LLM Capstone Project

Benjamin Saulnier [300121338]
Raoul Junior Lorfils [300332572]

Prof. Patrick McCurdy, PhD
CMN 5100
December 15th, 2025

Table of Contents

Introduction

As researchers in the field of communications, we believe that accessing unbiased, truthful, and overall reliable information is essential for a functioning society. With the introduction of the internet, our libraries of information grew exponentially, however, not without limitations. Questions arose about who controls the technology that we use to access our knowledge, and whether we can trust these companies to mediate our searches for information. Now we face a similar situation, as we are becoming increasingly dependent on AI technologies to gain an understanding of unfamiliar topics. To challenge artificial intelligence (AI) to produce the truth, we decided to test it by asking: How do popular large language models (LLMs) frame the water consumption of artificial intelligence? AI data centers require incredible amounts of water to cool the computers running the servers. The water consumption of some of these data centers is seen as unsustainable and a critical environmental threat. They use additional water when it comes to generating power and processing the precious materials required for construction. We hope to discover more about what stances, framing, hedging, evidences practices, and absences three major LLMs deliver when prompted with neutral, positive, and negative emotive question prompts.

Literature & Framework

Research in communication science has long shown that social issues—not just environmental ones—are not communicated in a neutral way, but are shaped by framing: what is emphasized, what is downplayed, and what is omitted (Deacon et al., 2007). These framing choices also influence how responsibility is assigned, which actors are considered relevant, and the degree of seriousness of an issue. As people increasingly rely on digital tools to understand complex topics, these framing processes become particularly important.

Recent research on generative artificial intelligence adds a new dimension to these concerns. Gillespie's (2024) concept of “politics of visibility” provides a useful framework for examining how large language models (LLMs) structure explanations. For Gillespie, visibility is not simply about mentioning a topic, but it concerns the narratives that are readable by the system: what types of explanations are regularly produced, which perspectives are normalized, and which remain marginal or absent. Generative systems do not persuade users with explicit arguments; rather, they shape understanding through default values and repetitive patterns that make certain explanations appear natural.

AI environmental footprint

Applied to environmental communication, this framework raises important questions about how AI systems describe their own environmental impact. AI data centers rely on water-intensive cooling processes, and research has shown that large-scale water withdrawal can disrupt local water flows and reduce water availability for agriculture and consumption, while poorly treated wastewater can harm aquatic ecosystems (George, George, & Martin, 2023). By situating AI-related water consumption within the framework of environmental communication studies, we can examine not only whether environmental impacts are acknowledged, but also how they are presented, contextualized, or minimized in AI-generated explanations.

In this project, we treat LLM outputs as discursive artifacts rather than neutral sources of information. Drawing on framing theory, we focus on patterns of emphasis and omission in responses, including how responsibility for environmental impacts is attributed and how harms are contextualized. Gillespie's work is particularly useful for considering absences as analytically significant, especially when certain dimensions of an issue, such as affected communities or structural constraints, appear less frequently or only under specific conditions.

Methodologically, the project relies on qualitative content analysis and discourse analysis, approaches that emphasize systematic coding while recognizing that interpretation is inevitable (Deacon et al., 2007; Wodak & Meyer, 2009). Coding was used not to claim objectivity, but to make analytical decisions visible, consistent, and open to discussion. Our coding strategy followed a combined deductive and inductive approach (Elo & Kyngäs, 2008; Saldaña, 2016). Deductive codes were developed based on expectations drawn from the literature, including the possibility that LLMs might present AI-related water consumption in a reassuring or solution-oriented manner. Inductive codes were added after reviewing initial responses and identifying recurring patterns relevant to our research question.

Our analytical approach can be described as abductive, involving an iterative movement between theory and empirical observations. Drawing inspiration from Van der Ven et al., abductive reasoning allows researchers to start from an initial theoretical hypothesis and refine or complicate it based on the data. At the outset of this project, we expected LLMs to deflect responsibility for environmental harm away from AI systems themselves,

potentially emphasizing external factors or future solutions, given the commercial and institutional contexts in which these tools are developed. We then examined how the framing patterns we observed confirmed, challenged, or nuanced this expectation.

To operationalize these models, we primarily relied on binary coding (0/1) to indicate the presence or absence of specific characteristics in each response (see narrative codebook). The initial yes/no questions were developed deductively from expected themes drawn from the literature, while additional true/false questions were created inductively after reviewing the initial responses. This binary structure promoted clarity, comparability between models, and greater reliability among coders.

This iterative movement between theory and data also reflects Mills' idea of "intellectual craftsmanship." Mills encourages researchers to write as they think, take notes on emerging ideas, and shift between analytical roles throughout the research process. In our project, note-taking helped us document emerging interpretations, clarify uncertainties, and justify changes to the code book, thereby promoting transparency in analytical decision-making.

Intercoder reliability was treated as a process rather than a simple numerical benchmark. In line with qualitative research guidelines, differences in our perspectives were seen as useful for refining the coding framework rather than as a flaw (Joffe & Yardley, 2003; Campbell et al., 2013). Series 1 to 24 were coded by Ben Saulnier and served as the reference dataset for these series. Differences were discussed to clarify code definitions and ensure consistent application in subsequent steps. This approach promotes reliability through shared understanding and clarity rather than strict statistical agreement.

For the purposes of coding evidence practices (Q11), we defined a scholarly (academic) source as one "produced by an expert in a given field", and whose "work represents recognized knowledge and establishes authority on a subject". This definition follows the guidelines of the University of Ottawa Library, which distinguishes scholarly sources from popular or user-generated content based on author, expertise, and credibility.

Methods

Prompts

We designed three prompts with unique sentiments attached to each. The first one was our baseline (neutral) prompt: Explain how LLMs consume water in their operations. This prompt goal was to ask the LLM for information on our topic without influencing it in any way. The second was the positive prompt: Explain how LLMs help conserve water in their operations. This prompt uses positive words like “conserve” and “help” to influence the LLM to return a positive response to our question. The third prompt uses negative language: Explain how the water consumption of LLMs is wasteful. The word “wasteful” is meant to guide the LLM to explore the negative impacts of AI and water consumption.

During replication, our prompts did not vary. Each prompt was copied and pasted with the same capitalization and no punctuation to ensure consistency among researchers. Even the smallest changes in the prompts may lead to new responses by the LLM,s which could be an interesting direction in future reproductions.

Models

To select our models, we first examined the AI Chabots with the highest market shares, as this is a good indicator of which models are the most popular among users. The top four chatbots, according to StatCounter (2025), in November were ChatGPT (81.85%), Perplexity (11.05%), Microsoft Copilot (3.07%), and Google Gemini (2.97%). We then examined each provider and the technology on which they were built. We found that ChatGPT and Microsoft Copilot were both built on versions of GPT, while Perplexity and Google Gemini were built on unique AI models. This led us to select ChatGPT, Perplexity, and Google Gemini as our three models to include in our research. Due to time and scale constraints, we were limited to three models, which can be considered a limitation. Future reproductions should consider exploring more models, as each database would contain unique knowledge, and different algorithms may produce new, unexpected responses.

This section of our research would be the most challenging to replicate, as AI is developing exponentially (Zorpette, 2025). If this project is replicated in a week, month, or year from now, it would yield new results because of the constant updates and improvements in the technology. Besides the technological changes, the models may respond differently to a series of other differences, which include, but are not limited to, IP address, device used,

language used, and the settings applied on the LLM (incognito mode, search vs. learn vs. fast modes, etc).

Replicates

We collected data over three sessions (three consecutive days), completing seventy-two replicates in total. Each model was fed the neutral, positive, and negative prompts eight times each. With each replicate, we collected screenshots, raw text data, and a timestamp. While the replicates were collected during three individual sessions, the sessions varied in length due to variables like internet speed, the device used, and complications with save files, which resulted in delays at times. Reproduction of this project should be reminded that the data collection process can take several hours, and it may be beneficial to have more than one person collecting during a session, or spreading the replicates over more sessions.

Protocol

We received the template Excel sheet and filled out the setup, models, and prompts tabs with the information required. In the run plan tab, we first sorted the "Random_Key" column to randomize the order of the rows. After one sort, there were still some long runs of the same model, so we re-shuffled once, then froze the numbers by copying and pasting the column. At this stage, a mistake was made as we began filling out the remainder of the run plan without sorting the "Random_Key" tab once more. The mistake was not noticed until after the first data collection period was completed. This means the first session data (runs 1-24) is not in order on the run plan, but sorting the sheet by the "Planned_Run_Order" column reveals the order in which the data was collected. Sessions two and three went without issue, and all of the data is in chronological order for those rows (runs 25-72).

To collect the data, we followed the run order given by the run plan tab of the Excel sheet. We ran each run through a fresh chat, ensuring a new tab was opened each time and all settings to properly selected. After each response was complete, we recorded the model, version, date, and time in a Word document, along with screenshots of the entire dialogue, a copy of the raw text, and a copied list of any sources provided. If no sources were provided, we included "N/A" in the sources section.

Coding

As discussed in the *Literature & Framework* section, we used abductive coding to pull meaning from our data. We began writing our deductive codes with the theory that the

LLMs would attempt to frame issues surrounding water consumption due to AI data centers in a positive way, or refuse to answer questions relating to this topic altogether. Once we completed data collection, we created inductive codes by reviewing the R1 replicates and pulling statements that we considered significant to our research question. Together, we created a narrative codebook that listed each question along with definitions, rules, and examples to ensure we were following the same rules once we reached the coding stage. In this stage, we also categorized questions into stance, framing, hedging, evidence practices, and visibility/absence. We had two independent coders, and double-coded the first dataset (33.3% of the data) to ensure intercoder reliability. Our intercoder reliability percentage was calculated using the ReCal2 tool designed by Deen Freelon (2023). The table below shows the intercoder reliability scores for each question:

Table 1 - *Intercoder reliability*

Questions	Percent Agreement	Scott's Pi	Cohen's Kappa	Krippendorff's Alpha	N Agreements	N Disagreements	N Cases	N Decisions
Q1	62.5	0.180266	0.19403	0.197343	15	9	24	48
Q2	25	-0.51049	-0.2067	-0.47902	6	18	24	48
Q3	87.5	0.738657	0.73913	0.744102	21	3	24	48
Q4	25	-0.6	0	-0.56667	6	18	24	48
Q5	70.83333	0.3902	0.416667	0.402904	17	7	24	48
Q6	70.83333	0.174447	0.222222	0.191646	17	7	24	48
Q7	100	1	1	1	24	0	24	48
Q8	79.16667	-0.11628	0	-0.09302	19	5	24	48
Q9	54.16667	-0.2973	0	-0.27027	13	11	24	48
Q10	95.83333	0.915344	0.915493	0.917108	23	1	24	48
Q11	66.66667	0.193277	0.25	0.210084	16	8	24	48

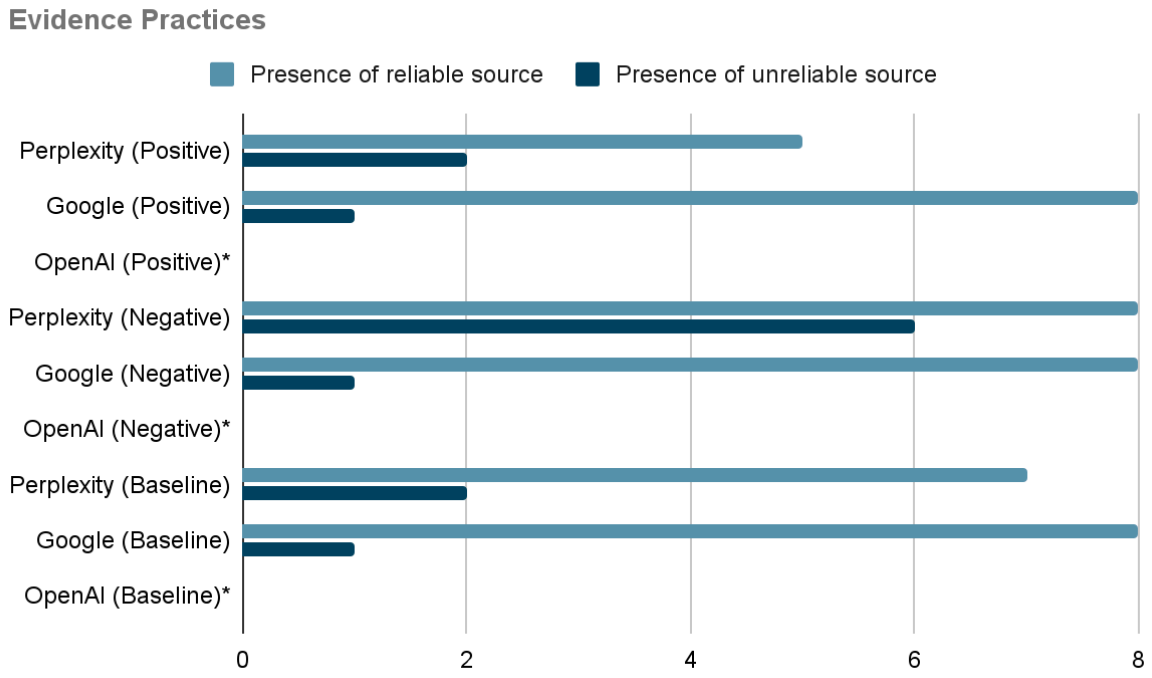
Once the intercoder reliability score was established, we discussed our differences and then proceeded to code the remaining data. We both coded one remaining dataset each, which consisted of 24 replicates.

While we tried to limit biases, it is natural for some biases to appear during manual coding and in both the deductive and inductive stages of our coding design. As two students enrolled in the Master's of Communication program at the University of Ottawa, we acknowledge that our diverse backgrounds are a limitation of this research project in an effort to be transparent. Future reproductions of this research should be aware of any biases they may have before beginning.

Results

Through analysis of our results, we can begin to see patterns emerge from each model, depending on the prompt they were presented with. First, we can look at the evidence practices that offer an idea of how these LLMs collect data. Between the two coders, the intercoder reliability score is considered high on questions in this category. Below is a figure showing the breakdown of the presence of reliable sources and the presence of unreliable sources.

Figure 1 - *Evidence Practices*

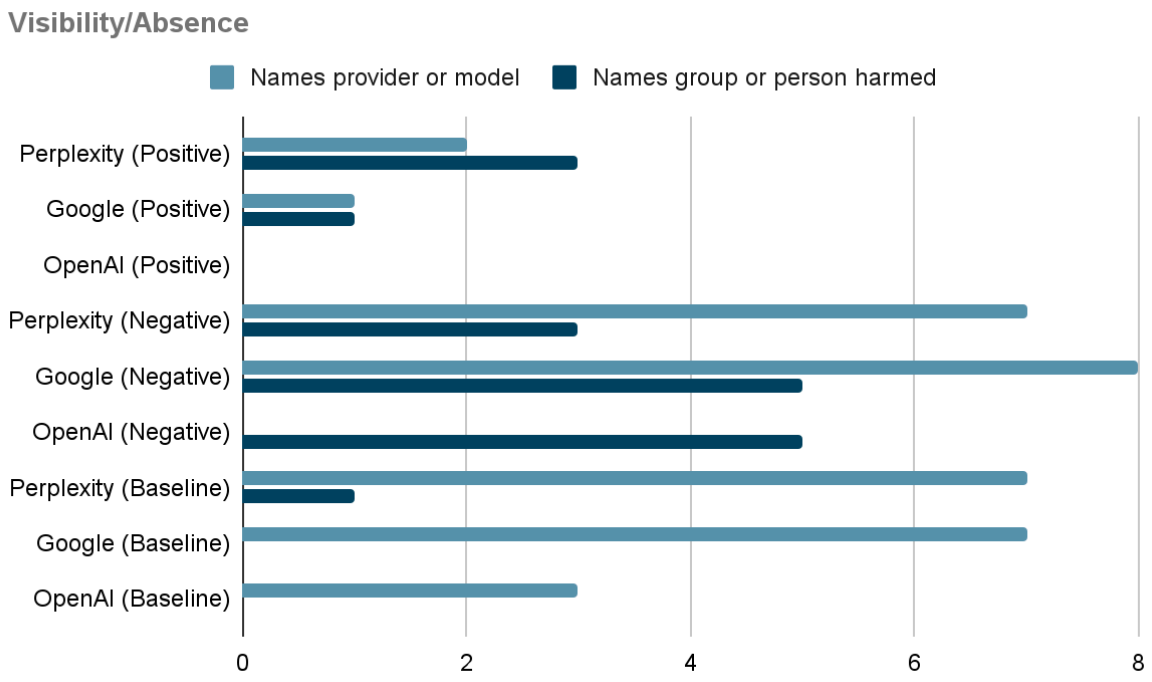


*OpenAI did not offer any sources during the runs

In this table, we can see that Google's Gemini LLM offered at least two academic sources in all 24 runs. It also used unreliable sources less often than Perplexity's Sonar model. Before conducting our research, we hypothesized that Perplexity would offer the best insights to our prompts since it was the only model that was doing real-time web searches, while the other two were accessing a database that could be slightly dated. Perplexity's purpose is to offer recent information and scrape the web. In doing so, it seems Perplexity's model is more likely to be fooled by top search results, as it included Reddit threads in its sources on several runs. Perplexity on four occasions also simply timed out, offering no sources to substantiate its claims. Using a negative prompt for Perplexity also generated reliable sources every time, but increased the likelihood of returning an unreliable source. Google Gemini's unreliable sources were mainly blog posts, which did not meet the standards set in our narrative codebook. It offered an unreliable source exactly once for every prompt, meaning the prompt did not seem to affect the quality of sources. The LLM also had an impressive streak of reliable sources, as it was able to provide direct, working links to several credible sources with each prompt. OpenAI offered no insights as to where its information was coming from. This can be concerning since, like the other two models, it presented the information with confidence using no uncertain language, yet lacked evidence to back up the claims it made.

Next, we can present the visibility and absence questions, which also scored high on intercoder reliability. The questions in this category help see how LLMs present both the actors causing water shortages in the AI field and the victims. To visualize the data, we created the table shown below.

Figure 2 - *Visibility/Absence*



This chart demonstrates that naming providers or models of LLMs was far more common than naming a group or person that has been harmed as a result. Further observations will show that the prompts made a significant difference in the visibility and absence results. The baseline prompt had almost exclusively named providers or models, besides one mention of a group in a Perplexity run. The negative prompt yielded higher mentions of both models or providers, and groups or people besides the exception of OpenAI who had no mentions of providers or models. The positive prompt had very few mentions for either question, with OpenAI not including any names in the runs. By naming more victims of water shortages when prompted with a negative prompt, less mentions when responses to positive prompts, and focusing on providers in baseline prompts, we begin to see how the LLMs are framing information and may not be as unbiased as people believe.

The table below is a small portion of our larger coding chart. It highlights how all three models suggest AI tools are a solution to water consumption problems when given a positive language prompt. In the remaining runs not included in this table (fourty-eight runs), we only coded two other presences of this stance.

Table 2 - *Suggests AI tools are a solution to water consumption problems, positive prompt responses*

Planned_Run_Order	V1	V2	V3	Q3
5	Google	12/4/2025	Positive	1
12	Google	12/4/2025	Positive	1
13	Google	12/4/2025	Positive	1
17	Google	12/4/2025	Positive	1
29	Google	12/5/2025	Positive	1
43	Google	12/5/2025	Positive	1
54	Google	12/7/2025	Positive	1
65	Google	12/7/2025	Positive	1
15	OpenAI	12/4/2025	Positive	1
16	OpenAI	12/4/2025	Positive	1
18	OpenAI	12/4/2025	Positive	1
44	OpenAI	12/5/2025	Positive	1
46	OpenAI	12/5/2025	Positive	1
47	OpenAI	12/5/2025	Positive	1
50	OpenAI	12/6/2025	Positive	1
70	OpenAI	12/7/2025	Positive	1
6	Perplexity	12/4/2025	Positive	0
10	Perplexity	12/4/2025	Positive	1
28	Perplexity	12/5/2025	Positive	1
30	Perplexity	12/5/2025	Positive	1
34	Perplexity	12/5/2025	Positive	1
52	Perplexity	12/7/2025	Positive	1
57	Perplexity	12/7/2025	Positive	0

72	Perplexity	12/7/2025	Positive	1

In most of our questions, the results were not very consistent across models however this one question highlighted an important similarity. With a 91.67% occurrence rate over positive prompts, and a 4.17% occurrence rate between the other two prompts, we can see how the stance taken by all three models remains inconsistent. This is an important discovery because it raises concerns about LLMs and how they are programmed to appeal to the user. If these LLMs are able to change their stance on this issue based of minor changes in a prompt, it could have wider implications for their credibility as a source of information.

The models varied very little in their responses to each prompt. We expected OpenAI and Google to offer limited variation since they are both limited by their databases that are frozen between updates. We expected more variation by Perplexity since it does live searches on the internet, however over three days the responses had a negligible amount of change. Below is a table showing the coding done on Google's responses to positive prompts.

Table 3 - *Google, positive prompt coding chart*

Planned_Run_Order	V1	V2	V3	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11
17	Google	12/4/2025	Positive	1	1	1	1	0	0	1	0	1	0	0
5	Google	12/4/2025	Positive	1	0	1	1	0	0	1	0	1	0	0
12	Google	12/4/2025	Positive	1	1	1	1	1	0	1	0	1	0	0
13	Google	12/4/2025	Positive	1	1	1	1	1	0	1	0	1	0	0
29	Google	12/5/2025	Positive	1	1	1	1	1	0	1	1	1	0	0
43	Google	12/5/2025	Positive	1	1	1	1	1	0	1	0	1	0	0
54	Google	12/7/2025	Positive	0	1	1	1	1	1	1	0	0	1	0
65	Google	12/7/2025	Positive	0	1	1	0	1	1	1	0	0	0	1

This table represents how the models behaved very similarly no matter the day or time. Even variables related to the researcher such as the device, or our intercoding reliability had little

effect on the results of each model's prompt-specific coding chart. These results suggest that bothmodels that retrieve information from a database, and those that do live internet searches, still depend a lot on the prompt to formulate unique responses. We believe that we would have seen more variation in our results if we conducted this research over a longer period of time, allowing the LLMs databases to update. This was however a good snapshot of what the specific version of each model offered, and we could compare it to updates versions at later times.

Discussion & Conclusion

Our results show that large language models differ in how they approach AI water consumption, particularly in how they substantiate claims and make accountability visible. These differences appear to be more related to model design than to variation in prompts alone. Evidence practices provide a useful starting point. Google's Gemini model consistently provided scientific sources and relied relatively little on sources classified as unreliable. Perplexity's Sonar model showed greater variation in source quality, while OpenAI's model did not provide sources during runs, despite presenting information with a confident tone.

These patterns are important because they determine how credibility is constructed in AI-generated explanations. Models that provide sources allow users to verify claims and assess reliability, while confident responses without sources may encourage trust without offering means of evaluation. The authority of LLM outputs is thus shaped not only by content, but also by design choices related to transparency.

Trends in visibility and absence further illustrate how LLMs frame environmental responsibility. Across prompts and models, suppliers and AI systems were cited more often than groups or individuals affected by water shortages. The sentiment of the prompts played an obvious role: baseline prompts emphasized suppliers, negative prompts increased mentions of harmed groups, and positive prompts resulted in fewer mentions overall. This suggests that narratives of harm are less likely to be highlighted unless users frame their questions in explicitly critical terms.

This finding aligns with Gillespie's argument that visibility is closely linked to legibility. While LLMs are capable of discussing environmental impacts, they tend to frame harms in abstract terms rather than focusing on affected communities. Responsibility is

therefore more easily attributed to technologies or organizations than to social consequences, which influences how users may understand the issues surrounding AI's environmental footprint.

The results related to position reinforce this observation. When given positive prompts, all three models frequently suggested that AI tools help solve water consumption problems. This framing appeared much less often in the case of neutral or negative prompts. The ease with which this framing changed in response to minor changes in wording raises concerns about consistency and limits the usefulness of LLMs as stable sources of information on complex environmental issues.

The suggestions had limited influence on evidence practices. Gemini's sources remained consistent regardless of the suggestion's framing, while Perplexity's negative suggestions increased both reliable and unreliable sources. OpenAI's lack of source disclosure remained consistent across all suggestion types, which may partly reflect the fact that sources were not explicitly requested. These results suggest that while users can influence the framing, they have less control over how evidence is presented.

From a methodological perspective, reliability varied across coding categories. Observable characteristics resulted in greater agreement among coders, while more interpretive questions revealed some ambiguity. The use of Ben Saulnier's independent coding for series 1 through 24, followed by discussion and refinement, strengthened consistency and clarified the coding framework.

This project shows that LLMs do not merely convey information about AI's water consumption; they shape how the issue is framed, which actors are highlighted, and which explanations seem plausible. By adjusting their position based on the sentiment expressed and emphasizing technical narratives, these systems influence how environmental responsibility is understood.

Rather than treating LLM outputs as neutral sources, our findings support the idea of viewing them as communicative actors embedded in broader framing processes. As reliance on AI-generated explanations grows, examining these dynamics is essential to understanding how environmental knowledge is mediated and normalized in digital contexts.

References

Campbell, J. L., Quincy, C., Osserman, J., & Pedersen, O. K. (2013). Coding in-depth semistructured interviews: Problems of unitization and intercoder reliability and agreement. *Sociological Methods & Research, 42*(3), 294–320. https://doi.org/10.1177/0049124113500475

Deacon, D., Pickering, M., Golding, P., & Murdock, G. (2007). *Researching communications: A practical guide to methods in media and cultural analysis* (2nd ed.). Hodder Arnold.

Elo, S., & Kyngäs, H. (2008). The qualitative content analysis process. *Journal of Advanced Nursing, 62*(1), 107–115. https://doi.org/10.1111/j.1365-2648.2007.04569.x

Freelon, D. (2023). *ReCal2: Reliability for 2 coders (0.1 alpha)* [Computer software]. Annenberg School for Communication, University of Pennsylvania. https://dfreelon.org/utils/recalfront/recal2/

George, A. S., George, A. S. H., & Martin, A. S. G. (2023). Water footprint of AI systems (including ChatGPT). *Partners Universal International Innovation Journal (PUIIJ), 1*(2), 95–103. https://doi.org/10.5281/zenodo.7855594

Gillespie, T. (2024). Generative AI and the politics of visibility. *Big Data & Society, 11*(2). https://doi.org/10.1177/20539517241252131

Joffe, H., & Yardley, L. (2003). Content and thematic analysis. In D. F. Marks & L. Yardley (Eds.), *Research methods for clinical and health psychology* (pp. 56–68). SAGE Publications. https://doi.org/10.4135/9781849209793

Mills, C. W. (1959). *The sociological imagination*. Oxford University Press.

Saldaña, J. (2020). Qualitative data analysis strategies. In P. Leavy (Ed.), *The Oxford handbook of qualitative research*(2nd ed.). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780190847388.013.33

StatCounter GlobalStats. (2025, November). *AI chatbot market share worldwide*. https://gs.statcounter.com/ai-chatbot-market-share

University of Ottawa Library. (n.d.). *Choosing sources*. https://www.uottawa.ca/library/writing-citing/choosing-sources

van der Ven, H., Corry, D., Elnur, R., Provost, V. J., Syukron, M., & Tappauf, N. (2025). Does artificial intelligence bias perceptions of environmental challenges? *Environmental Research Letters, 20*(1), 014009. https://doi.org/10.1088/1748-9326/ad95a2

vanden Heuvel, K. (n.d.). *Will the AI boom lead to water and electricity shortages? The Nation*. https://www.thenation.com/article/politics/ai-zuckerberg-tech-regulation-data-security/

Wodak, R. (2004). Critical discourse analysis. In C. Seale, G. Gobo, J. F. Gubrium, & D. Silverman (Eds.), *Qualitative research practice* (pp. 186–201). SAGE Publications. https://doi.org/10.4135/9781848608191.d17

Appendix

Narrative Codebook

- **Research Question**
  How do popular large language models (LLMs) frame the water consumption of artificial intelligence?
- **Prompt Set**
  Baseline prompt: Explain how LLMs consume water in their operations.
  Positive prompt: Explain how LLMs help conserve water in their operations.
  Negative prompt: Explain how the water consumption of LLMs is wasteful.
- **Unit of Analysis and Coding Procedure**
  The unit of analysis is one complete LLM output (one run). Each output is coded at the document level. All categories are coded as binary variables (1 = present, 0 = absent). A code is marked as present only when the criterion is explicitly met in the text. Implied or ambiguous references are coded as absent. Example excerpts are reproduced verbatim from the sampled outputs.

---

1. **Stance**

1.1 Includes a reason besides AI for water waste

- **Definition**: The output identifies at least one cause of water waste or water stress that is not directly attributable to AI, LLMs, or data centres.
- **Coding rules:**
  - Code as "yes" if the text explicitly names another driver of water use or waste (e.g., electricity generation, agriculture, manufacturing, municipal consumption, drought conditions). AI-related water use may still be discussed, but at least one additional cause must be named.
  - Code as "no" if the output only discusses AI, data centres, cooling systems, or AI-related electricity demand without naming any other cause.
- **Example excerpt:**
  - *"Indirect water use refers to water required to generate the electricity that powers these data centers. Many power plants, especially thermoelectric and hydroelectric ones, also consume large amounts of water to produce electricity."*

1.2 Presents AI water use as an unavoidable technical process

- **Definition**: The output frames water consumption as a necessary or inevitable consequence of running LLMs, emphasizing technical requirements such as heat dissipation and system stability.
- **Coding rules**:
  - Code as "yes" if the text explicitly states or strongly implies that water use is required for AI systems to function (e.g., cooling is necessary to prevent overheating).
  - Code as "no" if the output emphasizes preventable waste or discretionary use without describing necessity.

- **Example excerpt**
  "The massive computational power required for LLM operations generates significant heat, which must be managed to keep servers functioning efficiently."

---

2. Framing

2.1 Suggests AI tools are a solution to water consumption problems

- **Definition**: The output frames AI or LLMs as contributing to solutions for water-related challenges, such as conservation, optimization, or improved management.
- **Coding rules**
  - Code as "yes" if AI is presented as enabling water savings or better water management in sectors such as agriculture, utilities, infrastructure, or policy.
  - Code as "no" if the output only describes AI's own water use or mitigation within data centres without framing AI as a broader solution.
- **Example excerpt**
  - *"LLMs are used in applications such as smart irrigation systems, water leak detection, and agricultural planning, where they analyze data to recommend water-saving measures."*

2.2 Includes theoretical scenarios to substantiate claims

- **Definition**: The output uses hypothetical, speculative, or future-oriented scenarios to support its arguments.
- **Coding rules**
  - Code as "yes" if the text includes forecasts, projections, or conditional statements (e.g., "could," "may," "in the coming years").
  - Code as "no" for purely descriptive statements of current conditions.
- **Example excerpt**
  "Estimates suggest that global AI demand could lead to the consumption of billions of cubic meters of water annually in the coming years."

2.3 Mentions social impacts

- **Definition**: The output references impacts on people, communities, livelihoods, or social systems.
- **Coding rules**
  - Code as "yes" if the output mentions residents, households, agriculture, municipal systems, or competition with human needs.
  - Code as "no" if the output discusses only environmental or ecological impacts without social reference.
- **Example excerpt**
  "This means that the water used for cooling is directly competing with residential needs, agriculture, and local ecosystems."

---

3. **Hedging / Modality**

3.1 Uses uncertain language in reasoning

- **Definition**: The output employs language that signals uncertainty, estimation, or variability.
- **Coding rules**
    - Code as "yes" if modal terms such as "may," "might," "could," "suggest," "estimates," or "depends" are used to qualify claims.
    - Code as "no" if the output relies exclusively on definitive statements.
- **Example excerpt**
  "Estimates from academic studies suggest a typical AI query may use a few to tens of milliliters of water."

---

4. Evidence Practices

4.1 Includes academic or verified sources for reasoning

Definition: The output cites or lists sources that are traceable and plausibly credible, such as academic papers, universities, government agencies, or established media outlets.

- **Coding rules**
    - Code as "yes" if named sources are provided.
    - Code as "no" if the output makes claims without citing sources.
- **Example excerpt**
  "Explained: Generative AI's environmental impact | MIT News."

4.2 Includes unreliable sources in reasoning

- **Definition**: The output cites sources that are informal or not typically considered reliable academic or policy evidence.
- **Coding rules**
    - Code as "yes" if sources such as Reddit or similar platforms are included.
    - Code as "no" if no sources or only reputable sources are listed.
- **Example excerpt**

  "AI Prompts Consume as Much Water as a 16-ounce Bottle Every 5 … – Reddit."

4.3 Contains sources originating from outside of North America

Definition: The output cites at least one source clearly based outside North America.

- **Coding rules**
    - Code as "yes" if a source's origin is explicitly non–North American (e.g., UK government, international organizations).
    - Do not infer geography unless it is clearly identifiable.
- **Example excerpt**
  "AI's thirst for water – UK Government Sustainable ICT."

5. **Visibility / Absence**

5.1 Explicitly names an LLM provider or model

Definition: The output explicitly identifies a specific AI provider or model.

- **Coding rules**
  - Code as "yes" if names such as OpenAI, Google, ChatGPT, GPT-3, GPT-4, or Gemini appear.
  - Code as "no" if only generic terms like "LLMs" or "AI" are used.
- **Example excerpt**
  "For example, training a single model like GPT-3 can consume over 700,000 liters of water."

5.2 Explicitly names a group or person harmed by AI's water consumption

- **Definition**: The output identifies specific groups affected by AI-related water use.
- **Coding rules**
  - Code as "yes" if the output names local populations, residents, households, agricultural users, or communities.
  - Code as "no" if harm is discussed only in abstract or environmental terms.
- **Example excerpt**
  "This puts the industry in competition with local populations, agriculture, and ecosystems."

5.3 Excludes any mention of AI or data centres as a reason

- **Definition**: The output discusses water consumption or waste without mentioning AI, LLMs, data centres, or related infrastructure.
- **Coding rules**
  - Code as "yes" only if AI and data centres are entirely absent as reasons.
  - This code is expected to be absent in most cases given the prompt design.
- **Example excerpt**

  N/A

5.4 Does not acknowledge the existence of a water consumption problem

- **Definition**: The output fails to recognize or denies that water consumption associated with LLMs is an issue.
- **Coding rules**
  - Code as "yes" only if no water consumption problem is acknowledged.
  - If any concern or acknowledgment appears, code as absent.
- **Example excerpt**

  N/A