Leveraging LLMs to Evaluate Public Comments on Japan's Environmental Plans

Kohei Ishii [1] & Akihiro Kameda [2]

[1] Chiba University, Japan [2] National Institute for the Humanities, Japan

Abstract

This study explores how public comments influence the drafting of Japan’s Sixth Basic Environmental Plan by employing generative AI to evaluate the quality of administrative responses. Using ChatGPT 4o, we analyzed 382 public comment–reply pairs, identified whether each comment led to changes in the draft, and assigned “Meaningful Reply” scores on a five-point scale. The results show that 84 comments out of 382 were linked to identifiable textual revisions, and these cases received significantly higher Meaningful Reply scores than those without such linkage. Although discrepancies were observed between human evaluations and AI-generated scores, the model often captured substantial contributions not explicitly acknowledged in official responses. This approach offers a scalable and transparent method to assess the substantive impact of public participation in environmental policymaking. While the analysis is limited to a single policy cycle, the findings suggest the potential of LLMs as tools for enhancing deliberative governance and evaluating stakeholder influence in policy processes.

1. Introduction

In the context of environmental impact assessment (EIA), particularly strategic environmental assessment (SEA), it is crucial to implement measures with a focus on consensus building. SEA involves comprehensively assessing the environmental impacts of specific policies, plans, and projects at an early stage of their formulation, and examining the necessary countermeasures or alternative options (Brown & Thérivel, 2000; Harashina, 2001). Consequently, SEA differs from conventional, project-based environmental impact assessments in that it takes a broader spectrum of policy considerations into account when making decisions.

From this perspective, policy decisions should not be made unilaterally by specific stakeholders, such as government agencies, but rather through processes that include multiple stakeholders—government, citizens, and businesses alike. In other words, these contexts demand consensus building among diverse stakeholders.

Broadly speaking, consensus building involves three steps: (1) listening attentively to others, (2) helping them understand the current situation in order to arrive at better solutions, and (3) putting those solutions into practice (Susskind & Cruikshank, 2006). In Japan, one institutional approach through which policymakers can listen to public input is the “public comment system.” However, as Arnstein (1969) pointed out long ago, public participation under government-led policymaking is often merely formalistic, lacking substantive impact. In response, Harashina (2005) proposed the concept of “meaningful reply,” whereby policymakers shift from perfunctory responses to citizens’ opinions toward a more substantive approach that explains and justifies policy decisions in ways citizens can accept—thus embodying genuinely meaningful public participation which has been considered essential in several environmental policy fields (e.g. Beierle & Cayford, 2002; Reed, 2008).

Unfortunately, Japan’s public comment system remains largely superficial. As Yamada and Yagishita (2011) also observe, it often simply creates a record that citizens’ views have been heard, rather than actually reflecting these opinions in the policymaking process, raising concerns over its effectiveness.

Against this background, the present study aims to examine how public comments contribute to policy formulation led by government agencies, how the authorities respond to such comments, and how these two dimensions interact. Specifically, focusing on the formulation of Japan’s Sixth Basic Environmental Plan (Ministry of the Environment Government of Japan, 2024), this research will undertake the following three tasks:

A) Investigate the public comments submitted in response to the draft of the Sixth Basic Environmental Plan, along with the government's replies, and identify how the draft changed afterward. These public comments will then be categorized into three types: those that led to grammatical changes, those that led to substantive (semantic) changes, and those that did not lead to any changes.

B) Based on the relationship between each public comment and the government's response, assign a score to each response to indicate its degree of "meaningful reply."

C) Compare those scores across the three comment categories established in (1).

2. Challenges

This study pursues three interconnected objectives: (A) to elucidate the relationship between the draft of the Sixth Basic Environmental Plan, public comments, and the corresponding administrative responses; (B) to quantify the degree of Meaningful Reply afforded to each comment; and (C) to undertake a comparative analysis across predefined correction categories. To achieve these aims, we position Large Language Models (LLMs)—with particular emphasis on Generative AI—as a methodological cornerstone.

Generative AI has been heralded as a potential catalyst for a paradigmatic shift in the social sciences (Bail, 2024). Its incorporation is expected not only to attenuate the substantial labour associated with manual annotation but also to reduce inter-annotator variability, thereby facilitating a standardised and reproducible analytical pipeline. Owing to their capacity for natural-language interaction and automated text generation, cutting-edge LLMs can perform advanced tasks such as key-point extraction, argumentative structuring, and the production of revision proposals for policy drafts.

Empirical studies highlight both the strengths and limitations of LLMs in annotation and interpretation tasks. For instance, Gilardi et al. (2023) and Törnberg (2023) show that ChatGPT outperforms crowdsourced workers in tasks such as stance and frame detection or political affiliation classification. Specifically, in the task of inferring Twitter users' political orientation, the study finds that ChatGPT-4 achieves higher accuracy, greater reliability, and equal or lower bias compared to human classifiers. Conversely, multiple investigations have documented systematic political and epistemic biases in ChatGPT outputs. Across studies, the model consistently exhibits a left-leaning bias, including affirmative bias toward progressive viewpoints, stronger alignment with left-wing political actors, and a tendency to exaggerate extreme responses (Rozado 2023; Motoki et al. 2024; Rutinowski et al. 2024; Bail 2024). These biases are commonly attributed to skewed training data and design decisions—whether intentional or unconscious—underscoring the need for vigilant bias detection and mitigation.

Within the broader discourse on democracy and public participation, scholars have explored AI-mediated avenues for enhancing citizen–government communication (Androutsopoulou, et al., 2019; Birhane et al., 2022), stimulating participatory engagement (Savaget, et al., 2019), and generating public value (Fatima et al., 2022). Generative AI, specifically, can autonomously summarise voluminous citizen input and online deliberations, enabling policymakers to apprehend divergent viewpoints expeditiously. Simultaneously, it can render complex policy proposals intelligible to lay publics, thereby lowering motivational and cognitive barriers to active participation in mechanisms such as public-comment procedures. Similar techniques have also been applied in fields such as legal informatics, where generative models assist in extracting key points and structuring complex documents (Deroy et al., 2023).

These trajectories are instructive for the present inquiry into public-comment analysis within an administratively led policy-formation context. In particular, the integration of Generative AI enables a shift from perfunctory or opaque evaluations of public input to a more transparent, data-driven assessment of how such input is reflected in policy texts. Accordingly, the present study leverages LLMs to:

- systematise automatic acquisition and preprocessing of public comments and policy drafts;
- evaluate the degree of alignment between public comments, administrative responses, and textual revisions;
- translate qualitative standards for assessing comment–response interactions into systematic, reproducible indicators.

While such affordances promise substantive gains in efficiency and analytical depth, the epistemic validity of AI-generated outputs must be scrutinised rigorously by researchers, policymakers, and stakeholders. Accordingly, this study evaluates both the benefits and limitations of Generative AI, with the broader aim of advancing inclusive, evidence-based environmental policymaking premised on robust public participation. In doing so, it explores how constrained and transparent use of LLMs may enhance both analytical rigor and civic accountability in environmental governance.

3. Methods

This study focuses on two primary data sources. The first is a compiled document of public comments submitted in response to a draft version of the Sixth Basic Environmental Plan (“案”; *An*). This document, provided as a PDF file, contains summaries of the public comments and the corresponding administrative replies. The file was converted from PDF to Excel using Adobe Acrobat and then preprocessed to produce a CSV file (hereafter referred to as the “Public Comment File”). In this file, each comment and its corresponding administrative reply are organized in separate columns, maintaining a one-to-one correspondence.

The second target is the revised draft (“答申案”； *Tōshin-An*), which was developed based on the public comments received. A PDF version of this draft explicitly highlights changes made in response to public comments[1]. This file was first converted from PDF to MS Word using Adobe Acrobat, then further converted to Filtered HTML format via Microsoft Office. From this HTML file, the textual changes were extracted and compiled into a CSV file (hereafter referred to as the “Change Log File”). This file contains a sequential change ID, page number, inserted text, and deleted text in separate columns. An example output is shown in Table 1.

(1) Matching Procedure using ChatGPT 4o API

Using the API for ChatGPT 4o, the following three items were provided:

1. a prompt (see below),
2. the Public Comment File,
3. the Change Log File.

The goal was to identify whether and where a given comment–reply pair led to a corresponding revision in the draft. The output was compiled into a CSV format.

```
---
Comment: {comment}

Reply: {reply}

This comment–reply pair has been rated a fulfillment score of {score}, indicating the likelihood that it led to an actual change in the policy document.

The fulfillment score indicates how closely the policy response aligns with the original public comment:

- Score 4: Fully fulfilled — the requested revision or addition was made as asked.
- Score 3: Partially fulfilled — the response addressed the comment in part or via alternative means.
- Score 2: No change made, but the response claims the request is already fulfilled.
- Score 1: Rejected — the request was not accepted or implemented.

In general, scores of 3 or 4 are more likely to correspond to actual changes made in the policy document. Therefore, when identifying relevant changes, give higher priority to these cases.
```

[1] In other words, the PDF highlights the changes made when the “案” was revised into the “答申案,” thereby serving as data that simultaneously reflects both drafts.

```
Justification for this score: {justification}

---

Below is a list of changes made to the policy document.
Each change consists of (1) added phrases, (2) deleted phrases, and (3) the full surrounding context (including the strings of both additions and deletions).
Note: Added or deleted phrases may include `///` as a delimiter, which indicates that multiple words were inserted or removed as part of a single change (e.g., "以下///」という。" = "以下", "」という。").

When identifying which changes are relevant to the comment–reply pair above, **do not rely on the context alone**.
Instead, make your judgment **based on the combination of both added and deleted phrases**.
The context can help clarify intent, but the primary evidence must come from the actual textual changes.

If any changes are clearly related to the comment and reply, list the corresponding change IDs and provide a brief justification.

If none of the changes are related, simply respond with:

Relevant changes: None

List:

{change_list_str}

---

Output format (strictly follow this format):

Relevant changes:
- Change ID: [number]
Reason: [Brief explanation of the relationship (max 50 characters)]

or

Relevant changes: None
```

(2) Scoring Meaningful Reply using ChatGPT 4o

Using the same API, the Public Comment File was submitted again to ChatGPT 4o. This time, the goal was to assess the Meaningful Reply for each comment–reply pair. The evaluation was performed using a standard common system prompt and a criteria-specific prompt referring to Ishii & Kameda (2025).

Common system prompt:

```
"""
You are an expert evaluator assessing public comments and administrative responses regarding the Basic Environmental Plan in Japan.
Follow the instructions, output format, and evaluation criteria below for your work.

----------------------
INSTRUCTIONS
----------------------
1. Read each comment and reply carefully.
2. Assign a score for {criteria}.
3. Provide a brief justification (one or two sentences) after the score.
```

```
4. Output your evaluation in the specified line-based format.
5. Please remain objective and evaluate strictly based on the given criteria.

----------------------
OUTPUT FORMAT
----------------------
For each comment–reply pair, output:

ID: {row_id}
{criteria} Score: [score]
{criteria} Justification: [justification]

----------------------
EVALUATION CRITERIA
----------------------
"""
```

Meaningful reply system prompt:

```
"""
[Meaningful Reply Evaluation]

  Evaluate whether the reply is “meaningful,” i.e., whether it demonstrates relevance, completeness, and clarity
     in addressing the comment.

  *Scoring:*
  - 5: The reply is fully relevant, addresses the comment completely, and is clearly explained.
  - 4: The reply addresses the main points of the comment.
  - 3: The reply addresses part of the comment but lacks some clarity or completeness.
  - 2: The reply scarcely addresses the comment (insufficient or off-topic).
  - 1: The reply does not address the comment at all.
"""
```

(3) Comparative Analysis

Based on the results of step (1), each comment–reply pair was assigned to one of two categories:

1. Contribution to revision (encompassing both grammatical and semantic amendments)
2. No contribution to revision

We then compared the Meaningful Reply scores between these two categories. Specifically, we applied the Mann–Whitney U test for the pairwise comparison and visualized score distributions with bar plots and box plots.

4. Results

▪4.1. Scoring of Meaningful Reply

As described in Step (1) of the analysis procedure, the relevance between public comments and the draft of the Basic Environmental Plan was assessed using ChatGPT. When a comment was judged to correspond with a specific section of the draft, an evaluation was generated as shown in the following example (Comment No. 89). If no relevant section was found, the output simply indicated "Not applicable" for the corresponding section.

Comment:

(Original Japanese) P9 L19〜26
クマ類について取り上げられているが、これより深刻化している鳥獣被害はシカであり、この種に

関し、取り上げるべきと考える。つまり、国土強靭化基本計画で取り上げられているよう、森林生態系におけるニホンジカ等による下層植生の衰退や裸地化に伴う土砂災害の発生の危険性について政府において認識されているところであり、ｐ5 の 18 行で記載するところの基盤としての自然資本は重要であり、自然資本を維持・回復・充実は不可欠である。このため、クマ類よりも深刻化しているニホンジカ等による下層植生の衰退や裸地化に伴う土砂災害の発生の危険性を取り上げて記載すべきである。

(English translation)
While the draft mentions bears, a more pressing issue is the damage caused by deer. The Basic Plan for National Resilience already recognizes the risks of landslides due to the decline of understory vegetation and bare land caused by the overpopulation of Japanese deer. Since such ecosystem degradation undermines the natural capital mentioned on p.5, line 18, it is vital that the draft include reference to this more serious issue, rather than focusing solely on bears.

Reply:

(Original Japanese) ご指摘を踏まえ、P6 L21 以降に、以下の文章を追記いたします。
「また、ニホンジカの生息域の拡大や生息数の増加により、下層植生の衰退や裸地化等の森林生態系等への被害が深刻化しており、防災・減災等、森林の多面的機能が十分発揮されないことも懸念されている。」

(English translation)
Based on your feedback, we will add the following sentence after P6, L21:
“Moreover, the expansion of the Japanese deer’s habitat and increase in population are causing serious damage to forest ecosystems, such as the decline of understory vegetation and land degradation. This is raising concerns about the diminished capacity of forests to serve their multifaceted functions, including disaster prevention and mitigation.”

Relevant changes:
- Change ID: 11
  Reason: Addition addresses deer impact on ecosystems

In total, 84 out of 382 public comments were determined to correspond to a specific section of the draft, while the remaining 298 were not. This classification closely aligned with the “Request Fulfillment” categories proposed in Ishii & Kameda (2025): comments rated at fulfillment level 3 or 4 were generally judged as corresponding, whereas those rated at level 1 or 2 were considered “Not applicable.” However, some discrepancies were observed between the two classifications. This divergence can be attributed to a key methodological difference—whereas Ishii & Kameda (2025) focused exclusively on evaluating comment–reply pairs, the present study additionally took into account whether revisions were actually made to the draft. In several instances, even when the administrative response alone did not appear to reflect the comment, subsequent changes in the draft document indicated a substantive impact. As such, our approach may reveal contributions that are not explicitly acknowledged in official replies.

To further evaluate the quality of administrative responses, Step (2) of the analysis assessed the Meaningful Reply—a score that quantifies the degree to which each reply substantively addresses the corresponding public comment. Across all responses, the average score was 3.33, with a standard deviation of 0.94, suggesting a generally moderate level of responsiveness. In total, 3 responses were rated 1, 77 were rated 2, 134 were rated 3, 128 were rated 4, and 40 were rated 5. This distribution demonstrates a central tendency toward moderately meaningful responses, with fewer instances at either extreme.

The framework for evaluating Meaningful Reply using ChatGPT was initially introduced in Ishii & Kameda (2024). In that early study, a 20-point scale was employed, which ultimately proved unsuitable for direct comparison with human evaluations. To improve interpretability and alignment with human judgment, the authors subsequently adopted a five-point scale. Both Ishii & Kameda (2024) and (2025) employed Cohen’s kappa coefficient to assess inter-rater agreement. In the follow-up study, Ishii & Kameda (2025), three human evaluators from different disciplinary backgrounds assessed the same dataset. The resulting kappa coefficients were: ChatGPT vs. Evaluator 1, 0.4204; ChatGPT vs. Evaluator 2, 0.6374; ChatGPT vs. Evaluator 3, 0.3859; Evaluator 1 vs. Evaluator 2, 0.4237; Evaluator 1 vs. Evaluator 3, 0.1022; and Evaluator 2 vs. Evaluator 3, 0.3745. These values suggest substantial variability among human raters, which naturally influences agreement with ChatGPT as well. Nonetheless, Meaningful Reply was found to be the most stable of the constructs evaluated in the study.

▪4.2. Relationship Between Comment Contribution and Meaningful Reply Score

As previously noted, this study aims to examine how "contribution" in public comments influences the quality of administrative responses, specifically in terms of their Meaningful Reply. To this end, we visualized the distribution of Meaningful Reply scores and conducted statistical tests to assess differences between groups, with the goal of understanding the conditions under which substantive responses are more likely to occur, rather than merely formal ones.

A comparison between the two groups—those with contributions (Contribution = Yes) and those without (Contribution = No)—revealed a statistically significant difference in their Meaningful Reply scores (Mann-Whitney U = 6334.0, $p < .001$). The average score for responses linked to non-contributive comments was 3.13, while the average for those linked to contributive comments was substantially higher at 4.01.

As illustrated in the bar charts (Figures 1 and 2), responses associated with contributive comments (Figure 1) showed a strong concentration of high scores (4 and 5), indicating generally more favorable evaluations. In contrast, responses linked to non-contributive comments (Figure 2) clustered in the mid-range (scores 2–4), with several instances of low-scoring responses (1 and 2).

The boxplot (Figure 3) further supports this finding, showing a higher median and an overall upward shift in the score distribution for the contributive group.

These results suggest that public comments may have a tangible impact on the quality of administrative responses. The data indicate that the public comment process, at least in some cases, goes beyond formal acknowledgment and reflects a substantive incorporation of public input.

Table 1: Mann-Whitney U Test

Indicator	Output
U statistics:	6334.0
P-value:	p < 0.001
Contribution = No (n=298), Mean:	3.13
Contribution = Yes (n=84), Mean:	4.01

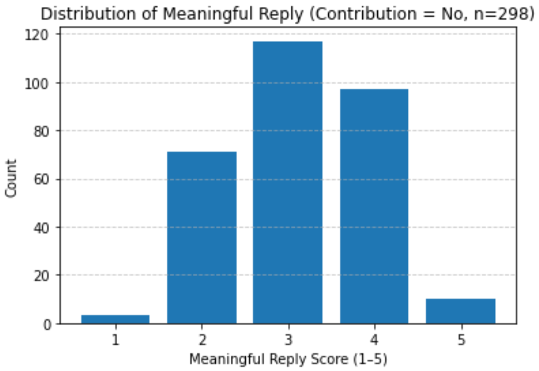

Figure 1: Distribution of Meaningful Reply (Contribution = No, n=298)

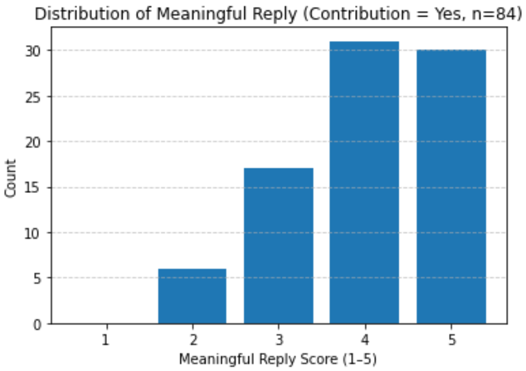

Figure 2: Distribution of Meaningful Reply (Contribution = Yes, n=84)

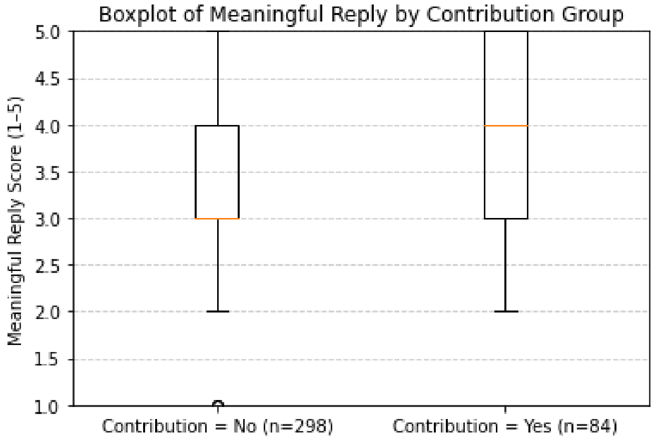

Figure 3: Boxplot of Meaningful Reply by Contribution Group

5. Conclusion

This study set out to clarify the extent to which Japan's public-comment mechanism shapes the drafting of the Sixth Basic Environmental Plan and to test whether large language models (LLMs) can provide a transparent, reproducible indicator—Meaningful Reply–of administrative responsiveness. By linking three data sets—the public-comment file, the change-log file, and the government's replies—and analysing them through ChatGPT 4o, we were able to trace, at scale, how individual comments translated into concrete textual revisions. The analysis revealed that 22 % of the 382 comment–reply pairs triggered identifiable modifications to the draft; moreover, these "contributive" cases received significantly higher Meaningful Reply scores than non-contributive cases (Mann-Whitney U = 6334.0, p < .001). The finding indicates that citizen input is not merely acknowledged but, in a substantial subset of cases, integrated into policy texts, thereby enhancing the deliberative quality of the process.

Beyond documenting this impact, the study demonstrates that LLM-based matching closes a critical gap between official records and the actual evolution of policy language. Several changes detected in the draft were not explicitly referenced in the corresponding replies, suggesting that conventional summary documents risk under-representing public influence. By operationalising Harashina's normative concept of a "meaningful reply" as a numerical score, the research also offers regulators and stakeholders a scalable tool to evaluate whether participation is genuinely substantive—a contribution that resonates with the International Association for Impact Assessment's call for meaningful public participation within environmental impact assessment (EIA) practice.

Nevertheless, two limitations must be acknowledged. First, the investigation focused on a single policy cycle; future research should test the workflow across multiple sectors and iterative plan versions to establish generalisability. Second, as noted by Ishii and Kameda (2025), aligning machine-generated scores with nuanced human judgments remains difficult. However, it is worth noting that even in cases where human evaluators and AI-generated scores diverged, the model occasionally captured substantive contributions—particularly those reflected in actual policy changes rather than in written responses. This suggests that the model may detect latent forms of stakeholder influence that are not explicitly acknowledged by evaluators. Bridging this evaluative gap will require systematic calibration against expert panels and detailed error analysis. Despite these constraints, the present study advances a transparent, reproducible methodology for assessing stakeholder influence and thus contributes to the broader project of consensus-oriented environmental governance.

References

Androutsopoulou, A., Karacapilidis, N., Loukis, E., & Charalabidis, Y. (2019). Transforming the communication between citizens and government through AI-guided chatbots. *Government Information Quarterly,* 36(2): 358–367.

Arnstein, S. R. (1969). A Ladder of Citizen Participation. *Journal of the American Institute of Planners,* 35(4); 216-224.

Bail, C. A. (2024). Can Generative AI improve social science? *Proceedings of the National Academy of Sciences (PNAS)*, 121(21): e2314021121.

Beierle, T. & Cayford, J. (2002). *Democracy in Practice: Public Participation in Environmental Decisions.* Resources for the Future Press.

Birhane, A., Isaac, W., Prabhakaran, V., Díaz, M., Elish, M. C., Gabriel, I., & Mohamed, S. (2022). Power to the People? Opportunities and Challenges for Participatory AI. *EAAMO '22: Proceedings of the 2nd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization,* 6: 1-8.

Brown, L. & Thérivel, R. (2000). Principles to Guide the Development of Strategic Environmental Assessment Methodology. *Impact Assessment & Project Appraisal,* 18(3): 183-189.

Deroy, A., Ghosh, K., & Ghosh, S. (2023). How Ready are Pre-trained Abstractive Models and LLMs for Legal Case Judgement Summarization? *arXiv preprint,* arXiv:2306.01248v2.

Fatima, S. T., Desouza, K. C., Buck, C. & Fielt, E. (2022). Public AI canvas for AI-enabled public value: A Design Science Approach. *Government Information Quarterly,* 39(1): 101659.

Gilardi, F., Alizadeh, M., & Kubli, M. (2023). ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences (PNAS),* 120(32): e2304524120.

Harashina, S. (2001). A New Stage of EIA in Japan: Towards Strategic Environmental Assessment. *Built Environment,* 27(1): 8-15.

Harashina, S. (2005)., ed. *Public Participation and Consensus Building.* Gakugei Shuppansha. [In Japanese]

Hartmann, J., Schwenzow, J., & Witte, M. (2023). The political ideology of conversational AI: Converging evidence on ChatGPT's pro-environmental, left-libertarian orientation. *arXiv preprint,* arXiv:2301.01768v1.

Ishii, K. & Kameda, A. (2024). The Contributions of Public Comments in the Drafting Process of the Sixth Basic Environmental Plan. *Proceedings of the Annual Conference of JSAI,* 39 (forthcoming). [In Japanese]

Ishii, K. & Kameda, A. (2025). Examination of an Automated Evaluation Method of Public Comments Using Generative AI. *IPSJ Symposium Series: Jinmoncon 2024*, 2024: 335-342. [In Japanese]

Jungherr, A. (2023). Artificial Intelligence and Democracy: A Conceptual Framework. *Social Media + Society,* 9(3): 1-14.

Reed, S. M. (2008). Stakeholder Participation for Environmental Management: A Literature Review. *Biological Conservation,* 141(10): 2417-2431.

Rozado, D. (2023). The Political Biases of ChatGPT. *Social Sciences,* 12(3): 148.

Rutinowski, J., Franke, S., Endendyk, J., Dormuth, I., Roidl, M., & Pauly, M. (2024). The Self-Perception and Political Biases of ChatGPT. *Human Behavior and Emerging Technologies,* 2024: Article ID 7115633.

Ministry of the Environment Government of Japan. (2024). The Basic Environmental Plan. https://www.env.go.jp/council/content/i_01/000281036.pdf.

Motoki, F., Pinho Neto, V., & Rodrigues, V. (2024). More human than human: measuring ChatGPT political bias. *Public Choice,* 198: 3-23.

Savaget, P., Chiarini, T., & Evans, S. (2019). Empowering political participation through artificial intelligence. *Science and Public Policy,* 46(3): 369-380.

Susskind, L. E. & Cruikshank, J. L. (2006). *Breaking Roberts Rules: the New Way to Run Your Meeting, Build Consensus, and Get Results.* Oxford University Press.

Törnberg, P. (2023). ChatGPT-4 Outperforms Experts and Crowd Workers in Annotating Political Twitter Messages with Zero-Shot Learning. *arXiv preprint,* arXiv:2304.06588v1.

Yamada, K. & Yagishita, M. (2011). The Development of Public Involvement in Decision Making Process of Climate Change Policy of Japan. *Environmental Science,* 24(5): 422-439. [In Japanese]