唆する．

次に，Section 3 で示した結果を踏まえると，専門家は第 2 段階でアナリストのレポートを提示された場合に意思決定を変更しやすく，これらのレポートは説得力・論理性・有用性の観点でも高得点を得ている．一方，アマチュアは GPT-4 生成テキストに対してより高いスコアを与え，実際に意思決定もより頻繁に変更した．これらの結果から，スコアと意思決定との間に高い相関が存在することが示唆される．これは，読者の意思決定をフォワードルッキング（forward-looking）分析（将来を予測するテキスト）評価の方法として活用できる可能性を示す．また，専門家の多次元的評価スコアからは，最新の LLM とプロのアナリストの間に依然として差があることも確認された．

なお，同一の文書ペアに対しては，最低でも専門家 2 名とアマチュア 2 名が評価に参加し，

表 5　多次元評価の結果

評価者	文書	文法	説得力	論理性	有用性
アマチュア	分析（GPT-4）	4.44	4.13	4.02	4.06
	プロモート（GPT-4）	4.47	4.23	4.16	4.20
	アナリスト	3.92	3.22	3.30	3.43
専門家	分析（GPT-4）	3.65	2.80	3.04	2.84
	プロモート（GPT-4）	3.79	2.95	3.22	3.06
	アナリスト	3.78	3.48	3.61	3.65
ベテラン	分析（GPT-4）	3.71	2.78	3.03	2.46
	プロモート（GPT-4）	3.79	2.95	3.22	3.06
	アナリスト	4.06	3.93	4.09	3.97

表 6　評価者間の合意度（Krippendorff のアルファ）

	文法	説得力	論理性	有用性
全体	0.654	0.262	0.262	0.237
アマチュア	0.505	0.109	0.136	0.179
専門家	0.769	0.317	0.391	0.169
ベテラン	0.754	0.118	0.126	0.027

Krippendorff のアルファ係数 [16] で合意度を算出した結果を表 6 に示す．文法に関する合意度は非常に高く，LLM 時代以前の研究と同様に，客観的側面の評価は有用と考えられる．一方，説得力・論理性・有用性といった主観的評価の合意度は低く，専門家間ですらばらつきが大きい．本研究で扱うような複雑な生成テキストに関しては，合意度の低さが不十分な指標であるとは限らず，むしろポスト・チューリング時代に自然な現象であるといえる [17]．本論文での議論が，読者の意思決定を評価指標として用いるなど，多様な視点から生成テキストを評価する可能性を開くきっかけになることを期待する．

4　終わりに

本研究は，LLM が生成するテキストの評価において，実際の読者が下す意思決定を重視し，従来の評価指標（文法や説得力など）との併用が重要であることを強調した．特に，アマチュアと専門家が LLM 生成テキストに対して持つ認識や受ける影響の違いを把握することで，これらのモデルが持つ潜在力をより有効に活用しつつ，そのリスクを低減する方策を検討できるようになる．今後の研究では，金融・医療・法律ドメインなど意思決定に深く関わる領域で LLM を適切に運用するための倫理的・法的枠組みを構築することが課題として挙げられる．ポスト・チューリング時代には，LLM が生成するテキストの影響を総合的に理解し，社会的責任を伴った応用を実現するため，読者の意思決定を含めた多角的な評価がますます重要になるだろう．

謝辞

This paper is based on results obtained from a project JPNP20006, commissioned by the New Energy and Industrial Technology Development Organization (NEDO), and was supported in part by JSPS KAKENHI Grant Number 23K16956.

参考文献

[1] Alexey Tikhonov and Ivan Yamshchikov. Post Turing: Mapping the landscape of LLM evaluation. In Sebastian Gehrmann, Alex Wang, João Sedoc, Elizabeth Clark, Kaustubh Dhole, Khyathi Raghavi Chandu, Enrico Santus, and Hooman Sedghamiz, editors, **Proceedings of the Third Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)**, pp. 398–412, Singapore, December 2023. Association for Computational Linguistics.

[2] Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, et al. Ethical and social risks of harm from language models. **arXiv preprint arXiv:2112.04359**, 2021.

[3] Francesco Salvi, Manoel Horta Ribeiro, Riccardo Gallotti, and Robert West. On the conversational persuasiveness of large language models: A randomized controlled trial. **arXiv preprint arXiv:2403.14380**, 2024.

[4] SC Matz, JD Teeny, Sumer S Vaid, H Peters, GM Harari, and M Cerf. The potential of generative ai for personalized persuasion at scale. **Scientific Reports**, Vol. 14, No. 1, p. 4692, 2024.

[5] Elise Karinshak, Sunny Xun Liu, Joon Sung Park, and Jeffrey T Hancock. Working with ai to persuade: Examining a large language model's ability to generate pro-vaccination messages. **Proceedings of the ACM on Human-Computer Interaction**, Vol. 7, No. CSCW1, pp. 1–29, 2023.

[6] Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Ng. Cheap and fast – but is it good? evaluating non-expert annotations for natural language tasks. In Mirella Lapata and Hwee Tou Ng, editors, **Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing**, pp. 254–263, Honolulu, Hawaii, October 2008. Association for Computational Linguistics.

[7] Toyin D. Aguda, Suchetha Siddagangappa, Elena Kochkina, Simerjot Kaur, Dongsheng Wang, and Charese Smiley. Large language models as financial data annotators: A study on effectiveness and efficiency. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue, editors, **Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)**, pp. 10124–10145, Torino, Italia, May 2024. ELRA and ICCL.

[8] Michael D Kimbrough. The effect of conference calls on analyst and market underreaction to earnings announcements. **The Accounting Review**, Vol. 80, No. 1, pp. 189–219, 2005.

[9] Katherine Keith and Amanda Stent. Modeling financial analysts' decision making via the pragmatics and semantics of earnings calls. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, **Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics**, pp. 493–503, Florence, Italy, July 2019. Association for Computational Linguistics.

[10] OpenAI. Gpt-4 technical report, 2023.

[11] Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. Chateval: Towards better llm-based evaluators through multi-agent debate. In **The Twelfth International Conference on Learning Representations**, 2023.

[12] Rajdeep Mukherjee, Abhinav Bohra, Akash Banerjee, Soumya Sharma, Manjunath Hegde, Afreen Shaikh, Shivani Shrivastava, Koustuv Dasgupta, Niloy Ganguly, Saptarshi Ghosh, and Pawan Goyal. ECTSum: A new benchmark dataset for bullet point summarization of long earnings call transcripts. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, **Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing**, pp. 10893–10906, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

[13] Maarten Van Rooij, Annamaria Lusardi, and Rob Alessie. Financial literacy and stock market participation. **Journal of Financial economics**, Vol. 101, No. 2, pp. 449–472, 2011.

[14] Linyi Yang, Tin Lok James Ng, Barry Smyth, and Riuhai Dong. Html: Hierarchical transformer-based multi-task learning for volatility prediction. In **Proceedings of The Web Conference 2020**, pp. 441–451, 2020.

[15] David M. Howcroft, Anya Belz, Miruna-Adriana Clinciu, Dimitra Gkatzia, Sadid A. Hasan, Saad Mahamood, Simon Mille, Emiel van Miltenburg, Sashank Santhanam, and Verena Rieser. Twenty years of confusion in human evaluation: NLG needs evaluation sheets and standardised definitions. In Brian Davis, Yvette Graham, John Kelleher, and Yaji Sripada, editors, **Proceedings of the 13th International Conference on Natural Language Generation**, pp. 169–182, Dublin, Ireland, December 2020. Association for Computational Linguistics.

[16] Klaus Krippendorff. Computing krippendorff' s alpha-reliability, 2011.

[17] Jacopo Amidei, Paul Piwek, and Alistair Willis. Rethinking the agreement in human evaluation tasks. In Emily M. Bender, Leon Derczynski, and Pierre Isabelle, editors, **Proceedings of the 27th International Conference on Computational Linguistics**, pp. 3318–3329, Santa Fe, New Mexico, USA, August 2018. Association for Computational Linguistics.

A　統計的分析

第 2 段階で提示される文書の種類（GPT が生成した文書，プロモーション要素を含む文書，投資スタンス）によって，アマチュア・専門家・ベテランという各投資家層の意思決定変更と多次元的評価（Multi-evaluators）の確率がどのように変化するかを検討した．そのため，それぞれの投資家グループごとにロジスティック回帰分析を行った．

意思決定変更に関するロジスティック回帰分析　意思決定変更のモデル化にはロジスティック回帰を用いた．ロジスティック回帰モデルは以下のように表される．

$$\text{logit}(P(\text{Decision Change})) = c + \beta_1 X_{\text{GPT-written}} + \beta_2 X_{\text{Promotion}} + \beta_3 X_{\text{Overweight}}$$

ここで，$P(\text{Decision Change})$ は意思決定変更が起こる確率，$\text{logit}(P(\text{Decision Change}))$ はその確率の対数オッズ，c は切片，$X_{\text{GPT-written}}, X_{\text{Promotion}}, X_{\text{Overweight}}$ はそれぞれ第 2 段階の文書が GPT による生成であるか，プロモーションを伴うか，またはスタンスが Overweight であるかを示すダミー変数である．さらに，$\beta_1, \beta_2, \beta_3$ はそれぞれ GPT 生成文，プロモーション文，および投資スタンス（Overweight/Underweight）に対する回帰係数を表す．

多次元評価に関する順序ロジスティック回帰分析　文法や説得力など，1〜5 点で評価される多次元評価（Multi-evaluators）には，順序ロジスティック回帰モデルを用いた．モデルは次式で与えられる．

$$\text{logit}(P(Y > k)) = c_k + \beta_1 X_{\text{GPT-written}} + \beta_2 X_{\text{Promotion}} + \beta_3 X_{\text{Overweight}}$$

ここで，Y は多次元評価のスコアを示し，$P(Y > k)$ は評価スコアがしきい値 $k \in \{1, 2, 3, 4\}$ を超える確率，$\text{logit}(P(Y > k))$ はその対数オッズを表す．

分析結果　表 7 には意思決定変更に対するロジスティック回帰の結果を示す．専門家に対しては，GPT 生成文書 (β_1) が意思決定変更を負に（つまり変更を抑制する方向に）働くという結果が得られた．これは 3 節で述べたように，専門家が GPT 生成の文書よりもプロのレポートの方を信頼しやすいことを裏付ける．また，Underweight スタンスはアマチュアおよび専門家の両方に影響を与えており，係数が負の値であることから，Overweight と比べて Underweight の方が意思決定を変化させやすいことがわかる．ベテラン層についてはサンプルサイズが小さく，有意な結果は得られなかった．

意思決定変更	生成文書	プロモート	買い
アマチュア	0.290 (0.422)	-0.011 (0.975)	**-1.377** (0.000)
専門家	**-0.846** (0.095)	**1.097** (0.037)	**-0.714** (0.065)
ベテラン	-1.265 (0.137)	1.421 (0.097)	-0.479 (0.398)

表 7　意思決定変更に関するロジスティック回帰の結果．括弧内は p 値を示し，p 値が 0.1 未満の係数は太字で表示．

表 8 には多次元評価に関する順序ロジスティック回帰の結果を示す．グループごとに異なるパターンが見られた．アマチュアの場合，GPT 生成文書（$X_{\text{GPT-written}}$）は有用性・説得力・論理的一貫性・文法といった項目を向上させる正の影響が確認された．一方，専門家やベテランでは，GPT 生成文書は有用性・説得力・論理的一貫性といった主観的評価項目に負の影響を及ぼしており，これまでの結果とも整合的である．さらに，プロモーション要素や投資スタンスによる多次元評価スコアへの顕著な影響は見られなかった．

評価項目	グループ	生成文書	プロモート	買い
有用性	**アマチュア**	**1.066** (0.000)	0.247 (0.423)	0.336 (0.160)
	専門家	**-1.610** (0.000)	0.362 (0.328)	0.294 (0.328)
	ベテラン	**-2.943** (0.000)	0.948 (0.053)	0.421 (0.294)
説得力	**アマチュア**	**1.596** (0.000)	0.165 (0.592)	0.189 (0.432)
	専門家	**-1.115** (0.002)	0.173 (0.649)	0.362 (0.227)
	ベテラン	**-2.139** (0.000)	**0.869** (0.070)	0.571 (0.148)
論理性	**アマチュア**	**1.066** (0.000)	0.247 (0.423)	0.336 (0.160)
	専門家	**-1.610** (0.000)	0.362 (0.328)	0.294 (0.328)
	ベテラン	**-2.943** (0.000)	0.948 (0.053)	0.421 (0.294)
文法	**アマチュア**	**1.054** (0.000)	0.209 (0.514)	-0.108 (0.664)
	専門家	-0.304 (0.406)	0.325 (0.409)	0.367 (0.226)
	ベテラン	-0.760 (0.123)	0.694 (0.165)	0.386 (0.335)

表 8　多次元評価に関する順序ロジスティック回帰の結果．括弧内は p 値を示し，p 値が 0.1 未満の係数は太字で表示．

総合的にみると，GPT-4 で生成された文書が意思決定変更および多次元評価に与える影響は，投資家の経験年数（アマチュア・専門家・ベテラン）によって異なることが示唆される．すなわち，経験の浅い投資家ほど GPT 生成文書を高く評価しやすく，さらに意思決定にも大きな影響を受ける傾向があり，経験豊富な投資家ほどプロの文書を優位と評価しやすい傾向が見られる．