Advancing geologic document digitalization and information retrieval with generative AI

Song Hou[1], Tianhao Dong[1], Ojasvi Sancheti[1], and Hao Liu[2]

https://doi.org/10.1190/tle44020108.1

Abstract

This paper demonstrates how generative artificial intelligence (AI) enhances geoscientific document processing by improving text analysis, table extraction, and figure classification. Traditional workflows struggle with domain-specific terminology, poor-quality inputs, and rare formats. To address these challenges, we employ domain fine-tuned bidirectional encoder representations from transformers (BERT) models to enhance text processing. Additionally, we utilize multimodal large language models for precise table recognition and context-aware image classification. Finally, a domain-optimized retrieval system, GeoRAG, improves the relevance and accuracy of information retrieval. These AI-driven advancements streamline digitalization, enhance data extraction, and enable efficient handling of complex geoscientific documents. While challenges such as hallucinations, interpretability, and output consistency remain, this study highlights the transformative potential of generative AI for geoscience workflows and decision-making processes.

Introduction

Geoscience activity generates vast amounts of data including geologic reports, well logs, core-sample analyses, and geologic maps. Much of this valuable information is contained within legacy formats and low-quality scans, making it difficult to extract and process meaningful insights. Traditional methods of extracting this information rely heavily on manual data entry or rule-based approaches, which are inefficient and prone to errors.

With advancements in machine learning (ML) technologies, large-scale processing of geoscientific documents has become more feasible (Lun et al., 2022a). ML has automated many key workflow steps for analyzing the text, tables, and images embedded in these documents. However, in real-world production environments, these solutions still face challenges:

- Domain adaptation: ML models often face challenges in processing geoscience content due to its specific and complex nature, which requires customized preprocessing, tailored feature engineering, and expert domain knowledge.
- Data collection: The process of collecting and labeling geoscience data is labor intensive and usually results in imbalanced training data sets with potential human error, which impacts model performance.
- Data quality: Low-quality inputs and out-of-distribution cases in real production settings limit model reliability. More sophisticated solutions and quality-control measures are required.

In recent years, generative artificial intelligence (AI) has emerged as a promising tool for transforming document and knowledge processing, especially in specialized fields like geoscience. Generative AI refers to AI systems that can create new content such as text, images, or even videos. This technology has the potential to revolutionize how we handle and interpret large volumes of data. Vision foundation models are a subset of generative AI designed to process and understand visual data. These models, such as SAM (Kirillov et al., 2023) and SegGPT (Wang et al., 2023), are trained on vast amounts of image data, enabling them to perform tasks such as grain segmentation and mineral classification in thin section images (Vellappally et al., 2024). The term "foundation model" refers to their ability to generalize across a wide range of tasks after being trained on diverse data sets.

Language models, such as bidirectional encoder representations from transformers (BERT) (Devlin et al., 2019), generative pretrained transformer (GPT) (Radford et al., 2018), and, more recently, large language models (LLMs) and multimodal LLMs like GPT-4v and Claude 3.5, are another type of generative AI that has significantly improved the automation of information extraction, classification, and retrieval from unstructured data. These models excel at handling domain-specific language and generating structured outputs with minimal manual intervention, reducing the need for extensive data labeling and enabling faster more scalable solutions for geoscientific workflows. For example, Hou et al. (2024) integrated multimodal LLMs for well header analysis and curve digitalization on rasterized well-log documents, significantly streamlining the workflow and enhancing accuracy on the low-quality scans.

This paper aims to demonstrate how generative AI models can be effectively integrated into geoscientific workflows, focusing on geologic documents. First, we will provide a brief overview of the ML-based workflow for document processing and curation. Next, we will demonstrate how generative AI enhances three key components of this workflow: (1) text analysis and enrichment, (2) table recognition and extraction, and (3) figure classification and description. Finally, we will introduce a domain-aware retrieval system specifically optimized for geoscientific data. This study underscores the transformative potential of generative AI in addressing the unique challenges of geoscientific document processing, emphasizing the scalability and flexibility these models bring to real-world applications.

Traditional geologic document ML workflow and its challenges

Before exploring the potential of generative AI in geoscientific workflows, it is essential to review the general ML workflow for

Manuscript received 5 October 2024; revision received 1 December 2024; accepted 13 December 2024.
[1]Viridien, London, UK. E-mail: song.hou@viridiengroup.com; tianhao.dong@viridiengroup.com; ojasvi.sancheti@viridiengroup.com.
[2]Viridien, Calgary, Canada. E-mail: hao.liu@viridiengroup.com.

processing geologic documents, as illustrated in Figure 1. First, a fine-tuned layout detection model is applied to segment and classify document elements such as text, tables, and figures. Each element is then processed through its respective pipeline. Detected text is handled by optical character recognition (OCR) followed by named entity recognition (NER) to extract domain-specific terms such as geologic entities and formations. The text is further classified by geoscience-related topics to enhance searchability. Tables are digitalized, categorized, and stored in SQL databases for further analysis, while figures such as maps, thin sections, and core samples are categorized and processed according to their types. All extracted and processed information is then aggregated and indexed into a graph database with a predefined schema. Graph databases are designed for managing interconnected data, making it easy to navigate relationships and handle complex queries that can be challenging in SQL databases. Unlike SQL, which may require multiple JOIN operations across many tables, graph databases efficiently handle these queries by directly connecting related data. This is particularly valuable in geoscience workflows, where understanding detailed relationships between geologic entities is crucial.

While the workflow has significantly streamlined geologic document processing, the traditional ML still faces challenges when handling the complex domain-specific nature of geoscientific documents. These ML models often struggle with nonstandard formats, low-quality inputs, and rare cases, which require extensive user intervention. The following sections will explore how generative AI can either replace or complement traditional methods, enhancing both robustness and accuracy in document processing.

Advancing document processing with generative AI in geoscience

Text analysis and enrichment. Traditional text analysis primarily relied on rule-based methods such as string matching and grammatical analysis. However, with the advent of transformer-based models, text processing has been revolutionized, enabling more advanced capabilities. In our work, we utilize both BERT and LLMs to enhance geoscience-related text processing.

We employed the original BERT-based architecture and further pretrained it on an extensive collection of geoscience-related text corpora. Our training corpus included hundreds of millions of words from diverse sources including academic papers, government geologic reports, and extensive in-house proprietary geologic studies. This additional pretraining utilized the masked language model approach, enhancing the model's understanding of domain-specific terminology.

For NER, BERT performs word-level tagging to identify entities within the text. For example, given the input

```
"Sequence NSPg30 is lithostratigraphically equiva-
lent to the upper part of the Maureen Formation.
Examples of Sequence NSPg30 are provided by wells
UK3/25a-4, N35/12-1, N35/8-3."
```

BERT generates the following word-level tagging output:

```
["B-SEQUENCE", "I-SEQUENCE", "O", "O", "O", "O",
"O", "O", "O", "O", "O", "B-FORMATION",
"I-FORMATION", "O", "O", "O", "B-SEQUENCE",
"I-SEQUENCE", "O", "O", "O", "O", "B-WELL", "O",
"B-WELL", "O", "B-WELL", "O"]
```

This output indicates the start ("B-") and continuation ("I-") of entities such as "Sequence," "Formation," and "Well." Note that it requires additional postprocessing to aggregate these tags into usable formats.

While the original BERT model, pretrained on general text corpora, provides a strong foundation for language understanding, it initially struggles with domain-specific contexts where common words may have specialized meanings (e.g., "well," "migration," and "formation"). However, our extensive domain-specific pretraining significantly addresses this limitation. By exposing the model to large volumes of geologic literature, we enable it to effectively capture specialized terminology and context. As illustrated in Figure 2, we use t-distributed stochastic neighbor embedding (t-SNE), a dimensionality reduction technique, to visualize

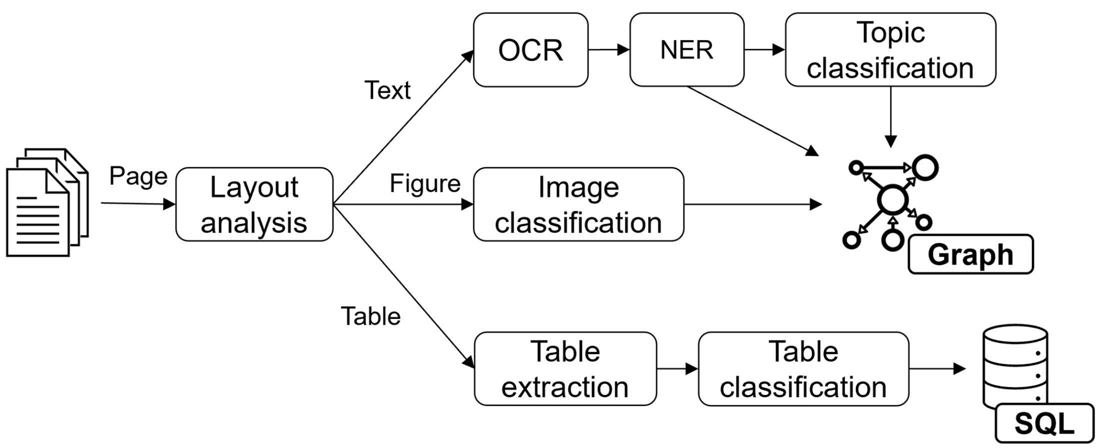


Figure 1. Traditional geologic document ML workflow.

how the model represents geologic terms in a 2D space. The t-SNE plot reduces BERT's high-dimensional word embeddings (768 dimensions) to two dimensions while preserving the relative distances between points, where x- and y-axes represent the arbitrary dimensions that best preserve the relationships between terms in the original space. The visualization demonstrates significantly improved clustering of geologic terms after pretraining, with semantically similar geologic concepts appearing closer together.

While our domain-specific pretrained BERT model handles most geoscience terminology effectively, it may occasionally encounter extremely rare or novel entities not present in the pretraining corpus. In such exceptional cases, LLMs can provide complementary support through their zero-shot and few-shot learning capabilities, though these instances are limited, given our comprehensive pretraining approach.

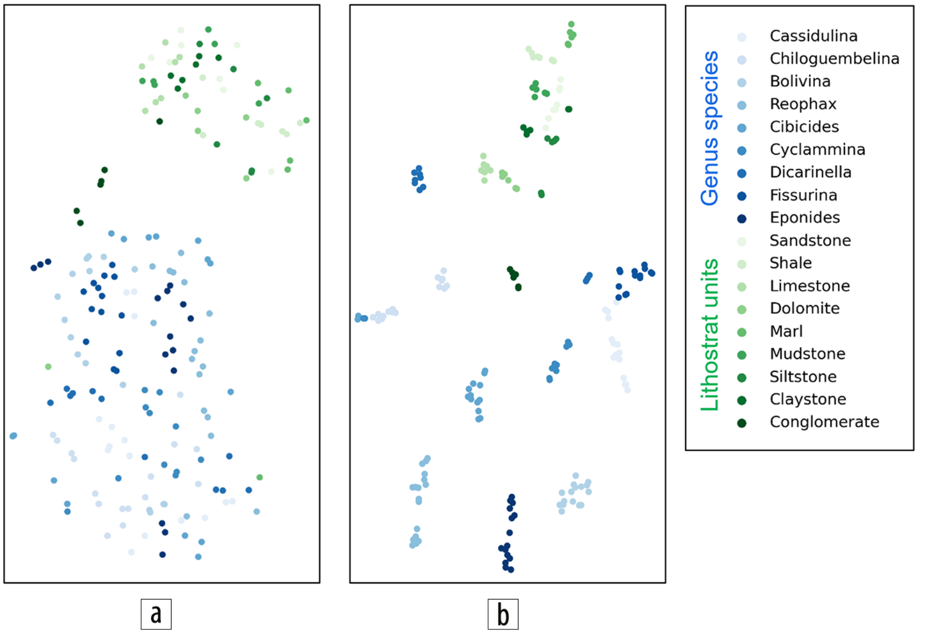


Figure 2. t-SNE visualizations of word embeddings from (a) a pretrained BERT model and (b) the same BERT model after additional pretraining on geoscience-related documents.

For example, when given the same input, LLMs can directly generate structured data in JSON format:

```
{
    "Sequence": ["NSPg30"],
    "Formation": ["Maureen Formation"],
    "Well": ["UK3/25a-4", "N35/12-1", "N35/8-3"]
}
```

In summary, our approach leverages three key methods: string matching, BERT, and LLMs. For entities with a well-defined taxonomy, we employ string matching for straightforward extraction. When sufficient training data are available, we fine-tune a BERT model to handle domain-specific tasks. For rare cases or when data are limited, we rely on LLMs with zero-shot and few-shot learning, providing a versatile solution for managing complex or out-of-distribution entities.

Table recognition and extraction. Tables in geoscientific documents often contain critical data such as well-log measurements, reservoir fluid analyses, and core-sample details. However, conventional ML methods for digitalizing these tables, which rely on table structure detection and OCR (Lun et al., 2022b; Smock et al., 2022), often struggle to handle challenging cases such as rare formats and templates, low-quality scans, and poor handwriting. For example, when digitalizing the table in Figure 3, conventional ML methods recognized the column names and indexes as `["`**`ca. a.`**`", "WELL DEPTH `**`of`**` SAMPLE `**`Na`**`", "LITHOLOGY", `**`None`**`, "FLUO. 3 ML"]` and the first column as `["`**`,`**`","2","`**`2`**`","4","`**`s`**`","`**`&`**`"]`, with errors highlighted in bold. This was due to the poor resolution of the scanned table and handwriting, which made it challenging even for subject-matter experts (SMEs) to interpret accurately, as shown in Figure 3b.

However, multimodal LLMs such as LLaVA (Liu et al., 2023) were able to correctly interpret and extract the data by combining contextual information, prior knowledge, and the table itself. In this example, it accurately identified the column names as `["EQ. NO.", "WELL DEPTH or SAMPLE NO.", "LITHOLOGY", "CUT COLOR", "FLUOR. 3ML"]` and the first column indexes as `["1", "2", "3", "4", "5", "6"]`. Comparing the results to those of conventional ML methods, the multimodal LLM demonstrates superior OCR capabilities. For example, it correctly recognized the column name "CUT COLOR," which is challenging even for SMEs to accurately interpret from the handwriting. Furthermore, the multimodal LLM recognized "FLUO" as "FLUOR," showcasing its advanced reasoning and interpretation abilities by understanding that this is an abbreviation for "fluorescence" in the given context. Note that while OCR struggles with poor-quality inputs, it can still be preferable for faster lightweight text extraction for high-quality documents with minimal domain-specific terminology.

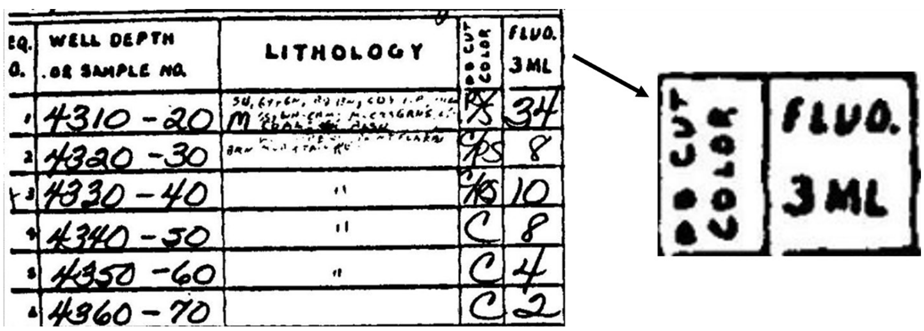
Figure classification and description. Previously, we developed a convolutional neural network (CNN)-based image classifier to categorize figures and images in geoscientific documents (Lun et al., 2022a). Trained on a data set of 20,000 labeled images across nine categories and 29 subcategories, this classifier achieved an impressive average F1 score of 0.97 on validation. However, challenges persist in further enhancing accuracy due to imbalanced data sets and the extensive effort required for data collection and labeling. Common figures, such as maps and satellite images, are relatively easy to classify, whereas rarer types, such as wellbore schematics and engineering diagrams, present more difficulties. Additionally, visual ambiguities frequently arise; sometimes even human experts struggle to make correct classifications without the accompanying context. For example, as shown in Figure 4, the classifier labeled all four images as core plug photos. While Figures 4a and 4b were correctly classified, Figure 4c should have been labeled as a thin section and Figure 4d as a hand-held core sample. Figure 4d is a rare type of core-sample photo, and even experts found it difficult to classify Figure 4c without textual context, highlighting the difficulty in accurately classifying less common geoscientific images by working on the image alone.

To overcome the challenges, we integrate multimodal LLMs into our workflow through a two-step process. First, the multimodal LLM generates detailed figure descriptions that capture relevant details and context. Then, these descriptions, along with the original images, captions, and surrounding text, are fed into an LLM-based classifier. For example, in Figure 4c, the LLM detected the phrase “thin section” in the text, correctly classifying the image as a thin section. Similarly, in Figure 4d, the multimodal LLM recognized the presence of a hand, inferring the scale and correctly identifying the sample as a hand-held core sample. By leveraging multimodal LLMs to extract and integrate textual information, we significantly enhance the accuracy and robustness of our image classification workflow, particularly in complex scenarios where visual cues alone may be insufficient. An additional benefit of this approach is the ability to leverage the generated figure descriptions for semantic search, allowing users to query images using natural language descriptions and significantly enhancing the discoverability and accessibility of complex geoscientific imagery.

Domain-aware information retrieval

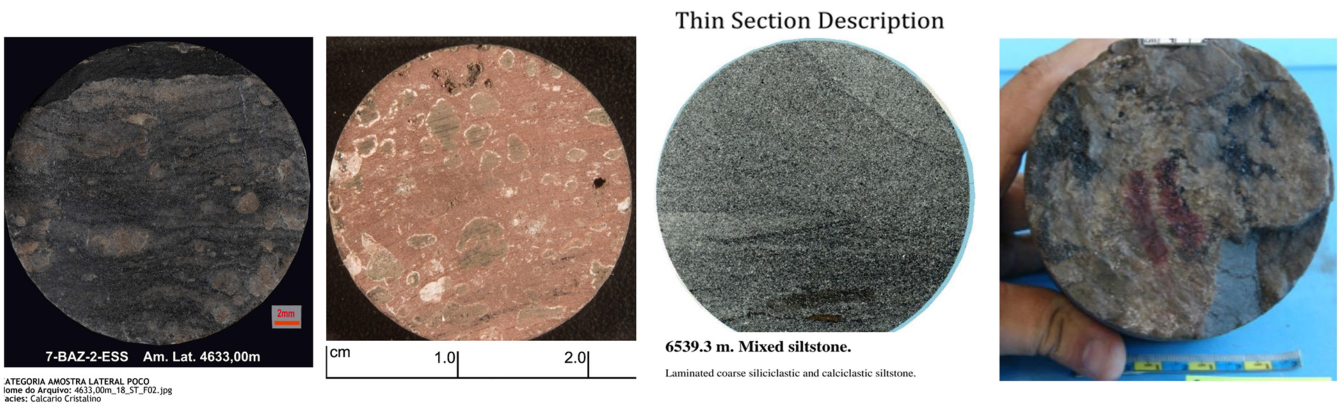
Retrieval-augmented generation (RAG), introduced by Lewis et al. (2020), has transformed semantic information retrieval by employing vector embeddings to represent semantic meaning in text. Relevant texts are retrieved through a similarity search of the embeddings and without the need for predefined database schemas. This approach is particularly valuable in geoscience, where data are often unstructured and diverse. However, standard RAG systems and LLMs struggle with retrieving irrelevant information and generating hallucinations, leading to unreliable outputs. For example, in Table 1, when queried about licenses related to the Nise Formation, a general-purpose LLM like ChatGPT can only provide vague information about the formation’s location, lacking specific license details due to limited training data.

To address these challenges, we developed a geoscience-specialized RAG (GeoRAG) system (Dong et al., 2024) with two key innovations: (1) domain-aware information enrichment via geoscience-specific NER and text classification, and (2) a fine-tuned domain-specific language model optimizing vector embeddings for geoscience context, improving mean reciprocal rank in similarity searches. These



a b

Figure 3. (a) Example of a challenging table with poor scan quality and handwritten entries. (b) Zoom in of difficult-to-read column headers.



a b c d

Figure 4. These four images were all classified as core plug photos by the traditional ML model. While (a) and (b) are correctly classified, (c) should be a thin section and (d) a hand-held core sample.

	(a)	(b)	(c)	(d)
CNN	Core plug	Core plug	Core plug	Core plug
LLM	Core plug	Core plug	Thin section	Core sample
Ground truth	Core plug	Core plug	Thin section	Core sample

enhancements enable more effective filtering and increase retrieval relevance and answer accuracy. When queried about Nise Formation licenses, GeoRAG accesses specialized Norwegian relinquishment documents in the database, providing specific details on licenses like PL1071 and PL1124, including award dates, work programs, and technical evaluations. This capability empowers users to extract meaningful insights from complex geoscientific data.

Further work and exploration could focus on the scalability and robustness of this implementation in a production environment, especially one that contains multimodal data. One challenging aspect is ensuring that semantic meaning is represented consistently across different modalities (e.g., text, images, tables, etc.). Additionally, in cases where vector search is not enough to retrieve the most relevant documents, a cross-encoder model can be used to perform reranking (Reimers and Gurevych, 2019). Finally, user feedback is crucial in providing insights into user behavior and expectations and in helping us to resolve gaps in our understanding of some domain-specific terms and entities. This information could then be incorporated into further improvements across the various tasks mentioned earlier.

Challenges and limitations

As illustrated by the previous examples, generative AI has demonstrated significant potential in streamlining overall geoscience document processing workflows by improving each component with a tailored domain-specific solution. However, scaling these solutions to meet the accuracy and robustness required for production environments presents several challenges. One primary concern is maintaining coherent outputs, which can be inconsistent due to the probabilistic nature of language modeling. This often results in hallucinations and difficulty in reliably following complex instructions, such as adhering to specific formats or structured outputs. These issues can significantly undermine the reliability and precision that are critical in geoscientific applications.

Fine-tuning LLMs with domain-specific knowledge offers a promising solution, enabling them to better handle specialized geoscientific tasks and terminology. However, this approach comes with its own challenges, particularly the need for vast amounts of high-quality domain-specific data and the significant computational resources required for effective fine-tuning. Furthermore, developing high-quality validation data sets tailored to geoscientific applications is essential. These data sets not only optimize the fine-tuning process but also ensure comprehensive evaluation of model performance, enhancing accuracy and reliability before these systems are deployed in real-world production environments.

Conclusion

In this paper, we explored the transformative potential of generative AI for geologic document processing. Our study covered text analysis, table extraction, and figure classification, demonstrating significant improvements in efficiency and accuracy for document digitalization and information retrieval in geoscience. We also discussed the challenges of applying generative AI in production environments and proposed directions for future work, emphasizing the need for domain-specific optimizations and more robust context-aware models.

Table 1. Comparison of responses obtained from ChatGPT and GeoRAG on a domain-specific question, "What licenses have information about the Nise Formation?"

Question	What licenses have information about the Nise Formation?
ChatGPT	The Nise Formation is a geological formation located offshore Norway in the Norwegian Sea. To provide accurate information about licenses related to this formation, I would need to check the most up-to-date records from the Norwegian Petroleum Directorate (NPD) or other relevant authorities. As of my last update in April 2024, there were likely several production licenses covering areas that include parts of the Nise Formation. However, without access to current data, I can't specify exact license numbers or operators. …
GeoRAG	Based on the provided documents, the following licenses have information about the Nise Formation: • PL1071: This license was awarded in February 2020 with an initial work program of G&G studies and seismic reprocessing. … The decision to relinquish is based on the latest technical evaluations using the reprocessed seismic data, which indicate significant risk on trap integrity for Nise prospects and only prove small upside in the Springar Formation. • PL1124: This license was evaluated using …

In conclusion, the application of generative AI has generated valuable insights such as enhanced document digitalization and improved information retrieval processes. By addressing the identified challenges and continuing the development of specialized AI solutions, generative AI will play a pivotal role in advancing geoscientific research and enhancing decision-making processes in the field. TLE

Acknowledgments

We would like to thank Viridien for permission to publish this work. We would also like to thank our colleagues at Viridien AI Lab and Data Hub for valuable discussions.

Data and materials availability

Data associated with this research are confidential and cannot be released.

Corresponding author: song.hou@viridiengroup.com

References

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova, 2019, BERT: Pre-training of deep bidirectional transformers for language understanding: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics, 4171–4186.

Dong, T., C. Subia-Waud, and S. Hou, 2024, Geo-RAG: Gaining insights from unstructured geological documents with large language models: Fourth Digitalization Conference and Exhibition, EAGE, https://doi.org/10.3997/2214-4609.202439068.

Hou, S., O. Sancheti, R. Diaz, M. Phipps, and M. Iskandar, 2024, AI-powered digitization of extensive scanned well log archives: Presented at the Fourth Digitalization Conference and Exhibition, EAGE.

Kirillov, A., E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, et al., 2023, Segment anything: arXiv: 2304.02643.

Lewis, P., E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, et al., 2020, Retrieval-augmented generation for knowledge-intensive NLP tasks: Proceedings of the 34th International Conference on Neural Information Processing Systems, 9459–9474.

Liu, H., C. Li, Q. Wu, and Y. Lee, 2023, Visual instruction tuning: Proceedings of the 37th International Conference on Neural Information Processing Systems.

Lun, C. H., T. Hewitt, and S. Hou, 2022a, A machine learning pipeline for document extraction: First Break, **40**, no. 2, 73–78, https://doi.org/10.3997/1365-2397.fb2022016.

Lun, C. H., T. Hewitt, and S. Hou, 2022b, Machine learning for table cell classification: 83rd Annual Conference and Exhibition, EAGE, https://doi.org/10.3997/2214-4609.202210426.

Radford, A., K. Narasimhan, T. Salimans, and I. Sutskever, 2018, Improving language understanding by generative pre-training: OpenAI.

Reimers, N., and I. Gurevych, 2019, Sentence-BERT: Sentence embeddings using siamese BERT-networks: Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, 3982–3992.

Smock, B., R. Pesala, and R. Abraham, 2022, PubTables-1M: Towards comprehensive table extraction from unstructured documents: Proceedings of the Conference on Computer Vision and Pattern Recognition, 4634–4642, https://doi.org/10.1109/CVPR52688.2022.00459.

Vellappally, A., S. Hou, and J. Emmings, 2024, Automated grain segmentation and mineral classification in rock thin sections: 85th Annual Conference and Exhibition, EAGE, https://doi.org/10.3997/2214-4609.2024101405.

Wang, X., X. Zhang, Y. Cao, W. Wang, C. Shen, and T. Huang, 2023, SegGPT: Segmenting everything in context: arXiv: 2304.03284.