



OPEN ACCESS

EDITED BY

Yen-Chiang Chang,
Dalian Maritime University, China

REVIEWED BY

Nadia Khadam,
Fatima Jinnah Women University, Pakistan
Wen Duan,
Hainan University, China

*CORRESPONDENCE

Yitong Chen
✉ chenyitong@ouc.edu.cn

RECEIVED 10 January 2024

ACCEPTED 26 March 2024

PUBLISHED 08 April 2024

CITATION

Wang X, Zhang M, Liu H, Ma X, Liu Y and
Chen Y (2024) ChatBBNJ: a question–
answering system for acquiring knowledge on
biodiversity beyond national jurisdiction.
Front. Mar. Sci. 11:1368356.
doi: 10.3389/fmars.2024.1368356

COPYRIGHT

© 2024 Wang, Zhang, Liu, Ma, Liu and Chen.
This is an open-access article distributed under
the terms of the [Creative Commons Attribution
License \(CC BY\)](https://creativecommons.org/licenses/by/4.0/). The use, distribution or
reproduction in other forums is permitted,
provided the original author(s) and the
copyright owner(s) are credited and that the
original publication in this journal is cited, in
accordance with accepted academic
practice. No use, distribution or reproduction
is permitted which does not comply with
these terms.

ChatBBNJ: a question– answering system for acquiring knowledge on biodiversity beyond national jurisdiction

Xiaowei Wang¹, Mingdan Zhang², Hao Liu¹, Xiaodong Ma¹,
Yingchao Liu¹ and Yitong Chen^{2*}

¹College of Computer Science and Technology, Ocean University of China, Qingdao, China, ²Law
School, Ocean University of China, Qingdao, China

The marine biodiversity in Areas beyond national jurisdiction (ABNJ), encompassing approximately two-thirds of the global ocean, is persistently declining. In 2023, the agreement on the Conservation and Sustainable Use of Marine Biodiversity of Areas Beyond National Jurisdiction (BBNJ) was officially adopted. Implementing the BBNJ Agreement has the potential to effectively meet global needs for preserving marine biodiversity. Nevertheless, the implementation requires dealing with thousands of legal clauses, and the parties participating in the process lack adequate means to acquire knowledge connected to BBNJ. This paper introduces ChatBBNJ, a highly efficient question-answering system that combines a novel data engineering technique with large language models (LLMs) of Natural Language Processing (NLP). The system aims to efficiently provide stakeholders with BBNJ-related knowledge, thereby facilitating and enhancing their comprehension and involvement with the subject matter. The experimental results demonstrate that the proposed ChatBBNJ exhibits superior expertise in the BBNJ domain, outperforming baseline models in terms of precision, recall, and F1-scores. The successful deployment of the suggested system is expected to greatly assist stakeholders in acquiring BBNJ knowledge and facilitating the effective implementation of the BBNJ Agreement. Therefore, this is expected to contribute to the conservation and sustainable use of marine biodiversity in ABNJ.

KEYWORDS

BBNJ agreement, ABNJ, LLMS, nlp, intelligent question-answering

1 Introduction

Areas beyond national jurisdiction (ABNJ) face persistent degradation of marine biodiversity (Humphries and Harden-Davies, 2020). A United Nations agreement on the conservation and sustainable use of marine biodiversity in areas beyond national jurisdiction (the BBNJ Agreement) was formally adopted in August 2023 (United Nations, 2023). The

BBNJ Agreement regulates four key elements concerning ocean governance: marine genetic resources; area-based management tools, including marine protected areas; environmental impact assessments; and capacity building and marine technology transfer (Tessnow-von Wysocki and Vadrot, 2020). It will act as a governance mechanism to achieve conservation and sustainable use of marine biodiversity in ABNJ (Tiller et al., 2023).

However, the implementation of the treaty is facing a lot of resistance. First, as a package deal, the BBNJ Agreement provides a framework for reaching consensus; it still lacks specifics, which needs to be discussed in future Conference of Parties (COP) (Deasy, 2023). Second, learning from the implementation of past international law (Bodansky, 2011), industry may seek to weaken implementation measures in order to reduce its adjustment costs. To solve these problems, countries, organizations, and other stakeholders need to submit implementation reports and discuss these issues in the COP. However, many of the stakeholders were not involved in the BBNJ negotiations. They needed to rapidly comprehend thousands of clauses and four interdisciplinary key knowledge points before the discussion. This presents a significant challenge to the stakeholders.

Even though stakeholders can acquire BBNJ-related knowledge through search engines, existing search engines provide information retrieval based on keywords within relevant documents. Users have to deal with the burden of browsing and filtering out results to find the candidate passages (Lau et al., 2005). Thus, providing a convenient information acquisition system is necessary, which is already called for by the BBNJ Agreement (United Nations, 2023). Knowledge question-answering (Q-A) systems take in natural language questions and provide accurate answers, which can reduce the burden of users reading a large number of irrelevant documents to obtain answers (Zhu et al., 2021). At present, it has been widely used in many fields (Zhong et al., 2020; Dai et al., 2022; Lee et al., 2023), but few Q-A systems popularize professional knowledge of international law of the sea. Therefore, this study aims to develop a Q-A system and to efficiently provide stakeholders with BBNJ-related knowledge, thereby facilitating and enhancing their comprehension and involvement with the subject matter, prompting the effective implementation of the BBNJ Agreement.

Researchers can use various methods to develop a Q-A system for BBNJ knowledge popularization. Early Q-A systems heavily relied on rule-based methods (Riloff and Thelen, 2000), in which linguistic experts needed to manually formulate rules based on the characteristics of BBNJ texts. This method lacks generalization and makes it difficult to cover all scenarios. With rapid advancements in artificial intelligence technologies, several Q-A models have been developed applying statistical language models (Rosenfeld, 2000). These models do not require manual rules, they automatically learn statistical language patterns to predict the correct answers to questions. However, this method cannot obtain semantic information. The BBNJ Agreement contains many repetitive terminologies, which makes it challenging for most statistical language models to differentiate subtle situations. Neural language models (Bengio et al., 2003) characterize the probability of word sequences generated by neural networks. These models generate a contextual representation that encodes semantic and syntactic information. However, in addition to repetition, the language of

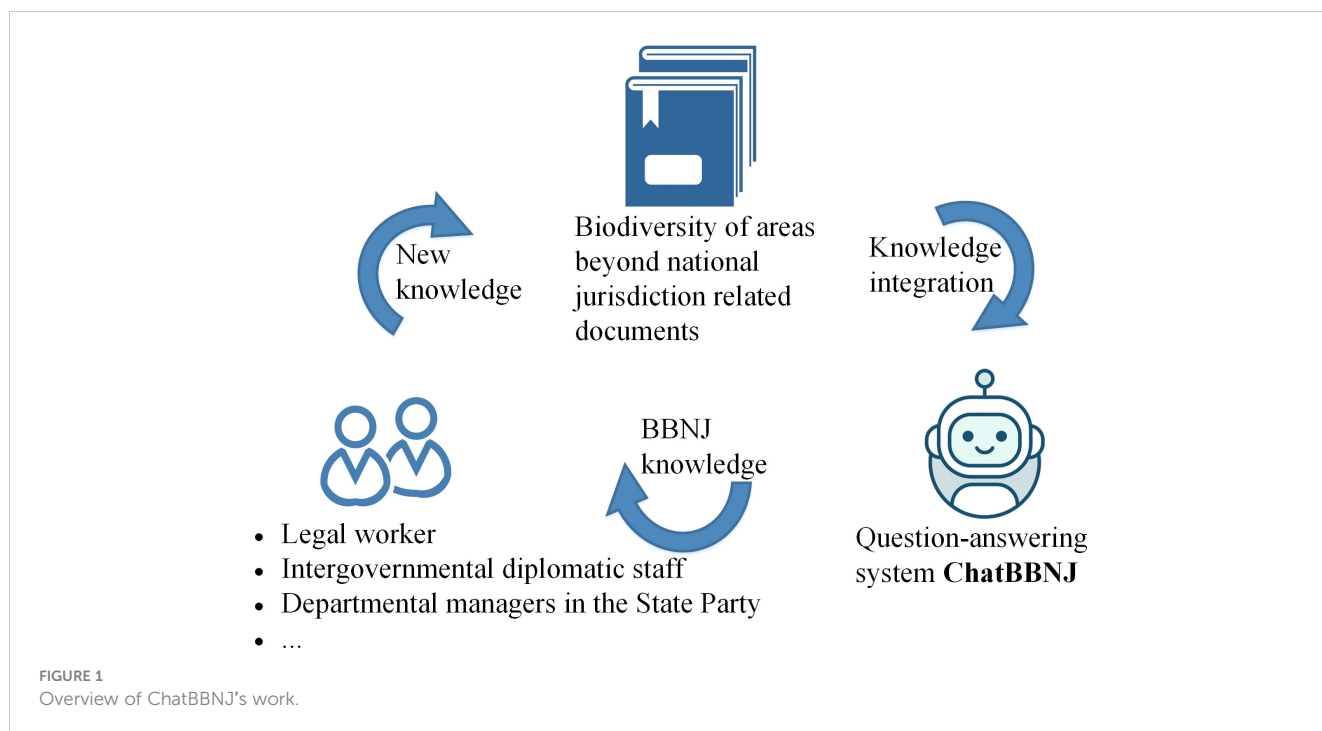
the BBNJ Agreement shows long-term dependency. Limited by the context window, these models cannot model global semantics. Pretrained language models (Devlin et al., 2019) capture context-aware word representations by fine-tuning the networks according to specific downstream tasks. Q-A systems built upon these models exhibit better performance.

Recently, researchers find that scaling pretrained language models often leads to improved model capacity on downstream tasks (i.e., following the scaling law (Kaplan et al., 2020)). With the significant success of large language models (LLMs) like ChatGPT (Ouyang et al., 2022) in tasks related to understanding and generating human-like responses (Eloundou et al., 2023), applying LLMs to Q-A systems has become a popular choice among researchers (Cui et al., 2023; Huang et al., 2023; Li et al., 2023; Vaghefi et al., 2023; Wang et al., 2023; Xiong et al., 2023; Yang et al., 2023). At present, LLMs have significantly improved their performance in open-domain Q-A (Li et al., 2023). However, when applying LLMs to the BBNJ domain, it is difficult to fully utilize its advantages (Wang et al., 2023). LLMs are trained on general corpora, such as Common Crawl (Common Crawl, 2023) and Wikipedia (Wikipedia, 2023). BBNJ-related knowledge is complex and multidisciplinary, involving legal, scientific, and international relations considerations (Humphries et al., 2021). These specialized areas differ significantly from the pre-trained data of LLMs. Therefore, efficiently adapting LLMs to the BBNJ domain, fully utilizing LLMs' understanding abilities remains a challenging problem.

Recent studies have shown (Amer-Yahia et al., 2023) that fine-tuning LLMs by using high-quality domain-specific data can improve their domain adaptation ability. However, developing a Q-A system for BBNJ knowledge popularization by fine-tuning LLMs with domain-specific data presents challenges. First, there is a lack of available Q-A datasets for fine-tuning LLMs in the BBNJ domain. Second, LLMs suffer from outdated information after fine-tuning has concluded. Ensuring stakeholders rapidly acquire the latest implementation situations and recommendations through the Q-A system is urgently needed in practice. Hence, providing accurate and up-to-date responses is paramount. Such accurate responses can help stakeholders understand the complex and dynamically updated BBNJ knowledge and prompt the implementation of the BBNJ Agreement. Thus, this study (Figure 1) aims to solve these problems with the following contributions:

(1) We proposed a data engineering method, called PDGC, to generate a higher-quality Q-A dataset for the BBNJ domain. PDGC contains two-stage data-generation and iterative correction. The two-stage data-generation method enables the model to generate higher-quality data based on BBNJ Q-A examples. Moreover, our iterative correction improves correction quality by following the human learning pattern based on easy-to-difficult.

(2) We developed a BBNJ domain language model called ChatBBNJ, which is fine-tuned by utilizing the United Nations Convention on the Law of the Sea (UNCLOS) and its annexes. BBNJ Agreement is developed under the UNCLOS. Furthermore, we introduced a domain knowledge-based prompt engineering. The domain knowledge base is constructed by utilizing the BBNJ



Agreement. The agreement offers the latest regulations for the management of BBNJ. It ensures that ChatBBNJ obtains the latest BBNJ domain information quickly.

2 Methods

An overview of the proposed method framework for BBNJ-related knowledge popularization is shown in Figure 2. The method framework is composed of three parts: dataset construction, model fine-tuning, and domain knowledge-based prompt engineering. At the dataset construction stage, we applied PDGC to construct a high-quality dataset for fine-tuning. Then, at the model fine-tuning stage, we applied the LoRA (Hu et al., 2021) method to fine-tune the base model, which improved the domain adaptation ability of the base model. Finally, at the domain knowledge-based prompt engineering stage, a domain knowledge base was used when constructing prompts, which ensured the timeliness of the model's answers.

2.1 Q-A dataset construction

The proposed PDGC method comprises three modules: text preprocessing, data generation, and data correction (Figure 3). During the text preprocessing, we performed reference completion as the BBNJ-related documents have many reference expressions. Then, during the data generation, since the BBNJ-related documents are multidisciplinary, involving legal, scientific, and international relations considerations (Humphries et al., 2021), single data generation will result in a significant amount of noisy data. To solve this problem, a two-stage data generation method was applied. Finally, during the data correction, existing consistency

validation methods reduce data quantity and diversity. Therefore, we proposed a similarity-based data division and iterative correction method.

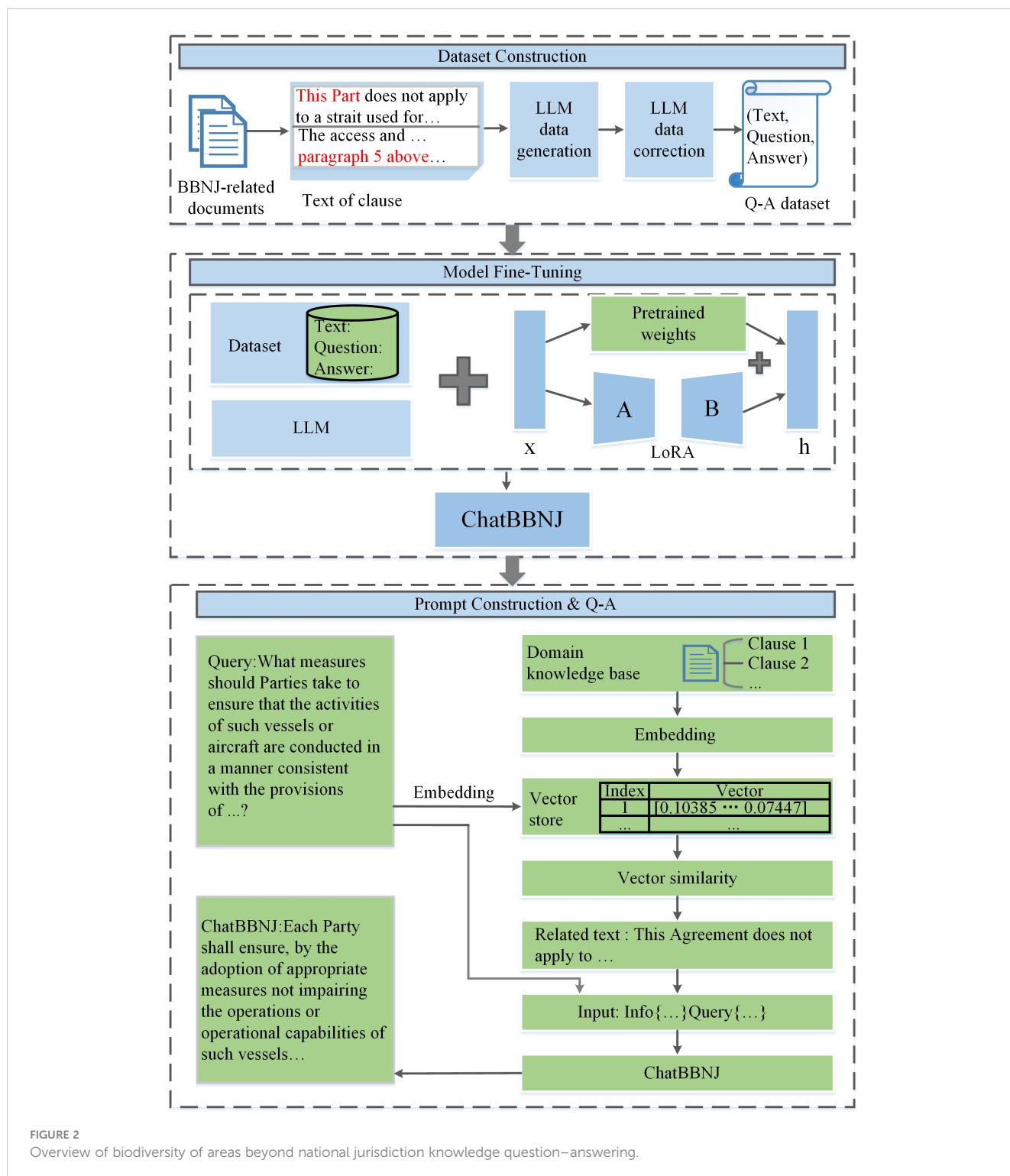
2.1.1 Text preprocessing

BBNJ-related documents were collected to build our dataset. The details are shown in Table 1. In our study, text preprocessing consists of two steps: (1) division of regulations; and (2) reference completion. Division of regulations divides regulations into multiple paragraphs according to their clauses. As some clauses list multiple contents when describing "requirements" and "steps", they often go beyond the input limit of the data generation model. Therefore, we took each part of the clause as a paragraph and supplemented Q-A pairs to ensure completeness. In addition, all the section titles were merged with any sentence within the corresponding section to retain semantic information. Reference completion ensures that each clause retains complete semantic information. Lots of clauses use referential terms such as "above" and "this section" in their descriptions, they lose complete semantic information after the division of regulations. Therefore, we completed the reference to the terms according to the context. Finally, 3,089 BBNJ-related paragraphs were obtained and used for Q-A data generation.

2.1.2 Data generation

Preliminary evaluation using GPT-4 as a judge shows that Vicuna-13B achieves more than 90% quality of OpenAI's ChatGPT (Chiang et al., 2023). Since ChatGPT (Ouyang et al., 2022) is in a non-open source state, we applied open source Vicuna-13B as the base model for data generation to minimize experimental costs.

Prompting is a method for guiding the LLMs toward desired outputs. To achieve the best performance of LLMs in data generation,



proper design of prompts is essential. We designed prompts for generating data at different stages. In the first stage, we used BBNJ-related paragraphs as input to generate Q-A pairs based on the content of the Phase 1 data generation prompt shown in Figure 4. In the second stage, following the concept of in-context learning (Dong et al., 2023), BBNJ-related paragraphs and Q-A pairs generated in Phase 1 were used as input to generate higher-quality Q-A pairs based

on the content of the Phase 2 data generation prompt shown in Figure 5. Although it has been clearly stated in the prompt that LLMs need to generate new questions different from the examples, it was found that there were still occurrences of repetitive Q-A pairs during the experiment. Therefore, after the question generation process, it is necessary to carry out a deduplication operation on the generated Q-A pairs. After deduplication, we obtained 29,273 Q-A pairs, (q, a).

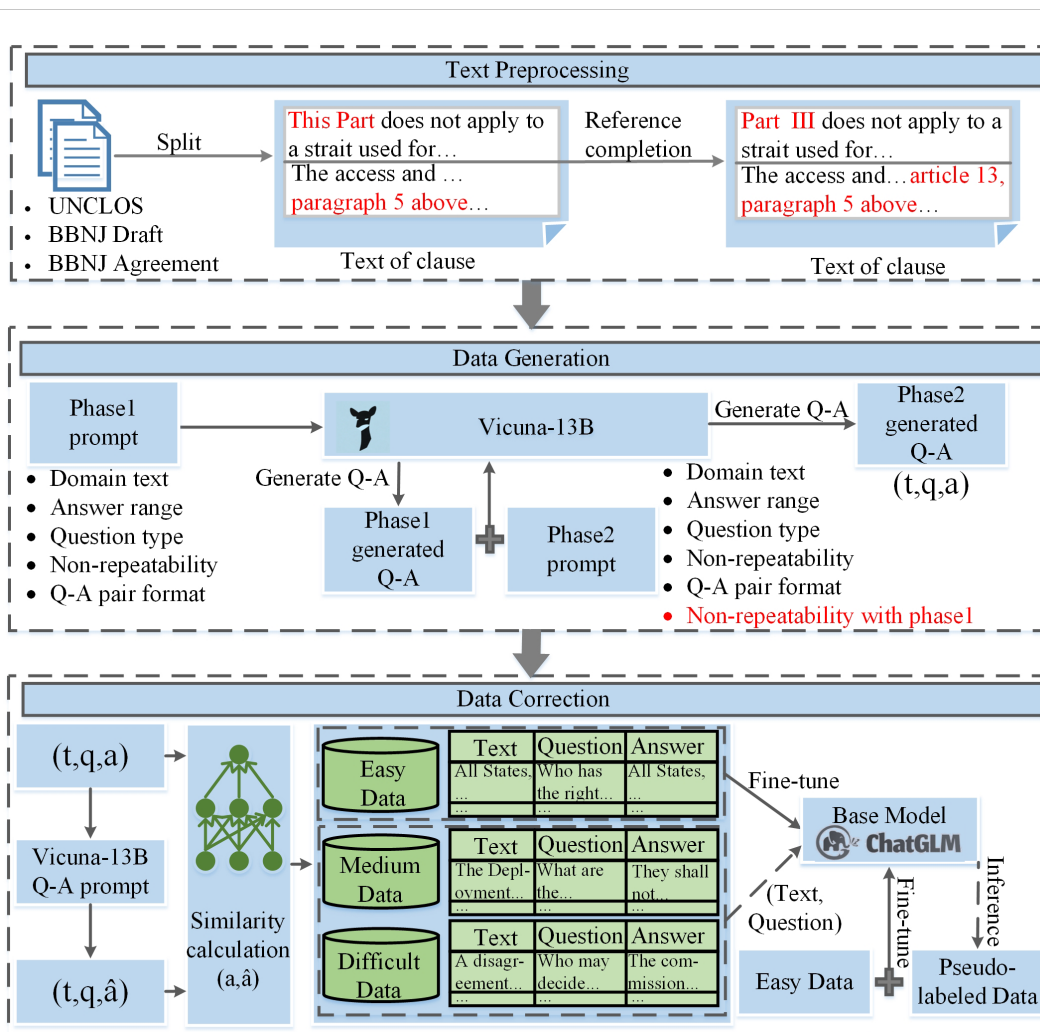


FIGURE 3

Dataset construction method framework. “UNCLOS” represents the United Nations Convention on the Law of the Sea. “BBNJ Draft” represents the draft agreement under the United Nations Convention on the Law of the Sea on the conservation and sustainable use of marine biological diversity of areas beyond national jurisdiction. “BBNJ Agreement” represents the agreement under the United Nations Convention on the Law of the Sea on the conservation and sustainable use of marine biological diversity of areas beyond national jurisdiction. “Vicuna-13B” is a large language model used to generate data. “ (t, q, a) ” contains three parts: “ t ” is the text of clause obtained in text preprocessing, “ q ” and “ a ” represent the question and answer generated by Vicuna-13B in data generation respectively. “ (t, q, \hat{a}) ” contains three parts: “ t ” and “ q ” are the same as “ t ” and “ q ” in “ (t, q, a) ”, but “ \hat{a} ” represents the new answer generated using “Vicuna-13B Q-A prompt” for data correction. “ChatGLM” is a large language model used to revise data.

2.1.3 Data correction

To further improve the quality of the generated data, the data correction module depicted in Figure 6 is applied to revise the generated Q-A pairs. Since the LLMs will generate out-of-scope answers, we obtained new model-generated answers \hat{a} based on the content of the Vicuna-13B Q-A prompt shown in Figure 7. Then, we removed Q-A pairs with “no answer” and obtained new pairs (q, \hat{a}) . The data were divided into three levels: easy, medium, and difficult by calculating the similarity between a and \hat{a} .

To minimize the computational resources used for model fine-tuning during the data correction, ChatGLM (Zeng et al., 2022) is used as the base model for data correction, which has only 6.2 billion parameters. Inspired by (Wang et al., 2021), we optimized the data correction model iteratively by increasing the difficulty of the Q-A pairs fed to the model gradually. First, we

used easy Q-A pairs to conduct the initial fine-tuning of ChatGLM, and used the fine-tuned model to reannotate the medium Q-A pairs. Then, we further fine-tuned the model using the reannotated data and used the model obtained after this fine-tuning to reannotate the difficult Q-A pairs. Thus, the abilities of Vicuna-13B are transferred to ChatGLM for data correction in low-resource settings. Finally, 18,296 annotated Q-A pairs of BBNJ domain were obtained.

2.2 Model fine-tuning

To improve the domain adaptation ability of LLMs applied to BBNJ domain, we fine-tuned the LLMs by using the BBNJ domain Q-A dataset constructed in Section 2.1. ChatGLM is applied as the

TABLE 1 Documents contained in the dataset.

No.	BBNJ-related documents
1	Draft agreement under the United Nations Convention on the Law of the Sea on the conservation and sustainable use of marine biological diversity of areas beyond national jurisdiction
2	Agreement under the United Nations Convention on the Law of the Sea on the conservation and sustainable use of marine biological diversity of areas beyond national jurisdiction
3	United Nations Convention on the Law of the Sea
4	United Nations Convention on the Law of the Sea Annex I
5	United Nations Convention on the Law of the Sea Annex II
6	United Nations Convention on the Law of the Sea Annex III
7	United Nations Convention on the Law of the Sea Annex IV
8	United Nations Convention on the Law of the Sea Annex V
9	United Nations Convention on the Law of the Sea Annex VI
10	United Nations Convention on the Law of the Sea Annex VII
11	United Nations Convention on the Law of the Sea Annex VIII
12	United Nations Convention on the Law of the Sea Annex IX

base model. The open-source nature of the model is an important consideration.

To minimize the computational resources used for model fine-tuning, we applied the commonly used LoRA technique (Hu et al., 2021), which has been shown to effectively adapt LLMs to specific domain tasks and improve their performance (Lukichev et al., 2023). LoRA applies a simple linear design that allows the trainable matrix to be combined with frozen weights during model deployment. Compared with fully fine-tuned models, this approach does not create inference latencies, which is necessary for a knowledge Q-A system.

Data processing is necessary when fine-tuning ChatGLM. Research shows that ChatGLM and other LLMs can generalize

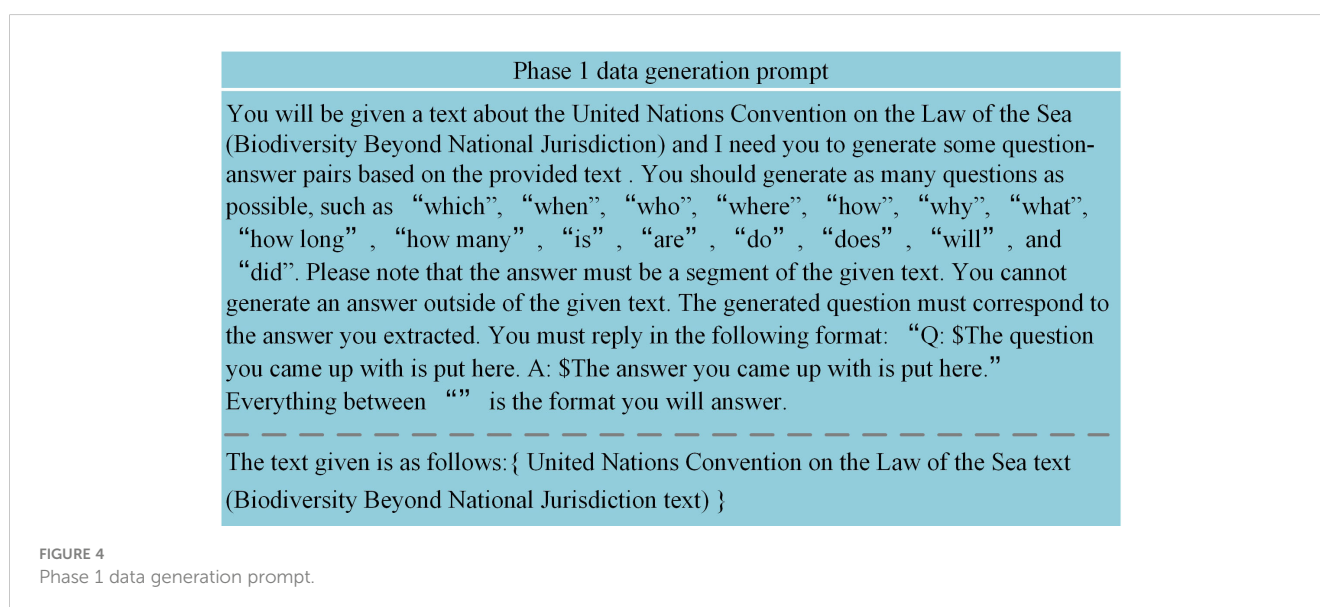
well to unseen tasks and follow task descriptions after instruction tuning (Wei et al., 2021). However, LLMs have some weaknesses when lacking instructions, such as repetitive output and difficulty in fulfilling researchers' expected task types. Therefore, it is necessary to construct the training set based on the organization of data in instruction tuning before fine-tuning. Each training data used for instruction tuning consists of three parts: instruction, question, and answer. Table 2 shows an example of the BBNJ domain Q-A dataset.

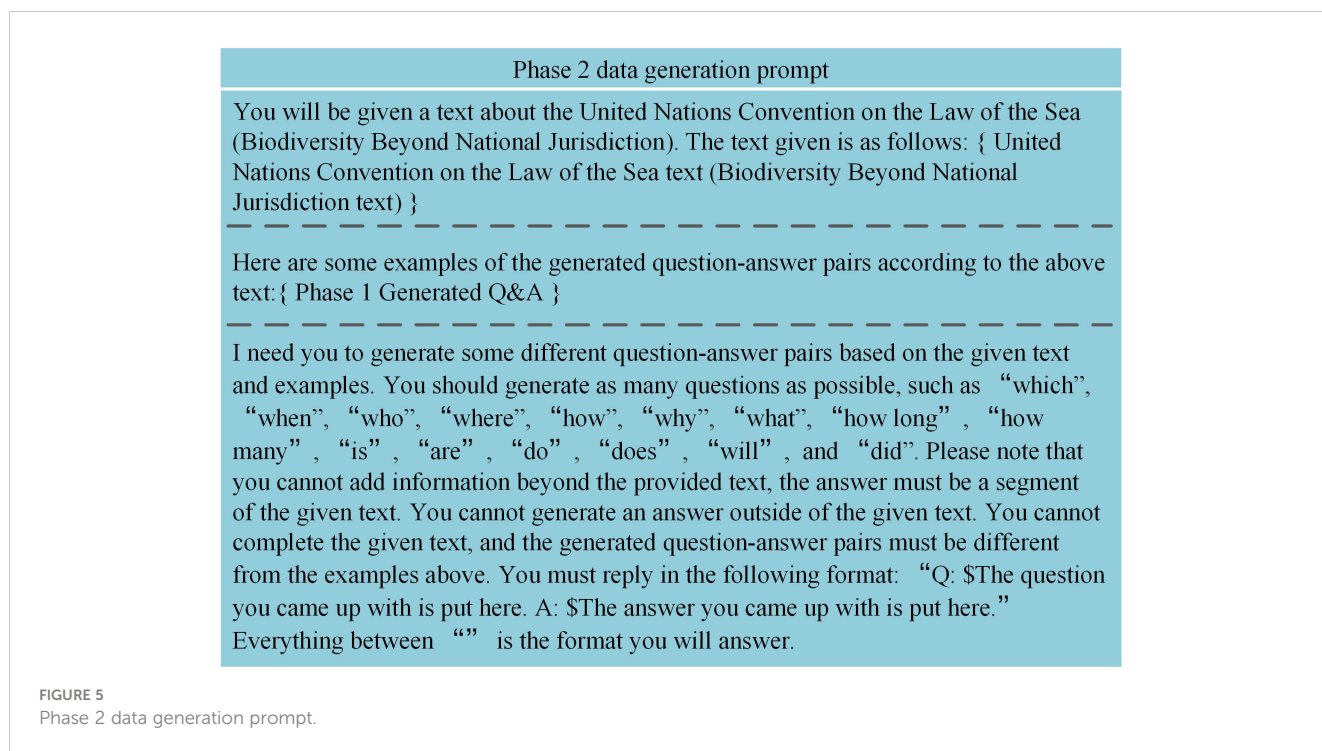
The ChatBBNJ model was obtained by using the LoRA technique to fine-tune ChatGLM for the BBNJ knowledge Q-A task.

2.3 Domain knowledge-based prompt engineering

The training data for the ChatBBNJ model is limited to a specific time period, the model cannot provide time-sensitive knowledge. To solve this problem, we applied the framework depicted in Figure 8 to the ChatBBNJ model.

First, we create the BBNJ domain knowledge base using BBNJ Agreement and its draft. Second, we applied ERNIE 2.0 (Sun et al., 2020) to obtain vectorized representations of the domain knowledge, which was stored in our vector database. When a user poses a question, it is first embedded and then indexed using semantic similarity to find the top-k nearest vectors corresponding to the inquiry. The dot product of two vectors is utilized to analyze the similarity between vector embeddings, which is obtained by multiplying their respective components and summing the results. After identifying the nearest vectors to the query vector, we decode the numeric vectors to text and retrieve the corresponding text from the database. The textual information and user question are used to improve ChatBBNJ's prompt. This method enables users to receive reliable and up-to-date answers. The framework can be extended to other domains that require periodic knowledge updates.





3 Experiment

3.1 Baseline

To assess the performance of the proposed ChatBBNJ, we performed a comparative analysis using its base model ChatGLM and two additional language models.

LLaMA (Touvron et al., 2023a) is a collection of foundation language models ranging from 7 to 65 billion parameters, these models are trained using publicly available data. The experimental results demonstrate that LLaMA-7B outperforms GPT-3 in several natural language processing benchmark tests without relying on domain datasets.

Vicuna-13B (Chiang et al., 2023) is an open-source chatbot trained by fine-tuning LLaMA on user-shared conversations collected from ShareGPT. Preliminary evaluation using GPT-4 as a judge showed that Vicuna-13B achieved more than 90% quality of OpenAI’s ChatGPT and Google’s Bard.

3.2 Evaluation metrics

The models used in this study provide knowledge answers using a generative method. Traditional generative task evaluation metrics only consider word matching, they cannot provide a reasonable assessment for expressions with the same semantics. Thus, we applied the BERTScore (Zhang et al., 2019) to evaluate semantic equivalence. The precision, recall, and F1-scores were computed. Equations (1-3) provide the calculation methods for the evaluation metrics, where $x = \langle x_1, \dots, x_k \rangle$ represents the ground-truth answers in the test set and $\hat{x} = \langle \hat{x}_1, \dots, \hat{x}_l \rangle$ represents the answers from

ChatBBNJ and baselines.

$$P_{BERT} = \frac{1}{|\hat{x}|} \sum_{x_i \in \hat{x}} \max_{x_i \in x} x_i^T \hat{x}_i \quad (1)$$

$$R_{BERT} = \frac{1}{|x|} \sum_{x_i \in x} \max_{x_i \in \hat{x}} x_i^T \hat{x}_i \quad (2)$$

$$F_{BERT} = \frac{2 * P_{BERT} * R_{BERT}}{P_{BERT} + R_{BERT}} \quad (3)$$

3.3 Results

We demonstrated the effectiveness of our method framework, the efficacy of its individual modules, and the Q-A performance of ChatBBNJ quantitatively through experiments. We divided the BBNJ domain dataset constructed in Section 2.1 into training and testing sets with ratios of 77 and 23%, respectively.

First, we compared ChatBBNJ with several other LLMs in Q-A tasks. To quantitatively evaluate the performance of ChatBBNJ, we calculated the metrics for both baselines in Section 3.1 and ChatBBNJ. The results are shown in Figure 9. The results show that ChatBBNJ achieves significantly higher precision, recall, and F1-scores, compared to baselines. The results show that ChatBBNJ achieves 0.097 precision, 0.025 recall, and 0.062 F1-scores improvement over its base model ChatGLM. These results demonstrate that ChatBBNJ exhibits superior expertise in the BBNJ domain.

Second, we conducted ablation experiments to demonstrate the efficacy of individual modules in our method framework. To

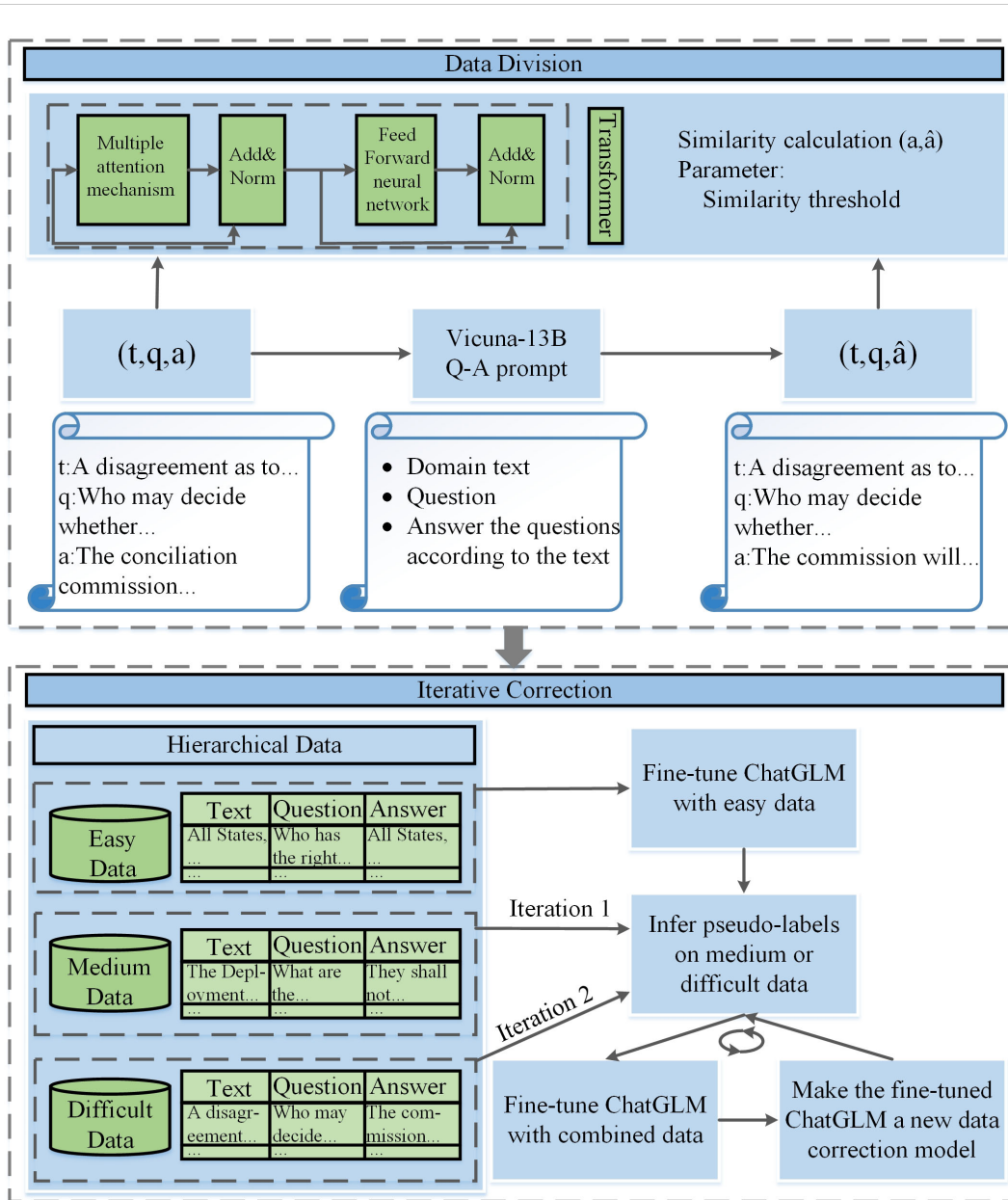


FIGURE 6

Data correction module. “ (t, q, a) ” is obtained from data generation. “ t ” and “ q ” in “ (t, q, \hat{a}) ” are the same as “ t ” and “ q ” in “ (t, q, a) ”, but “ \hat{a} ” is the new model-generated answer using the “Vicuna-13B Q-A prompt”. During data division, first, we use the “Vicuna-13B Q-A prompt” to obtain a and \hat{a} ; finally, we obtain the divided data including easy, medium and difficult levels. During iterative correction, first, we fine-tune the ChatGLM using only easy data; then, we use the supervised fine-tuned model to infer pseudo-labels on medium data and form the pseudo-labels in memory; then, we use the model, which accesses the pseudo-labels in memory to infer pseudo-labels on difficult data; finally, we obtain an annotated biodiversity of areas beyond national jurisdiction domain dataset.

quantitatively evaluate the effectiveness of the three proposed modules, namely PDGC for BBNJ domain dataset construction, model fine-tuning, and domain knowledge-based prompt engineering, we also calculated the metrics under different scenarios. The complete evaluation results are shown in Table 3. The quality of training data influences how well the model learns. Therefore, to evaluate the effectiveness of PDGC, we fine-tuned the model for Q-A task before and after using PDGC, then tested the model’s performance. The results indicate that the model’s Q-A performance improves after fine-tuning the model with data

generated by PDGC. Additionally, fine-tuning the model helps enhance its performance. We find that the model achieves improvement over ChatGLM. To evaluate the effectiveness of the domain knowledge-based prompt engineering, we also compared the Q-A performance before and after using this module. The results indicate that the model performs better after using the domain knowledge-based prompt engineering when other modules are the same. The results indicate that the model incorporating with each of our module achieves higher precision, recall and F1-scores in all cases.

Vicuna-13B Q-A prompt

I will provide a text and a question about the United Nations Convention on the Law of the Sea (Biodiversity Beyond National Jurisdiction). Please answer the question according to the provided text. Requirements: Please answer the question strictly according to the provided text. If there is no answer to the question in the text, answer with “no answer.” Please be careful not to add additional information before answering the question. The text is as follows: { United Nations Convention on the Law of the Sea text (Biodiversity Beyond National Jurisdiction text) }

The question is as follows: { Phase 1 and Phase 2 Generated Questions }

FIGURE 7
Vicuna-13B Q-A prompt.

Finally, we demonstrate the effectiveness of ChatBBNJ through specific Q-A examples. The examples generated by ChatBBNJ and baseline models are listed in [Table 4](#), [Table 5](#) and [supplementary materials](#). In the tables, “Q” represents the question provided to the models. “ChatGLM”, “LLaMA-7B”, “Vicuna-13B”, and “ChatBBNJ” represent the answers generated by the models. “Answer” represents the ground-truth answer. We underline the key points of the models’ answers.

In [Table 4](#), we provide a question related to BBNJ domain terminology. ChatGLM considers “area-based management tool” to be a software application, LLaMA-7B categorizes as manager and employee resource management, their answers deviate from the BBNJ domain. Vicuna-13B’s answer is more inclusive, but it lacks domain specificity. Our ChatBBNJ provides an answer that reflects the BBNJ domain specificities and includes most of the key points from the ground-truth answer.

In [Table 5](#), we provide a time-sensitive question related to the BBNJ domain. This question is regulated first in the BBNJ Agreement, which was formally adopted in August 2023. The answers from ChatGLM, LLaMA-7B, and Vicuna-13B are related to the COP but not to the BBNJ domain, and ChatGLM displayed language inconsistencies. It can be seen that ChatGLM’s answer is a mixture of Chinese and English format, the answer includes “会议室”(conference room), “会议设施”(conference facilities) and other Chinese format. ChatGLM’s knowledge is cut off in 2022. However, ChatBBNJ, which is based on ChatGLM, provides correct BBNJ domain answers based on knowledge in 2023.

TABLE 2 Example for instruction tuning.

Instruction	The Conference of the Parties shall ordinarily meet at the seat of the secretariat or at United Nations Headquarters. Answer the following question based on the text given.
Question	Where should the Conference of the Parties meet?
Answer	At the seat of the secretariat or at United Nations Headquarters.

In [Supplementary Table 1](#), we provide two Q-A examples for complex application scenario. The answers of ChatGLM, LLaMA-7B and Vicuna-13B use terms from the BBNJ Agreement rarely, especially ChatGLM expresses ABMT as ABM tools in Q2, which is not conform to the treaty. Besides, ChatGLM’s answer is a mixture of Chinese and English format, it includes “寻求国际协作, 以解决这一问题” (seek international cooperation to solve this problem), “采取可持续的措施” (take sustainable measures) and other Chinese format. However, ChatBBNJ not only provides the original text from the Agreement in Q2, but also emphasizes the necessity of the relevant reports and the role of the United Nations in Q1, which make the answers more comprehensive. It can be seen that compared to other models, our ChatBBNJ can provide answers more aligned with the BBNJ Agreement in complex scenarios, especially in Q2, where ChatBBNJ provides almost all requirements related to ABMTs in the BBNJ Agreement.

In [Supplementary Table 2](#), we provide two Q-A examples for the interpretation of the BBNJ Agreement under Article 31 of the Vienna Convention on the Law of Treaties (VCLT). Article 31 of the VCLT can be summarized as: be interpreted in good faith, be interpreted in the light of treaty’s object and purpose, be interpreted with supplementary means of interpretation, be interpreted in accordance with the ordinary meaning and be interpreted with the terms of the treaty in their context. In Q1, ChatBBNJ’s response highlights the importance of international cooperation in marine scientific research and technology development in first sentence, which complies with the principle of good faith interpretation. And the first sentence also links to the objective of the treaty. Moreover, ChatBBNJ’s response provides relevant contents from Article 143 of the UNCLOS, reflecting supplementing interpretation based on external materials. In Q2, ChatBBNJ provides an ordinary explanation of “transparency”, supplemented it with interpretations of the term in different contexts, and provide the role of transparency as well as the challenges it faces. This response reflects an ordinary meaning interpretation. Besides, ChatBBNJ provides some measures involved in the BBNJ Agreement, including making decisions and documents open to the public, open meeting practices, publishing and maintaining a public record of decisions and publishing decision information. These measures

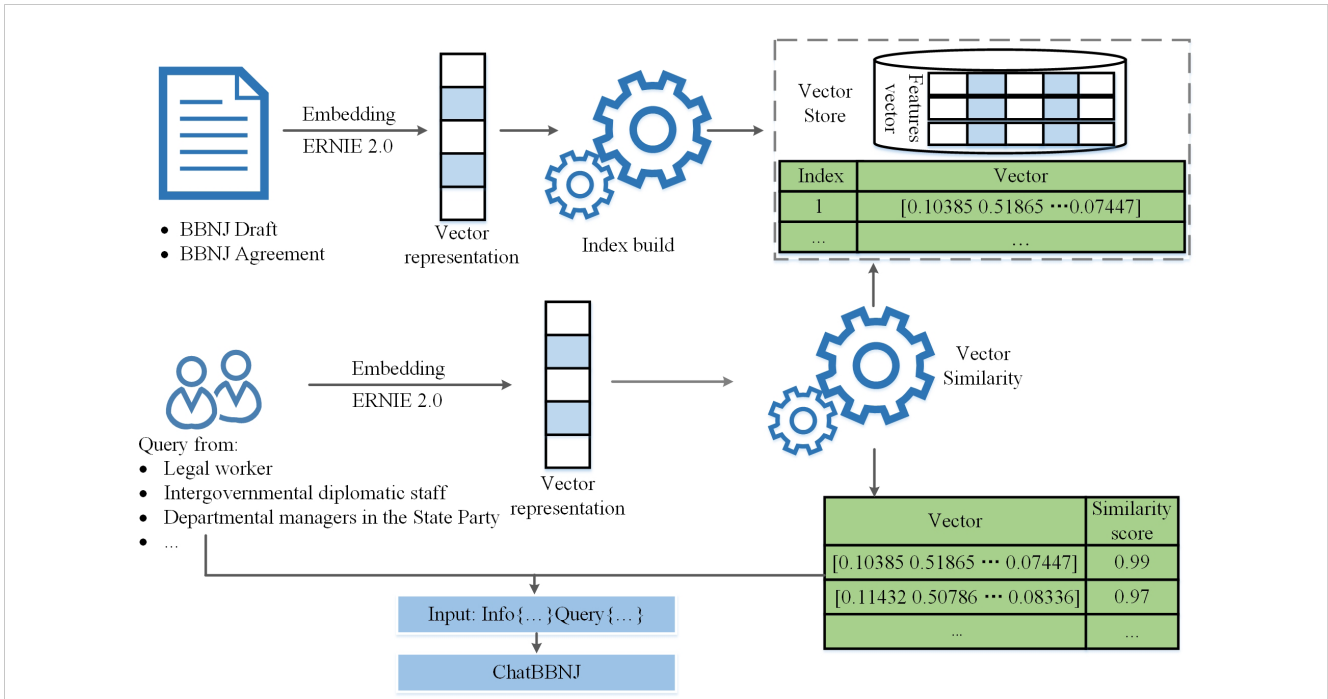


FIGURE 8 Domain knowledge-based prompt engineering framework. “BBNJ Draft” represents the draft agreement under the United Nations Convention on the Law of the Sea on the conservation and sustainable use of marine biological diversity of areas beyond national jurisdiction. “BBNJ Agreement” represents the Agreement under the United Nations Convention on the Law of the Sea on the conservation and sustainable use of marine biological diversity of areas beyond national jurisdiction.

come from different articles in the agreement, which reflects interpreting with the terms of the treaty in their context.

The latest versions of the baseline models (e.g., ChatGLM2 and LLaMA2-7B (Touvron et al., 2023b)) had improved abilities over ChatGLM and LLaMA-7B used in this study. Since we applied the ChatGLM as a base model, we compared ChatBBNJ with the latest baseline models, ChatGLM2 and LLaMA2-7B. The results are shown in Table 6. The results show that our ChatBBNJ outperforms the more advanced ChatGLM2 and LLaMA2-7B.

These findings suggest that our method framework for BBNJ domain Q-A, significantly enhance the performance of the LLMs in the BBNJ domain.

4 Discussion

Over the past 50 years, the degradation of marine biodiversity in ABNJ persists, with a lack of effective governance mechanisms to

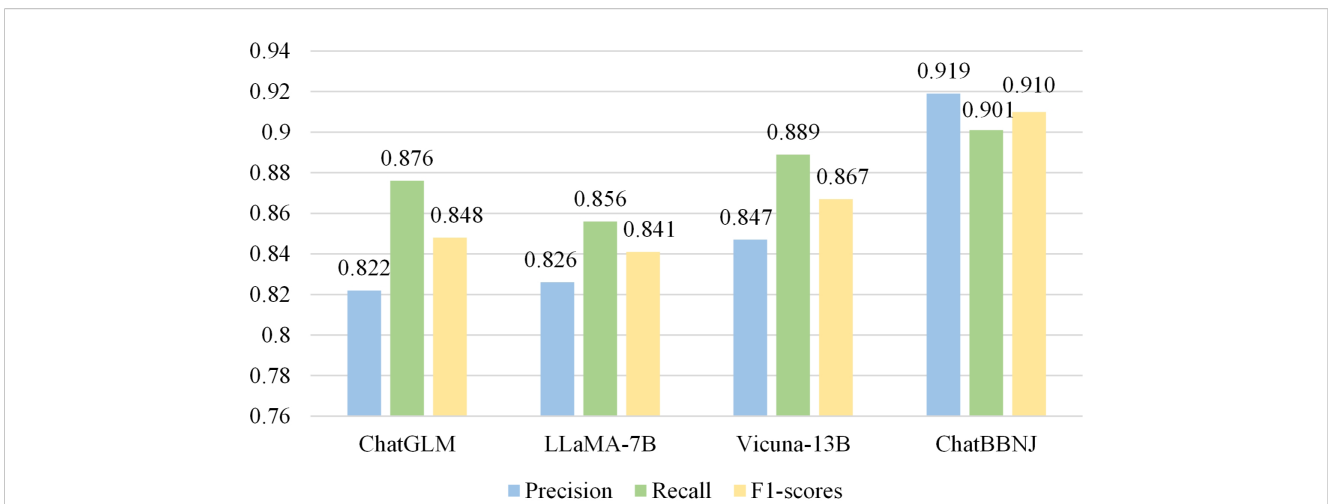


FIGURE 9 Comparison of the experimental results for ChatBBNJ.

TABLE 3 Ablation experiment results for the three modules of ChatBBNJ.

Models	Precision	Recall	F1
ChatGLM (without PDGC, model fine-tuning, and DKB prompt engineering)	0.822	0.876	0.848
ChatBBNJ (without PDGC, with model fine-tuning, without DKB prompt engineering)	0.865	0.891	0.873
ChatBBNJ (with PDGC, with model fine-tuning, without DKB prompt engineering)	0.872	0.894	0.888
ChatBBNJ (without PDGC, with model fine-tuning and DKB prompt engineering)	0.897	0.900	0.898
ChatBBNJ (with PDGC, model fine-tuning and DKB prompt engineering)	0.919	0.901	0.910

DKB prompt engineering represents domain knowledge-based prompt engineering. The bold values represent the best results.

halt this decline (Ward et al., 2022). Industrial fishing affects about 50% of oceans (Sala et al., 2018), leading to overexploitation of 31% of marine fish stocks (FAO, 2016), and ecosystem-level alterations in high seas (Ortuño Crespo and Dunn, 2017). The international maritime order is dynamic and evolving, and UNCLOS' authoritative, comprehensive, and extended nature does not imply its perfection. Due to the game of interests and compromise among countries, many UNCLOS rules are principled and articulated, creating legal ambiguities that often need to be further addressed in practice. The current international ocean order is in a state of rapid transition, necessitating a reaction to several earth system changes such as sea-level rise, plastic pollution of the seas, acidification, and destruction of marine biodiversity (Chen and Liu, 2023). BBNJ agreement is intended to serve as a governance mechanism for the protection and sustainable use of marine biodiversity in the ABNJ.

The present study introduces ChatBBNJ, a question-answering system, that designed to enhance the treaty implementation. First, we proposed PDGC, a data engineering method, that constructs a high-quality Q-A dataset specific to the BBNJ domain. Second, we applied this dataset to fine-tune the ChatGLM to obtain the BBNJ domain model, ChatBBNJ. Finally, we introduced a domain knowledge-based prompt engineering. We demonstrated improvements by testing ChatBBNJ on Q-A data related to the BBNJ Agreement and its draft. Experiment results demonstrate that ChatBBNJ outperforms baseline LLMs across three Q-A metrics. Additionally, hybrid ChatBBNJ, which introduces a domain knowledge-based prompt framework outperforms standalone ChatBBNJ. The main findings of our work are summarized as follows:

- (1) The quality of LLMs text generation can be enhanced through appropriate prompt engineering and data correction. The effectiveness of model training is to some extent related to the quality of training data. Therefore, Table 3 compares the Q-A performance of ChatBBNJ

TABLE 4 Q-A example in BBNJ domain terminology.

Q: What is the purpose of area-based management tool?
<p>ChatGLM: An area-based management tool is a software application that allows organizations to manage their resources and operations across different locations, regions, or areas. The purpose of an area-based management tool is to help organizations more effectively manage their resources, improve communication, and increase productivity. One of the main benefits of using an area-based management tool is that it allows organizations to manage their resources and operations across different locations, regions, or areas more efficiently. This can help to reduce costs by reducing the need for travel and accommodation, and improving communication by allowing for real-time collaboration between teams in different locations. Additionally, area-based management tools can help to improve productivity by allowing organizations to manage their resources and operations more effectively across different locations. This can help to reduce the time it takes for teams to work together, and improve the overall performance of the organization. Overall, area-based management tools are an important tool for organizations looking to improve their resource management, communication, and productivity across different locations.</p>
<p>LLaMA-7B: Area based Management Tool (ABMT) was developed to help managers and employees in making decisions about how best to manage their resources, such as people or money. It helps them identify which areas are most important for improvement by identifying what needs attention first. The ABM system can be used with any type of resource but it works particularly well when you have limited time and/or budget available so that you need to prioritize your efforts effectively. The main steps involved include: 1. Identify all possible options; 2. Priority ranking each option according to its potential impact on performance goals; 3. Selecting one or more priority actions from those ranked highest; 4. Implement selected action(s); 5. Evaluate results against original objectives; 6. Repeat process if necessary until desired level of effectiveness has been achieved.</p>
<p>Vicuna-13B: Area-based management tools are used to help organizations and individuals manage specific areas or regions. These tools can be used for a variety of purposes, such as to track and manage resources, monitor activity, and identify trends and patterns. Some examples of area-based management tools include geographic information systems (GIS), which can be used to create maps and visualizations of an area, and resource management software, which can be used to track and manage resources such as fish stocks or forests. Area-based management tools can be useful for a wide range of applications, including natural resource management, emergency response, and urban planning.</p>
<p>ChatBBNJ: The purpose of an area-based management tool is to <u>manage a Geographically defined area through which one or several sectors or activities are managed with the aim of achieving particular conservation.</u> The tool may include <u>a marine protected area</u> as one of its components. The establishment of area-based management tools, including marine protected areas, is to prevent the inclusion of any areas within national jurisdiction in order to prevent asserts or denying any claims to sovereignty.</p>
<p>Answer: "Area-based management tool" means a tool, including a marine protected area, for a geographically defined area through which one or several sectors or activities are managed with the aim of achieving particular conservation and sustainable use objectives in accordance with this Agreement.</p>

ChatGLM', 'LLaMA-7B', 'Vicuna-13B', and 'ChatBBNJ' represent the answers generated by the models. 'Answer' represents the ground-truth answer. The underline words represent the key points of the models' answers.

before and after using PDGC. It can be seen that the Q-A performance of ChatBBNJ improves when the model is fine-tuned with data generated by PDGC.

- (2) The domain adaptation ability of LLMs in Q-A tasks can be improved by fine-tuning the model on domain-specific Q-A datasets. Analyzing the experimental results in Table 3, regardless of the method used for data generation, ChatBBNJ's Q-A performance improves after model fine-tuning.

TABLE 5 Time-sensitive Q-A example in the BBNJ domain.

Q: Where should the Conference of the Parties meet?
<p>ChatGLM: The Conference of the Parties (CP) is an international organization that致力于 developing and implementing policies to reduce climate change. It meets regularly, usually in locations that are convenient for its members. The CP's current meeting, which took place in Madrid, Spain in 2020, was the 11th meeting of the CP since its formation in 1995. The next meeting of the CP will take place in late 2021 or early 2022, and the location of the meeting will be announced at a later date. The CP's meetings are held in a variety of locations, including 会议室 (会议室是指会议所需的会议室、会议桌、椅子等会议设施), which are designed to provide a comfortable and productive environment for conference attendees. The location of the CP's next meeting will be determined by the needs and preferences of its members, as well as the availability of 会议室 and other 会议设施. The CP meets regularly to discuss and develop policies to reduce climate change, including policies to reduce carbon emissions, promote renewable energy, and protect the environment. The organization is committed to working towards a goal of reducing climate change by 2C by the end of the century.</p>
<p>LLaMA-7B: The conference will be held in Paris, France because it has been suggested by many countries and they have agreed to host this meeting there for 2 weeks starting on December 1st until January 3rd at Hotel de Ville (City Hall). This location was chosen due to its centrality within Europe which makes traveling easier as well as being able to accommodate all delegates from around the world with ease. The city also offers great food options such as baguettes or crepes along with other delicious cuisines making sure everyone's taste buds are satisfied! It would not only benefit those who attend but also help boost tourism throughout the country bringing more money into their economy.</p>
<p>Vicuna-13B: The Conference of the Parties (COP) is a meeting of the parties to the United Nations Framework Convention on Climate Change (UNFCCC). The COP is the supreme decision-making body of the UNFCCC, and it is responsible for reviewing the implementation of the Convention and taking decisions to further its objectives. The COP is typically held annually, and it is attended by representatives from the governments of the parties to the Convention, as well as experts, civil society organizations, and the media. The location of the COP is typically decided by the UNFCCC secretariat and the government of the host country. The COP has been held in a variety of locations, including developed and developing countries, and it has typically been hosted by countries that are parties to the Convention and have the capacity to host a large international meeting. In the past, the COP has been held in countries such as Germany, France, and South Africa. The next COP is scheduled to be held in Glasgow, Scotland in 2021.</p>
<p>ChatBBNJ: <u>At the seat of the secretariat or at United Nations Headquarters.</u></p>
<p>Answer: The Conference of the Parties shall ordinarily meet at the seat of the secretariat or at United Nations Headquarters.</p>

ChatGLM, 'LLaMA-7B', 'Vicuna-13B', and 'ChatBBNJ' represent the answers generated by the models. 'Answer' represents the ground-truth answer. The underline words represent the key points of the models' answers.

- (3) The outdated issues of LLMs can be refined by giving the model access to the knowledge beyond its fine-tuning phase time and instructing LLMs on how to utilize that knowledge. In Table 5, the model was asked a time-sensitive question, with relevant information not within

TABLE 6 Supplementary comparisons of experimental results for ChatBBNJ.

Models	Precision	Recall	F1
ChatGLM2	0.846	0.883	0.864
LLaMA2-7B	0.852	0.889	0.870
ChatBBNJ	0.919	0.901	0.910

The bold values represent the best results.

the model's training data. However, our ChatBBNJ, equipped with the domain knowledge-based prompt engineering, accessed external knowledge and provided the correct answer.

- (4) This study not only demonstrates the feasibility of BBNJ knowledge acquisition by using Q-A system but also provides a method framework to apply Q-A system in BBNJ domain. The method framework has three parts. Firstly, domain datasets are acquired using data engineering method. Secondly, LLMs are fine-tuned for Q-A task based on the generated datasets. Finally, a domain knowledge-based prompt engineering is employed to assist the model in accessing the latest information, ensuring the timeliness of its responses.

In the future, the successful application of ChatBBNJ can bring several benefits. From the perspective of the representatives who have not participated in BBNJ negotiations, they can acquire BBNJ knowledge and avoid reading a large number of irrelevant documents. Simultaneously, governments can benefit from the Q-A system. The BBNJ Agreement calls for parties to take the necessary legislative measures to ensure the implementation of this agreement (United Nations, 2023). ChatBBNJ can provide popularization and interpretation of the agreement, which is helpful to assist governments in enacting implementation legislation. The successful implementation of the proposed ChatBBNJ will make a substantial contribution towards prompting the effective implementation of the BBNJ Agreement, which will hopefully result in the conservation and sustainable use of marine biodiversity in ABNJ.

However, the limitations of this study remain. First, the accuracy of ChatBBNJ needs to be improved. The present system has achieved reasonable accuracy in the current research field, but there is still room for improvement. Second, the sources of knowledge need to be expanded. The present study mainly focused on promoting the implementation of the BBNJ Agreement, but the methods can be applied to more implementation in international agreements in the law of the sea domain. In future research, more international law of the sea data will be added to the study to obtain richer knowledge and expand the scope of applications.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Author contributions

XW: Conceptualization, Data curation, Formal Analysis, Investigation, Methodology, Project administration, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. MZ: Data curation, Formal analysis, Writing –

review & editing. HL: Conceptualization, Resources, Supervision, Writing – review & editing. XM: Project administration, Validation, Writing – original draft. YL: Project administration, Writing – original draft. YC: Conceptualization, Data curation, Formal Analysis, Funding acquisition, Resources, Supervision, Writing – review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This research was funded by the National Social Science Fund of China, Grant No. 20CFX082.

Acknowledgments

We would like to thank the computing resources and the professional technical guidance provided by Hao Liu. Thanks to Mittens and Panda.

References

- Amer-Yahia, S., Bonifati, A., Chen, L., Li, G. L., Shim, K., Xu, J. L., et al. (2023). From large language models to databases and back: A discussion on research and education. *SIGMOD Rec.* 52, 49–56. doi: 10.1145/3631504.3631518
- Bengio, Y., Ducharme, R., Vincent, P., and Jauvin, C. (2003). A neural probabilistic language model. *JMLR* 3, 1137–1155. doi: 10.1162/153244303322533223
- Bodansky, D. (2011). Implementation of international environmental law. *Jpn. Yearb. Int. Law.* 54, 62–96.
- Chen, Y. T., and Liu, H. R. (2023). Critical perspectives on the new situation of global ocean governance. *Sustainability* 15, 10921. doi: 10.3390/su151410921
- Chiang, W.-L., Li, Z., Lin, Z., Sheng, Y., Wu, Z., Zhang, H., et al. (2023). *Vicuna: an open-source chatbot impressing GPT-4 with 90%* chatGPT quality*. Available online at: <https://lmsys.org/blog/2023-03-30-vicuna/> (Accessed November 4, 2023).
- Common Crawl (2023). Available online at: <https://commoncrawl.org/> (Accessed November 4, 2023).
- Cui, J. X., Li, Z. J., Yan, Y., Chen, B. H., and Yuan, L. (2023). ChatLaw: open-source legal large language model with integrated external knowledge bases. *arXiv*. doi: 10.48550/arXiv.2306.16092
- Dai, F., Zhao, Z., Sun, C., and Li, B. (2022). “Intelligent audit question answering system based on knowledge graph and semantic similarity,” in *2022 11th International Conference of Information and Communication Technology (ICTech)*, Wuhan, China: IEEE. 125–132. doi: 10.1109/ICTech55460.2022.00033
- Deasy, K. (2023). What we know about the new high seas treaty. *NPJ Ocean Sustain.* 2, 7. doi: 10.1038/s44183-023-00013-x
- Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: pretraining of deep bidirectional transformers for language understanding. *arXiv*. doi: 10.48550/arXiv.1810.04805
- Dong, Q. X., Li, L., Dai, D. M., Zheng, C., Wu, Z. Y., Chang, B. B., et al. (2023). A survey on in-context learning. *arXiv*. doi: 10.48550/arXiv.2301.00234
- Eloundou, T., Manning, S., Mishkin, P., and Rock, D. (2023). GPTs are GPTs: an early look at the labor market impact potential of large language models. *arXiv*. doi: 10.48550/arXiv.2303.10130
- FAO (2016) *The state of world fisheries and aquaculture 2016. contributing to food security and nutrition for all* (Rome, Italy: Food and Agriculture Organization of the United Nations). Available online at: <https://www.fao.org/3/i5555e/i5555e.pdf> (Accessed March 4, 2024).
- Hu, E. J., Shen, Y. L., Wallis, P., Allen-Zhu, Z., Li, Y. Z., Wang, S. A., et al. (2021). LoRA: low-rank adaptation of large language models. *arXiv*. doi: 10.48550/arXiv.2106.09685
- Huang, Q. Z., Tao, M. X., An, Z. W., Zhang, C., Jiang, C., Chen, Z. B., et al. (2023). Lawyer LLaMA technical report. *arXiv*. doi: 10.48550/arXiv.2305.15062

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher’s note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fmars.2024.1368356/full#supplementary-material>

Humphries, F., and Harden-Davies, H. (2020). Practical policy solutions for the final stage of BBNJ treaty negotiations. *Mar. Policy.* 122, 104214. doi: 10.1016/j.marpol.2020.104214

Humphries, F., Rabone, M., and Jaspars, M. (2021). Traceability approaches for marine genetic resources under the proposed ocean (BBNJ) treaty. *Front. Mar. Sci.* 8. doi: 10.3389/fmars.2021.661313

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., et al. (2020). Scaling laws for neural language models. *arXiv*. doi: 10.48550/arXiv.2001.08361

Lau, G. T., Law, K. H., and Wiederhold, G. (2005). “Legal information retrieval and application to e-rulemaking,” in *Proceedings of the 10th International Conference on Artificial Intelligence and Law, ICAIL’05* (Association for Computing Machinery, New York, NY), 146–154. doi: 10.1145/1165485.1165508

Lee, S.-H., Choi, S.-W., and Lee, E.-B. (2023). A question-answering model based on knowledge graphs for the general provisions of equipment purchase orders for steel plants maintenance. *Electronics* 12, 2504. doi: 10.3390/electronics12112504

Li, Y. X., Li, Z. H., Zhang, K., Dan, R. L., Jiang, S., and Zhang, Y. (2023). ChatDoctor: A medical chat model fine-tuned on a large language model meta-AI (LLaMA) using medical domain knowledge. *Cureus* 15, e40895. doi: 10.7759/cureus.40895

Li, J. L., Zhang, Z. S., and Zhao, H. (2023). Self-prompting large language models for zero-shot open-domain QA. *arXiv*. doi: 10.48550/arXiv.2212.08635

Lukichev, D., Kryanina, D., Bystrava, A., Fenogenova, A., and Tikhonova, M. (2023). “Parameter-efficient tuning of transformer models for Anglicism detection and substitution in Russian,” in *Proceedings of the International Conference “Dialogue 2023.”* Available at: <https://www.dialog-21.ru/media/5911/lukichevpluseta1042.pdf> (Accessed April 3, 2024).

Ortuño Crespo, G., and Dunn, D. C. (2017). A review of the impacts of fisheries on open-ocean ecosystems. *ICES J. Mar. Sci.* 74, 2283–2297. doi: 10.1093/icesjms/fsx084

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., et al. (2022). Training language models to follow instructions with human feedback, in *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 27730–27744. Available at: https://proceedings.neurips.cc/paper_files/paper/2022/file/b1efde53be364a73914f58805a001731-Paper-Conference.pdf (Accessed April 3, 2024).

Riloff, E., and Thelen, M. (2000). “A rule-based question answering system for reading comprehension tests,” in *ANLP-NAACL 2000 workshop: reading comprehension tests as evaluation for computer-based language understanding systems*, (Seattle, WA, USA: ACL). Vol. 6. 13–19. doi: 10.3115/1117595.1117598

Rosenfeld, R. (2000). Two decades of statistical language modeling: where do we go from here? *Proc. IEEE* 88, 1270–1278. doi: 10.1109/5.880083

Sala, E., Mayorga, J., Costello, C., Kroodsmas, D., Palomares, M. L. D., Pauly, D., et al. (2018). The economics of fishing the high seas. *Sci. Adv.* 4, eaat2504. doi: 10.1126/sciadv.aat2504

- Sun, Y., Wang, S. H., Li, Y. K., Feng, S. K., Tian, H., Wu, H., et al. (2020). "ERNIE 2.0: a continual pre-training framework for language understanding," in *Proceedings of the AAAI Conference on Artificial Intelligence*, (New York, USA: AAAI). Vol. 34. 8968–8975. doi: 10.1609/aaai.v34i05.6428
- Tessnow-von Wysocki, I., and Vadrot, A. B. M. (2020). The voice of science on marine biodiversity negotiations: a systematic literature review. *Front. Mar. Sci.* 7. doi: 10.3389/fmars.2020.614282
- Tiller, R., Mendenhall, E., De Santo, E. D., and Nyman, E. (2023). Shake it off: negotiations suspended, but hope simmering, after a lack of consensus at the fifth intergovernmental conference on biodiversity beyond national jurisdiction. *Mar. Policy*. 148, 105457. doi: 10.1016/j.marpol.2022.105457
- Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., et al. (2023a). LLaMA: open and efficient foundation language models. *arXiv*. doi: 10.48550/arXiv.2302.13971
- Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., et al. (2023b). Llama 2: open foundation and fine-tuned chat models. *arXiv*. doi: 10.48550/arXiv.2307.09288
- United Nations (2023) *Agreement under the united nations convention on the law of the sea on the conservation and sustainable use of marine biological diversity of areas beyond national jurisdiction*. Available online at: <https://undocs.org/en/A/77/L.82> (Accessed August 10, 2023).
- Vaghefi, S. A., Stambach, D., Muccione, V., Bingler, J., Ni, J. W., Kraus, M., et al. (2023). ChatClimate: Grounding conversational AI in climate science. *Commun. Earth Environ.* 4, 480. doi: 10.1038/s43247-023-01084-x
- Wang, X., Chen, Y., and Zhu, W. (2021). A survey on curriculum learning. *Proc. IEEE* 44, 4555–4576. doi: 10.1109/TPAMI.34
- Wang, H. C., Liu, C., Xi, N. W., Qiang, Z. W., Zhao, S. D., Qin, B., et al. (2023). HuaTuo: tuning LLaMA model with Chinese medical knowledge. *arXiv*. doi: 10.48550/arXiv.2304.06975
- Wang, Z. Z., Yang, F. K., Zhao, P., Wang, L., Zhang, J., Garg, M., et al. (2023). Empower large language model to perform better on industrial domain-specific question answering. *arXiv*. doi: 10.48550/arXiv.2305.11541
- Ward, D., Melbourne-Thomas, J., Pecl, G. T., Evans, K., Green, M., McCormack, P. C., et al. (2022). Safeguarding marine life: conservation of biodiversity and ecosystems. *Rev. Fish Biol. Fisheries*. 32, 65–100. doi: 10.1007/s11160-022-09700-3
- Wei, J., Bosma, M., Zhao, V. Y., Guu, K., Yu, A. W., Lester, B., et al. (2021). Finetuned language models are zero-shot learners. *arXiv*. doi: 10.48550/arXiv.2109.01652
- Wikipedia (2023). Available online at: <https://en.wikipedia.org/wiki/MainPage> (Accessed November 4, 2023).
- Xiong, H. L., Wang, S., Zhu, Y. T., Zhao, Z. H., Liu, Y. X., Huang, L. L., et al. (2023). DoctorGLM: fine-tuning your Chinese doctor is not a herculean task. *arXiv*. doi: 10.48550/arXiv.2304.01097
- Yang, H. Y., Liu, X.-Y., and Wang, C. D. (2023). FinGPT: open-source financial large language models. *arXiv*. doi: 10.2139/ssrn.4489826
- Zeng, A. H., Liu, X., Du, Z. X., Wang, Z. H., Lai, H. Y., Ding, M., et al. (2022). GLM-130B: an open bilingual pre-trained model. *arXiv*. doi: 10.48550/arXiv.2210.02414
- Zhang, T. Y., Kishore, V., Wu, F. L., Weinberger, K. Q., and Artzi, Y. (2019). BERTScore: evaluating text generation with BERT. *arXiv*. doi: 10.48550/arXiv.1904.09675
- Zhong, B., He, W., Huang, Z., Love, P. E. D., Tang, J., and Luo, H. (2020). A building regulation question answering system: a deep learning methodology. *Adv. Eng. Inform.* 46, 101195. doi: 10.1016/j.aei.2020.101195
- Zhu, F. B., Lei, W. Q., Wang, C., Zheng, J. M., Poria, S., and Chua, T.-S. (2021). Retrieving and reading: a comprehensive survey on open-domain question answering. *arXiv*. doi: 10.48550/arXiv.2101.00774