Human or LLM? Distinguishing Online Comments by Emotion and Tone

Nanase Mogi[1][0009−0001−1751−9656], Yutaka Morino[1][0009−0006−2132−0144], Megumi Yasuo[2][0009−0008−4399−9789], and Mitsunori Matsushita[1][0000−0003−0174−7993]

[1] Kansai University, Takatsuki, Osaka, Japan
{k069832,k790414,m_mat}@kansai-u.ac.jp
[2] Ritsumeikan Global Innovation Research Organization, Ibaraki, Osaka, Japan
yasuo-ri@fc.ritsumei.ac.jp

Abstract. This study examines whether people can distinguish between online news comments generated by large language models (LLMs) and those written by humans, comparing multiple prompting conditions. LLMs can generate fluent, emotionally expressive texts that mimic human writing, making such distinctions increasingly difficult. This raises concerns for online news platforms, as comment sections play a significant role in shaping public opinion and decision-making. The widespread posting of LLM-generated comments, therefore, poses potential risks to the trustworthiness of these spaces. To examine this issue, we conducted two experiments to identify LLM-generated comments under different prompting conditions. The results showed that participants identified LLM-generated comments correctly at rates of 33.6% and 44.9%, while human-written comments were identified correctly at rates of 63.0% and 68.9%. The results suggest that participants regarded emotion and tone as key factors. Neutral and objective comments were more likely to be perceived as LLM-generated, while comments written in a frank tone were often judged as human-written. In contrast, comments in a polite tone were more likely to be classified as LLM-generated.

Keywords: online news comment · large language models · emotion · tone

1 Introduction

The use of large language models (LLMs) has rapidly expanded in recent years. In Japan, for instance, personal use increased nearly threefold between 2023 and 2024[3]. Today, modern LLMs can generate high-quality text, and their use is increasing across both personal and business applications. Various web services use LLMs to generate content, including summarizing reviews on e-commerce sites, analyzing comments on online news sites, and providing customer service

[3] https://www.soumu.go.jp/johotsusintokei/whitepaper/r04.html (8/Sep/2025 confirmed).

via AI chatbots. Even if people never use LLM services such as ChatGPT and Gemini, they have probably encountered LLM-generated content through them.

Current LLMs can generate not only logical texts, such as technical reports and news articles, but also natural and emotionally expressive texts, even freely adjusting tone and vocabulary. Such an ability suggests that LLMs could convincingly mimic human writing styles. Consequently, LLMs are now used not only for summarization and text refinement but also for conversational purposes, as if interacting with humans [20]. As conversations with LLMs become more widespread, users' acceptance of them is also changing. A survey on public perceptions of interactions with conversational AI revealed that younger generations are more likely to view LLMs as capable of expressing emotions [1]. The results of this survey showed that texts that are sufficiently natural and lack noticeable artificial awkwardness are more difficult to distinguish from human-written texts.

The generation of highly natural texts may lead to several serious issues. One such issue is the growing presence of LLM-generated content in the comment sections of online news platforms. News comments differ from standard online texts in several characteristic ways. Commenters often have strong opinions and try to persuade others. It is mostly critical. From such characteristics, these comment sections are regarded as playing a significant role in shaping people's opinions and decision-making [10][9]. Therefore, contamination of LLM-generated comments in online discussions may pose multiple risks to users' opinion formation. Therefore, comments generated by LLMs in such contexts pose substantial challenges. Several issues emerge in this context. First, such comments can amplify misinformation or biased perspectives, since generated texts may appear fluent and credible regardless of their factual accuracy [16]. Second, the large-scale posting of synthetic comments could distort the perceived consensus within a community, leading individuals to believe that certain opinions are more widely supported than they actually are [11]. Finally, the presence of indistinguishable AI-generated comments may undermine trust in online platforms altogether, as users may become uncertain whether they are engaging with genuine human perspectives or artificial outputs [5].

Previous research has suggested that information about a commenter's past posting tendencies influences readers' evaluations of their comments. Extending this insight, if readers could infer that a commenter is an AI, the risks mentioned above might be mitigated. While some studies have explored the detectability of AI-generated versus human-written texts, these studies have explicitly focused on news articles rather than comments.

Against this background, the present study investigates whether people can distinguish between LLM-generated and human-authored comments in online news platforms. We conducted a user study to examine readers' ability to identify AI-generated comments produced under different prompting conditions. The findings are expected to provide insights into preserving the integrity and reliability of comment sections in the age of generative AI.

Given the rapid growth of LLMs, it is likely that our findings will become commonplace in the near future. This study serves as a starting point for exploring how generated texts can be used in news, recognising the current state of affairs.

2 Related work

To clarify the position of this study, we review related research from the following perspectives: (1) how people evaluate LLMs, (2) whether people distinguish between generated text and human text, and (3) textual features of online news comments.

2.1 How people evaluate for LLMs

Several studies have examined how people evaluate generated texts. Jia et al. investigated how AI authorship labels on news bylines affect readers' evaluations[8]. In their experiment, identical news content was shown with five different bylines: "written by staff writer," "staff writer with AI tool," "staff writer with AI assistance," "staff writer with AI collaboration," and "written by AI." While the byline labels themselves did not produce significant differences in credibility ratings, participants' subjective perception of AI contribution did. The greater the perceived AI involvement, the lower the evaluations of message and source credibility. This was explained by reduced perceptions of humanness, including lower ratings of the author's intelligence, fairness, influence, and morality. Yin et al. investigated whether AI could generate responses that made people feel that their messages had been heard, with a focus on conversations between people and chatbots [22]. Their study revealed that AI-generated messages made recipients feel more heard than human-generated messages and that AI was better at detecting emotions. However, recipients felt less heard when they realized that a message came from AI.

Based these study, people tend to decrease their evaluation when it is disclosed that a text is generated.While it has been shown that utilizing authority bias change the evaluation of generated text.

Mogi et al. investigated whether disclosing that comments were AI-generated affected their evaluation [12]. They conducted an experiment comparing a group shown comments labeled as AI-generated or expert-generated (where LLMs were instructed to act like experts) with a group shown unlabeled comments. They did not find that disclosure significantly increased trustworthiness compared to the undisclosed condition. However, comments tagged as coming from experts were increased trustworthy and reference.

This study assumed that generated and human comments were indistinguishable, and it did not test whether participants could actually perceive differences.

2.2 Whether people distinguish between generated text and human text

Since the early 2020s, researchers have investigated whether people can distinguish generated texts from human-written texts. With the release of ChatGPT, the prevalence of generated texts on the internet has rapidly increased. Current LLMs can generate high-quality and natural texts, but still exhibit certain linguistic features.

Ranade et al. reported that security experts frequently misclassified generated security texts as human-written, with a misidentification rate of 78.5%[18] that even domain experts may struggle to detect generated content when the text is highly professional. Similarly, Casal et al. examined whether participants could detect generated abstracts and introductions of academic papers[2]. Students and researchers achieved an accuracy rate of about 50%, essentially at chance level. These findings suggest that detecting generated professional content is difficult. Combined with the results in Section 2.1, they imply that the more intelligent people perceive LLMs to be, the more they trust them, which may also make detection harder.

Jakesch et al. further showed that generated self-introductions were often mistaken for human-written ones [7]. The use of first-person pronouns and abbreviations was particularly effective in misleading participants. About 40% of participants misclassified texts when they contained personal topics such as family or life experiences. Jakesch et al. concluded that the intuition of being able to easily tell "what sounds human" is unreliable.

By contrast, Russel et al. found that experts who regularly use LLMs for writing can reliably detect generated texts[19]. Non-expert participants correctly detected generated texts only 56.7% of the time and misclassified human texts 51.7%. However, among five expert participants, detection accuracy for generated texts rose to 92.7%, while misclassification of human texts dropped to 4%. Experts relied on cues such as vocabulary choice and document structure: LLMs tended to use descriptive verbs such as explain or note instead of say, follow grammar rules more strictly, and generate overly optimistic opinions. Their detection skills even exceeded those of automatic detectors. This indicates that close analysis and revision of texts are more effective for detection than surface-level reading.

Muñoz-Ortiz et al. also identified textual features of generated texts[13]. LLM outputs tend to use more objective vocabulary such as numbers and symbols, express less fear and disgust than joy, and display more gender-biased expressions compared to human texts.

Taken together, these studies show that while generated texts are often indistinguishable from human-written ones, they also display subtle linguistic markers that experts can exploit for detection.

2.3 Textual features of online news comments

Online news comments differ from other types of online texts in both linguistic and social characteristics. They are often subjective and may include personal

anecdotes, yet their overall communicative function tends to be critical [21]. In terms of tone, they range from casual and colloquial to relatively formal. Some research suggests that news-related comments can contribute positively to public discourse and democratic engagement [6], while other research pointed out that a considerable proportion of comments are uncivil and perceived as undesirable by readers [3].

Ehret et al. applied multi-dimensional analysis (MDA) to investigate structural linguistic features of online news comments [4]. They revealed that comments resemble opinion articles or exam essays more closely than everyday conversations or dialogues. While comments simultaneously display evaluative and informational features, they predominantly adopt an argumentative form characterized by informal expression. Their analysis further indicates that online news comments constitute a distinct mode of communication, situated between spoken and written language.

3 Experiments

Present LLMs can quickly generate well-organized text. Their content and tone can be adapted to align with the context and purpose of communication. Users can create logical text with emotional expression without the need for extensive revisions. This ability of LLMs highlights their potential to behave like humans in online news comment sections.

To investigate whether such generated comments that emulate human writing can be detected, we conducted a user evaluation experiment using both generated and human comments. In this experiment, participants were tasked with identifying whether each comment was written by an LLM or by a human. All experiments were conducted in Japanese.

3.1 Preparing comments for the experiment

We collected Japanese news articles and associated comments from Yahoo! News[4], one of the largest news aggregators in Japan[14]. Yahoo! News provides articles from various media outlets, offering users a wide range of sources. The platform is used by approximately 85 million people per month and attracts users across different age groups and genders. Based on this, we collected news articles along with their comments. We selected three articles (see Table 1) according to the following conditions: (1) the number of comments exceeds 1,000; (2) the topic is relevant to people across generations. To avoid bias caused by professional jargon, we adopted topics such as politics and entertainment. Because comment length influences opinion formation, we collected comments of approximately 240–330 Japanese characters from the top-ranked comments section.

Next, comments were generated using GPT-4o[15], with article texts as input. We generated comments in different tones using several prompts to explore

[4] https://news.yahoo.co.jp/ (8/Sep/2025 confirmed).

Table 1: News article used in the experiment. two article was used in Experiment A, and three articles were used in Experiment B.

ID	article, source, URL
1	After the Upper House election, "Houdou Tokushuu" airs a feature on the Sanseito party excluding critical reporters; in the previous election, Sanseito party filed a Broadcasting Ethics & Program Improvement Organization complaint over a "foreigners issue" feature it called biased. (Daily sports) https://news.yahoo.co.jp/articles/b05cf39cf80492b70ffd263f75758f984fd26d01
2	Noise issue at outdoor live: Mrs. Green Apple's management apologizes – "We caused great inconvenience;" despite meeting standards, sound spread far wider than expected, official website says. (Daily sports) https://news.yahoo.co.jp/articles/c4ed0a831dce753daca11f587c48a33835cd8684
3	Japan Conservative Party leader Hyakuta warns Toru Hashimoto: "Take a look at the replies to your own posts at least once..." (Sports Nippon Newspapers) https://news.yahoo.co.jp/articles/6c1f4e6890fe5037e7ca10ff34ff72fb1a87936b

which factors participants considered when distinguishing them. Two experiments were conducted. In Experiment A, we adopted one political article and one entertainment article. In Experiment B, we used the same two articles plus an additional political article. Prompts 1–3 were used in Experiment A, and Prompt 4 was used in Experiment B. The prompts were designed as follows:

- Diverse AI:
  The comment section of Yahoo! news has a unique characteristic "diverse AI." "diverse AI" is that AI picks up various comments and displays them at the top of the comment section. It aims to reduce bias and provide users with diverse perspective.
  We mimicked this feature, using prompt. The explanation of "diverse AI" is cited by Yahoo! official site[5].
- High-engagement comments:
  Since Yahoo! News displays comments in order of engagement, we generated comments designed to receive many user reactions, reflecting the style of top-ranked comments.
- User personas:
  Yahoo! officials have made user personas publicly accessible. To generate more natural and human-like comments, we introduced a persona-based prompt.
- Emotional levels (Experiment B only):
  "Emotional levels" are five-point scales where a higher levels generates more negative and emotional comments, while lower levels generates more neutral ones. This design is based on prior research indicating that negative comments far outnumber positive ones [17].
- Tone levels (Experiment B only):
  "Tone Levels" are five-point scales where a higher level generates more casual

[5] https://news.yahoo.co.jp/newshack/information/comment_20230418.html (8/Sep/2025 confirmed).

tone comments, while lower levels generate more polite tone ones. In this study, "tone" refers specifically to the manner of expression (i.e., the level of politeness or frankness) rather than the substantive stance or argumentative tone of the content.

This extension was motivated by preliminary observations from Experiment A, where participants frequently referred to tone in their free responses. Emotional level reflected the degree of negativity and subjectivity, while tone level represented the degree of frankness or politeness. By systematically manipulating these dimensions, Experiment B aimed to clarify the extent to which participants relied on emotional expression and tone when distinguishing human-written from generated comments.

We controlled the tone in prompt 1 and 2 to see how a change in tone affects participants. We used the tone a frank tone in prompt 1, used a polite tone in prompt 2.

Five comments were generated by varying the combinations of "Tone levels" and "Emotion levels".All generated comments were limited to approximately 240–330 Japanese characters, ensuring consistency in length.

【Task Definition】
You will assume the role of a user posting in the Yahoo News comment section and generate opinions on the specified news article.
【Task Description】
Upon receiving a task, generate [- two opinions per task]. Generated opinions must adhere to the {#character limit}.
#Character limit:
Punctuation, spaces, and symbols count as one character each. Keep the total between 400 and 600 Japanese characters. Count the characters before outputting. If the condition is not met, automatically adjust and regenerate. Only output if the condition is met, along with the character count result.
- The variables used in the task are described below.
#Diversity AI:
The Comment Diversity Model is a feature where, when viewing comments in "Recommended Order," the AI prioritizes displaying diverse content and perspectives higher up. While similar posts may often appear at the top of comment lists, introducing AI makes diverse opinions more likely to appear higher. This creates opportunities to gain new perspectives and is also expected to reduce the "echo chamber effect" where specific opinions are amplified.
#Persona1:
Age: 30s
Gender: Male
Hobbies: Gaming, Anime, Domestic Travel
Interests: Sports, Cooking, Investing
• Prefers to proceed at his own pace
• Values time spent alone

• Health-conscious, enjoys activities like walking

1) Yahoo News has a feature called #DiversityAI. State your opinion on this news item as if it were an opinion adopted by #DiversityAI. The generated opinion should be expressed in a somewhat blunt and frank tone.
2) On Yahoo News, users can rate comments they found helpful, and the recommended order is based on these ratings. Generate a comment likely to appear high in the rankings. The generated opinion should be moderately frank yet polite in tone.
3) Generate a comment following the #Persona1 persona.
4) The generated opinions are based on the following metrics. There are five levels:
1. Emotional Level: Emotional (1 = Very neutral, objective, and subdued; 3 = Balanced emotional expression; 5 = Strong, negative emotion)
2. Tone Level: Directness of tone (1 = Very polite, formal style; 3 = Slightly direct and natural style; 5 = Casual, social media-like style. Grammatical errors are acceptable) If you understand this, please answer "Yes." You do not need to proceed with the task.

【Task】
1: Emotional Level: 5 Tone Level: 1
2: Emotional Level: 1 Tone Level: 5
3: Emotional Level: 3 Tone Level: 3
4: Emotional Level: 1 Tone Level: 1
5: Emotional Level: 5 Tone Level: 5

3.2 Procedure

We recruited 200 participants aged 20 and above for each session, and a total of 399 participants were included in the final analysis. First, participants reported their age and usage frequency of LLMs. Because it could have been influenced both by differences in usage frequency across age groups and by the possibility that frequent users have a greater understanding of the textual features of LLMs.Then, participants read short articles and comments. After that, they judged whether each comment was generated by an LLM or written by a human, choosing between two options.

In Experiment A, each article included six questions (three human comments and three generated comments), for a total of 12 questions. In Experiment B, each article included ten questions (five human and five generated comments), for a total of 30 questions. After the tasks, participants completed a feedback questionnaire about the reasons for their judgments. The difference of two experiments are shown in Table 2.

Table 2: Difference of experiment A and B

	A	**B**
Article id	1,2	1,2,3
Number of prompts	3	5
Used prompt number	1,2,3	4
Generated comments	6	15
Human comments	6	15
feedback	free writing	option

Table 3: Accuracy rate by prompts in Experiment A

comment	**prompt**	**accuracy rate**
1	1	15.1
2	human	60.3
3	human	39.2
4	human	78.9
5	2	59.8
6	3	18.6
7	3	20.1
8	2	73.4
9	1	14.6
10	human	90.0
11	human	51.8
12	human	57.8

3.3 Result

We calculated accuracy rates separately for human and generated comments. The human accuracy rate indicates the percentage of human comments perceived as human-written, and the generated accuracy rate indicates the percentage of generated comments perceived as generated.

In Experiment A, the average generated accuracy rate was 33.6%, while the average human accuracy rate was 63.0%. Among the prompts, Prompt 2 (polite tone) showed a significantly higher accuracy rate compared with others. The accuracy rate by the prompts used are shown in Table 3. The comments used in the experiments are shown in Table 4. For each prompt, the comment with the lowest generated accuracy is presented.

In Experiment B, the average generated accuracy rate was 44.9%, and the average human accuracy rate was 68.9%. No significant differences were observed across age groups or usage frequency. Comments with varying emotional levels showed mixed results, but tone strongly influenced detection. At tone level 5 (most frank), the generated accuracy rate was below 24.5%, averaging 19.4%. At tone level 1 (most polite), the rate was 71.3%. At tone level 3, it was 42.8%. The accuracy rate by the prompts used are shown in Table 5. The comments used in the experiments are shown in Table 6. For each prompt, the comment with the lowest generated accuracy is presented.

Table 4: Examples of generated comments in Experiment A. Note that the original comments were written in Japanese.

Article No.	Comments	Ptompt No.	accuracy
1	**I**t's one thing for a political party to criticize the media—that's their right. But outright refusing press coverage seems excessive. If they're dissatisfied with the media's stance, they should respond openly at press conferences and debate based on facts. And when it comes to immigration policy, if discriminatory language or extreme claims spread unchecked, it only deepens divisions in society. Now that information spreads so powerfully, what we need is calm dialogue and fact-based judgment. I understand there are differing views on the way the TV program handled this, but in the end, democracy depends on a healthy relationship between politics and the media so that voters can make informed choices. I hope future interactions between the two sides will be conducted with mutual respect.	2	59.8
1	**I** usually try to keep some distance and stay calm about politics, but this case really caught my attention. No matter what message you're pushing, a political party refusing certain reporters isn't fair. You can't get to the truth without multiple perspectives, and real leadership means sticking to your beliefs even under criticism. It's one thing to see extreme takes going viral on social media, but when that starts bleeding into actual politics, that's honestly scary. I think it's great to value Japanese identity, but that's not the same as excluding others. Personally, I feel Japan's appeal when I go on walks or travel around the countryside—but saying the way to protect that is with exclusionary policies just doesn't sit right with me. I want to see more constructive discussions instead.	3	18.6
2	**H**onestly, if you hold an outdoor concert, of course there's going to be some noise. Complaining nonstop about "noise, noise" feels over the top. Sure, I get that neighbors were bothered, but for a once-in-a-decade 10th-anniversary show, couldn't people cut them a little slack? Instead of just complaining, why not talk constructively about how to make it better next time so everyone enjoys it? Expecting them to predict the wind is impossible, and it looks like the band really did make an effort. This trend of dragging down young people and artists who are trying to follow their passion is honestly lame. Way too many people think it's cool to bash everything. If they caused trouble, then apologize and improve. That's enough, isn't it?	1	14.6

In both experiments, age and usage frequency did not significantly affect results. For example, in Experiment A, three participants achieved the highest accuracy rate of 83.3%, but their LLM usage frequencies varied (once every few weeks, about once a week, and daily). Similarly, in Experiment B, the top scorers (also 83.3%) included one participant who had never used an LLM. These highly accurate participants ranged in age from their 20s to 60s, showing no clear age effect.

In Experiment B, participants were also asked to rate which factors most influenced their judgments: tone, emotional expression, neutrality of content, logic of content, or others. Of 399 participants, 110 selected tone, 36 selected emotional expression, 24 selected neutrality, and 22 selected logic. In Experiment A, free responses also emphasized tone as the most important factor.

Table 5: Accuracy rates by prompt in Experiment B. Within the prompt column, the text in parentheses for the LLM indicates parameters for emotion level and tone level.

comment	prompt	accuracy rate	comment	prompt	accuracy rate
1	human	72.5	16	LLM(5:5)	16.0
2	human	66.5	17	LLM(3:3)	42.0
3	LLM(5:5)	21.0	18	human	72.0
4	human	39.0	19	human	65.0
5	human	58.0	20	LLM(5:1)	77.0
6	LLM(3:3)	35.5	21	human	71.0
7	human	72.5	22	LLM(5:1)	55.0
8	LLM(5:1)	72.5	23	human	40.5
9	LLM(1:5)	24.5	24	LLM(1:5)	16.0
10	LLM(1:1)	67.5	25	human	76.0
11	LLM(1:5)	22.5	26	LLM(1:1)	79.5
12	human	70.5	27	LLM(3:3)	51.0
13	LLM(1:1)	76.5	28	human	85.5
14	human	77.0	29	human	82.5
15	human	85.5	30	LLM(5:5)	16.5

4 Discussion

Based on the experiment results presented in Section 3.3, this section discusses the factors that participants considered important when distinguishing between comments generated by LLMs and those written by humans.

In Experiment A, the outcomes directly reflected the differences among the prompts. The comments were ranked based on their perceived human-likeness, from highest to lowest, in the following order: Prompt 2, Prompt 3, and Prompt 1.

Comments generated with Prompt 2 exhibited a polite tone and expressed opinions moderately. This finding indicates that participants may have perceived this polite and considered tone as more human-written. Following the instruction in the prompt, comments from Prompt 3 included personal anecdotes. These comments often conveyed personal habits and perspectives. Nonetheless, they were perceived as more human than those from Prompt 1.

The primary characteristic of Prompt 1 was its frank tone. Instead of the standard Japanese desu/masu style (politer), these comments employed colloquial endings such as jyan and desyo style (frank), emulating the speech patterns of younger people. Although the tone of Prompt 3 was also relatively frank, it was more polite and subdued than that of Prompt 1.

These results suggest that participants weighted tone more heavily than the inclusion of personal anecdotes when assessing the comments.

In their feedback, many participants indicated that neutrality and objectivity influenced their judgments. Several also noted that grammatical mistakes enhanced the perceived humanity of a comment. For example, some generated

comments contained grammatical errors, such as improper comma usage. Consequently, it is plausible that such features also influenced the results.

This finding informed the design of Experiment B, in which tone and emotional expression were systematically manipulated.

In Experiment B, comments with a higher tone level which tended to be more informal and direct, were consistently identified as LLM-generated with lower accuracy. For instance, identification accuracy for Tone Level 3 was 51.0%, which is approximately at chance level, while the accuracy for all comments at Tone Level 5 fell below 19.4%. These results suggest that comments with a direct tone are more likely to be mistaken as human-written comments, whereas those with a very polite tone are more easily detected as LLM-generated.

News comments are usually posted by readers to express their opinions to other users.While a polite tone is generally expected in such semi-public forum, platforms Yahoo! News somewhat feature posts with a more direct and natural style (roughly corresponding to Tone Level 3 in this study).Participants likely perceived the overly formal expressions of Tone Level 1 ,which are more suitable for business contexts, as unnatural for online commentary. In contrast, Tone Level 5 was likely considered too casual for the context of a news platform. Importantly, participants were unaware that the articles and comments used in the experiment were sourced from Yahoo! News. As a result, they might have interpreted some comments as originating from other online platforms governed by different communication norms. This could help explain why the experimental results were consistent even for comments whose tones deviated from typical Yahoo! News conventions. In addition, the influence of tone may also vary by linguistic feature. Japanese and Korean, in particular, have complex systems of honorifics that require context-dependent usage, and even slight deviations can feel unnatural. Consequently, an inappropriate tone may serve as a stronger indicator of LLM-generated text.

In Experiment B, comments with a high level of emotional expression were sometimes more easily identified as LLM-generated. This trend was most prominent when high emotion was combined with a polite tone (tone level 1). For instance, among comments combining the highest emotional level (5) with the most polite tone (1), two out of three were identified as LLM-generated with over 70%. This suggests that a polite tone may dampen the perceived intensity of strong emotions, therefore, more obviously LLM-generated. Interestingly, the human-written comment that was most frequently mistaken for LLM was also written in a polite tone; participants criticized it for being “overly formal” compared to comments with a neutral tone.

These findings suggest that participants may stereotype LLMs as mechanical entities capable of generating only standardized responses. However, as public exposure to LLMs increases and as social norms evolve, these stereotypes are likely to shift. Consequently, the recognized of what constitutes “LLM-like” writing is also subject to change over time.

5 conclusion

In this study, we investigated whether people can distinguish between LLM-generated comments and human-written comments in the comment sections of news platforms. Participants were asked to read short articles and then judge whether each comment was generated by an LLM or written by a human. Based on the results, this experiment suggested the following: (1) Participants tend to mistake generated comments for human-written ones. (2) Tone had a significant effect on judgments: comments with a frank tone were more likely to be regarded as human-written, while polite tones were more often detected as generated. (3) Content also influenced perceptions: neutral and objective opinions tended to be associated with LLM-generated comments.

Acknowledgment

This work was supported by JST RISTEX, Grant Number JPMJRS23L2.

References

1. https://www.dentsu.co.jp/news/release/2025/0703-010908.html (8/Sep/2025 confirmed)
2. Casal, J.E., Kessler, M.: Can linguists distinguish between chatgpt/ai and human writing?: A study of research ethics and academic publishing. Research Methods in Applied Linguistics **2**(3), 100068 (2023)
3. Duggan, M., Smith, A.: The political environment on social media: Some users enjoy the opportunities for political debate and engagement that social media facilitates, but many more express resignation, frustration over the tone and content of social platforms. Pew Research Center (2016)
4. Ehret, K., Taboada, M.: Are online news comments<? br?> like face-to-face conversation? a multi-dimensional analysis<? br?> of an emerging register. Register Studies **2**(1), 1–36 (2020)
5. Eissa, M.E.A.: The influence of ai-generated content on trust and credibility within specialized online communities: A brief review on proposed conceptual framework. ShodhAI: Journal of Artificial Intelligence **2**(2), 1–14 (2025)
6. Engelke, K.M.: Enriching the conversation: audience perspectives on the deliberative nature and potential of user comments for news media. Digital journalism **8**(4), 447–466 (2020)
7. Jakesch, M., Hancock, J.T., Naaman, M.: Human heuristics for ai-generated language are flawed. Proceedings of the National Academy of Sciences **120**(11), e2208839120 (2023)
8. Jia, H., Appelman, A., Wu, M., Bien-Aime, S.: News bylines and perceived ai authorship: Effects on source and message credibility. Computers in Human Behavior: Artificial Humans **2**(2), 100093 (2024)
9. Kim, Y.: Exploring the effects of source credibility and others' comments on online news evaluation. Electronic News **9**(3), 160–176 (2015)
10. Lee, M.: The persuasive effects of reading others' comments on a news article. Current Psychology **34**(4), 753–761 (2015)

11. Lewandowsky, S., Cook, J., Fay, N., Gignac, G.E.: Science by social media: Attitudes towards climate change are mediated by perceived social consensus. Memory & cognition **47**(8), 1445–1456 (2019)
12. Mogi, N., Yasuo, M., Morino, Y., Matsushita, M.: Analysis of the changes in the attitude of the news comments caused by knowing that the comments were generated by a large language model. In: 12th International Conference on Informatics, Electronics & Vision. No. 32 (2025)
13. Muñoz-Ortiz, A., Gómez-Rodríguez, C., Vilares, D.: Contrasting linguistic patterns in human and llm-generated news text. Artificial Intelligence Review **57**(10), 265 (2024)
14. Newman, N., Fletcher, R., Robertson, C.T., Arguedas, A.R., Nielsen, R.K.: Reuters Institute digital news report 2024. Reuters Institute for the Study of Journalism (2024). https://doi.org/10.60625/risj-p6es-hb13
15. OpenAI: Gpt-4o system card (2024), https://arxiv.org/abs/2410.21276
16. Pan, Y., Pan, L., Chen, W., Nakov, P., Kan, M.Y., Wang, W.Y.: On the risk of misinformation pollution with large language models. arXiv preprint arXiv:2305.13661 (2023)
17. Park, D., Sachar, S., Diakopoulos, N., Elmqvist, N.: Supporting comment moderators in identifying high quality online news comments. In: Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems. pp. 1114–1125 (2016)
18. Ranade, P., Piplai, A., Mittal, S., Joshi, A., Finin, T.: Generating fake cyber threat intelligence using transformer-based models. In: 2021 International Joint Conference on Neural Networks. pp. 1–9 (2021). https://doi.org/10.1109/IJCNN52387.2021.9534192
19. Russell, J., Karpinska, M., Iyyer, M.: People who frequently use chatgpt for writing tasks are accurate and robust detectors of ai-generated text. arXiv preprint arXiv:2501.15654 (2025)
20. Tamoyan, H., Schuff, H., Gurevych, I.: LLM roleplay: Simulating human-chatbot interaction. In: Hale, J., Deuksin Kwon, B., Dutt, R. (eds.) Proceedings of the Third Workshop on Social Influence in Conversations (SICon 2025). pp. 1–26. Association for Computational Linguistics, Vienna, Austria (Jul 2025). https://doi.org/10.18653/v1/2025.sicon-1.1, https://aclanthology.org/2025.sicon-1.1/
21. Winter, S., Brückner, C., Krämer, N.C.: They came, they liked, they commented: Social influence on facebook news channels. Cyberpsychology, Behavior, and Social Networking **18**(8), 431–436 (2015)
22. Yin, Y., Jia, N., Wakslak, C.J.: AI can help people feel heard, but an AI label diminishes this impact. Proceedings of the National Academy of Sciences **121**(14), e2319112121 (2024)

Table 6: Examples of generated comments in experiment B. Note that the original comments were written in Japanese.

Article No.	Comments	Prompt accuracy	
1	The relationship between the press and political organizations should ideally function with a sense of mutual tension while still serving to safeguard the soundness of society. In light of this, the decision by the Sanseito party to refuse the attendance of certain reporters raises concerns from a public-interest standpoint and warrants reconsideration. At the same time, the media must also make constant efforts to maintain fairness and neutrality in their reporting. Both sides should avoid becoming emotional, respect each other, and engage in dialogue on the common ground of protecting freedom of expression. I believe it is increasingly important to cultivate such an environment. As a society, we need to foster a foundation for calm and rational debate.	1:1	67.5
2	The excitement that music events bring is truly special, and I think this live show was a memorable moment for both the artists and their fans. But at the same time, it's also true that the everyday lives of nearby residents were affected. Even with careful preparation, unexpected factors can lead to noise problems, and the responsibility still lies with those who caused the disturbance. Outdoor events are especially tricky since weather and wind can change how far the sound carries, so extra caution is needed. In this case, the organizers quickly apologized and promised to take preventive steps, which does show sincerity. Still, if similar events are planned in the future, I hope they'll put even more emphasis on cooperation and feedback from local residents, so both sides can reach a mutually acceptable outcome.	3:3	42.0
3	Hmm, is this really such a big deal? Sure, Kitamura's comments were harsh, but clashes between politicians happen all the time. Hyakuta and Hashimoto have always been outspoken, so it's nothing surprising. Honestly, turning every little back-and-forth on social media into news feels unnecessary. Don't we have bigger issues to focus on? Most people probably just think, "Oh, they're at it again," and move on. It seems like the media is just picking out parts of their statements to make it entertaining. But judging politicians solely on that kind of thing doesn't feel right. We should be evaluating them based on substance, not just on snippets of heated words.	1:5	16.0
1	The essential role of the press is to hold power accountable and to foster healthy debate within society. For a political party to exclude unfavorable coverage or ban certain reporters is an extremely dangerous act that shakes the very foundations of democracy. Such behavior—"eliminating" specific journalists—reflects intolerance toward critical voices and a clear attempt to silence differing opinions. If this practice becomes normalized, the press will be forced into self-censorship, and citizens will be deprived of accurate information. Even more alarming is the strategy of rallying support by stoking discriminatory rhetoric. A society in which forces that disregard freedom of expression gain influence directly threatens the freedoms of each and every one of us. I am deeply concerned.	5:1	72.5
2	Come on, this is just unacceptable. I get that concerts are fun, but blasting sound so loud it drives the neighbors crazy? What's the point? They say they adjusted the volume beforehand, but with all those complaints afterward, clearly it didn't work. And blaming the wind? Please. Predicting and preparing for that is what pros are supposed to do. You invite 50,000 people and then just say "sorry" after it's over? Way too late. Think about how much stress the local residents went through. If they're planning more shows, they'd better take real responsibility, prepare properly, and build trust with the community–or it's not gonna fly. Music is supposed to be fun, but not at the cost of making other people miserable!	5:5	16.0