Chat GDE: Can Large Language Models Evaluate Sustainability Plans for Groundwater Dependent Ecosystems?

Seneth Waterman[1], Ryan Bernstein[1],
Kirk Klausmeyer[2], M.A., Cody Carroll[1], Ph.D.

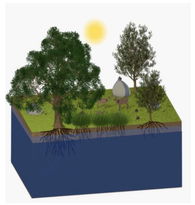
[1]Data Institute, University of San Francisco, [2]The Nature Conservancy

Background and Problem Overview

What are Groundwater-Dependent Ecosystems?

- **Definition:** Natural ecological communities that rely on groundwater to meet their water needs.
- **Examples:** Springs, wetlands, rivers, and certain types of forests, where the presence of groundwater is crucial for their survival.
- **Why are they important?** GDEs help maintain water quality by filtering pollutants and sediments, providing clean water for humans, plants, and animals. They support unique biodiversity, including species that are not found elsewhere. They also play a vital role in carbon storage, helping to mitigate climate change impacts.
- **Threats:** Because GDEs are often sensitive to changes in groundwater levels and quality, their management is critical for environmental sustainability.[1]

Plants within these ecosystems require groundwater to be close to the surface so their roots can access it, especially during dry periods. Intensive groundwater pumping can lower these levels significantly, causing some plants to lose access to groundwater as it falls below their root zones, potentially leading to habitat loss.



How do we Protect them?

In response to these challenges, the California legislature enacted the Sustainable Groundwater Management Act (SGMA) in 2014. This statewide framework mandates local agencies to establish Groundwater Sustainability Agencies (GSAs) and develop Groundwater Sustainability Plans (GSPs).[2] These plans aim to prevent groundwater overdraft. To assist with this management effort, The Nature Conservancy conducted an extensive review of 60+ GSPs and provided detailed comments on how they could be improved for nature and disadvantaged communities.

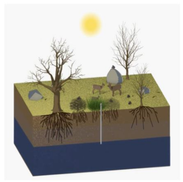


Figure 1: Effect of Overdraft On Ecosystems [5]

Recent advances in artificial intelligence (AI) have created tools like ChatGPT, which has been proven to excel at text summarization; it performs extremely well in the benchmark domain of news summarization.[3] It has also taken, and passed, the Bar Exam.[4]

Our goal in this study is to run a series of experiments to test if extensive document review can be performed with significantly greater efficiency via AI tools and large language models (LLMs) like ChatGPT.

Data Description

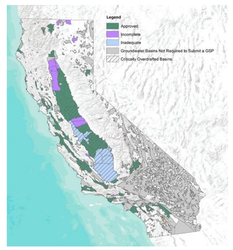


Figure 2: Current Status of GSPs [6]

Dataset

Our dataset consists of:

- **111 Groundwater Sustainability Plans**: PDFs of draft and final plans and appendices submitted as of December 2022.
- **Scoring System**: Each plan is scored by a human reviewer using a detailed rubric.[6]
- **Rubric Details**: The rubric encompasses 70 questions, each including:
  - ✓ A response (Yes, Somewhat, No)
  - ✓ Excerpts of relevant text from the GSP documents
  - ✓ Indexed references to relevant pages and sections

Experimental Subset

We decided to focus on a random subset of 5 GSPs when testing our model.

- **GSPs Selected**: Big Valley, Butte, East Coast Contra, Sonoma, and San Margarita.
- **Performance Metric**: Model accuracy was evaluated by comparing the proportion of responses (Yes/No/Somewhat) from the model to those from human evaluators for each question.

Vector Database Creation

To facilitate sophisticated question-answering (Q&A) capabilities with LLMs, such as ChatGPT, and to address its token limit, we needed a robust method to handle lengthy documents like Groundwater Sustainability Plans (GSPs). To make these documents accessible to the LLM, we converted the GSPs from their original PDF format into a machine-readable form. This process involved:

1. Loading each GSP into a document loader.
2. Splitting and transforming the text into manageable segments.
3. Converting these segments into numerical representations using vector embeddings.
4. Storing these embeddings in a vector database, creating a structured external knowledge base for our model.

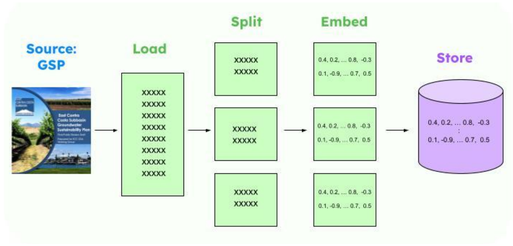


Figure 3: Transforming GSPS to Vector Database[7]

Model Architecture

With our vector database in place, we developed a Retrieval-Augmented Generation (RAG) model. This AI framework improves the quality of LLM-generated responses by grounding the model on this external source of knowledge, supplementing the LLM's internal representation of information.

The RAG model operates in two phases:

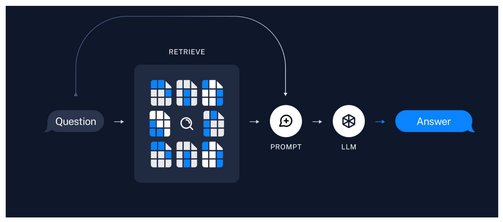


Figure 4: RAG Architecture[8]

Retrieval Phase:

1. A user inputs a prompt.
2. The prompt is converted into a vector representation.
3. This vector is matched with the 10 most similar vectors in our database.
4. The 10 most relevant pages related to the GSP and the prompt are identified.

Augmented Generation Phase:

1. The RAG model augments the user's input by incorporating the relevant retrieved data into the context.
2. These ten pages, along with the original prompt, are fed into an LLM, in our case ChatGPT.
3. The LLM then generates a response based on this augmented input.

Refining Model Output with JSON-Formatted Outputs

The output from our model was consistently in formats that were difficult to analyze. To resolve this, we sent the model's output through an alternative version of ChatGPT designed to produce outputs exclusively in JSON format.

Adjusting Model Parameters

Experiment 1: Vector Split Size for Database Storage

Objective: Investigate the impact of different vector sizing strategies on model accuracy.

Methods:

- **Page-Based Vector Size**: Vector size adjusted to match the entire page length.
- **Chunk-Based Vector Size**: Vector size set with a fixed chunk size of 1000 and chunk overlap of 200.

Results: Vector sizing based on page length showed better results.

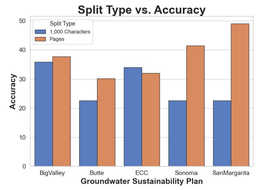


Figure 5: Split Type vs. Accuracy

Experiment 2: Influence of Vector Match and Retrieval Quantity

Objective: Assess how changes in the 'k' parameter, which controls the number of similar vectors retrieved, affect the accuracy of the model.

Methods:

- **Parameter Settings**: 'k' values of 5, 10, and 15

Results: Ten retrieved vectors (k=10), yielded the highest accuracy across most datasets.

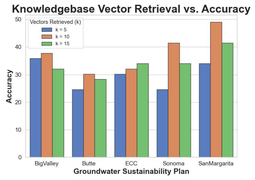


Figure 6: Knowledgebase Vector Retrieval vs. Accuracy

Experiment 3:Comparing GPT-3.5 Turbo and GPT-4

Objective: Compare the performance of GPT-3.5 Turbo and GPT-4 to determine which model delivers superior accuracy.

Methods:

- **Model Versions**: Utilized GPT-3.5 Turbo and GPT-4 to process the same set of conditions.

Results: GPT-4 consistently outperformed GPT-3.5 Turbo in accuracy across all GSPs, though the degree of improvement varied.

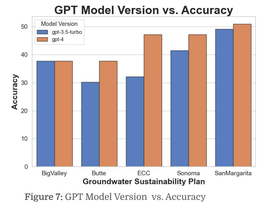
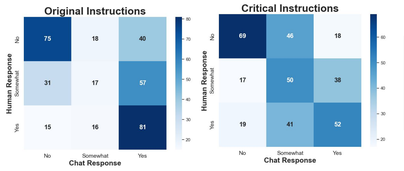


Figure 7: GPT Model Version vs. Accuracy

Prompt Engineering

Experimenting with Critical Language in Chatbot Instructions

Objective: Refine the models response criteria to ensure more evaluative and critical answers by adjusting the instructions to emphasize "skeptical evaluation" and "critical assessment."



Instruction Type	Accuracy
Original	38.12%
Critical	43.39%

Table 1: Accuracy by Instruction Type

Figure 8: Confusion Matrix of AI and Human Responses

Results: Modifying instructions significantly influenced the distribution of the models response and resulted in a modest increase in accuracy.

Experimenting with Standardization of Prompts

Objective: Assess whether standardizing chatbot prompts, by using a consistent format and spelling out acronyms, improves performance.

Results: Standardizing prompts did not significantly impact model performance.

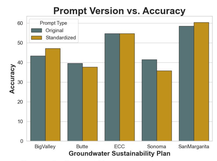


Figure 9: Prompt Version vs. Accuracy

Simplifying the Task

Experimenting with Binary Classification

Objective: Explore how consolidating "No" and "Somewhat" into one response option impacts chatbot accuracy compared to a three-point spectrum (Yes, No, Somewhat).

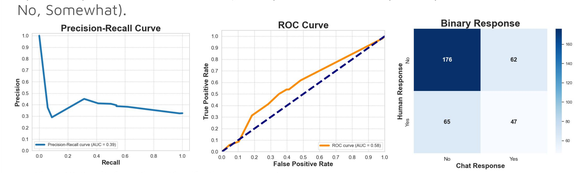


Figure 10: Results from Binary Classification Experiment

Results: Merging the "No" and "Somewhat" categories enhanced accuracy while weakening the ability to identify correct instances accurately.

- → **Accuracy**: Increased from 44% to 64%
- → **Precision**: Decreased from 48% to 43%
- → **Recall**: Decreased from 46% to 42%

Concerns:

- **Inconsistency Across Tests**: Model outputs varied significantly during experiments conducted on different days, even with the temperature parameter set to zero.
- **Response Variability**: Over 25% of responses shifted between "Yes" and "No" in repeated tests.
- **Stability Issues**: These variations highlight significant stability issues, exacerbated by updates from OpenAI.

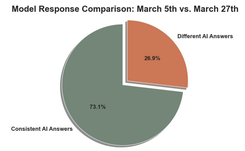


Figure 11: Model Response Comparison: March 5th vs. March 27th

Discussion

Evaluating ChatGPT's Role in GSP Analysis

- **Complimenting Rather than Replacement**: ChatGPT generally offers accuracy superior to random guessing. However, its results do not consistently match those of human evaluators, indicating it is not yet ready to replace human judgment.
- **Potential for Improvement**: Performance could improve if ChatGPT were allowed to analyze a larger context window, as it currently focuses on too narrow a section of text.
- **Inconsistency Issues**: Output varies even under consistent parameters and zero temperature setting, highlighting the stochastic nature of LLM responses due to transformer architecture, developer updates and model variations.
- **Need for Stability**: Effective implementation of LLMs as official evaluators requires addressing and controlling these inconsistencies.

Next Steps for Improvement

- **Further Refine the Existing Model**: Experiment with OpenAI's new fine-tuning API features to enhance accuracy and relevance.
- **Experiment with other LLMs**: Run our experiments with Claude, Llama, or Gemini instead of ChatGPT.

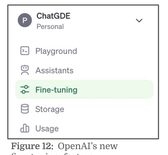


Figure 12: OpenAI's new fine-tuning feature

Acknowledgements

This work was supported by The Nature Conservancy.

References

1. Kreamer, David K., Lawrence E. Stevens, and Jeri D. Ledbetter. "Groundwater dependent ecosystems—Science, challenges, and policy directions." Groundwater 205 (2015): 230.
2. California Department of Water Resources. "SGMA Groundwater Management." Ca.gov, 2014, water.ca.gov/Programs/Groundwater-Management/SGMA-Groundwater-Management.
3. Goyal, Tanya, et al. "News Summarization and Evaluation in the Era of GPT-3." (2023) arXiv: 2209.12356
4. Bommarito, Michael, et al. "GPT Takes the Bar Exam." (2022): 0011249.s001.
5. "Introduction to Terrestrial Vegetation." The Nature Conservancy, 2023. https://www.youtube.com/watch?v=LJUHQJWMIrY
6. "Groundwater Sustainability Plans." Ca.gov, 18 Apr. 2018, water.ca.gov/Programs/Groundwater-Management/SGMA-Groundwater-Management/Groundwater-Sustainability-Plans.
7. Perrone, D., Rohde, M.M., Hammond Wagner, C. et al. Stakeholder integration predicts better outcomes from groundwater sustainability policy. Nat Commun 14, 3793 (2023).
8. "Syncing Data Sources to Vector Stores." LangChain Blog, 6 Sept. 2023, blog.langchain.dev/syncing-data-sources-to-vector-stores/. Accessed 24 Apr. 2024.
9. "Question Answering | 🦜🔗 Langchain." Python.langchain.com, 2024, python.langchain.com/docs/use_cases/question_answering/.