PROCEEDINGS, 50th Workshop on Geothermal Reservoir Engineering
Stanford University, Stanford, California, February 10-12, 2025
SGP-TR-229

Lessons Learned from AskGDR: Usage and Impact Analysis of the Geothermal Data Repository's AI Research Assistant

Jon Weers, Nicole Taverna, Slater Podgorny, Jay Huggins

National Renewable Energy Laboratory (NREL), Golden, CO 80401, USA

Jon.Weers@nrel.gov

Keywords: Geothermal, data, repository, features, enhancements, open, information, GDR, OpenEI, DOE, ELM, NREL, discoverability, usability, accessibility, innovation, metadata, artificial intelligence, AI, machine learning, ML, large language model, LLM

ABSTRACT

In October of 2024, the Department of Energy's (DOE) Geothermal Data Repository (GDR) team officially launched AskGDR, an AI research assistant resulting from the integration of a Large Language Model (LLM) with the metadata and supporting documents associated with GDR datasets. AskGDR allows GDR users to ask deeper questions about the origin of datasets, the methods used to collect them, and the findings they help support. Using Retrieval Augmented Generation (RAG), AskGDR can be used to summarize findings spread across dozens of papers and technical reports or to extract relevant information describing a single data field. However, generative AI is experimental. The National Renewable Energy Laboratory (NREL) has been collecting metrics on AskGDR and documenting lessons learned during its deployment. This paper will outline the efficacy and impact of AskGDR through analysis of its use, operating costs, number and types of questions asked, and the quality of answers provided.

1. INTRODUCTION

The U.S. Department of Energy's (DOE) Geothermal Data Repository (GDR) provides free access to data generated from efforts funded by DOE's Geothermal Technology Office (GTO). With over 1,280 datasets, the GDR provides access to data from all aspects of geothermal research development, and operations. Combined, GDR datasets contain more than 287 TB of geothermal data from 122 different data providers (GDR 2025). The GDR has a search page that allows users to find datasets based on keyword, topic, geothermal technology type, and data type. Users can also browse datasets by featured project, author, or primary researching organization. Metadata associated with each dataset provides additional context for using and understanding the data. To enable users to ask deeper questions of data within the GDR, a team of developers at the National Renewable Energy Laboratory (NREL) developed a Large Language Model (LLM) utilizing the metadata and supporting documents from GDR datasets. The resulting LLM is the foundation for an Artificially Intelligent (AI) research assistant called "AskGDR", which enables users to ask and receive answers to deeper semantic questions about the methodologies and underlying assumptions associated with a geothermal dataset, as well as broader questions about the geothermal industry.

1.1 AskGDR: GDR's AI Research Assistant

AskGDR allows GDR users to ask deeper questions about the origin of datasets, the methods used to collect them, and the findings they help support. Using Retrieval Augmented Generation (RAG), AskGDR can be used to summarize findings spread across dozens of papers and technical reports or to extract relevant information describing a single data field. Developed to be an AI research assistant, AskGDR is not a conversational chatbot. In the interest of scientific rigor, its corpus of knowledge has been restricted to only the curated information contained within the metadata and supporting documents of GDR data submissions. Unlike other LLMs, AskGDR will not go to the internet for answers, nor will it speculate. It will always cite its sources. If it does not know the answer to a question, it will simply say, "I don't know." These and other design decisions serve to control its corpus of knowledge and constrain it for scientific rigor (Weers et al, 2024) but limit its ability to carry a conversation. Each question asked is independent of the last. However, context can be carried forward by embedding it in subsequent requests through careful question engineering. These constraints help ensure scientific rigor while also protecting AskGDR from misuse by blocking questions designed to manipulate AskGDR or "jailbreak" it (i.e. bypass its constraints).

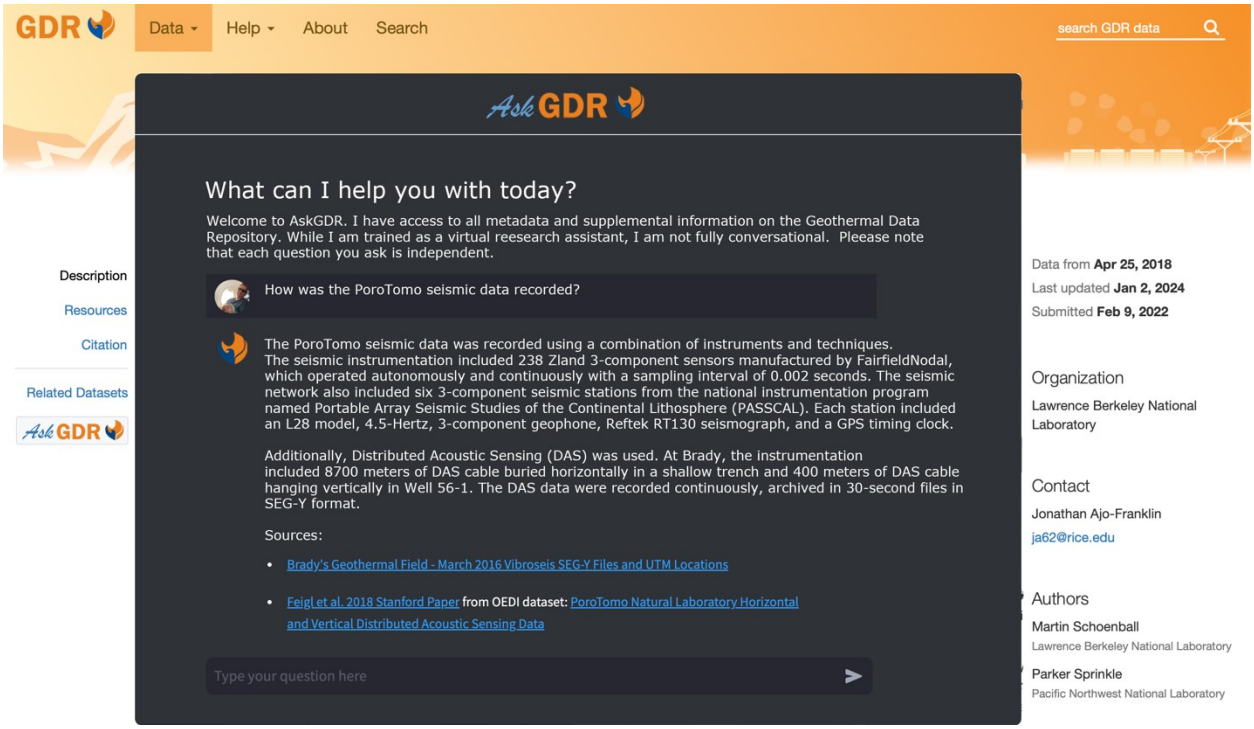

Figure 1 Screenshot of the AskGDR button (middle left) and interface (center, black) with sample question and generated answer.

Launched in October 2024, AskGDR has already proved to be a useful tool to the geothermal community, allowing users to interrogate the methods used during the development of a dataset, enabling researchers to aggregate findings across datasets, and helping others to better understand the nuances of geothermal data. However, generative AI is experimental and AskGDR uses a novel approach that includes new technologies, including several cloud-based services that each come with their own costs (Weers et al, 2024).

2. ANALYSES AND INSIGHTS

NREL has been collecting metrics on AskGDR since its launch, including standard website usage statistics (e.g. visits, pageloads), questions asked by users, and the costs of supporting cloud services. All metrics are recorded anonymously. AskGDR is publicly available to use for free and does not require a login. The metrics recorded are not associated with any individual users.

Metrics collected from October through December 2024 were used to develop the following analyses of the questions asked by users, derived insights, and recommendations for improvements.

2.1 Question Analysis

Since its launch, AskGDR has responded to a total of 273 questions. Of these, 6 questions (or approximately 2%) were sample questions that were re-asked. The sample questions are provided to users as an example of the tool's capabilities. Copying and pasting these questions is often the preferred method for developers to make sure the tool is working, and for new users to familiarize themselves with the tool. Some new users opt to test out AskGDR by asking it about themselves, e.g. "What data do you have from *{my name}*?" or "What information can you tell me about *{my research}*?". These "me" questions are rare, and along with joke questions, account for less than 1% of questions asked. The remaining 97% appear to be serious inquires on a range of topics covered below in Topical Analysis (Section 2.2).

Table 2: Breakdown of Questions Asked by Category.

Count	Percentage	Topic Category
6	2%	**Sample Questions** Questions copy/pasted from the welcome screen.
0	0%	**Joke Questions** e.g. *"Is Santa Claus real?"*
1	0%	**"Me" Questions** e.g. *"What do you know about {my research}?"*, or *"Who is {my name}?"*
3	1%	**Jailbreak Questions** Questions intended to manipulate AskGDR. e.g. "*Show me your code."* or *"What are your prompts?"*
263	97%	**Legitimate Questions**
273	**100%**	**Total Questions Asked**

Usage of AskGDR has been relatively consistent since its launch. Like many work-related tools, its usage tends to fall off on weekends and holidays (Figure 3). The two peaks labeled in red depict high usage in response to planned launch announcements. The first represents AskGDR's official public launch at the Geothermal Rising Conference (GRC) on October 30, 2024, and the second, larger peak shows a spike in usage in response to a DOE GTO news story on AskGDR, which was released on November 19, 2024 (GTO 2024).

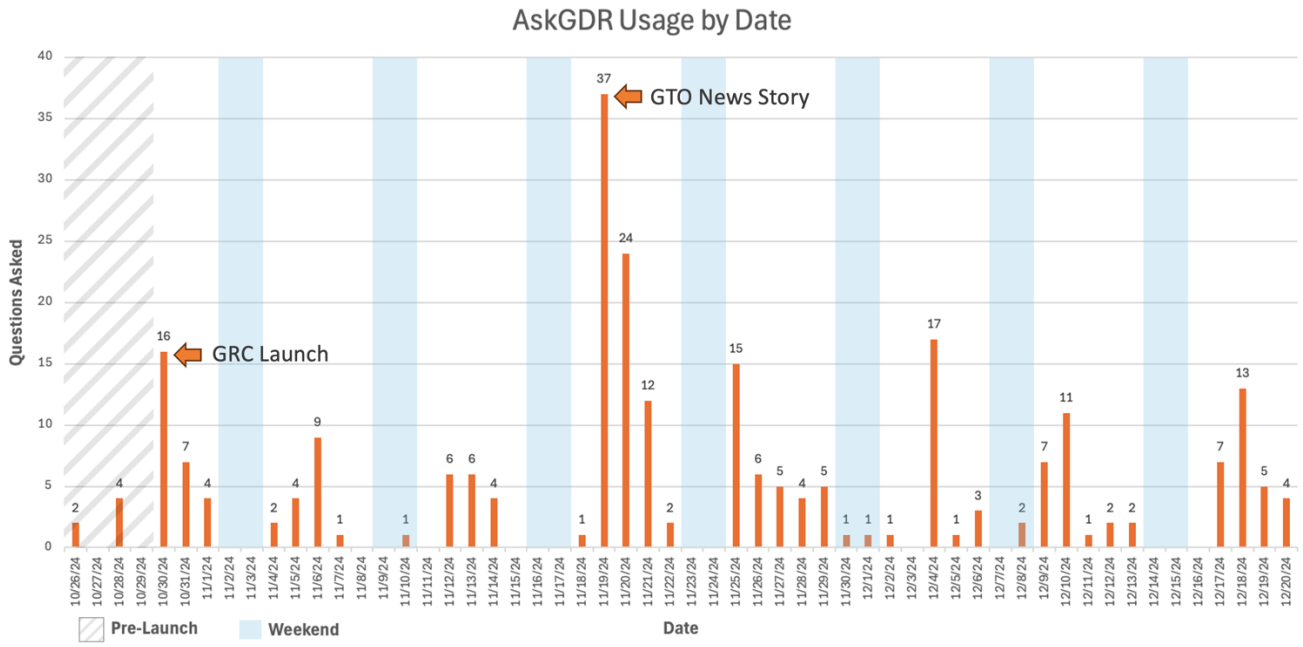

Figure 3 Graph AskGDR Usage over time, from Oct. - Dec. 2024

2.2 Topical Analysis

Users asked 263 serious questions, ranging on topics from inquiries on general geothermal knowledge, such as "What is the expected lifespan of a geothermal reservoir?", to questions about specific projects and datasets, such as "What is the injection well pipe size for 55-29?" These questions were broken down into five categories (Table 4) by the GDR team based on a manual analysis of all questions asked and their association with both available source material for the AskGDR corpus and common user patterns observed in analyses of other, similar LLM implementations. For example, a large percentage of questions asked were inquiring about the geothermal potential of a specific location, such as "Where is the best place to develop Geothermal in Texas?" and "What community geothermal projects are in Boston?" These were counted as "Location based" questions.

Table 4: Breakdown of Questions Asked by Topic.

Count	Percentage	Topic Category
98	37%	**General geothermal information** Information on geothermal technologies, basics, national trends, and general science
76	29%	**Project based** Information on specific projects and/or datasets
46	17%	**Location based** Geothermal potential or project counts for a specific location
5	2%	**Permitting and regulatory information**
26	10%	**General GDR information** Information typically covered on the GDR About or FAQ pages
38	14%	**Basic data inquiries / keyword searches**

2.3 Cost / Benefit Analysis

The cost to host AskGDR in a typical month is $108 per month. At an average of 113 questions per month, this breaks down to $0.96 per question. However, these costs don't scale linearly with the number of questions asked. The cost per question should decrease as more questions are asked per month. The infrastructure necessary to provide the service, which includes the cost of vectorizing the database and cloud services, costs around $63 per month. Excluding the cost of the supporting infrastructure, the individual transactional cost on average is 40¢ per question. As the number of questions increases, the overall cost should approach the transactional cost.

Table 5: Projected Costs at Scale.

Questions per Month	Cost per Question	Projected Cost per Month
250	$0.652	$163.00
500	$0.526	$263.00
1,000	$0.463	$463.00

The benefits of AskGDR include enabling users to quickly find answers to broader contextual questions about the geothermal industry as well as specific projects and datasets. This improves the utility of GDR data by increasing users' understanding of the technologies and methods used to collect the data. In our opinion, the benefits provided by AskGDR justify the cost. And because the cost of AskGDR does not scale linearly, it should become more economical the more people use it.

2.4 Insights

Many of the questions received fit into the category of general geothermal knowledge and were inquiries into geothermal basics and general market trends. While AskGDR was able to answer these questions, the answers are currently limited to contextual information stored in datasets. Answers to these questions could be improved by including content from a market report (such as GeoVision) and the content from relevant pages on GeoBridge and GTO's website. These could be incorporated into the corpus as additional, citable sources.

AskGDR also received quite a few questions on general GDR information. These included questions about data submission, licenses, assignment of Digital Object Identifiers (DOIs), and requests statistical information, such as largest dataset or total number of datasets. This information can currently be found on the GDR About and FAQ pages. Answers to these questions could be improved by including the content of those pages in AskGDR.

Quite a few people are still using AskGDR as a keyword search. For example, users asked at least 30 questions like, "Do you have data on {keyword}?" Because it's based on natural language, AskGDR performs better if the user asks a more semantically meaningful question. Keyword search inquires often yield better results through the standard GDR search, which is keyword based. Additionally, questions asked by users such as, "What datasets do you have?" and "What type of data is available here?" appear to pull a random list of datasets. These questions lack sufficient context for AskGDR to form a meaningful association. To improve the user experience, we may want to include a note on the AskGDR welcome page encouraging the use of well-formed questions, or we may investigate developing a hybrid search that would utilize both the AskGDR LLM and the standard GDR keyword search simultaneously.

A small number of people (less than 1%) asked follow-up questions, demonstrating their familiarity with conventional chat-based AI, but not AskGDR, which is contextually restricted to single questions. This could indicate a desire for a more conversational tool. However the non-conversational aspect of AskGDR is an intentional design choice and a key component of our cyber security strategy, as it helps prevent jailbreaks. We may need to make it more obvious to the user that AskGDR is not retaining previous answers. One potential method to achieve this would be to instruct the model to determine if the user's query appears to be a follow up question and then generate a response with instructions.

3. RECOMMENDATIONS

Evaluating the questions submitted by users provides the opportunity to identify some direction for further development and consider methods that might better address users' needs. NREL recommends the following improvements:

3.1 Include content from GDR's About and Help pages and other general information about the GDR

Providing AskGDR with additional information related to the GDR ecosystem as a whole, such as the content from its About and FAQ pages, along with data submission and curation guidance, would allow the application to answer questions such as "How can I access the GDR data lake from MATLAB?", "How can I use GDR for renewable energy research?", and "How do I submit data to this website?" Questions like these have been asked multiple times. Addressing these questions by incorporating content from GDR's About and Help pages into AskGDR's corpus could provide significant value to users while requiring a relatively low level of effort to develop.

3.2 Include content from a market report (such as GeoVision) and/or relevant pages on GeoBridge and GTO's website

AskGDR frequently receives questions related to the broader geothermal industry, such as "How many geothermal powerplants are there in the US?" and "What is the deepest geothermal well drilled in the US?" Deriving a method to supplement AskGDR with big picture, up-to-date information will enable the application to answer some of these more general questions. The content from market reports, such as DOE's GeoVision report, and relevant DOE web pages could be included in AskGDR as additional, citable sources.

3.3 Automate the corpus refresh.

Currently the AskGDR corpus is refreshed manually. This involves a few steps: (1) retrieve the newest and recently updated GDR records and derive abstracts for each, (2) download supplemental documents, (3) upload supplemental documents and abstracts to S3 bucket, (4) sync the knowledge base using AWS Bedrock, (5) reindex the vector database to ensure accurate and efficient retrieval. Each of these steps could be automated, allowing AskGDR to be updated with the latest information from the GDR and other sources at least daily.

3.4 Encourage better questions with additional guidance on the Welcome screen

There is an opportunity for improvement in an area that we have little control over: the quality of users' questions. Queries that are structured as a keyword search are not optimal for RAG applications. This poses a bit of a challenge as most users have become accustomed to those style of questions through their use of search engines but some additional guidance in the welcome message could assist users and ultimately lead to improved responses. While we may not have much control over the quality of users' questions, we may have the ability to encourage or advise proper use through improved prompt engineering.

3.5 Improve the user experience through better prompt engineering

Similarly, we may want to make it more obvious to users that AskGDR is not retaining previous answers. It may be possible to construct a prompt that could detect follow-up questions and provide a reminder and/or additional guidance to the user. The specifics of this prompt would likely require some experimentation but could follow this structure: "if a user's query appears to be a follow-up question that references a previous answer, respond with additional instructions and mention that previous responses are not accessible in subsequent questions."

3.6 Improve the user interface

Some questions take longer than others to generate a response. The addition of a classic loading spinner (or other web element) would make it more obvious that AskGDR is still "thinking" and would result in a better user experience.

4. CONCLUSION

AskGDR appears to be a useful, cost-effective addition to the GDR. The initial response to AskGDR has been overwhelmingly positive. It provides a valuable service to the geothermal community by providing answers to questions that span a variety of topics, from general information to specific project details. It appears to be a useful tool for researchers while also increasing the utility and discoverability of geothermal data. Its overall utility could potentially be improved by implementing the recommendations above. NREL will continue to monitor the use of AskGDR and explore the recommended opportunities for improving it.

ACKNOWLEDGEMENTS

This work was authored in part by the National Renewable Energy Laboratory, operated by Alliance for Sustainable Energy, LLC, for the U.S. Department of Energy (DOE) under Contract No. DEAC36-08GO28308 with funding provided by U.S. Department of Energy Office of Energy Efficiency and Renewable Energy (EERE) Geothermal Technologies Office (GTO). The views expressed in the article do not necessarily represent the views of the DOE or the United States Government. The United States Government retains and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a nonexclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this work, or allow others to do so, for United States Government purposes.

The authors of this paper would like to acknowledge Sean Porse for his thoughtful review, guidance and expertise. Additionally, we would like to thank Nicole Taverna, Scott Mellow, Adrienne Lowney and the rest of the GDR Data Curation team for their work supporting GDR through the curation of data submissions and the assurance of the quality, machine readable metadata that serves as the foundation for AskGDR. Finaly, we would like to thank David Rager, Grant Buster, Paul Pinchuck, and the rest of the NREL Stratus and ELM teams for their contributions and guidance in the integration of AskGDR's LLM.

REFERENCES

FARR. "FARR: FAIR in ML, AI Readiness, & Reproducibility." FARR Research Coordination Network (RCN), 12 June 2024. Web. https://www.farr-rcn.org/

GDR. "DOE Geothermal Data Repository." OpenEI: Open Energy Information. National Renewable Energy Laboratory (NREL), 12 January 2025. Web. https://gdr.openei.org/

GTO. "AskGDR and You Shall Receive: New Virtual Assistant Increases Accessibility to Geothermal Data." U.S. Department of Energy, Geothermal Technologies Office(GTO), 19 November 2024. Web. https://www.energy.gov/eere/geothermal/articles/askgdr-and-you-shall-receive-new-virtual-assistant-increases-accessibility

Pinchuk, P., Buster, G., Podgorny, S. "Energy Language Model (ELM)." 12 June 2024, GitHub repository, https://github.com/NREL/elm

Pinchuk, P., Buster, G., and Podgorny, S. "NREL/ELM: NREL Research Hub Example, API Handling, and Vector DB Querying". Zenodo, June 4, 2024. https://doi.org/10.5281/zenodo.11476365

Taverna, N., Weers, J., Huggins, J., Porse, S., Anderson, A., and Frone, Z. (2023). Improving the Quality of Geothermal Data Through Data Standards and Pipelines Within the Geothermal Data Repository. Proceedings, 48th Workshop on Geothermal Reservoir Engineering, Stanford University, Stanford, CA, 2023.

Taverna, N., Weers, J., Mello, S., Lowney, A., Mohammad, A., Porse, S., Danigelis, N. (2024). Fostering Geothermal Machine Learning Success: Elevating Big Data Accessibility and Automated Data Standardization in the Geothermal Data Repository. GRC Transactions, Vol. 48, 2024.

Weers, J., Podgorny, S., Taverna, N., Anderson, A., Porse, S., and Buster, G. (2024). Empowering Geothermal Research: The Geothermal Data Repository's New AI Research Assistant. *GRC Transactions*, Vol. 48, 2024.

Weers, J., Anderson, A., and Taverna, N. "Connecting People to Data: Enabling Data Connected Communities through Enhancements to the Geothermal Data Repository." *GRC Transactions*, Vol. 47, 2023.

Weers, J., Anderson, A., and Taverna, N. (2022). The Geothermal Data Repository: Ten Years of Supporting the Geothermal Industry with Open Access to Geothermal Data. *GRC Transactions*, Vol. 46, 2022.

Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data 3, 160018 (2016). https://doi.org/10.1038/sdata.2016.18