Application of generative artificial intelligence in the petroleum field—Automatic generation of seismic acquisition design reports

Wei Jin[a,b], Xiangyun Guo*[c], Shijia Gu[b]
[a]Polytechnic Institute, Zhejiang University, Hangzhou 310015, Zhejiang, China; [b]Kunlun Digital Technology Co., Ltd., Beijing 102200, China; [c]PipeChina Digital Co., Ltd., Beijing 100013, China

ABSTRACT

In the field of geophysical exploration, writing seismic acquisition design reports is an important yet tedious task. How to improve the efficiency while ensuring the quality has become a significant challenge for seismic acquisition designers. This paper proposes a seismic acquisition design report generation technology solution based on generative artificial intelligence. First, it divides documents into multi-level headings to create a repository of historical design reports and supports convenient report retrieval functions. Then, leveraging the powerful learning and reasoning abilities of the large language model, in-depth analysis of historical design reports is conducted, and new technical reports are automatically generated based on this. This innovative solution greatly improves the work efficiency of seismic acquisition designers. The information collection and report writing work, which originally took more than a week to complete, can now generate a high-quality report in just a few minutes. This technical solution can not only be used for seismic acquisition design but also provides effective exploration and practice for the digital transformation of related industries. In the future, with the continuous development and improvement of technology, this technological solution will demonstrate greater application potential and value in various fields.

Keywords: Seismic acquisition design, automatically generate reports, large language model (LLM), generative artificial intelligence (GAI)

1. INTRODUCTION

In the process of oil and gas exploration and development, seismic acquisition design is crucial. At present, the writing of seismic acquisition design reports largely relies on engineers' research and understanding of historical material and personal experience accumulation. When writing acquisition design report for new work area, due to the different experiences and reference materials of each engineer, there are significant differences in the quality and efficiency of report writing, which undoubtedly brings many challenges and troubles to seismic acquisition design work. Therefore, developing an efficient and accurate technical solution has become an important task that urgently needs to be solved. This solution will use automated methods to generate seismic acquisition design reports to assist seismic acquisition designers in improving writing efficiency and quality.

The technology of automatic report generation has been widely studied and applied in fields such as healthcare[1-11], construction[12-14], water conservancy[15,16], finance[17-19], etc. However, due to the unique characteristics of seismic acquisition design reports—which differ from the abundance of image samples in healthcare or the fixed indicators and templates in finance—these existing methodologies cannot be directly applied to the field of geophysical exploration. In view of this, this paper proposes a seismic acquisition design report generation technology solution based on generative artificial intelligence technology. This solution combines multidisciplinary knowledge such as artificial intelligence and natural language processing, aiming to assist seismic acquisition designers in completing report writing work more efficiently and accurately.

2. TECHNOLOGY SOLUTION

The solution proposed in this paper for generating seismic acquisition design reports based on generative artificial intelligence comprises five main modules, encompassing data preprocessing, repository construction, information retrieval, large language model learning, and report generation. An illustration of the technical solution is depicted in Figure 1.

*kelly1141@126.com

International Conference on Optics, Electronics, and Communication Engineering (OECE 2024), edited by Yang Yue, Proc. of SPIE Vol. 13395, 133953Z · © 2024 SPIE · 0277-786X · Published under a Creative Commons Attribution CC-BY 3.0 License · doi: 10.1117/12.3048541

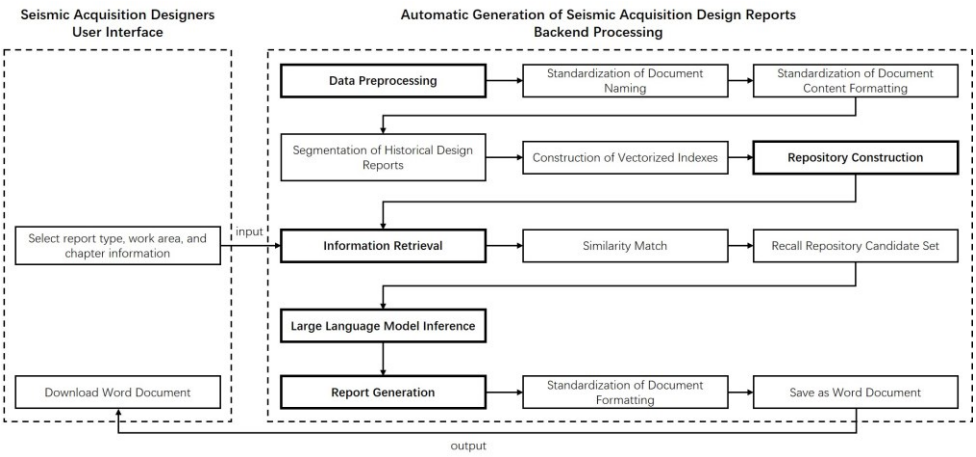

Figure 1. Technical solution diagram for automatically generating seismic acquisition design reports.

2.1 Data preprocessing

The development of artificial intelligence model relies heavily on a substantial amount of annotated data for training and learning. Data preparation consumes over 80% of the time in model development, and the quality of data determines the upper limit for model accuracy[20]. Data serves as the foundation for the entire technical solution; therefore, it is important to construct of a standardized dataset. This module focuses on the initial preprocessing of historical seismic acquisition design reports, aiming to establish a high-quality dataset with standardized formatting to facilitate subsequent model training and learning requirements.

(1) Standardization of document naming

Given that engineers' individual work habits result in inconsistent naming conventions for historical design reports, this step adopts the guidelines outlined in the "Technical Regulations for Onshore Petroleum Seismic Exploration Data Acquisition" and the "Technical Regulations for Onshore Longitudinal Wave Seismic Data Acquisition". Reports are renamed following a standardized format, such as "XX Year XX Basin (XX Region) 2D (or 3D) Seismic Data Acquisition Technical Design/Construction Design/Construction Summary", saved in .docx Word format. This naming convention ensures unity and consistency across the dataset, providing an accurate and clear basis for data analysis and model training.

(2) Standardization of document content formatting

To effectively manage and leverage the content of historical design documents, their formatting must be standardized. This involves ensuring a clear hierarchical structure of headings within the documents and consistently applying built-in heading styles provided by Word (e.g., "Heading 1", "Heading 2", "Heading 3", etc.). Adopting such a practice not only enhances the document's overall appearance, making it look more organized and professional, but also facilitates subsequent automated processing using Python scripts, tasks like content segmentation and index generation become more streamlined and efficient.

2.2 Repository construction

To enhance the efficiency and convenience of information retrieval from historical design reports, this paper proposes a unified report repository. By detailed dividing and organizing the design reports, they can be systematically stored in the repository, enabling efficient centralized management.

(1) Segmentation of Historical Design Reports

Reference[21] utilizes VBA for Word document segmentation[22], uses Python for Excel document handling. In order to improve the automation level of the later process, a multi-level headings splitting strategy is adopted. The built-in docx library in Python is used to split the title and body of the Word version of the historical design report, and these segmentation results are orderly stored in an Excel file for subsequent processing work. An example of historical design report segmentation is shown in Table 1.

Table 1. Example table of historical design report segmentation.

No.	Title	H1	H2	H3	Text	Year	Area	Type
1	2013 Year T basin 3D seismic data acquistion technical design	Overview of the work area	Seismic geological conditions	Surface seismic geological conditions	The surface of the work area is mainly covered by loess, and desert……	2013	T Basin	Acquistion technical design

(2) Construction of Vectorized Indexes

When processing text data, the bag of words model and embedding are two completely different representation methods. Reference[23] uses embedding vectors to label images helps capture deep features in the image and convert them into computable vectors, while[24] uses a bag of words model (BoW) to vectorize the text.

The bag of words model mainly constructs feature vectors based on the frequency of words appearing in the text, ignoring the contextual information and order of words, resulting in generated representation dimensions that are usually high and sparse. This feature makes the bag of words model effective in handling simple text classification tasks, but it appears inadequate in dealing with complex semantic understanding tasks. Taking historical design reports as an example, there are many tags that are semantically similar but have different expressions, such as “seismic exploration deployment”, “seismic deployment”, “exploration deployment”, etc., all of which express the same concept. However, due to its inherent limitations, the bag of words model is difficult to accurately identify the semantic connections between these labels.

In contrast, embedding technology trains a large amount of text data to map each word to a low dimensional dense vector space, which can capture the semantic and syntactic relationships between words. When dealing with tags with significant semantic differentiation and diversity, embedding technology can more accurately construct the vectorized index of tags. Therefore, this paper chooses to use embedding vector technology to construct the vectorized index of titles to ensure the accuracy and efficiency of subsequent information retrieval.

(3) Repository Construction

The construction of the historical design report repository is completed by storing the segmentation design report and the constructed vectorized index together in the database. When user need to retrieve reports, the system provides a flexible query mechanism. Users can not only perform precise queries based on keywords such as report type and work area, but also fully utilize vectorized indexing technology to efficiently retrieve report information similar to query conditions, thereby providing users with more comprehensive and accurate search results.

2.3 Information retrieval

By vectorizing the chapter information fields and indexing them, similarity calculations are performed between these indexes and the pre-existing vectorized indexes within the repository. This process filters out the most similar titles (TopN) to the queried chapter as a candidate set. Subsequently, leveraging the report type, work area, and the selected title candidates, a highly efficient information retrieval process combining both exact and fuzzy matching is conducted against the established Elasticsearch repository. Ultimately, the pertinent information retrieved from the query is returned. Below is an example of an Elasticsearch information retrieval code snippet.

```
body={
   “query”: {
      “bool”: {
         “must”: [
            {“match”: {“area”: area}},
            {“terms”: {Hi: [ES_list]}}]
            }
      }
}
```

2.4 Large language model inference

Through information retrieval of the repository, historical design document information that matches the query criteria can be obtained. Next, process this information through the prompt engineering to construct a format that meets the input requirements of the large language model. This step aims to ensure that the large language model can accurately understand and absorb the key information in these historical design documents.

Once the information is correctly input into a large language model, the model will utilize its powerful learning and reasoning abilities to conduct in-depth learning on these historical design documents. Through this process, the large language model will be able to capture key elements in the document and return the learned content in an easily understandable way, aiming to help users better understand and utilize the knowledge in historical design documents.

2.5 Report generation

After integrating and standardizing the format of the content returned by the large language model, these materials will be stored in Word document format to form a newly generated report. This report will be provided as auxiliary materials to seismic acquisition designers to support their subsequent work and decision-making.

3. KEY TECHNICAL POINTS

3.1 Generative artificial intelligence (GAI)

In the field of artificial intelligence, generative artificial intelligence, as an important technological branch, is gradually showing its enormous potential and broad application prospects. Generative artificial intelligence can learn and simulate the inherent patterns and distributions of data, generating new content that is similar but not entirely identical to the original data. Generative artificial intelligence model can be roughly divided into two categories: probability based model and neural network-based model.

The technical solution for automatically generating seismic acquisition design reports proposed in this paper is mainly based on generative artificial intelligence.

3.2 Large language model (LLM)

The large language model is a natural language processing technique based on neural network, which can learn and predict the patterns of natural language texts. In simple terms, a large language model is an AI program that can understand and generate natural language.

This paper inputs the candidate set of historical design reports retrieved from the repository into a large language model, which learns and infers information to generate technical report content.

3.3 Python

Python is a concise and easy to learn programming language that provides efficient high-level data structures and supports simple and intuitive object-oriented programming. The clear syntax, dynamic typing system, and interpretive language features of Python make it an ideal choice for cross platform scripting and rapid application development.

The technical solution proposed in this paper is mainly implemented through Python language, such as report normalization, report segmentation, large language model learning, report generation, etc.

3.4 Embedding

Embedding is a powerful technique used to transform high-dimensional data (such as text, images, etc.) into low dimensional, continuous vector representations. This transformation process can capture key features of data, making it more efficient in processing, analysis, and machine learning tasks. Embedding technology reduces the complexity of data and computational resource requirements by mapping raw data to low dimensional space, while improving the training and inference efficiency of the model.

This paper mainly uses Embedding to vectorize keywords for easy retrieval of chapter information.

3.5 Elasticsearch (ES)

Elasticsearch[25-30] is an open-source distributed full-text search engine that can quickly and real-time store, search, and analyze big data. It supports multiple data types and has high scalability and fault tolerance. It is a powerful tool for handling massive data search and analysis.

This paper mainly uses ES for repository construction and information retrieval.

3.6 Similarity calculation

Similarity algorithms play a core role in fields such as information retrieval, recommendation systems, and data mining, with the goal of quantifying the degree of similarity between objects. In vector space model, objects are typically represented as vectors composed of a set of eigenvalues. Once an object is converted into a vector, the similarity between objects can be measured by calculating the distance or similarity between these vectors. Several commonly used vector-based similarity calculation methods include Euclidean distance, cosine similarity, Manhattan distance, etc.

This paper uses Euclidean distance to calculate the similarity between the chapter information to be generated and the repository label index, and selects the most similar information as the candidate set for the queried chapter.

3.7 Prompt

Prompt engineering is a key technology in the field of natural language processing, which involves designing precise text or statement fragments (i.e. Prompt) to trigger and guide artificial intelligence language model to produce specific types of output. These prompts can be short words, phrases, questions, or complete sentences used to guide the model in generating specific answers, texts, abstracts, etc.

This paper uses Prompt for large-scale model learning, such as “You are a seismic acquisition design report generator, and you perform the following tasks through relevant descriptions in historical solutions...”

4. PRACTICAL APPLICATION EXAMPLE

This paper takes “Intelligent Assistant for Geophysical Design” as an example to demonstrate the solution of automatically generating seismic acquisition design technology reports based on generative artificial intelligence. The geophysical design intelligent assistant is built using the Streamlit framework, while also utilizing the Elasticsearch database to build a historical design reports repository. Users can easily obtain the required seismic acquisition design report by selecting information such as report type, work area, and desired report structure. The usage process is as follows:

(1) **Set Report:** On the user interface, user can select information such as report type, work area, and report structure.

(2) **Generate Report:** The user can click the “Generate Report” button below to enter the report generation status. During the report generation process, the generation progress will be displayed, for example, when generating information for a certain chapter, the corresponding chapter will be displayed as being generated. A complete technical report is expected to be generated in 3-5 minutes.

(3) **Download Report:** When the report document link is displayed on the page, the report is generated. Click to download the generated report locally.

An example of the interface of geophysical exploration design intelligent assistant is shown in Figure 2.

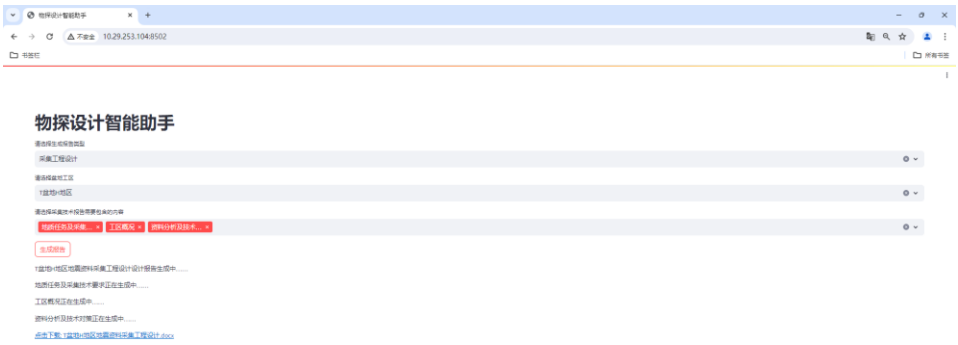

Figure 2. Interface of geophysical exploration design intelligent assistant.

The newly generated report is preceded by the name of the report. According to the navigation bar, it can be seen that the report has generated three chapters of content according to the structure selected in the example, with corresponding section information for each chapter.

For historical design reports with similar content in similar blocks, relevant report chapters will be formed through a large model, and reference information sources will be noted after the chapters. For those without historical reference information, a prompt will be given that there is currently no reference information available.

An example of generating new seismic acquisition design reports based on geophysical design intelligent assistant is shown in Figure 3.

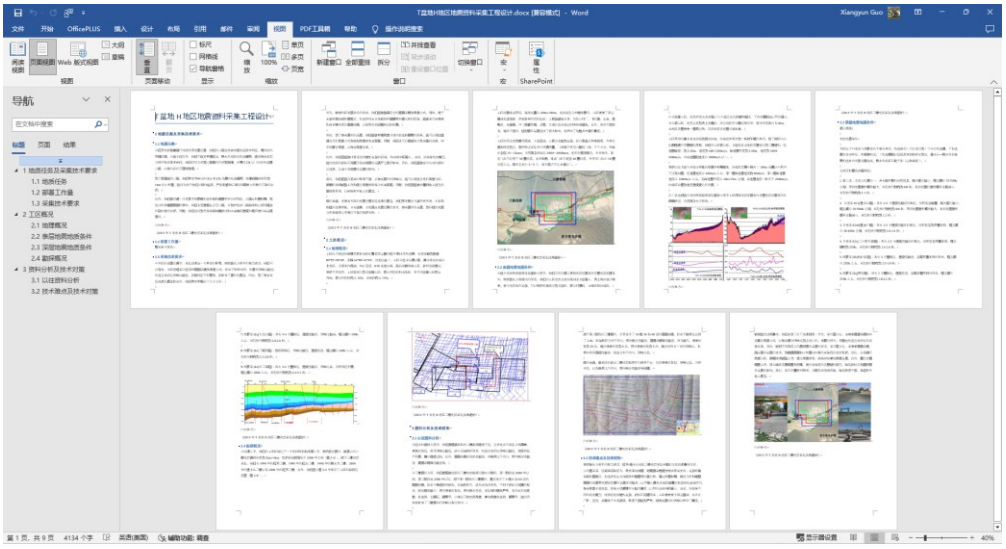

Figure 3. Generating new seismic acquisition design reports based on geophysical design intelligent assistant.

5. CONCLUSION

This paper introduces an innovative seismic acquisition design report automatic generation technology solution, which is based on generative artificial intelligence. Through deep mining and analysis of historical seismic acquisition design reports, combined with the learning and reasoning abilities of large language model, this technology can efficiently and accurately generate technical reports. This solution not only greatly improves the work efficiency of seismic acquisition design personnel, but also ensures the professionalism and quality of the report. In addition, this technological solution also provides valuable practical experience and reference for the digital transformation of related industries, and has enormous application potential and broad development prospects in more fields in the future.

REFERENCES

[1] Han, Q., Zhang, S. J., Tan, L. W. and Li, J. S., "Medical report generation method based on multi-scale feature fusion and cross-training," Journal of Computer-Aided Design & Computer Graphics, 36, 1-11 (2024).

[2] Jiang, W., [Research on Medical Report Generation Based on Transform], Hangzhou Dianzi University, Hangzhou, Zhejiang, China, Master's Thesis, (2023).

[3] Zhang, J. S., Cheng, M., Shen, X. X., Liu, Y. X. and Wang, L. Q., "Diversified label matrix for medical image report generation," Computer Science, (2023).

[4] Xing, S. X., Fang, J. Z., Qu, Z. H., Guo, Z. and Wang, Y., "Research on automatic generation of multimodal medical image reports based on memory driven," Journal of Biomedical Engineering, 41, 60-69 (2024).

[5] Du, X., [Research on Medical Report Generation Method based on Cross-modal Feature Fusion], Harbin Engineering University, Harbin, Heilongjiang, China, Master's Thesis, (2023).

[6] Chen, W. P., [Research on Medical Report Generation Method based on External Knowledge and Visual Memory], Harbin Engineering University, Harbin, Heilongjiang, China, Master's Thesis, (2023).

[7] Xie, W. Q., Ye, Q. Y., Wang, Y. N., Ge, S., Wang, E. L., Li, H., Qi, H. F., Zhang, H. B. and Yuan, K. H., "Application of artificial intelligence in breast ultrasound report generation," Beijing Biomedical Engineering, 42, 483-487 (2023).

[8] Zhang, J. W., [Research on Automatic Generation Model of Radioactivity Report for Patients with Heart Failure], Taiyuan University of Technology, Taiyuan, Shanxi, China, Master's Thesis, (2023).

[9] Zhang, J. W., [Technology and Application of Knowledge-Driven Multimodal Medical Diagnostic Report Generation], Shandong University, Jinan, Shandong, China, Master's Thesis, (2023).

[10] Cao, Y. M., [Research on Key Techniques for Medical Diagnosis Report Generation Driven by Multi-Modal Data and Knowledge], Shandong University, Jinan, Shandong, China, Ph.D. Thesis, (2023).

[11] Shi, J. Y., Zhang, C., Wang, Y. Q., Luo, Z. J. and Zhang, M. H., "Generation of structured medical reports based on knowledge assistance," Computer Science, 51, 1-13 (2024).

[12] Qin, S. Z., Zheng, Z., Gu, Y. and Lu, Z. X., "Exploring and discussion on the application of large language model in construction engineering," Industrial Construction, 53, 162-169 (2023).

[13] He, Y. B., Chen, S. and Cai, X. Y., "Design of an automatic generation system for drilling core method reports based on Java," Research on Urban Construction Theory, (4), 92-94 (2024).

[14] Wen, X. F., Guo, R. X. and Chen, S., "Application of automatic report generation technology in foundation pile drilling core inspection," Intelligent City, (11), 123-125 (2023).

[15] Dai, G. X., Liang, C. J. and Yin, J. B., "Construction and application of an automatic generation system for water conservancy engineering inspection reports," Shandong Water Conservancy, (3), 7-9 (2023).

[16] Dai, L., Li, S. L., Peng, S. C., Liu, G. B. and Ji, C. B., "Automatic generation technology of water conservancy engineering safety monitoring reports based on template customization," Water Resources and Hydropower Express, 44, 117-121 (2023).

[17] Zhong, C. P., Gai, Z. H., Gu, R. Y. and Zhang, X., "Research and application of automatic generation mechanism for management accounting reports," Finance & Accounting, (3), 54-58 (2023).

[18] Cheng, P., Zhu, Z. Y. and Fu, Y. C., "Research on intelligent financial reporting based on ChatGPT," Finance and Accounting Monthly, 44, 64-69 (2023).

[19] Zhang, Z. G., Zhang, Z., Zhang, S. H. and Ji, F. J., "Exploration of the application of generative financial model in financial reporting system," Financial Supervision, (6), 99-104 (2024).

[20] Xie, X. H. and An, P., "Research and practice of CNOOC AI application," Petroleum Science and Technology Forum, 42, 22-29 (2023).

[21] Fan, X. Y. and Zhou, T. F., "Word document using VBA to achieve mass split and merge," Computer Knowledge and Technology, 7, 1554-1556 (2011).

[22] Zhou, Y. X., "Design and implementation of excel processing program based on python," China Computer & Communication, (23), 85-87 (2019).

[23] Qi, H. Y., Shi, W. M. and Li, X. L., "Exploration of technology application based on image search engine and vector database," China Newspaper Industry, (5), 38-40 (2024).

[24] Wang, Y., Chen, X. and Gao, Y. B., "Development and application of enterprise office document retrieval system based on python," China Computer & Communication, (2), 126-129 (2021).

[25] Tian, L. N., "Design and implementation of a search engine based on ElasticSearch," Wireless Internet Science and Technology, (23), 64-67 (2023).

[26] Chen, K. X. and Lin, N., “Design and development of web disk search system based on Elasticsearch,” Proceedings of the 2022 China University Computer Education Conference, (2022).
[27] Liu, Y., Liu, F. and Meng, J., “Design and implementation of a technology service recommendation system based on ElasticSearch,” Gansu Science and Technology, 40, 59-64 (2024).
[28] Wang, R., Hu, W. G., Hu, S. S. and Zhou, Y., “Design and implementation of medical data retrieval system based on ElasticSearch,” Information Technology, (4), 76-82 (2024).
[29] Li, M. K. and Wen, L., “Design and implementation of knowledge base retrieval engine system based on Elasticsearch,” Software, 44, 184-186 (2023).
[30] Dong, Y. H., Jia, Y., Zhu, Y., Li, E. Z. and Xue, X. H., “Research on information retrieval methods based on ElasticSearch distributed search engine,” Journal of Hubei Normal University (Natural Science), 43, 56-61 (2023).