13th World Congress on Water Resources and Environment (EWRA 2025)
"New challenges in understanding and managing water-related risks in a changing environment"
Palermo, Italy, 24 - 28 June 2025

Human-AI teaming for water quality documentation analysis with large language models

S. Conrad*, J. Rodriguez, G. Vizarreta Luna
Department of systems engineering, Colorado State University, 6029 Campus Delivery, Fort Collins, CO 80523
** e-mail: steve.conrad@colostate.edu*

Introduction

Water Quality management in urban systems is a complex endeavor reliant on extensive textual and quantitative documentation. Effective management involves testing and documenting the performance of water treatment, distribution, and environmental conditions, alongside regulatory and policy frameworks. Water managers must analyze an extensive array of water quality data, environmental regulations, test reports, and field documents to determine actionable requirements for managing water quality in urban settings. Quantitative data is routinely processed but documentation analysis is time-intensive and exacerbated by workforce limitations. Manual textual analysis by humans is also prone to notable inconsistencies and errors, especially when managing large datasets or working with multiple reviewers (González Canché, 2023). As a result, many textual data sets remain unexamined. Recent advances in large language models (LLMs) offer promising solutions to these challenges by creating a human-AI teaming environment for water quality documentation analysis.

Emerging research underscores the transformative potential of LLMs in qualitative data analysis. For instance, Tai et al. (2023) show applications of LLM with predefined definitions of content can replicate traditional document classification methods. Xiao et al. (2023) explored using GPT-3 for deductive coding in qualitative research. Similarly, Chew et al. (2023) evaluated GPT-3.5 within their LLM-Assisted Content Analysis framework, finding accuracy comparable to human coders. Our study advances this research by conducting a pre- and post-comparative analysis of using memory chunking approaches for coding water quality documentation, revealing how data localization impacts LLM performance. Our study comments on how LLM-assisted workflows have the potential to bring in additional content for water quality management.

Materials and methods

We used open and deductive coding approach to analyze water quality field and test observations. Laboratory tests included pH measurement, turbidity, biochemical oxygen demand, and contaminants such as heavy metals and nitrates. Field observations included flow rate measurements, visual notations on discoloration, and temperature readings. A team of researchers manually reviewed twenty documents discussing these aspects and were asked to code for water quality concerns, such as corrosion, pollutants, and or chemical imbalances. Coding was reviewed by collaborating water quality analysts. The same documents were provided to LLMs, and their performance was assessed using percent agreement, accuracy, precision, recall, and Fleiss' Kappa values.

OpenAI's gpt-4o, gpt-4o-mini, and o1-mini models were employed using a chunking text approach. Texts were pre-processed using optical character recognition to extract content, which was divided into 500-word chunks. We first asked each model to identify water quality concerns. We then provided the LLMs with detailed definitions of these concerns based on standard operating procedures and localized water quality content provided by collaborating water utilities. Results were classified as True if a minimum of one response indicated the presence of a concern. We used a consensus approach where we treated the LLM as an additional rater alongside manual coders to evaluate human-AI teaming. The performance of each approach was assessed and compared for how the process affected internal variance and false positives and negatives. These measures provide a comparison framework for assessing the efficacy of manual versus AI-assisted document analysis.

Results and concluding remarks

Figure 1 presents a comparison of the three LLM models—GPT-4o, GPT-4o-mini, and o1-mini with pre-localized and non-localized prompting strategies for lab data. Across all models, using localized chunked data improves overall accuracies over an open pre-localized approach. These findings suggest that including LLMs in the document analysis process can help with water quality management where large volumes of documentation and textual data must be analyzed. Using text chunking and localized data improves the accuracy of the LLMs. We also found that localized data improves agreement across multiple iterations of analysis and increased true positive rates. Using LLMs alongside human analysts can streamline the interpretive workload and give access to previously unexamined data. LLMs techniques can support water quality managers to identify water quality concerns and direct water quality efforts, ultimately contributing to more informed and efficient water quality management practices.

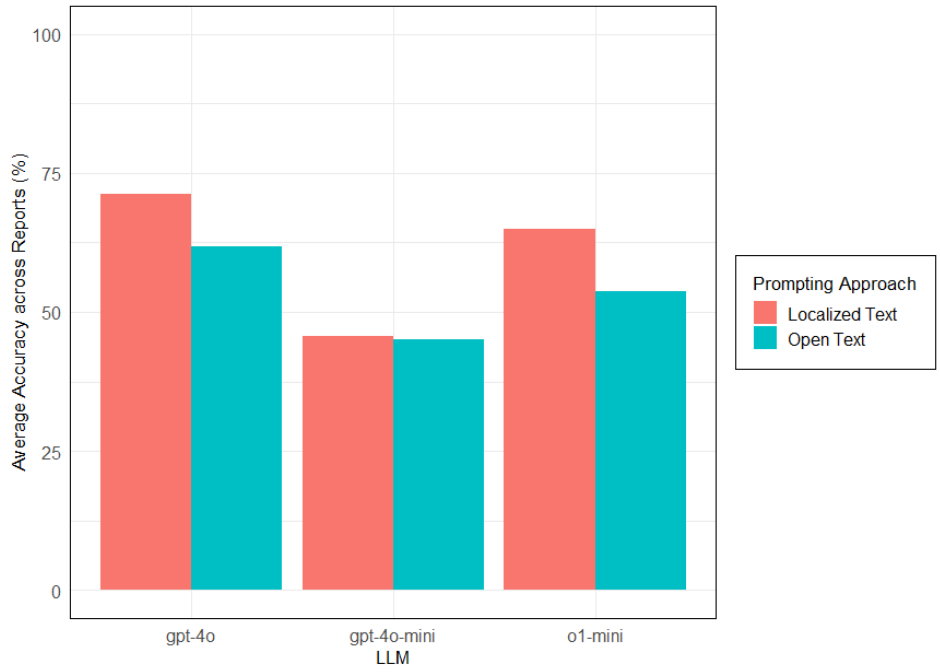

Figure 1. Comparison of Accuracy Across Models and Prompting Strategies

References

Chew, R., Bollenbacher, J., Wenger, M., Speer, J., & Kim, A. (2023). LLM-Assisted Content Analysis: Using Large Language Models to Support Deductive Coding (arXiv:2306.14924). arXiv. https://doi.org/10.48550/arXiv.2306.14924

González Canché, M.S. (2023). Latent Code Identification (LACOID): A Machine Learning-Based Integrative Framework [and Open-Source Software] to Classify Big Textual Data, Rebuild Contextualized/Unaltered Meanings, and Avoid Aggregation Bias. International Journal of Qualitative Methods

Tai, R.H., Bentley, L.R., Xia, X., Sitt, J. M., Fankhauser, S.C., Chicas-Mosier, A.M., & Monteith, B.G. (2024). An Examination of the Use of Large Language Models to Aid Analysis of Textual Data. International Journal of Qualitative Methods

Xiao, Z., Yuan, X., Liao, Q. V., Abdelghani, R., & Oudeyer, P.-Y. (2023). Supporting Qualitative Analysis with Large Language Models: Combining Codebook with GPT-3 for Deductive Coding. 28th International Conference on Intelligent User Interfaces