Improving Conservation Efficiency: Accelerating Groundwater Sustainability Plan Reviews Using Large Language Models

Ryan Bernstein , Seneth Waterman , Kirk R. Klausmeyer , Nicholas Murphy , Melissa M. Rohde , Cody Carroll

PII: S2667-0100(26)00169-1
DOI: https://doi.org/10.1016/j.envc.2026.101575
Reference: ENVC 101575

To appear in: *Environmental Challenges*

Received date: 10 November 2025
Revised date: 10 July 2026
Accepted date: 12 July 2026

Please cite this article as: Ryan Bernstein , Seneth Waterman , Kirk R. Klausmeyer , Nicholas Murphy , Melissa M. Rohde , Cody Carroll , Improving Conservation Efficiency: Accelerating Groundwater Sustainability Plan Reviews Using Large Language Models, *Environmental Challenges* (2026), doi: https://doi.org/10.1016/j.envc.2026.101575

Highlights

- LLMs accelerate conservation tasks like reviewing technically dense Groundwater Sustainability Plans.
- A fine-tuned GPT-4.1 model achieved 75.2% agreement with human reviewers in tests.
- Hybrid Human/LLM reviews can cut marginal GSP review time by a factor of 160x.
- We developed an automated LLM pipeline to review plans for $1.44 per plan.
- LLMs have the potential to *support but not replace* human environmental reviews.

Improving Conservation Efficiency: Accelerating Groundwater Sustainability Plan Reviews Using Large Language Models

Ryan Bernstein[1*], Seneth Waterman[1], Kirk R. Klausmeyer[2], Nicholas Murphy[2], Melissa M. Rohde[3,4], Cody Carroll[1*]

[1] *University of San Francisco*
[2] *The Nature Conservancy of California, San Francisco, CA*
[3] *Rohde Environmental Consulting, LLC, Seattle, WA*
[4] *SUNY College of Environmental Science and Forestry, Syracuse, NY*

Abstract

Background: The effective implementation of environmental policy relies on thorough review and public consultation, yet the analysis of lengthy, technical documents is a resource-intensive process that creates a significant barrier to entry for many civil society organizations and overburdened agencies. This analytical bottleneck can limit oversight and hinder the achievement of sustainability and environmental justice goals.

Objective & Methods: This study evaluates the potential for Large Language Models (LLMs) to augment human capacity for large-scale policy review. Using an existing assessment of 65 Groundwater Sustainability Plans in California as a case study, we compared the results of several cutting edge LLMs against a comprehensive, multi-year review previously conducted by expert human analysts.

Results: Our method led to a significant acceleration of the review process. The chosen LLM performed an initial analysis of a plan in under two minutes, a 160-fold increase in speed over the eight-hour average for a human expert. The LLM's qualitative assessments achieved a 75.2% agreement rate with the human-led benchmark, demonstrating substantial utility in identifying key policy components and deficiencies.

Conclusion: LLMs represent a transformative tool for the science-policy interface, not as a replacement for human expertise but as a powerful accelerator in a human-in-the-loop system. By drastically reducing

the initial effort of document review, this technology can enhance the capacity of organizations to participate in environmental decision-making, helping to overcome persistent bottlenecks in policy monitoring and evaluation. This approach offers a transferable methodological framework that can be adapted to other regions with similar policy review needs.

Keywords: Large Language Models; Groundwater Sustainability; Fine-Tuning

Introduction

Advances in large language models (LLMs) offer promise for increasing efficiency in many industries. LLMs are capable of human-like reading, writing, and communication skills, making them valuable for tasks like document summarization, classification, and policy analysis (Radford et al., 2018; Vaswani et al., 2017; Hadi et al., 2023). In fields like education, medicine, and finance, LLMs have already shown their utility in reducing costs and enhancing productivity (Mello et al. 2023; Biswas, 2023; Wu et al., 2023; Thawkar et al., 2023; Goyal et al. 2023). LLMs have also passed the Bar Exam and can prove mathematical theorems at the olympiad level, though not always consistently (Bommarito et al. 2022; Frieder et al. 2024). The use of LLMs is increasingly being explored for sustainable development and environmental conservation, including applications in energy management, emission forecasting, waste management, pollution classification, environmental monitoring and biodiversity conservation (Grief et al. 2025; Ullah et al. 2024). These new tools offer exciting opportunities for progress on difficult social and environmental challenges.

One significant environmental challenge is the decline of freshwater species abundance across the globe (WWF, 2024). This issue is most apparent in regions where human needs for water conflict with ecosystem needs. Groundwater is a major source of water in semi-arid and arid regions, where climate

change and growing human demand during droughts impact freshwater species. Groundwater-dependent ecosystems (GDEs), including rivers, wetlands, and springs, are crucial to maintain biodiversity, provide critical habitat, improve water quality, and store carbon (Eamus and Froend, 2006; Howard et al., 2023; Rohde et al., 2024; Yang et al., 2020). These ecosystems are highly sensitive to groundwater depletion, which has been exacerbated by unsustainable extraction practices (Rohde et al., 2021). Efficient groundwater management is pivotal for environmental sustainability and ecological well-being, particularly in heavily populated and agricultural regions like California.

California's Sustainable Groundwater Management Act (SGMA), enacted in 2014, requires Groundwater Sustainability Agencies (GSAs) to develop Groundwater Sustainability Plans (GSPs) for high- and medium-priority basins (Harrer et al., 2020). These GSAs are charged with both formulating and executing GSPs, and are required to meet groundwater sustainability goals by 2040. GSPs are vital for outlining groundwater management plans to avoid the six undesirable results of SGMA (chronic lowering of groundwater levels, reduction in groundwater storage, seawater intrusion, land subsidence, degradation of water quality, and depletion of interconnected surface water), and assessing the impacts of groundwater management on beneficial users (domestic, agricultural, municipal, environmental, tribes, and disadvantaged communities) (Harrer et al. 2020). GSP reviews by state regulators and interested parties are necessary to determine if GSAs have sufficiently met their regulatory requirements under SGMA; however they require significant time and financial investment by technical experts (Dumas Leslie, 2017). The evaluation of the 108 GSPs completed by The Nature Conservancy (TNC), partner organizations, and contractors during the public comment periods (issued by each GSA and the California Department of Water Resources) required significant financial and time investments to review and write comment letters (Perrone et al., 2023).

This study evaluates the ability of LLMs to review GSPs in a similar manner as human evaluators, through the lens of environmental review. Using a sample of 5 GSPs, we compared previous human-led

evaluations to LLM-assisted reviews, considering accuracy compared to human responses and efficiency as metrics. By investigating different LLMs and configurations through experiments with prompt engineering, retrieval-augmented generation, and fine-tuning, we assessed the LLM's ability to replicate human decision-making with the goal of reducing contractor review hours and monetary costs. We then scaled the most promising method up to the full set of 62 GSPs with scoring rubrics (Figure 1).

This work seeks to address the pressing need for scalable and cost-effective tools to facilitate and accelerate environmental review of resource management plans. By exploring the capabilities and limitations of LLMs in the context of GSPs, this study aims to illustrate their potential to support human-led environmental policy review and adaptive management to address the pressing environmental challenges facing the globe.

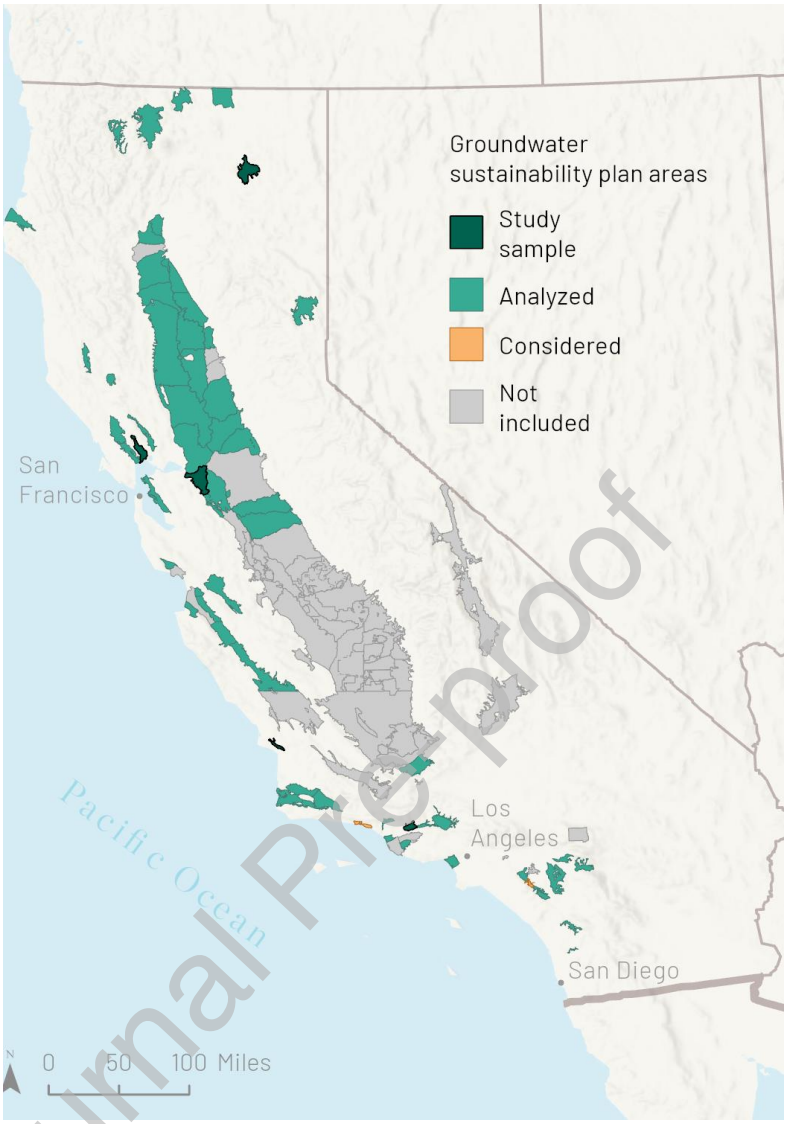


Figure 1. Locator map of California to highlight the groundwater sustainability plan (GSP) areas reviewed in this study. The 5 sample GSPs reviewed in most detail are in dark green and the 49 other GSPs with rubrics that we analyzed are in light green. We attempted, but were unable to analyze 3 additional GSPs due to missing the human scored rubric (shown in yellow). The remaining GSPs shown in grey were reviewed by humans using different methods and did not have a standardized scoring rubric so they were not included in this study. Any area not included in the legend did not have a GSP at the time of this study.

Approach

We acquired a sample of 65 GSPs and their corresponding human-generated scoring rubrics from a review led by TNC. Each human-generated scoring rubric was coded three times by five to seven individual coders, who concurrently worked on a subset of questions to ensure consistent application of the scoring rubric, and another individual for quality control. Intercoder reliability was assessed using Cohen's kappa and was above 0.7 representing substantial but not perfect agreement (McHugh, 2012; Perrone et al, 2023). Given the variance of human responses, we recognize the objectively "correct" answer for each question is difficult to determine and thus the term accuracy in this analysis refers to the accuracy as compared to the human reviewer response. For our experimentation, we looked at a subset of 5 representative GSPs (Big Valley, East Contra Costa, Fillmore, Sonoma, San Luis Obispo) and their 5 corresponding GSP scoring rubrics (Figure 1). We selected this subset for both its high-quality documentation and its balanced representation of coastal and inland regions. The sample GSPs come in the form of lengthy PDF documents, ranging from 212 to 459 pages long, with associated maps and scientific figures. For the scope of this paper, we only focused on leveraging the text of the PDFs and left the multimodal problem of incorporating image data for future research. GSP scoring rubrics each consist of 70 rows that have corresponding questions, one word answers as responses ("Yes/No/Somewhat"), relevant excerpts of text drawn from GSP entries, and references to relevant pages or sections of the plan. They are also divided into 5 Topics: Environment (17 entries), Disadvantaged Communities (12 entries), Climate Change (7 entries), Ecosystems (7 entries) and Tribes (6 entries). Seventeen of these rows were removed as they contained questions that necessitated the interpretation of map data for their answers, such as maps containing domestic well locations. Table 1 is an example of a row recreated from a rubric. It displays the evaluation criteria for Disadvantaged Community (DAC) stakeholder engagement in the GSP development process. The criteria include the level of engagement, spectrum of responses, and relevant search terms. The table also provides a detailed excerpt from the GSP document and the corresponding answer from the human reviewer, which in this case is classified as "Somewhat."

Topic	Criteria	Spectrum	Key Search Terms	Answer	Relevant Text from GSP
Disadvantaged Communities (DACs)	Does the GSP document how DAC stakeholders were given opportunities to engage in the GSP development process? If so, please describe the level of engagement (i.e., whether stakeholders were on an advisory committee, GSA board, working group) in the Notes. Answer according to level of engagement: Yes = Involve, Collaborate, Empower; Somewhat = Inform and Consult; No = No mention of engagement.	No = no reference to DACs, no reference to stakeholder engagement during development; Somewhat = no specific reference to DAC as a stakeholder, provides stakeholder engagement; Yes = references DACs and how they were engaged specifically during development beyond informing DACs	Disadvantaged Community, DAC, stakeholder, beneficial user	Somewhat	"Lassen and Modoc counties are fulfilling their unfunded, mandated roles as Groundwater Sustainability Agencies (GSAs) to develop this Groundwater Sustainability Plan (GSP) after exhausting its administrative challenges to the California Department of Water Resources' (DWR's) determination that Big Valley qualifies as a medium-priority basin. Both counties are disadvantaged, have declining populations, and have no ability to cover the costs of GSP development and implementation. These disadvantaged communities are on the losing end of the digital divide. While the GSAs made every attempt to conduct BVAC meetings with the ability for remote public participation, there were still major logistical and technical challenges with both conducting such meetings and members of the public participating. Those participants that had internet connectivity frequently could not hear or understand the dialogue in the Big Valley community venues and could not interact in the most effective way. However, the GSAs made the best of the circumstances and addressed all comments provided through the various means. The GSAs recognized the obstacles presented by the COVID pandemic early in the efforts to develop a GSP and were proactive in reaching out to both the Governor and Legislature to identify potential solutions. …"

Table 1: Example response from the human-review of GSP rubric. This example is taken from the review of the Butte GSP.

Our main scientific inquiry is to test whether a LLM-based application programming interface (API) is capable of answering the questions in the scoring rubric in the same way as a human reviewer.

Our feature engineering consisted of altering the format of the provided questions in a way that an LLM could understand, while also ensuring that we maintained the integrity of the questions' original content. The GSP criteria and spectrum were combined, the spectrum was converted to complete sentences, and all acronyms were defined. Table 2 illustrates an example of how we incorporated this feature engineering.

Criteria	Spectrum	Question
Does the Groundwater Sustainability Plan (GSP) document how Disadvantaged Community (DAC) stakeholders were given opportunities to engage in the GSP development process? If so, please describe the level of engagement (e.g., whether stakeholders were on an advisory committee, GSA board, working group) in the Notes. Answer according to level of engagement: Yes = Involve, Collaborate, Empower; Somewhat = Inform and Consult; No = No mention of engagement.	No = no reference to DACs, no reference to stakeholder engagement during development; Somewhat = no specific reference to DAC as a stakeholder, provides stakeholder engagement; Yes = references DACs and how they were engaged specifically during development beyond informing DACs	Does the Groundwater Sustainability Plan (GSP) document how Disadvantaged Community (DAC) stakeholders were given opportunities to engage in the GSP development process? Provide a one word answer to this question using the following Spectrum: No = no reference to DACs, no reference to stakeholder engagement during development; Somewhat= no specific reference to DAC engagement but lists DAC as a stakeholder, provides stakeholder engagement details; Yes= references DACs and how they were engaged specifically during development

Table 2: Example of prompt formation using the Criteria and Spectrum columns found in the GSP evaluation documents to create a "Question" to feed to the LLM.

Our analytic process consisted of three primary steps:

1. Prompt Engineering, which consisted of altering the format of the general instructions to the LLM to consider when answering the questions
2. Retrieval-Augmented Generation (RAG) Tuning, which consisted of preprocessing the GSP content in a way that allowed our LLM to focus on the pertinent subsections of the document when it formulated its response; and

3. Fine Tuning, in which we honed and refined higher-order feature representations by training and validating a custom GPT on the human-generated GSP rubrics.

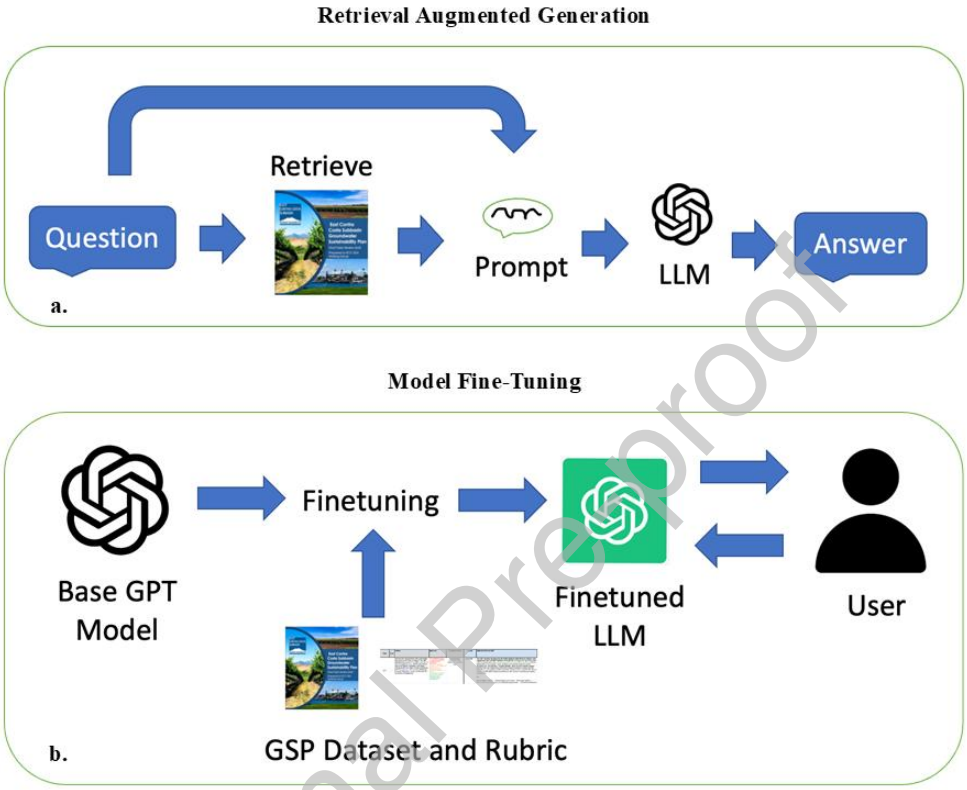


Figure 2. Flow Charts for the a. Retrieval Augmented Generation and b. Fine-Tuning Processes.

Methods

Large Language Models

The number and capabilities of foundational LLMs have increased substantially since the release of ChatGPT in 2022. We focused on the models with API access that were available at the time of this study, including OpenAI's GPT-3.5, GPT-4o, GPT-4.1, GPT-5.5, and o3 and Anthropic's Sonnet 4.6 and Opus 4.7 (vision) LLMs (Table 3). Other studies have shown that OpenAI and Anthropic's LLMs have high performance on general tasks (Radford et al., 2018; Vaswani et al., 2017; Hadi et al., 2023) and the

retrieval augmented generation tasks like the ones used in this study (Ke et al. 2025). Several of the LLMs provide fine-tuning customizations (discussed below) so those variants were tested as well (Table 3).

Model	API Model ID	Released	Input cost per 1M tokens (USD)	Output cost per 1M tokens (USD)	Est cost per GSP (USD)	Notes
GPT-3.5 FT	ft:gpt-3.5-turbo-0125:personal:gsp35v4:DbaumW68	2024-01-25	$3	$6	$1.44	Fine-tuned on gpt-3.5-turbo-0125; pricing at time of evaluation
GPT-4o (base)	gpt-4o-2024-08-06	2024-08-06	$2.5	$10	$1.35	Base model; no fine-tuning
GPT-4o FT	ft:gpt-4o-2024-08-06:personal:gsp4ov4:DbRURPJw	2024-08-06	$3.75	$15	$1.80	Fine-tuned on gpt-4o-2024-08-06
GPT-4.1 (base)	gpt-4.1-2025-04-14	2025-04-14	$2	$8	$1.20	Base model; no fine-tuning
GPT-4.1 FT	ft:gpt-4.1-2025-04-14:personal:gspv4:DbO4oSN8	2025-04-14	$3	$12	$1.44	Fine-tuned on gpt-4.1-2025-04-14; best overall model
o3	o3	2025-04-16	$15	$60	$14.09	Chain of thought model; Evaluated March 2026 at $15/$60 pricing; includes ~2000 estimated reasoning tokens billed at output rate
GPT-5.5 (base)	gpt-5.5	2026-04-24	$5	$30	$2.81	Base model; no fine-tuning
Claude Sonnet 4.6	claude-sonnet-4-6	2026-02-17	$3	$15	$1.72	Text only; base model
Claude Opus	claude-opus-4-7	2026-04-16	$5	$25	$4.65	Vision model; ~17000 input tokens/question

4.7 (vision)						due to image encoding

Table 3: LLM's selected for testing and fine-tuning to compare to human-reviewer responses.

Vector Store

A vector store is a specialized database that can store and manage our text embedding vectors. Our text embedding vectors are numerical vectors that represent the semantic meaning and context for each of the GSP documents. Vector stores are optimized for handling high-dimensional data, enabling rapid querying and retrieval (LangChain Team, 2024). Because our text embeddings have high dimensionality, and efficiency is a priority in this context, we choose to use a vector store to accelerate LLM response time and ensure future scalability.

Our study compared model performance between two candidate choices of vector stores: FAISS (Meta AI, 2024) and CHROMA (LangChain Team, 2024). We utilized the GPT-3.5 Turbo model to assess both vector stores' performance in terms of accuracy.

Prompt Engineering

Rewording the prompt in search of the optimal way to ask ChatGPT queries is known as prompt engineering. White et al. (2023) recently developed a catalog of prompt engineering techniques presented as patterns, offering solutions to common challenges encountered when interacting with LLMs. These prompt patterns serve as a method similar to software patterns, providing reusable solutions to common problems in output generation and interaction with LLMs within a specific context. To improve accuracy, we experimented with prompt engineering by refining both the prompts and the set of instructions given to the model.

We first experimented with adjusting the language used in the queries to encourage more critical responses and to prevent the model from being an excessively permissive judge. To refine the model's response and scrutiny, prompts were modified to emphasize "skeptical evaluation" and "critical assessment" in the instructions.

We also experimented with standardizing prompts to assess whether keeping question formats consistent and spelling out acronyms had an effect on performance. Given that human reviewers had only seen the original prompts, we had limited flexibility in changing their content and subject matter. However, many of the 70 prompts contained inconsistent formatting and numerous acronyms. We sought to determine if using a consistent format and spelling out acronyms would enhance performance.

Retrieval-Augmented Generation

To optimize the RAG step, we conducted a series of experiments using FAISS as the vector store while experimenting with other model parameters. We first investigated the impact of different vector sizing strategies on model accuracy. Since the total text of the GSPs were too big for the LLM to evaluate all at once, we needed a method to send the most relevant chunks of text to the model. Vector sizing strategies are ways to break up long text documents into smaller chunks. We compared page-based vector sizing, where the vector size was adjusted to match the entire page length, with chunk-based vector sizing, which used a fixed chunk size of 1,000 words and a chunk overlap of 200 words. Additionally, we evaluated how varying the number of context vectors retrieved during inference affected the model's accuracy. Specifically, we tested the impact of retrieving the top 5, 10, and 15 most similar vectors to the input query to determine the optimal amount of contextual information to include in the model's response generation. Token size limits were met for values exceeding 15. Performance was compared across GPT-3.5 Turbo and GPT-4. Finally, we tested performance when including the appendices of the GSPs along with the main body of the report.

Problem Simplification

We hypothesized that the three-point spectrum (Yes/No/Somewhat) might be too nuanced for a large language model trained on GSPs in which the Somewhat category is ambiguous, subjective, and evaluator-dependent. To avoid this complication, we simplified the task by merging the "No" and "Somewhat" categories into a single "No" category, effectively creating a binary classification problem. This consolidation is justified by the operational reality of the human-led review process. Both "No" and "Somewhat" responses uniformly signaled an insufficient aspect of the GSP, with both outcomes triggering the exact same policy action: a formal comment letter highlighting said deficiency. Consequently, merging these categories for the LLM evaluation creates no tangible divergence in regulatory engagement. Instead, it allows the model to serve as an efficient, actionable first-pass filter for compliance, which can then be safely audited and refined by subsequent human review. For reference, 105 of the 350 answers in our 5 sample GSPs were initially labeled as “Somewhat.” We also reported the results for the unsimplified 3 class problem (with “Yes/No/Somewhat” categories fully separated) in the appendix.

Confidence Levels

After we converted our task into a binary classification problem, we altered our prompts so the GPT also provided soft predictions, in order to compare Receiver Operator Curves and Precision-Recall Curves across models. In addition, we hypothesized that forcing the model to consider its confidence level regarding a prediction could improve accuracy. The model instructions that we used are replicated below: “You are a skeptical environmental scientist, tasked with answering questions about a section from a Groundwater Sustainability Plan (GSP) document. You are required to provide your confidence level along with the response. Format your response as 'X, Z' where 'X' is either
'Yes' or 'No' and 'Z' is your confidence level, which can be one of the following options: '["Extremely Confident, 100%", "Very Confident, 85%", "Fairly Confident, 75%", "Modest Confidence, 60%", “Random Guess, 50%"]. ' Answer the questions objectively, adhering to the provided spectrums. You

should only state you are "Extremely Confident, 100%" if it is irrefutably true that your answer is correct according to the Groundwater Sustainability Plan."

Fine-Tuned Models

Finally, we experimented with fine-tuning the GPT models used in GSP evaluation. LLM fine-tuning involves adapting a pre-trained large language model to a specific task or domain by training it further on a smaller, task-specific dataset, like our GSPs (OpenAI, 2024). This process leverages the model's existing knowledge while refining its performance on the desired application, enhancing its accuracy and relevance. We trained a fine-tuned GPT-3.5 model and a GPT-4o model (8 epochs, batch_size =1, learning rate =.5), using the 10 most similar pages of GSP text (determined via cosine similarity) as inputs. The fine-tuned GPT models were trained on the Yolo, San Jacinto, Modesto, Santa Rosa Plain, and Scott River Valley GSP rubrics, using a combination of the “Relevant Text from GSP” rubric column and our engineered prompts as input. To assess the accuracy of the fine-tuned models we excluded the five GSPs that were used in fine tuning.

Scaling Up

After completing experiments on a sample of GSPs, we applied the best performing model to the full set of 65 GSPs using python code and OpenAI's API to automate the process. 3 of the GSP Rubrics (Carpenteria, Montecito and Bedford) did not have a complete set of answers. This left 62 GSPs to score with the LLM and compare to the human review. Code for this pipeline is publicly available at: https://github.com/codycarroll/llms_for_gsps.

Results

Vector Stores, RAG Parameters, and Model Versions

FAISS was selected over CRHOMA as the preferred vector store for subsequent testing due to its superiority in handling and retrieving relevant information from the dataset (see Supplemental Information). Testing showed that accuracy was maximized by splitting documents into text chunks based on report pages rather than using arbitrary 1,000-character segments (Figure SI4). We also evaluated the number of chunks used to answer a question and found that 10 chunks provided the best accuracy. Detailed results of these experiments can be found in the Supporting Information section.

Figure 3 displays the classification accuracy across five distinct GSPs alongside their cumulative totals for the nine models: GPT-3.5 FT, GPT-4o, GPT-4o FT, GPT-4.1, GPT-4.1 FT, GPT-5.5, o3, Claude Sonnet 4.6, and Claude Opus 4.7 (vision). Looking at the overall performance summary (Figure 3 and Figure 4), GPT-4o FT demonstrates the highest overall accuracy at 77%, followed closely by GPT-4.1 FT at 76%, and both o3 and Sonnet 4.6 at 75%. GPT-4o has the lowest baseline performance, achieving an overall accuracy of 68%. This graph shows how consistently each model performs when evaluated against different localized datasets, illustrating that fine-tuned variants generally achieve higher accuracy and greater stability across both compliance categories than their base counterparts.

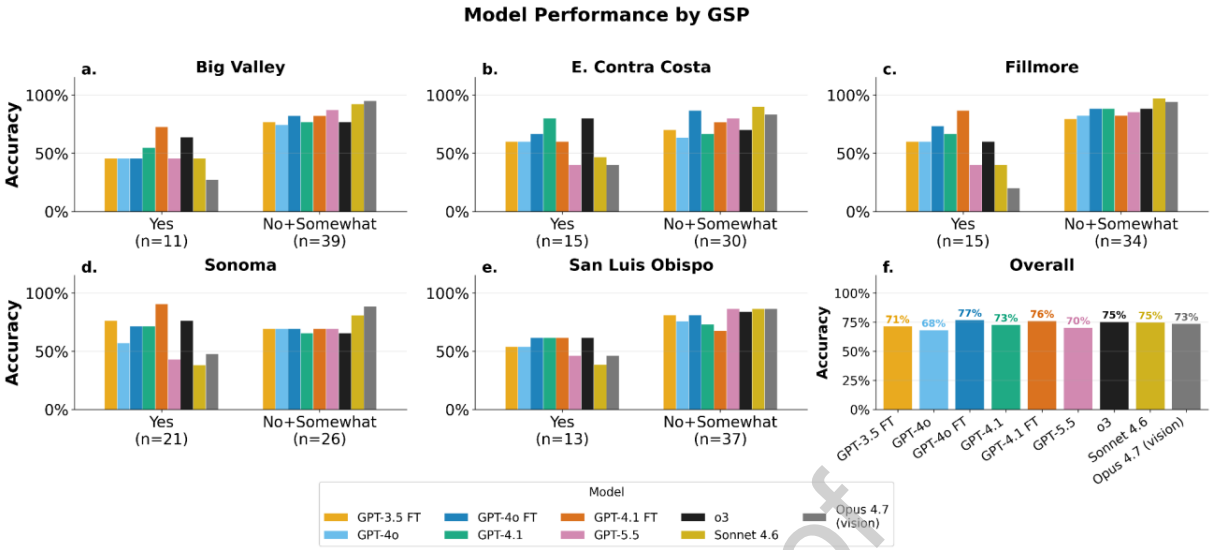


Figure 3: Performance Comparison on the 5 trial GSPs, separated by binary response type. Three class separation (Yes/No/Somewhat) can be found in the appendix.

Figure 4 displays the accuracy by response category ('Yes' and 'No+Somewhat') alongside the overall performance for the nine models: GPT-3.5 FT, GPT-4o, GPT-4o FT, GPT-4.1, GPT-4.1 FT, GPT-5.5, o3, Sonnet 4.6, and Opus 4.7 (vision). GPT-4.1 FT demonstrates the best balance across the categories, achieving an identical accuracy of 76% for both 'Yes' and 'No+Somewhat' responses, while also achieving the highest accuracy in the 'Yes' subset. Opus 4.7 (vision) and Sonnet 4.6 have the lowest performance on 'Yes' answers at 37% and 41% respectively, despite achieving the highest performance of 90% on 'No+Somewhat' answers. This graph shows how well each model handles class imbalances, with GPT-4.1 FT performing best in terms of maintaining consistent, high-level accuracy across distinct response distributions without severely skewing toward a single category.

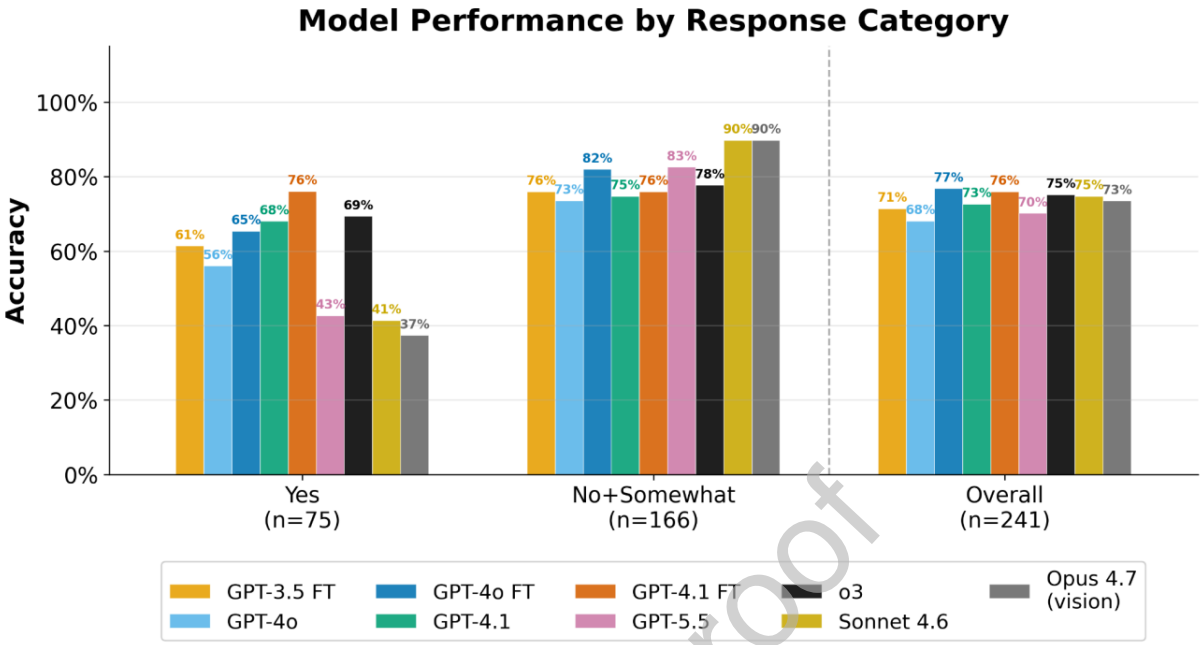


Figure 4: Overall accuracy comparison averaged across the 5 trial GSPs for each LLM, separated by binary response class (Yes/No+Somewhat).

Figure 5a displays the ROC curves for the nine models: GPT-4o Base, GPT-3.5 Fine-Tuned, GPT-4o Fine-Tuned, GPT-4.1, GPT-4.1 Fine-Tuned, GPT 5.5, GPT o3, Claude Sonnet 4.6 and Claude Opus 4.7. GPT-4.1 Fine-Tuned demonstrates the highest Area Under the Curve (AUC) at 0.773, followed by GPT-4o Fine-Tuned with an AUC of 0.770. GPT-3.5 Fine-Tuned has the lowest performance, with an AUC of 0.693. This graph shows how well each model balances precision and recall, with GPT-4.1 Fine-Tuned performing best in terms of having a high precision when also having a higher recall.

Figure 5b shows the Precision-Recall curves for the same models. Here, GPT-4.1 Fine-Tuned once again performs the best, achieving an AUC of 0.711, while GPT-4o Fine-Tuned has an AUC of 0.696. GPT-4o Base has the lowest AUC at 0.590. The ROC curve illustrates the trade-off between the true positive rate (sensitivity) and the false positive rate, with GPT-4.1 Fine-Tuned showing the best overall discrimination ability between positive and negative classes.

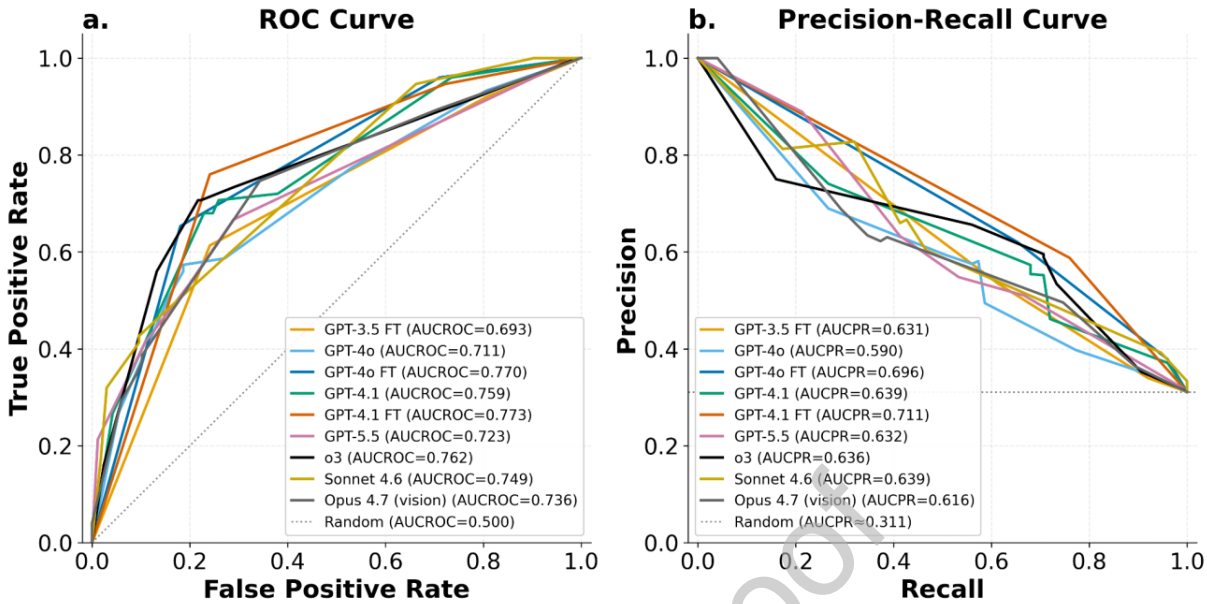


Figure 5: ROC Curves (a., left) and Precision-Recall Curves (b., right) for the 9 LLMs compared with baseline metrics from random guessing.

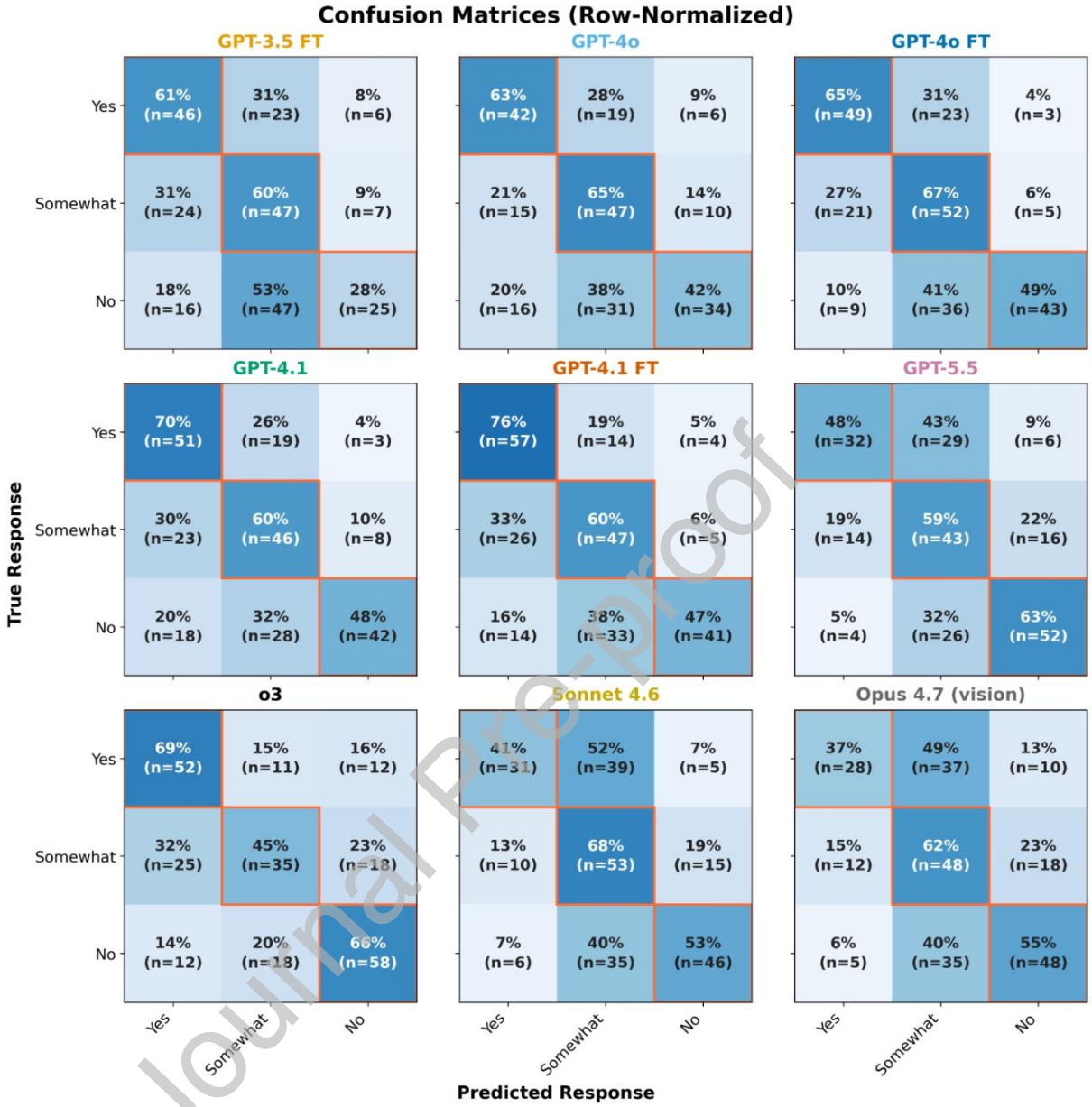


Figure 6

	Performance Metrics for 5 Trial GSPs					
Model	**Accuracy**	**AUCROC**	**AUCPR**	**Precision**	**Recall**	**F1 Score**
GPT-3.5 FT	0.714	0.693	0.631	0.53	0.61	0.57
GPT-4o	0.739	0.711	0.59	0.58	0.57	0.58
GPT-4o FT	**0.768**	0.770	0.696	0.62	0.65	0.64

GPT-4.1	0.730	0.759	0.639	0.55	0.69	0.62
GPT-4.1 FT	0.759	**0.773**	**0.711**	0.59	**0.76**	**0.66**
GPT-5.5	0.718	0.723	0.632	0.55	0.53	0.54
o3	0.755	0.762	0.636	0.59	0.71	0.64
Sonnet 4.6	0.755	0.749	0.639	**0.67**	0.43	0.52
Opus 4.7 (vision)	0.739	0.736	0.616	0.63	0.39	0.48

Table 4: Average classification metrics for the 5 case study GSPs across the nine models. The best model according to each metric is bolded in each column.

Comparing the nine models across accuracy, AUCROC, AUCPR, Precision, Recall, and F1 scores (Table 4), we find that GPT4.1 FT dominates on the majority, in particular ranking metrics (AUCROC, AUCPR, Recall, F1), while GPT4o FT performs slightly better (~1%) than GPT4.1FT on raw accuracy. Sonnet 4.6 has the highest precision but in contrast has one of the lowest recalls. Interestingly, including imaging capability with the vision model Opus 4.7 did not improve the scores at all. Similarly using more recent models like GPT 5.5 or reasoning models like o3 did not result in higher performance. Since GPT 5.5 and o3 are not fine-tunable, we were not able to assess whether fine-tuning would lift its performance above GPT 4.1 FT's. Because of this, we select GPT4.1 FT as our overall highest performing model according to these six metrics.

Scalability

We further validated our findings by testing our best-performing model, GPT-4.1 Fine-Tuned, on the full cohort of 62 GSPs pre-evaluated by TNC to assess its scalability (Table S2). To summarize performance, we generated average performance metrics for all GSPs (n=62) and the subset of GSPs not used in fine tuning (n=57) (Table 5). Excluding the GSPs used for fine tuning, the model demonstrated consistent but imperfect performance, achieving an average AUCROC of 0.761 and an accuracy of 75.2%. The

sustained level of performance from the five case study GSPs to the larger dataset strongly suggests our results are generalizable and not skewed by a small sample size.

Model	Total # GSPs	Accuracy	AUROC	AUCPR	Precision	Recall	F1 Score
GPT-4.1 FT	62	0.756	0.763	0.701	0.597	0.701	0.634
GPT-4.1 FT (excluding GSPs used for fine tuning)	57	0.752	0.761	0.695	0.586	0.673	0.627

Table 5: Average performance metrics for GPT-4.1 Fine-Tuned Model for all GSPs (n=62) and for the subset not used in fine tuning (n=57).

Discussion

We explored the application of a variety of current cutting edge LLMs for evaluating GSPs through a series of experiments. Specifically, we tested the impact of different context retrieval strategies, various prompt engineering strategies, and the choice of different pre-trained models on model accuracy. Furthermore, we investigated the effect of model fine-tuning on accuracy, precision-recall, and ROC metrics.

Based on our results, we conclude that while an LLM can offer accurate results that are significantly better than random guessing, it is not currently able to replace a human reviewer. We found that modifying system instructions to emphasize skepticism influenced both the distribution and accuracy of our model's responses. Giving the model greater amounts of context as opposed to smaller amounts of context, and using page length as opposed to chunk size gave us better accuracy. Context-window limits will almost assuredly be eliminated as LLMs continue to develop and scale, so we can assume that

accuracy will continue to improve. We also may see improved results if GSP drafters take AI into account by improving readability and by splitting the document into clearly organized sections. This should be built into the existing GSP drafting standards.

Our fine-tuning results suggest that an LLM specifically fine-tuned on GSPs will outperform a pre-trained LLM that has not explicitly learned from this domain-specific data. Our best-performing model, with regards to, AUCROC, and AUCPR, recall, and F1 score was a fine-tuned GPT-4.1 model. This model was tuned over 8 epochs with a batch size of 1 and a learning rate of 0.5, utilizing the 10 most similar pages of the text as RAG inputs. Crucially, when this model was applied to our entire cohort of 62 GSPs to evaluate scalability, it demonstrated strong generalization capabilities, with only a minor performance decrease in both AUC and accuracy compared to its performance on our smaller experimental subset.

A major concern discovered during the process of evaluating the LLMs was that model output varied substantially even under consistent parameters. The stochastic nature of LLM responses is due to transformer architecture, developer updates and model variations (Su 2025). If researchers and practitioners are going to effectively implement LLMs as official evaluators, there is a clear need for model developers and the companies behind them to address and control for these inconsistencies and provide more transparent version control in their APIs.

To quantify variability in an LLM's stochastic response, we also re-ran GPT-4.1 FT five times on each of the trial GSPs and reported mean, SD, range of per-GSP accuracy in addition to the majority vote accuracy and rate of unanimous agreement (see Table SI3). For an individual plan, single-pass binary accuracy fluctuated with a standard deviation of between 1.4 and 3.4 percentage points across runs. Pooled across all GSPs, the variability dropped further: accuracy varied by only a standard deviation of 0.5 percentage points. In the least stable case (Sonoma Valley) accuracy ranged from 70.2% to 78.7% depending only on which run was drawn. Across the five GSPs, between 84% and 92% of questions

received an identical answer in all five runs; the remaining questions are those where the underlying text is ambiguous and the model wavers.

We therefore suggest the following deployment protocol in practice: 1. run each rubric question three to five times, 2. report the majority answer together with whether the question's response was unanimous, and 3. route questions without unanimous agreement to human review. A split vote flags the questions where human judgment adds the most value. This approach turns a stochastic evaluation process into a reproducible and more easily auditable one while adding only a modest, parallelizable cost to the pipeline. Alternatively, one could implement a "Panel of Experts" approach to stabilize model performance. This process uses multiple unique LLMs (a "panel") to make predictions, and then makes a final prediction that is aggregated across models (Sourcery AI, 2024). Such an approach could further improve the robustness, consistency, and accuracy of our predictions. We conducted an ensemble experiment on our 5 trial GSPs to demonstrate this; see Table SI4 in the Supplemental Information for details.

Another concern is that the model performed poorly when we attempted to do 3-class classification as opposed to binary classification. Detailed results are presented in Figures SI1, SI2, and SI3 in the Supplemental Information. In our best performing model, accuracy dropped from 75% in a binary framework to 60% when the original three-class scheme (Yes/No/Somewhat) was retained. This gap indicates that while current LLMs can reliably distinguish clearly compliant items from those that are not, they struggle to completely differentiate between gradations of partial compliance. In practical terms, the pipeline therefore functions most effectively as a binary triage tool, capable of filtering clearly non-compliant items and flagging them for further human review, rather than as a fine-grained compliance scorer. Fortunately we only require a binary triage tool for this particular use case, but an inability to capture finer-grained distinctions could be an issue for other LLM document evaluation tasks. Future work should explore whether richer prompting strategies can improve three-class discrimination and enable more nuanced compliance assessments.

Overall our results suggest the possibility of a hybrid approach to GSP review in which human reviewers use an LLM-based application to help quickly identify relevant passages and give them a second opinion on evaluation. The implementation of groundwater sustainability plans in California has multiple future engagement points, as annual reports and five-year assessments and plan re-evaluations are required for medium and high-priority groundwater subbasins. LLM-assisted plan review has the potential to increase The Nature Conservancy's future engagement efficiency.

More broadly, this research explores the applicability of LLM-assisted content review. One of the main benefits of integrating LLMs is the time and cost savings associated with reviewing large volumes of text. For example, the time required for humans to manually review each GSP was approximately 8 hours, whereas the time required to review each GSP with the fine-tuned LLM review averaged just under 3 minutes and cost $1.44 per plan. TNC is already applying this LLM assisted review method to review municipal planning regulations in New York State to assess floodplain protections for individual cities and towns. If LLMs can significantly reduce the duration of human review processes, organizations that review and comment publicly on policy documents (e.g., environmental, advocacy, and research and policy institutes) would be able to save considerable amounts on future time and resource requirements. Organizations may be able to increase their scale of content review, stakeholder education and engagement practices, and policy evaluation as a result of LLM-assisted review workflows.

LLMs could also serve as an effective science education tool for less-technically-resourced stakeholder groups. LLMs have been identified as effective education tools, providing personalized topic instruction, adaptive feedback, and language translation capabilities (Bektik et al. 2024). Previous research has shown that increased representation of underrepresented stakeholder groups in environmental review processes improves outcomes for groundwater sustainability policy (Perrone, Rohde, Wagner et al. 2023). However, groundwater sustainability plans are often inaccessible to less-technically resourced stakeholder groups as

a function of both length (GSPs typically exceed 200 pages) and technical content (GSP subject matter assumes a background in foundational concepts of groundwater hydrology).

Additional work remains in terms of lifting evaluation metrics, but this research establishes the general framework and proof of concept for the utility of LLMs in evaluating GSPs and more generally, in the field of environmental reviews. Future research should explore complete tests of models with reinforcement learning from human feedback and multi-model ensemble approaches.

Conclusion

Our study demonstrates that LLMs can support the content review process for environmental advocacy, reducing time and monetary costs and achieving reasonable agreement with human reviewers. While not a total replacement for experts, these tools can support human capacity in addressing pressing international environmental policy challenges. With ongoing model improvements, better context retrieval, and careful and relevant fine-tuning, LLM-assisted reviews hold promise for enabling non-profits and agencies to do more with limited resources, ultimately enabling more efficient and sustainable natural resource management.

Acknowledgements

The authors would like to thank Alicia Canales for help with the cartography to generate the locator map. We would also like to thank two anonymous reviewers for their constructive comments and suggestions which have helped to improve the strength of this manuscript.

References

Bommarito, M, et al. (2022) GPT Takes the Bar Exam. https://arxiv.org/pdf/2212.14402.pdf

Bektik, D., Edwards, C., Whitelock, D., & Antonaci, A. (2024). Use of LLM tools within higher education: Report 1.

Biswas, S. S. (2023). Role of Chat GPT in Public Health. Annals of Biomedical Engineering, 51(5), 868–869.

Dumas, L. (2017). Implementing SGMA—An Update on California's Foray into Groundwater Regulation. World Environmental and Water Resources Congress 2017.

Eamus, D., & Froend, R. (2006). Groundwater-dependent ecosystems: the where, what and why of GDEs. Australian Journal of Botany, 54(2), 91-96.

Frieder, S., Pinchetti, L., Griffiths, R. R., Salvatori, T., Lukasiewicz, T., Petersen, P., & Berner, J. (2024). Mathematical capabilities of ChatGPT. Advances in neural information processing systems, 36.

Goyal, T, Li, J., Durrett, G. (2023). News Summarization and Evaluation in the Era of GPT-3. arXiv: 2209.12356.

Greif, L., Röckel, F., Kimmig, A., & Ovtcharova, J. (2025). A systematic review of current AI techniques used in the context of the SDGs. International Journal of Environmental Research, 19(1), 1.

Hadi, M. U., et al. (2023). Large language models: a comprehensive survey of its applications, challenges, limitations, and future prospects. *Authorea Preprints*.

Harrer, S., Howard, J., Rohde, M., et al. (2020). Groundwater Dependent Ecosystems under the Sustainable Groundwater Management Act. https://www.scienceforconservation.org/assets/downloads/GDEsUnderSGMA.pdf

Howard, J. K., et al. (2023). Ecosystem services produced by groundwater dependent ecosystems: a framework and case study in California. Frontiers in Water, 5, 1115416.

Ke, Y. H., Jin, L., Elangovan, K., Abdullah, H. R., Liu, N., Sia, A. T. H., ... & Ting, D. S. W. (2025). Retrieval augmented generation for 10 large language models and its generalizability in assessing medical fitness. npj Digital Medicine, 8(1), 187. https://www.nature.com/articles/s41746-025-01519-z.pdf.

LangChain Team. (2024) Vector Stores. Retrieved from https://js.langchain.com/v0.1/docs/modules/data_connection/vectorstores/.

McHugh, M. L. (2012) Interrater reliability: the kappa statistic. Biochem. Med. 22, 276–282.

Mello, R, et al. (2023) Education in the age of Generative AI: Context and Recent Developments. arXiv preprint arXiv:2309.12332.

Meta AI. (2024) Faiss. Retrieved from https://ai.meta.com/tools/faiss/.

OpenAI. (2024) Fine-tuning. Retrieved from https://platform.openai.com/docs/guides/fine-tuning.

Perrone, D., Rohde, M. M., Hammond Wagner, C., Anderson, R., Arthur, S., Atume, N., ... & Remson, E. J. (2023). Stakeholder integration predicts better outcomes from groundwater sustainability policy. Nature Communications, 14(1), 3793.

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training. OpenAI. https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf

Rohde, M. M., Biswas, T., Housman, I. W., Campbell, L. S., Klausmeyer, K. R., & Howard, J. K. (2021). A machine learning approach to predict groundwater levels in California reveals ecosystems at risk. Frontiers in Earth Science, 9, 784499.

Rohde, M. M., Albano, C. M., Huggins, X., Klausmeyer, K. R., Morton, C., Sharman, A., ... & Stella, J. C. (2024). Groundwater-dependent ecosystem map exposes global dryland protection needs. Nature, 632(8023), 101-107.

Sourcery AI. (2024) Better LLM Prompting using the Panel-of-Experts. Retrieved from https://sourcery.ai/blog/panel-of-experts/

Su, W. (2025). Do Large Language Models (Really) Need Statistical Foundations? arXiv preprint arXiv:2505.19145.

Thawkar, O, et al. (2023) Xraygpt: Chest radiographs summarization using medical vision-language models. arXiv preprint arXiv:2306.07971.

Ullah, F., Saqib, S., & Xiong, Y. C. (2024). Integrating artificial intelligence in biodiversity conservation: bridging classical and modern approaches. Biodiversity and Conservation, 1-21.

Vaswani, A, et al. (2017) Attention Is All You Need. arXiv preprint arXiv:1706.03762.

Wu, S, et al. (2023) Bloomberggpt: A large language model for finance. arXiv preprint arXiv:2303.17564.

White, J, et al. (2023) A Prompt Pattern Catalog to Enhance Prompt Engineering with ChatGPT. arXiv:2302.11382.

Wu, Y, et al. (2016). Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. arXiv preprint arXiv:1609.08144.

Yang, K, et al. (2020) Recent progress in the design and application of MOF-based materials for highly selective CO2 capture.ACS Energy Letters, 5(11), 3490-3496.

Supporting Information

Groundwater Dependent Ecosystems in California

Groundwater Dependent Ecosystems (GDEs) are natural ecological communities that rely on groundwater to meet their water needs (Eamus and Froend, 2006; Rohde et al, 2024). These ecosystems include several types of habitats, such as springs, wetlands, rivers, and certain types of forests, where the presence of groundwater is crucial for their survival. GDEs help maintain water quality by filtering pollutants and sediments, providing clean water for humans, plants, and animals (Howard, 2023). They support unique biodiversity, including species that are not found elsewhere, like the desert pupfish, which can only survive in desert springs fed by groundwater (California Department of Fish and Wildlife, 2012). They also play a vital role in carbon storage, helping to mitigate climate change impacts (Yang et al. 2020). Additionally, GDEs support agricultural productivity by ensuring the availability of groundwater for irrigation. They are often sensitive to changes in groundwater levels and quality, making their management critical for environmental sustainability (Kreamer et al. 2015).

To facilitate sustainable groundwater management in California, The Nature Conservancy, the California Department of Fish and Wildlife, and the California Department of Water Resources collaboratively developed a comprehensive database of indicators of groundwater dependent ecosystems. This tool is designed for the identification and mapping of Groundwater Dependent Ecosystems (GDEs) across the state. The database provides an essential, detailed, and user-friendly map of GDEs, serving as a key resource for the effective implementation of the Sustainable

Groundwater Management Act (SGMA) (Klausmeyer et al. 2018). This followed previous mapping efforts, which were done by independent volunteers (Howard et al. 2010).

Expanding on the mapping of Groundwater Dependent Ecosystems (GDEs) in California, Howard et al. (2023) explored their ecosystem functions and connections to various ecosystem services. This study found that GDEs significantly support pollination, with a third of pollinator-dependent crops within 1km of these ecosystems, enhancing potential agricultural yields. In large basins like the Central Valley, GDE's located between developed areas and bodies of water, are crucial for water quality regulation. Additionally, GDEs are key in climate regulation, storing 790 million tons of carbon dioxide, twice California's annual output Howard et al. (2023). These roles highlight GDEs' critical contribution to California's environmental and agricultural sustainability.

Building on the importance of GDEs, Rohde et al. (2021) used satellite remote sensing and machine learning to monitor groundwater levels from 1985 to 2019. Their findings revealed that 44% of GDEs experienced significant long-term groundwater level declines, with declines intensifying in recent decades. It also found that groundwater declines are most prevalent in areas lacking sustainable groundwater management. This research underscores the urgency for robust groundwater management policies in California, highlighting the critical relationship between groundwater levels and management policies.

The Sustainable Groundwater Management Act and Groundwater Sustainability Plans

The Sustainable Groundwater Management Act (SGMA), enacted by the State of California in 2014, represents a transformative approach to groundwater management, targeting issues like over-extraction and depletion with lasting ecological and economic repercussions (Harrer et al., 2020) . This legislation establishes a framework for the sustainable management of groundwater resources. Key to SGMA is the

formation of Groundwater Sustainability Agencies (GSAs) in designated high and medium priority basins. These GSAs are charged with the critical task of formulating and executing Groundwater Sustainability Plans (GSPs). GSPs are essential for setting sustainability goals, monitoring and managing groundwater usage, and addressing substantial and unreasonable reductions in interconnected surface water, and reducing the risks of overdraft (Harrer et al. 2020).

Upon their adoption, GSPs undergo a comprehensive review process by the State, including a 60-day public comment period, and are classified as either adequate, conditionally adequate, or inadequate. This third classification could lead to State intervention if plans are deemed inadequate. The SGMA compliance journey is intricate and prolonged, necessitating both technical measures, such as establishing basin budgets and sustainability criteria, and consistent stakeholder engagement. GSPs must periodically be updated and refined every five years, taking into account the economic, social, and environmental impacts of the management strategies employed. Meeting interim milestones is crucial to circumvent State control, with long-term monitoring being vital in tracking progress towards sustainability goals and facilitating adaptive management (Dumas Leslie, 2017).

For the five-year updates, constructing effective and inexpensive feedback mechanisms is essential. To improve these efforts, we sought to develop a large language model (LLM) designed to analyze a GSA's GSP and assess the plan according to a standardized rubric of evaluative questions. Development of such an LLM would enable a more efficient and insightful review process, ensuring that the GSAs receive comprehensive, data-driven recommendations to improve their groundwater management strategies, while avoiding prohibitive costs of consultant labor. This innovation could significantly enhance the adaptive management capabilities of GSAs, providing a dynamic and cost-effective tool to align with SGMA's objectives of sustainable groundwater management.

Large Language Models

Large language models (LLMs) were developed to emulate human-like reading, writing, and communication skills through advanced natural language processing. These models are trained on vast datasets to predict word sequences. The evolution of LLMs began with rule-based natural language processing in the 1950s, utilizing hand-crafted linguistic rules (Chomsky 1956). The 1980s saw the advent of statistical models that used probabilistic methods for word sequence prediction (Rosenfeld, 2018). Google's Neural Machine Translation system in 2015 marked a milestone as the first major neural language model, which employs deep learning to analyze extensive textual data (Wu et al. 2016). The Transformer model, introduced in 2017, revolutionized LLMs with its ability to learn long-term dependencies and parallel training capabilities (Vaswani et al. 2017). OpenAI's introduction of GPT-1 in 2018, featuring a transformer-based architecture and 117 million parameters, was a significant progression (Radford et al. 2018). This was followed by the release of GPT-3 in 2020, which, trained on a massive dataset, set new benchmarks in generating coherent and natural-sounding text (Hadi et al. 2023).

OpenAI's ChatGPT, a generative AI model for natural language tasks, leverages advanced deep learning techniques. Trained on a vast array of internet-sourced texts like books, articles, and websites (Hadi et al. 2023), it excels in generating human-like, coherent responses. ChatGPT predicts word sequences to produce contextually relevant text. This capability stems from its training on complex language patterns, enabling it to engage in meaningful conversations and respond accurately to various queries.

GPT models, as cutting-edge tools in the workplace, are revolutionizing various industries with their advanced language processing capabilities. In the medical sector, LLMs are expected to play a significant role in decision-making and patient care. For example, XrayGPT is used to automate X-ray image analysis, allowing users to interact and inquire about these analyses via chat (Thawkar et al. 2023). In education, ChatGPT has been instrumental in offering personalized learning experiences. Platforms like Khan Academy are integrating ChatGPT to enhance tutoring (Mello et al. 2023). In the finance sector,

models like BloombergGPT provide customer support and financial advice (Wu et al. 2023). In engineering, ChatGPT aids developers with code generation and debugging, showcasing the versatility and wide-ranging impact of these models in diverse fields. In Natural Language Processing research, ChatGPT has been proven to excel at text summarization; it performs extremely well in the benchmark domain of news summarization (Goyal et al. 2023). It has also taken, and passed, the Bar Exam (Bommarito et al. 2022).

Three Class (Yes/No/Somewhat) Evaluation Results

Figures SI1 and SI2 replicate Figures 3 and 4 from our main section. Figure SI1 displays the classification accuracy across five distinct GSPs alongside their cumulative totals for each model. Figure SI2 displays the accuracy by response category ('Yes' and 'No+Somewhat') alongside the overall performance for each model. The difference between Figures 3 and 4 and Figures SI1 and SI2 is that instead of combining “No” and “Somewhat”, we separate them. Our best performing model’s overall accuracy goes down from 77% to 60% with “No” and “Somewhat” separated.

Figure SI1: Performance Comparison on the 5 trial GSPs, separated by three response classes (Yes/No/Somewhat).

Figure SI2: Performance Comparison by three response classes (Yes/No/Somewhat).

Figure SI3: Confusion Matrices for all 9 Base GPT-4o, GPT-3.5 Fine-Tuned and GPT-4o Fine-Tuned Models.

Figure SI3 displays raw confusion matrices for evaluating each of the considered models. Corresponding metrics and interpretations are found in Table 3 in the main text.

Additional LLM Experiments

Vector Store, RAG Parameter, and Prompting Experiments

Preliminary experiments on vector stores and RAG parameters were performed on GPT 3.5, GPT4o, and GPT4o FT. For vector stores, GPT 3.5 Turbo's "Yes/No/Somewhat" results indicated that FAISS outperformed CHROMA, achieving an accuracy of 52.83% compared to 33.96%. Thus, FAISS was selected as the preferred vector store due to its superiority in handling and retrieving relevant information from the dataset.

We also ran a test that included the appendices for the 5 sample GSPs. The test with appendices answered 42% correct, whereas same model w/o appendices scored 51% correct. We found the model selected "Somewhat" 43% of the time when an appendix was included compared to 25% when the appendix was omitted. Humans selected "Somewhat" ~30% of the time on these sample GSPs. Thus, we concluded that adding the appendices did not improve accuracy and conducted further experiments without the appendices.

Initial testing of hyperparameter choices and models were evaluated on a test set of 255 questions from five sample GSPs using "Yes/No/Somewhat" results (Fig. SI3). Evaluating the GPT 3.5 Turbo model with different parameter choices showed that accuracy was maximized by using the report pages to split documents into text chunks instead of arbitrary chunks that were 1,000 characters long (Fig SI3a). We then evaluated the number of chunks used to answer a question, and found that 10 chunks provided the best accuracy (Fig SI3b). We also saw that depending on the GSP, GPT versions 4 or 4o outperformed the GPT-3.5 Turbo in terms of accuracy after optimizing hyperparameters (Fig. SI3c).

Figure SI4: Accuracy (in %) comparison across model and hyperparameter choices for non-fine-tuned models. Panels a. and b. depict accuracy for experiments on split type to define vector chunk size and vector retrieval (chunk count) size using GPT-3.5 Turbo, while panel c. compares accuracy across model versions.

Explicitly prompting the GPT 3.5 Turbo model to be a critical reviewer did not significantly influence model accuracy (67.8% original vs 67.1% critical).

Figure SI5: Performance comparison for different prompt engineering techniques.

Performance Comparisons across Models and GSPs

Despite GPT4.1FT being the favored model according to four of six metrics considered, most pairwise differences in accuracy were not found to be statistically significant according to McNemar's test for paired proportions, with the exception of the non-fine-tuned GPT4o model (Table SI1).

Table SI1: Pairwise accuracy comparison vs. GPT4.1FT on the 5 Trial GSPs using McNemar's test for paired proportions. Significance level notation: 0.1 (†), 0.05 (*), 0.01 (**), 0.001(***), not significant (ns).

Specific performance metrics for each of the 62 GSPs for the final GPT 4.1 FT model are displayed in Table SI2.

Notes:
*Used for fine tuning the LLM.
**Missing human coded rubrics prevented the accuracy assessment.

Table SI2: Accuracy (Binary and 3-class), AUCROC, and AUCPR by GSP after applying the GPT-4.1o Fine-Tuned model. Here N represents the number of relevant (i.e. non-NA) questions in each GSP's evaluation.

Stochasticity and Variability of LLM Responses

The stochasticity of LLM outputs has a direct practical consequence: a single pass of ChatGDE over a GSP is one draw from a distribution of possible scorings, and two runs on the same plan need not agree on every question. For formal policy monitoring, where a scoring may inform whether a plan is deemed compliant, reproducibility is essential. We therefore recommend that practitioners deploying this system in a regulatory setting run each query 3-5 times and adopt the majority vote answer across runs rather than relying on a single pass. The benefit of voting is reproducibility rather than a gain in accuracy. Checking for majority vote or unanimity collapses the run-to-run disagreement into a single auditable decision.

Table SI3. Run-to-run variability of GPT-4.1 FT over five repeated inference passes on the five trial GSPs, with retrieval held fixed so that all variation reflects intrinsic LLM variability (default temperature, no fixed seed). Columns report the mean, standard deviation, and range of single-pass binary accuracy across the five runs, the accuracy of the majority vote answer, and the fraction of questions answered identically across all runs. Note that per-plan accuracy varies by up to 3.4 points across runs while the pooled accuracy's standard deviation is only 0.5 percentage points.

Panel of Experts Evaluation

To explore whether combining model outputs could improve classification performance, we constructed an ensemble from the five highest-performing models evaluated on the trial GSPs: GPT-4o FT (76.8% accuracy), GPT-4.1 FT (75.9%), o3 (75.5%), Claude Sonnet 4.6 (75.5%), and Claude Opus 4.7 Vision (73.9%). Two aggregation strategies were compared: 1) score averaging and 2) majority vote.
The score averaging method averaged the continuous confidence scores produced by each model for a given question, then classified the response as "Yes" if the mean score exceeded a threshold of 0.5 and "No+Somewhat" otherwise; this approach preserves the full probabilistic signal from each model before committing to a binary decision. The majority vote method instead binarized each model's prediction independently at the 0.5 threshold, then assigned the ensemble's prediction by simple plurality (at least three of five models agreeing). Soft prediction scores for the majority vote method were taken as the fraction of majority vote, e.g. 3/5 = 0.6.

Score averaging achieved higher recall (70.7% vs. 60.0%) and F1 scores (67.1% vs. 63.8%), while majority vote achieved slightly higher precision (68.2% vs. 63.9%) and overall accuracy (78.8% vs. 78.4%) (Table SI4). Both ensemble methods outperformed any individual model on accuracy, but neither was ultimately adopted for the full 62 GSP evaluation, as the added cost and complexity of running five models per GSP was not justified by the marginal accuracy gains over GPT-4.1 FT alone.
Table SI4: Ensembled performance metrics on the top 5 performing models on the 5 trial GSPs.

References

Bommarito, M, et al. (2022) GPT Takes the Bar Exam. https://arxiv.org/pdf/2212.14402.pdf

California Department of Fish and Wildlife. (n.d.). Desert Pupfish (Cyprinodon macularis).https://wildlife.ca.gov/Regions/6/Desert-Fishes/Desert-Pupfish.

Chomsky, N. (1956). Three models for the description of language. *IRE Transactions on information theory*, *2*(3), 113-124.

Dumas, L. (2017). Implementing SGMA—An Update on California's Foray into Groundwater Regulation. World Environmental and Water Resources Congress 2017.

Eamus, D., & Froend, R. (2006). Groundwater-dependent ecosystems: the where, what and why of GDEs. Australian Journal of Botany, 54(2), 91-96.

Goyal, T, Li, J., Durrett, G. (2023). News Summarization and Evaluation in the Era of GPT-3. arXiv: 2209.12356.

Hadi, M. U., et al. (2023). Large language models: a comprehensive survey of its applications, challenges, limitations, and future prospects. Authorea Preprints.

Harrer, S., Howard, J., Rohde, M., et al. (2020). Groundwater Dependent Ecosystems under the Sustainable Groundwater Management Act. https://www.scienceforconservation.org/assets/downloads/GDEsUnderSGMA.pdf.

Howard, J. & Merrifield, M. (2010). Mapping groundwater dependent ecosystems in California. PLoS One, 5(6), e11249.

Howard, J. K., et al. (2023). Ecosystem services produced by groundwater dependent ecosystems: a framework and case study in California. Frontiers in Water, 5, 1115416.

Klausmeyer, K., et al. (2018). Mapping indicators of groundwater dependent ecosystems in California: Methods report.San Francisco, California.

Kreamer, D. K., Stevens, L. E., & Ledbetter, J. D. (2015). Groundwater dependent ecosystems—Science, challenges, and policy directions. Groundwater, 205, 230.

Mello, R, et al. (2023) Education in the age of Generative AI: Context and Recent Developments. arXiv preprint arXiv:2309.12332.

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training. OpenAI. https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf

Rohde, M. M., et al. (2021) A machine learning approach to predict groundwater levels in California reveals ecosystems at risk. Frontiers in Earth Science, 9, 784499.

Rohde, M. M., Albano, C. M., Huggins, X., Klausmeyer, K. R., Morton, C., Sharman, A., ... & Stella, J. C. (2024). Groundwater-dependent ecosystem map exposes global dryland protection needs. Nature, 632(8023), 101-107.

Rosenfeld, Roni (2018). Two Decades of Statistical Language Modeling: Where Do We Go From Here?. Carnegie Mellon University. Journal contribution. https://doi.org/10.1184/R1/6611138.v1

Thawkar, O, et al. (2023) Xraygpt: Chest radiographs summarization using medical vision-language models. arXiv preprint arXiv:2306.07971.

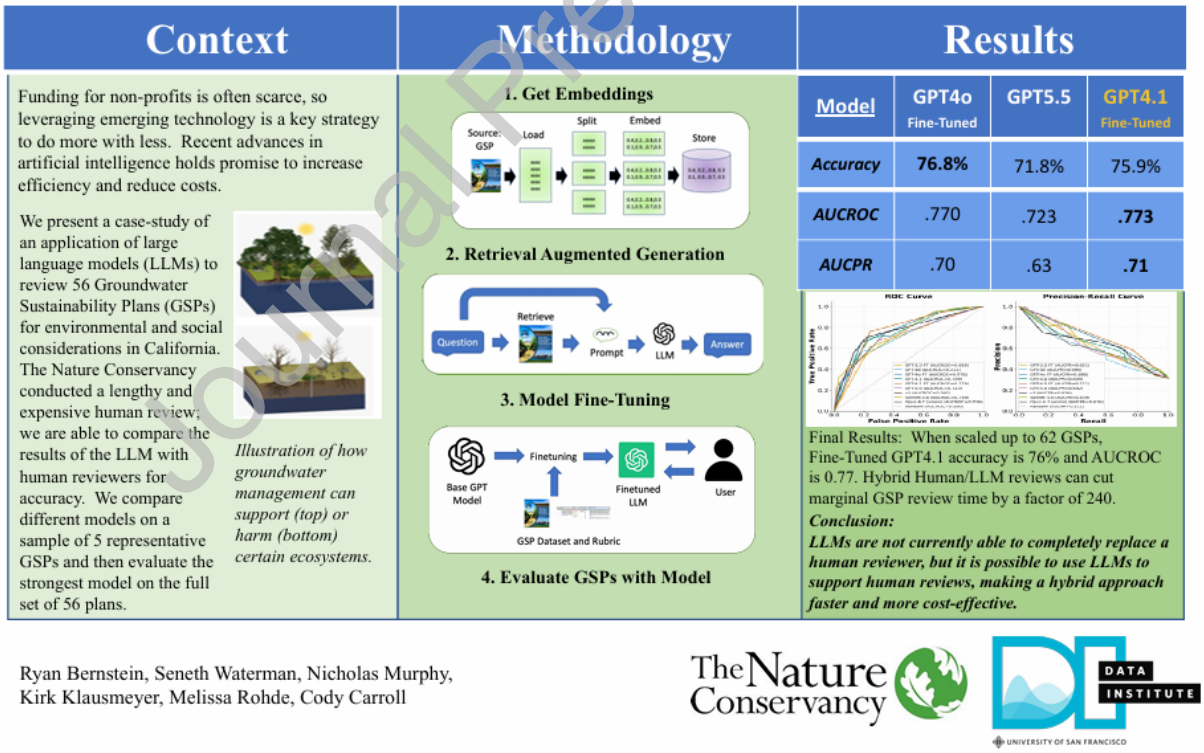
Vaswani, A, et al. (2017) Attention Is All You Need. arXiv preprint arXiv:1706.03762.

Wu, S, et al. (2023) Bloomberggpt: A large language model for finance. arXiv preprint arXiv:2303.17564.

Wu, Y, et al. (2016). Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. arXiv preprint arXiv:1609.08144.

Yang, K, et al. (2020) Recent progress in the design and application of MOF-based materials for highly selective CO2 capture. ACS Energy Letters, 5(11), 3490-3496.

Graphical Abstract



Declaration of interests

☒The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: