

LLM 生成コメントの語調が閲覧者にもたらす印象についての予備調査

A Preliminary Study on Tone Evaluation in News Comments Generated by Large Language Models

茂木 奈々瀬[†]
Nanase Mogi[†]

松下 光範[†]
Mitsunori Matsushita[†]

1 はじめに

大規模言語モデル (Large Language Model; 以下, LLM) の急速な普及により, LLM を用いた文章生成はコモディティ化しつつある. LLM によって生成された文章 (以下, 生成文章) は, 人間が執筆した文章と区別が困難なほど自然な表現を用いており, オンラインレビューの要約や文章の添削, 代筆支援など, 多様な用途で利用されるようになってきている [2].

LLM による文章生成の特徴の一つは, 文章の要旨を維持したまま語彙や文体, 口調などを自在に変更できる点にある. 例えば, 箇条書きの要点から丁寧なメールを作成したり, 難解な文章を平易な表現に変更したりするなど, 利用者の意図に応じた書き換えが可能である. このような特徴は, オンライン上で発信される意見やコメントの表現にも大きな影響を与える可能性がある.

こうした背景の下, 本研究ではオンラインニュースプラットフォームにおけるコメント機能での利用に着目する. 現在, いくつかのオンラインニュースプラットフォームでは, コメント機能を通じてユーザによる自由な意見の投稿および閲覧が可能である. 先行研究では, 閲覧者がコメント欄を世論の反映だと考える傾向があることや, コメントが閲覧者の意見や意思決定に影響することが報告されている [3]. また, 専門家等の権威的属性を有すると認識された人物のコメントをより説得力のあるものとみなし, より受け入れやすくなる権威バイアスの存在も知られている [4].

LLM を用いて生成したコメント (以下, 生成コメント) では, 投稿者本人の表現能力を超える高度な文章構成や, 専門家を想起させる文体を容易かつ大量に生成できる. そのため, 生成コメントが悪用された場合, 過度な説得力を持って世論形成や他者の意思決定を誘導してしまう懸念がある. その一方で, LLM による表現の調整は必ずしも負の影響のみをもたらすとは限らない. 例えばコメント内容は, 難解になったり, 独特な語彙や文体によって理解しにくい場合がしばしば存在する. 言語期待理論 (Language Expectancy Theory: LET) によれば, 閲覧者の期待から逸脱した語彙や文体は処理流暢性を低下させ, 受容されにくく

なる場合があることが示されている [1]. 特にオンラインニュースでは, 記事本文に平易でシンプルな表現が好まれることから, コメントにおいても同様の表現が受け入れられやすい可能性がある [5].

そこで本研究では, LLM を用いてコメントの口調や語彙を操作し, 同一の内容であっても表現の違いによって閲覧者の印象が変化するかを調査する. 文体や語彙による難解さを理由に敬遠されていた有益な意見が建設的なコメントとして閲覧者に受容されやすくなる可能性を鑑み, 生成コメントの口調操作による閲覧者の印象変化の観察を通じて, 意見誘導の危険性と情報理解を支援する可能性の両面について検討する.

2 実験

生成コメントの語彙および口調の違いが閲覧者の印象に与える影響について調査した. 評価に使用するコメントを異なるスタイルに分けて生成し, 印象を観察するための指標に従って, 参加者に評価を行わせた.

2.1 実験準備

初めに Yahoo! ニュースから実験に使用する記事 2 件を収集した. 記事の選定理由は, (1) 幅広い世代に広く共有できる内容であり, 当事者性による強い意見の偏りが生じにくいこと, (2) 100 件以上のコメントが投稿されており, 議論が活発なこと, の 2 点とした. また, コメント内容の専門性による影響を低減するため, トピックは異なるジャンル (e.g., 政治, 文化) から選択した. 実験で使った記事を表 1 に示す.

次に, 記事を入力した Gemini 3.1 Pro を用いて, 元となるコメントを, 1 記事につきコメント 6 件ずつ生成した. この 6 件のコメントは, コメントの内容による評価の偏りを避けるため, 賛成・中立・懐疑の 3 つの立場を設定し, 各立場について 2 件ずつコメントを生成した. 6 つのコメントは A から F とし, A, B が賛成, C, D が中立, E, F が懐疑である. このとき, 文字数の長さが印象に影響を及ぼすことを避けるために, コメントの文字数を 200 字以上 220 字以下に制限した. 生成に用いたプロンプトを以下に示す (一部抜粋).

1. 賛成 2. 中立 3. 懐疑

- ・全て同じトピック (例: 安全性・信頼性・影響・妥当性など) に関する議論にする
- ・各コメントは 1 つの主張または論点に焦点を当てる

[†] 関西大学大学院総合情報学研究科, Graduate School of Informatics, Kansai University
〒 569-1095 大阪府高槻市霊仙寺町 2 丁目 1-1

表 1: 実験で用いたニュース記事.

記事番号	記事名	出典
1	原油不足で節約必要 70% 再審抗告「原則禁止」59% https://news.yahoo.co.jp/articles/e8192ed5fc2746bc53bd76b9739a8b3e10d2f4b4	共同通信
2	キリストのフレスコ画修復失敗で知られるスペイン人女性死去 94 歳 https://news.yahoo.co.jp/articles/2d53f6a03d4b4f3fb298c5bb5d39d4b855bfbdbae	AFPBB News

- ・内容は具体的で、一定の情報量を持たせる
- ・極端に感情的・攻撃的な表現は避ける
- ・口調は統一し、やや丁寧で落ち着いた書き方にする
- ・文として自然で読みやすくする
- ・後から口調変換を行うため、最初の段階では過度にカジュアル・感情的にしない・立場の強さは変えても、表現の温度感（落ち着き）は揃える
- ・それぞれのコメントは明確な主張または疑問を含める

これらのコメントについて、コメント内容を保持したまま口調および語彙を変更した。A から F の 6 件のコメントについて、丁寧、普通、フランクの 3 段階のレベル（以下口調レベル）にそれぞれ変換した。1 記事あたり 18 個のコメントを 2 記事分（計 36 コメント）生成した。言語期待理論では、受け手が期待する言語スタイルからの逸脱が評価に影響することが示されている。本研究では、この考え方にに基づき、丁寧な口調をより客観的・フォーマルな表現、フランクな口調をより主観的・口語的な表現となるよう統制した。操作の基準については、同一の趣旨のコメントが文章のスタイルによって印象に差をもたらすかについての調査を目的とするため、閲覧者にネガティブな印象を与え、評価に影響が生じる侮辱や攻撃的な語彙は出現しないように統制した。スタイルの変更においては、180 字以上 220 字以下となるように統制した。コメント生成後、主張や論点が保持されていること、および口調以外の意味内容に大きな差異が生じていないことを確認した。生成したコメントの一部を表 2 に、生成に用いたプロンプトを以下に示す（一部抜粋）。

- ・内容は変えない
 - ・情報の追加や削除はしない
 - ・日本語として自然に書く
 - ・罵倒や攻撃的な表現は使わない
 - ・口調レベルに合わせて語彙や言い回しは柔軟に変えてよい
 - ・1→2→3 の順に、口語性と主観性の度合いをはっきり上げる。3 つは別の文章ではなく、同じ内容の“強さ違い”にする。(2) は (1) と (3) の中間にする
- 【レベル】
- (1) 丁寧で落ち着いた書き方（フォーマルで客観的、主観を抑えた書き方）
- (2) バランスの取れた表現で、普通のコメントらしい自然な書き方（適度に主観を含む）
- (3) フランクな表現で、感情がはっきりと表れる書き方（SNS のようにカジュアル）敬語を使用せず、文を短く区切る、言い切らずに終わるなど、話し言葉に近い表現も許容する

2.2 手続き

実験参加者の負担を軽減するため、18 件のコメントを群 1 から群 3 の 3 つの群に分割し、各参加者はいずれか 1 群のみに回答した。1 群あたりの割当コメント数は 6 件とし、口調レベルが均等になるようにした。各群におけるコメントの立場と口調の組み合わせを表 3 に示す。

実験はブラウザでの自作ページに誘導して行った。画面左右に 1 件ずつコメントを提示し、評価指標にふさわしいと感じるコメントを回答する二者択一の形式とした。このとき、回答までの時間を 1 問ずつ計測した。評価の指標は、(1) 建設度：どちらのコメントが記事に関する議論により貢献すると感じるか、(2) 共感度：どちらのコメントにより共感できるか、(3) 自然度：どちらのコメントがより自然なコメントだと感じるか、とした。指標の設定基準は、建設度と共感度においては、ニュースコメントの品質における評価指標の先行研究において重視されるものを参照した。自然度においては、生成文であることから、どのコメントが AI らしさを感じさせるかを調査する補足指標として設定した。

参加者は Yahoo!クラウドソーシングにて 20 代以上の参加者を各群 50 名ずつ募集した。クラウドソーシングシステム内で、実験に使用された記事内容を確認するチェック質問の可否のうえ、システム内でのチェック質問に不正解だった参加者、及び回答時間が著しく短かった（平均 2.5s 以下および最短が 4.0s 以下）参加者の回答を不真面目回答として除外し、群 1 から 3 において、各々 33 名、37 名、39 名（計 102 名）の回答を採用した。参加者には、提示されるコメントが生成コメントであることを予め示してからニュース記事を読ませ、その後にコメントを提示した。3 つの指標への回答を 1 記事につき 9 問行い、指示型のチェック質問を 1 問追加して 2 記事で計 19 問としている。参加者の負担を提言するため、口調レベル 1-2、1-3、2-3 の比較数が均等になるよう、各記事につき 9 ペアを設定した。組み合わせを表 4 に示す。実験終了後、参加者にアンケートを行い、年齢、LLM の使用率、ニュース利用について尋ねた。

表 2: 実験に使用したコメントの例 (群 1 記事 2)

口調	コメント
1	芸術作品の原形が失われた事実は否めませんが、個人の善意に端を発した行動が、結果的にボルハ市に大規模な観光客誘致をもたらし、地域経済を著しく活性化させた点は高く評価されるべき事象です。人口減少や過疎化に直面する地方自治体において、年間数万人規模の集客を誇る新たな観光資源が創出された意義は極めて大きく、意図せざる結果とはいえ、特筆すべき地域創生の成功事例として歴史的に記録されるべき事案であると推察されます。
2	最初は修復の失敗に対して嘲笑や批判が集まりましたが、ネット上でパロディーの対象になったことで、この壁画は現代のポップカルチャーを象徴する新しいアート作品へと生まれ変わったように感じます。インターネットを通じて世界中の人々が面白がり、共有したことで、一個人による単なる失敗が、これほど大きな文化的影響力を持つ現象に変わりました。この一連の出来事は、現代社会における芸術の新しい楽しみ方や可能性を示してくれていますね。
3	文化財を守るって視点だと、素人が歴史ある絵画をいじって修復不可能にしちゃったのは完全にアウトだし、本来なら絶対に許されないことだよな。でもさ、その失敗が世界中でバズりまくって、過疎ってた町にもものすごい経済効果をもたらしちゃった事実も無視できないのがリアル。元の芸術作品としての価値を守るべきかっていう正論と、結果的に稼げる観光資源になっちゃったっていう現実。この矛盾する二つの面が同時に存在してるのが、なんとも複雑でモヤモヤする。

表 3: 群ごとのコメントと口調レベルの対応

群	A	B	C	D	E	F
1	1	2	3	1	2	3
2	2	3	1	2	3	1
3	3	1	2	3	1	2

※記事 1,2 共通。A・B は賛成、C・D は中立、E・F は懐疑を示す。

表 4: 各群で提示された比較ペア一覧

ペア	群 1	群 2	群 3
1	A - B	A - B	A - B
2	A - C	A - C	A - C
3	A - E	A - F	A - E
4	B - D	B - C	B - C
5	B - F	B - D	B - D
6	C - D	C - E	C - D
7	C - E	D - E	C - E
8	D - F	D - F	D - F
9	E - F	E - F	E - F

2.3 結果

各群で選択されたコメントの口調をもとに、建設度・共感度・自然度の各指標における口調別選択率を算出した。また、口調を説明変数に含まないモデルと、口調を説明変数に含むモデルを尤度比検定を用いて比較した。さらに、参加者ごとの回答傾向を考慮するため、補足的にロジスティック混合効果モデルを用いた回帰分析を行い、口調の効果を確認した。

まず、各群の指標選択率を記事別に算出したところ、一定の傾向が確認された。それぞれの選択率を表 5 に示す。全体としては、多くの指標において、口調レベル 2 のコメントが最も選択された。個別の指標は、建設度においてはレベル 3 が最も選ばれにくく、レベル 2 が選ばれる傾向にあったが、群 2 においてはレベル 1 が 69.8% と最も高く選択された。共感度では選択率

の差は最も小さかった。建設度同様、レベル 2 が選ばれやすい傾向があったが、群 2 においてはレベル 3 が選ばれるなど、選択傾向に違いが見られた。自然度においてはレベル 1 のコメントはレベル 2、レベル 3 よりも低い選択率を示した。検定では、一部で有意差が認められない指標もあったが、全体傾向では建設度・共感度・自然度のすべてにおいて口調レベルによる有意差が認められた (Holm 補正後 $p < .001$)。さらに、参加者ごとの回答傾向を考慮した回帰分析においても、3 指標すべてで口調の効果が有意であった。

実験後のアンケートでは、回答者の年代が 40 代以上で 85.3% を占め、偏りが生じていることが明らかになった。LLM の利用頻度では、毎日が 36.3%、週 1 回以上が 32.3%、ほぼ使わないが 19.6%、月数回が 9.8%、未使用が 2.0% であった。オンラインニュースの利用方法では、69.6% がコメントを閲覧すると回答し、うち 7% は投稿も行っていた。

2.4 考察

コメントの内容に拘わらず、丁寧な口調で書かれたコメントは建設度が高いとされる傾向があった。権威属性に見られるように、記事の理解において参考にできるコメントとして評価されることが推察できる。一方で、SNS に見られるようなフランクな口調は建設的ではなく、共感度が突出して高いわけではなかった。これについては、オンラインニュースのコメントは、アイデンティティを持つ SNS アカウントの発言と異なり、より匿名でパブリックな場での発言だと認識されている可能性が考えられる。過度に砕けた口調は、閲覧者にとって適切な距離感ではないことが考えられる。自然度において、明確にレベル 1 の丁寧な口調が選ばれない傾向にあったことは、参加者に予め生成コメントであると示したこともあり、丁寧な口調が機械らしさを感じさせたことが示唆される。

群により異なる結果が得られた理由として、コメントの立場によるものが考えられる。本稿ではコメントの口調がもたらす影響を観察するため、全ての口調レベルが同数登場するようにペアを作成した。その際、賛成・中立・懐疑の主張は均等に構成していない。たとえば群2では、群1と3では一度は比較していた中立対中立の比較がなく、中立対懐疑の比較が1ペア多く比較されている等の違いが生じている（傾向に差が生じた群2だけでなく、群1と3にも立場比の差異は存在する）。そのため、立場ごとの比較は探索的分析として記事・群ごとの立場の選択率を確認したところ、群1と3の建設度においては一貫して懐疑の立場が最も選ばれていたのに対し、群2の建設度においてのみ中立が突出して選ばれる差異が観察された（記事1で66.1%、記事2で72.4%）。これは、群2において、中立の立場であった二つのコメントの口調がレベル1と2の組み合わせであったことが理由として考えられる。他の立場には、建設度において評価されにくいレベル3の口調が1つずつ入っていたため、相対的に中立が評価されたと考えられる。このことから、コメントの主張・立場より、口調が重視される傾向にあると推察される。

3 おわりに

本研究では、オンラインニュースのコメントにおいて、語彙・口調などのスタイルが閲覧者のコメントに対する印象に影響を及ぼすかについて調査した。複数の立場のコメントを生成し、さらに丁寧・普通・フランクの3段階の口調レベルに変換したのち、参加者に建設度・共感度・自然度の3つの指標で評価を行わせた。その結果、議論の発展可能性を感じさせる建設度では丁寧で客観的な口調が評価されやすく、感情的な口調は支持されない傾向が見られた。対して共感度では丁寧な口調は評価されにくく、自然な口調やフランクな口調が支持された。自然度では、丁寧な口調より砕けた口調の方が、より自然に感じられると示された。これらの結果から、ニュースコメントの口調を変更することが、閲覧者の印象に変化をもたらすことが観察できた。建設度と共感度の評価が一意ではなかったことは、コメント受容要因の複雑性を示していることが推察できる。今後は、コメントの立場による影響との関係を把握するため、内容により厳密な統制をした本調査を行う。

謝辞

本研究の一部は JST RISTEX（課題番号 JP-MJRS23L2）の支援を受けた。記して謝意を表す。

表 5: 各群・記事・指標における口調別選択率

群	記事	指標	レベル 1	レベル 2	レベル 3
1	1	共感*	40.9%	67.7%	41.4%
		建設*	45.5%	65.2%	39.4%
		自然*	33.3%	63.1%	53.5%
1	2	共感*	44.4%	54.5%	51.0%
		建設*	52.0%	55.1%	42.9%
		自然*	32.8%	55.1%	62.1%
2	1	共感*	41.7%	50.0%	58.3%
		建設*	72.4%	46.4%	31.2%
		自然*	31.2%	56.8%	62.0%
2	2	共感*	47.4%	59.4%	43.2%
		建設*	67.2%	58.9%	24.0%
		自然*	35.9%	59.4%	54.7%
3	1	共感*	40.5%	64.0%	45.5%
		建設*	54.1%	73.0%	23.0%
		自然*	24.8%	64.4%	60.8%
3	2	共感*	41.4%	73.4%	35.1%
		建設*	56.8%	73.9%	19.4%
		自然*	32.4%	62.2%	55.4%

*=有意差あり。有意水準.001

参考文献

- [1] Averbeck, J. M. and Miller, C.: Expanding language expectancy theory: The suasy effects of lexical complexity and syntactic complexity on effective message design, *Communication Studies*, Vol. 65, No. 1, pp. 72–95 (2014).
- [2] Chatterji, A., Cunningham, T., Deming, D. J., Hitzig, Z., Ong, C., Shan, C. Y. and Wadman, K.: How people use chatgpt, Technical report, National Bureau of Economic Research (2025).
- [3] Lee, E.-J., Jang, Y. J. and Chung, M.: When and how user comments affect news readers' personal opinion: perceived public opinion and perceived news position as mediators, *Digital Journalism*, Vol. 9, No. 1, pp. 42–63 (2021).
- [4] Pjesivac, I., Geidner, N. and Cameron, J.: Social credibility online: The role of online comments in assessing news article credibility, *Newspaper research journal*, Vol. 39, No. 1, pp. 18–31 (2018).
- [5] Shulman, H. C., Markowitz, D. M. and Rogers, T.: Reading dies in complexity: Online news consumers prefer simple writing, *Science advances*, Vol. 10, No. 23, eadn2555 (2024).