OceanGPT: A Large Language Model for Ocean Science Tasks

Zhen Bi[1,2,5], **Ningyu Zhang**[1,2,5*], **Yida Xue**[1], **Yixin Ou**[1], **Daxiong Ji**[2,3]
Guozhou Zheng[2,4], **Huajun Chen**[1,2*]
[1]College Computer Science and Technology, Zhejiang University   [2]Donghai Laboratory
[3]Ocean College, Zhejiang University [4]Zhoushan-Zhejiang University Ocean Research Center
[5]School of Software Technology, Zhejiang University
{bizhen_zju,zhangningyu,ouyixin,jidaxiong,guozhou,huajunsir}@zju.edu.cn
Project Website: https://zjunlp.github.io/project/OceanGPT/

Abstract

Warning: The model in this paper might produce hallucinations and reader discretion is recommended. Ocean science, which delves into the oceans that are reservoirs of life and biodiversity, is of great significance given that oceans cover over 70% of our planet's surface. Recently, advances in Large Language Models (LLMs) have transformed the paradigm in science. Despite the success in other domains, current LLMs often fall short in catering to the needs of domain experts like oceanographers, and the potential of LLMs for ocean science is under-explored. The intrinsic reason may be the immense and intricate nature of ocean data as well as the necessity for higher granularity and richness in knowledge. To alleviate these issues, we introduce **OceanGPT**, the first-ever LLM in the ocean domain, which is expert in various ocean science tasks. We propose DoInstruct, a novel framework to automatically obtain a large volume of ocean domain instruction data, which generates instructions based on multi-agent collaboration. Additionally, we construct the first oceanography benchmark, **OceanBench**, to evaluate the capabilities of LLMs in the ocean domain. Though comprehensive experiments, **OceanGPT** not only shows a higher level of knowledge expertise for oceans science tasks but also gains preliminary embodied intelligence capabilities in ocean technology[1].

1 Introduction

Ocean science, which delves into the intricacies of oceans that cover over 70% of our planet's surface, is essential not only for understanding the rich reservoirs of life and biodiversity but also for recognizing their pivotal role in regulating the global climate and supporting economies (Esaias et al., 1998; Falkowski, 2012; Visbeck, 2018; Jin et al., 2023). Recently, advances in Large Language Models (LLMs) (OpenAI, 2023; Jiang et al., 2023; Zha et al., 2023; Yin et al., 2023; Zhao et al., 2023) have transformed the paradigm in science domains such as medical science (Moor et al., 2023), molecular science (Fang et al., 2023), protein science (Lin et al., 2023) and geoscience (Deng et al., 2023). However, the potential for the LLM in ocean science is under-explored.

Despite remarkable success in general domain, current LLMs still do not fully meet the specific demand of oceanographers. This inadequacy is primarily due to: (1) The immense volume and intricate nature of ocean data. As ocean science research progresses, acquiring data becomes increasingly challenging, which makes enhancing the oceanic understanding both a golden opportunity and a significant hurdle. (2) The necessity for higher granularity and richness in knowledge. Note that the data requirements faced by researchers are becoming increasingly intricate and diverse. Ocean science encompasses various domains and subjects, each with its distinct data attributes and patterns.

To alleviate these issues, we introduce **OceanGPT**, the first-ever LLM in the ocean domain, which is expert in various ocean science tasks. Specifically, we propose DoInstruct, an efficient ocean

---

[*]Corresponding author.
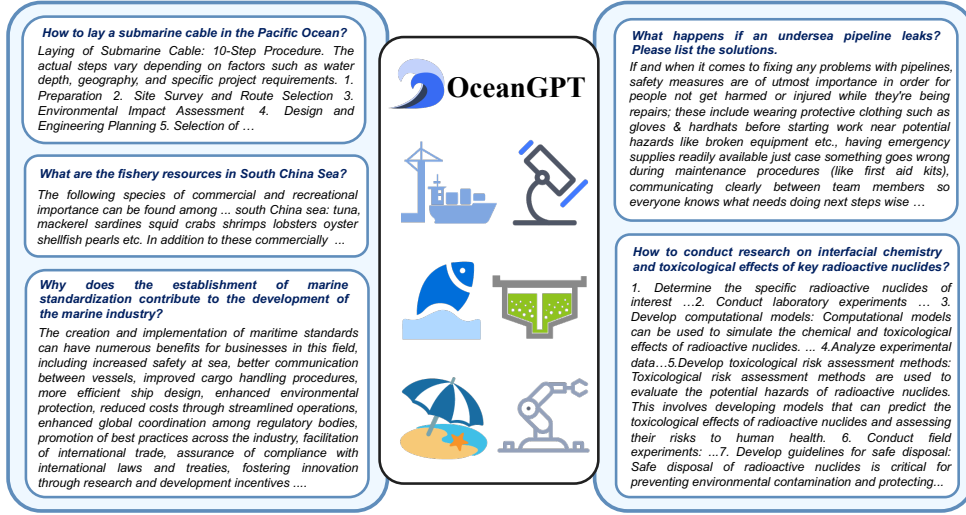[1]Codes, data and checkpoints will soon be available at https://github.com/zjunlp/KnowLM.

Figure 1: Capabilities of **OCEANGPT**.

science instruction generation framework that capitalizes on multi-agent collaboration. Each agent in our designed framework is considered as an expert in a specific domain (science and research, resources and development, ecology and environment etc.) and is responsible for generating the corresponding data. For the advancement of ocean science research using LLMs, we also create a benchmark called **OCEANBENCH** to evaluate the capabilities in ocean science tasks.

Through extensive experiments, **OCEANGPT** shows superiority for diverse ocean science tasks. Note that our benchmark data is based on criteria manually evaluated by ocean experts, and can accurately reflect the capabilities that LLMs possess in the field of ocean science. As depicted in Figure 1, our model can comprehensively answer questions according to the instructions of oceanographers, which demonstrates its expertise in oceanography. We further explore the potential of **OCEANGPT** from the perspectives of ocean engineering. Specifically, we integrate ocean robotics instructions into the training data and evaluate the ability of **OCEANGPT** via code or console commands. **OCEANGPT** not only demonstrates a higher level of knowledge expertise but also gains preliminary embodied intelligence capabilities in ocean technology.

Our contributions can be summarized as follows:

- We introduce **OCEANGPT**, the first ocean LLM, which shows superiority for various ocean science tasks. It can answer oceanographic questions according to the instructions of oceanographers, demonstrating expertise in oceanography.

- We propose DOINSTRUCT, an automated domain instruction evolving framework that constructs the ocean instruction dataset by multi-agent collaboration. Our framework effectively alleviates the difficulty of obtaining ocean domain data.

- Extensive experiments demonstrate the superiority of **OCEANGPT** in the **OCEANBENCH**. **OCEANGPT** not only demonstrates a higher level of knowledge expertise for oceans science tasks but also gains preliminary embodied intelligence capabilities.

2 RELATED WORK

Large Language Models. The landscape of LLM (Brown et al., 2020; Chowdhery et al., 2022; Touvron et al., 2023a;b) has rapidly evolved and achieved a series breakthroughs. Rae et al. (2021); Zhang et al. (2022); Thoppilan et al. (2022); Scao et al. (2022); Zeng et al. (2023) have explored the performance across a wide range of model scales and broadened the application scope. LLMs also exhibit capabilities in various aspects, including reasoning and knowledge interaction (Qiao et al., 2023a; Zhang et al., 2023a; Qiao et al., 2023b; Wang et al., 2023a; Xi et al., 2023). To obtain LLMs, instruction tuning (Wei et al., 2022; Zhang et al., 2023b; Ouyang et al., 2022) is a crucial technique

to alignment with user preferences and desired outputs. Taori et al. (2023) uses self-instruct (Wang et al., 2023d) to generate data from manually instructions. Chiang et al. (2023); Xu et al. (2023) uses user-shared conversations and AI-generated data for tuning. Different from those, we introduce an automatic and effective domain instruction generation framework via multi-agent collaboration.

Science Large Language Models. LLMs have emerged as cornerstone models in addressing challenges within scientific research. Singhal et al. (2022) explores the potential of clinical LLMs and introduces a human evaluation framework and instruction prompt tuning. Moor et al. (2023) proposes generalist medical AI that is capable of handling diverse medical tasks using self-supervised learning on large datasets. Kraljevic et al. (2021) introduces MedGPT, a model using EHR data and Named Entity Recognition tools for predicting future medical events. BioGPT (Luo et al., 2022) is a language model pre-trained on biomedical literature for improved text generation and mining. Theodoris et al. (2023) describes Geneformer, a model pre-trained on single-cell transcriptomes for making predictions with limited data in network biology. Lin et al. (2023) demonstrates the prediction of atomic-level protein structure from primary sequences using scaled-up language models. Deng et al. (2023) introduces the first LLM specifically designed for geoscience, including its training and benchmarking protocols. Different from previous works, we are the first LLM for ocean science tasks and explore its potential for ocean research.

3 OCEANGPT

To obtain **OCEANGPT**, we firstly construct the training corpus for ocean science and pre-train an ocean LLM based on LLaMA-2 (Touvron et al. (2023b) in Section 3.1. Then we propose DOIN-STRUCT, an automated framework for domain instruction generation to build an ocean domain-specific instruction dataset. Our framework leverages multi-agent collaboration and utilizes ocean literature to automatically generate a large volume of domain-specific instructions for ocean science tasks (Section 3.2). The overview training procedure of our **OCEANGPT** is shown in Figure 2.
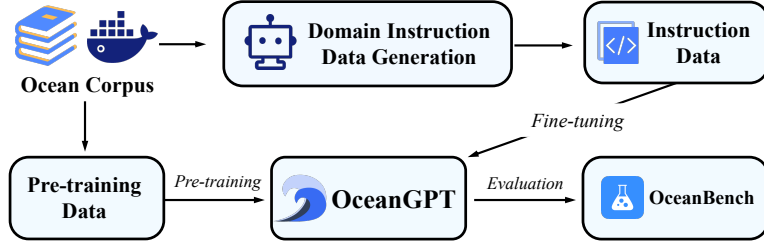

Figure 2: Overall training procedure of our **OCEANGPT**.

3.1 PRE-TRAINING STAGE

To pre-train the foundation model for ocean science tasks, it is essential to construct the pre-training corpus specific to ocean science. Therefore, we firstly collect a raw corpus of 67,633 documents from **open-access literature**. Specifically, we utilize the Python package *pdfminer* to convert the content of literature files into plain text. To ensure the quality and consistency of the data, further processing of the collected dataset is necessary. We apply regular expressions to filter out figures, tables, headers, footers, page numbers, URLs and references. Additionally, any extra spaces, line breaks, and other non-text characters are removed. The processed documents cover various aspects of ocean science such as ocean physics, ocean chemistry, ocean biology, geology, hydrology, etc. It is important to note that special characters, emoticons, and garbled characters are also replaced or eliminated during this process. We also employ *hash-based methods* to de-duplicate the data, which helps reduce the risk of over-fitting during pre-training and enhances its generalization capability. The detailed statistics is shown in Table 1.
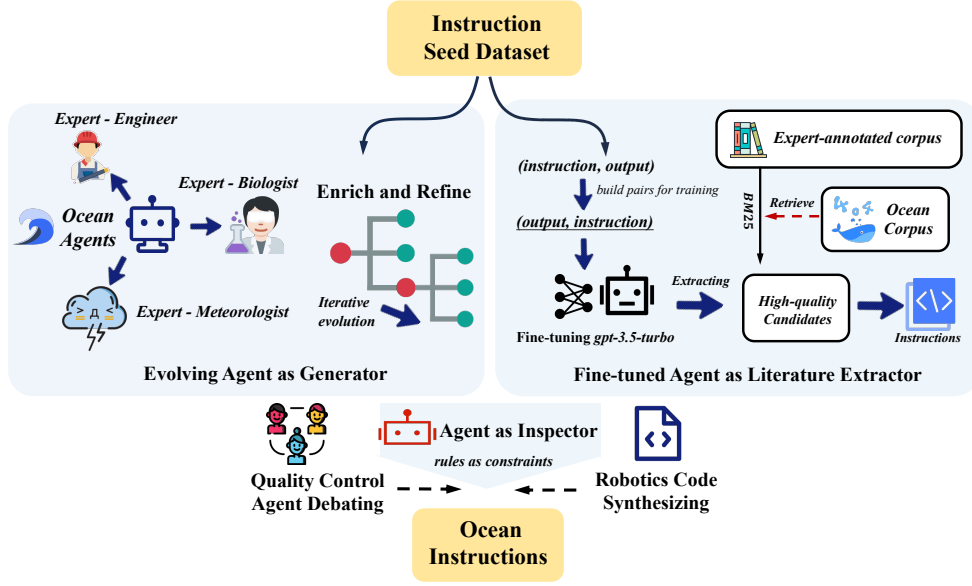
Figure 3: Procedure of our proposed DOINSTRUCT. We use agents as experts for each **ocean topic** and make them rapidly expand the instructions by collaboration. We design three agent roles: **evolving generator**, **fine-tuned literature extractor** and **inspector with rule constraints**.

3.2 DOMAIN INSTRUCTION DATA GENERATION

As ocean science research deepens, researchers are facing increasingly complex and diversified data demands. Ocean science corpus contains multiple fields and topics, and each topic has its unique data characteristics and patterns. To effectively simulate and obtain those data, we propose a domain instruction generation framework DOINSTRUCT to obtain ocean instructions H by multi-agent collaboration. Each agent is considered as an expert in a **specific domain (topic)** and is responsible for generating the corresponding data. It not only ensures the professionalism and accuracy of the data but also allows for the parallel and efficient generation of a large amount of data. Note that the proposed framework also has greater flexibility, allowing us to independently optimize and adapt to different science domains (e.g., astronomy).

Ocean Topic Definition. To provide researchers with a clear and organized resources, we manually categorize the data in ocean science into five major ocean topics, which are based on the expertise of experts in oceanography. The definitions of these five topics comprehensively cover all the main areas of ocean science and are relatively independent. The detailed explanation for the five major topics is described as follows:

- *Science and research* focuses on the fundamental scientific theories and research related to the ocean, such as ocean currents, sea temperatures and ocean biodiversity. This portion of data separately helps drive the advancement of pure scientific research and theories.

- *Resources and development* includes fisheries, minerals, oil and gas, as well as other sustainable development resources. It is set for a better examination and planning of the rational development of ocean resources.

- *Ecology and environment.* Environmental protection and ecological sustainability are currently global hot topics. It helps to address issues such as ocean pollution, ecological degradation, and the impact of climate change on the oceans in a more focused manner.

- *Technology and engineering* encompasses aspects ranging from ocean measurements, observational equipment, and ship engineering to ocean energy development. Such categorization aids in a more focused exploration of ocean engineering and technological needs, while also facilitating interdisciplinary research with other engineering disciplines.

- *Life, culture and others.* The ocean is not only a natural resource or a subject of scientific research; it is also an integral part of culture and lifestyle. This category consists of aspects ranging from history and culture to the mutual influences between the ocean and human societal activities, such as tourism, leisure, art, faith, and more.

While these five topics are distinct, there might be some overlap as well. For instance, some issues related to ocean environmental protection might also be associated with the technology of ocean engineering. For the sake of convenience in data analysis, in the actual construction of the dataset, we map each sample to the most relevant category it belongs to.

Algorithm 1 Domain Instruction Data Generation

Require:
- Seed dataset S with format (instruction, output);
- Ocean literature corpus O;
- Pre-defined rules R for filtering;

Ensure: High-quality instruction dataset H
    Initialize empty datasets: Step1Data $= \emptyset$, Step2Data $= \emptyset$, $H = \emptyset$
                                 ▷ Agent Collaboration as Domain Experts;
    for each sample in S **do**
        instruction, output $\leftarrow$ sample
        enriched_sample $\leftarrow$ Agent_Enrich(instruction, output)
        refined_sample $\leftarrow$ Agent_Refine(instruction, output)
        Step1Data $\leftarrow$ Step1Data $\cup$ refined_sample
    end for
                              ▷ Fine-Tuned Agent as Literature Extractor;
    RetrievedTexts $\leftarrow$ BM25_Retrieve(O)
    Model $M \leftarrow$ Agent_FineTuning(S)
    for each document in RetrievedTexts **do**
        output $\leftarrow$ document.content
        instruction $\leftarrow$ Model(output)
        Step2Data $\leftarrow$ Step2Data $\cup$ (instruction, output)
    end for
                             ▷ Agent as Inspector with Rule Constraints;
    MergedData $\leftarrow$ Agent_Inspector(Step1Data, Step2Data, R)
                                  ▷ Quality Control by Debating;
    for each sample in MergedData **do**
        instruction, output $\leftarrow$ sample
        debate_result $\leftarrow$ Agent_Debate(instruction, output)
        if debate_result is high-quality **then**
            $H \leftarrow H \cup$ sample
        end if
    end for
    return H

Agents as Domain (Ocean) Experts As shown in Figure 3 and Algorithm 1, we leverage agents as domain experts for each ocean topic and make them rapidly expand the instructions by collaboration. We first collect the seed instruction data and propose three strategies by using multiple agents acting as experts. The seed instruction dataset comes from annotations by ocean science researchers, which includes 5 major categories, over 500 sub-categories, and a total of more than 10,000 data entries.

- **Evolving Agent as the Generator**. We design an evolving approach that selects samples from the seed instruction dataset and simultaneously calls upon two agents (*gpt-3.5-turbo*) to evolve the selected samples. The evolution procedure includes two aspects: (1) we enrich the content of the sample by having the agent automatically add relevant background knowledge to it; (2) we guide the agent to refine the sample by conducting a more in-depth analysis of specific concepts or entities. Through multiple rounds of iterative execution, our method can rapidly expand the existing seed dataset, which allows for the rapid expansion of both the breadth and depth of information.

- **Fine-Tuned Agent as the Literature Extractor**. As shown in Figure 3, we collect a smaller expert-annotated corpus and use the *BM25* to retrieve high quality sentences in a larger ocean corpus. We use the retrieved texts as high-quality candidate samples. Meanwhile, we fine-tune *gpt-3.5-turbo* with the seed instruction dataset, regarding the fine-tuned agent as the literature ex-

tractor. In other words, it can automatically extract questions (*instruction*) from the unsupervised ocean science corpus (*output*). Therefore, we utilize the agent to automatically build pairs of *(instruction, output)* on external ocean science literature.

- **Agent as the Inspector with Rule Constraints**. For the massively generated instructions, we use the pre-defined rules as constraints and perform filtering on the data. These rules include syntactic and semantic constraints as well as some basic definitions in the ocean domain. We describe these rules using natural language because many constraints and norms related to ocean science cannot be directly represented with expressions. Therefore, we provide prompts to the *gpt-3.5-turbo* API as demonstrations, letting it play the role of an Inspector. Our approach ensures that the generated ocean instruction data is of higher quality. The detailed prompt agent is shown in Table 4.
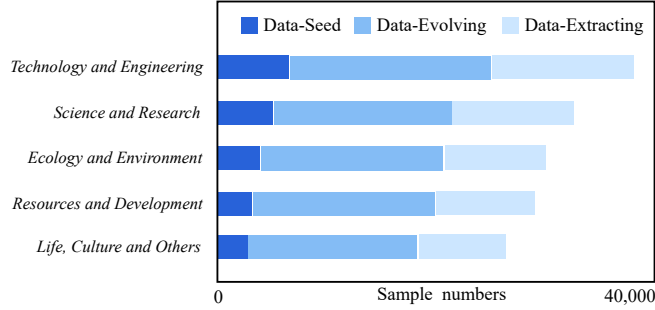

Figure 4: Statistics of our final instruction dataset. We use DOINSTRUCT to expand more than 15,000 instructions (*data-evolving*, *data-extracting*) from the seed data (*data-seed*).

Finally, we assign two extra *gpt-3.5-turbo* agents as roles to debate the quality of data and obtain high-quality instruction dataset. The designed framework is capable of rapidly constructing a ocean science dataset using multi-agents, and by incorporating external knowledge, it overcomes the limitations inherent to the agents themselves. Our framework can also be effectively applied to the instruction data construction in other scientific domains. It should be noted that we separately synthesize robot instructions to equip **OCEANGPT** with the capability to interact with the environment. The statistics of our dataset are shown in Figure 4.

Quality Control for the Dateset. We ask domain experts to carefully review and check data to ensure quality. Specifically, the human volunteers are first trained to make sure they have a comprehensive understanding of the task. Then we develop a platform that can help experts to sample instances from the instruction dataset. Next, the trained domain experts are asked to validate if there are potential errors in the sampled instances. The final IAA (inter-annotator agreement) score for our dataset is 0.82, which satisfies the research purpose.

4   BENCHMARKING OCEAN SCIENCE TASKS

We provide detailed explanations of the experimental setup and the baseline models in Section 4.1. In Section 4.2, we construct an ocean-specific benchmark **OCEANBENCH** to evaluate the capabilities of our **OCEANGPT**. We describe the automatic and human evaluation in Section 4.2.

4.1   IMPLEMENTATION DETAILS AND BASELINES

For the pre-training stage, we pre-train our **OCEANGPT** based on the LLaMA-2 (Touvron et al., 2023b) for seven days with six A800 Nvidia GPUs. For the instruction-tuning stage, we employ the LoRA method (Hu et al., 2021) to fine-tune the pre-trained model and choose three baseline models for comparison. We use the chat version of LLaMA-2 (*Llama-2-7b-chat-hf*) , which is a generative language model optimized for dialogue use cases. We also use *Vicuna-1.5* (Chiang et al., 2023), a chat model which fine-tunes LLaMA-2 on dataset collected from ShareGPT. We further use *ChatGLM2-6B*, the optimized version of GLM (Zeng et al., 2023). The detailed experimental settings are shown in Table 3 (Appendix A.3).

4.2 DATASETS AND METRICS

OCEANBENCH.   To evaluate the capabilities of LLMs for oceanography tasks, we design a benchmark called **OCEANBENCH**. Our benchmark includes a total of 15 type of ocean-related tasks, such as question-answering, extraction, and description tasks. Our evaluation samples are automatically generated from the seed dataset and have undergone deduplication and manual verification by domain experts. For the quality control, we further sample part of data and ask domain experts to evaluate the quality (those disagreed cases or bad cases will be manually fixed by domain experts.). The distribution of our **OCEANBENCH** and the statistics can be found in Table 2.

Automatic Evaluation.   To evaluate the performance and reduce reliance on manual evaluation, we leverage GPT-4 as the evaluator for the experimental results. Inspired by Wang et al. (2023c;b), we utilize an effective calibration method to evaluate the performance of two LLMs. For each testing question, given two responses and from two LLMs, we query the GPT4 to obtain the comparison result. We note that LLMs are sensitive to the position of responses, therefore, alleviating the positional bias is very important. To balance the position bias, we exchange the order of the responses to form the new prompt. The final evaluating result is the sum of the test results for the two prompts with their order swapped.

Human Evaluation.   To validate our proposed framework, we also collect the output data in different settings and evaluate it by human evaluation. We employ 5 students with an ocean science background as human annotators. For each evaluation setting, we sample a set of 200 examples and human annotators will rank the outputs they prefer. The total expense is about 500 US dollars.

5 EVALUATION
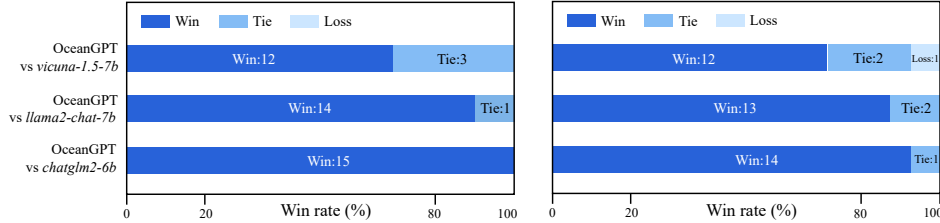
5.1 INSIGHTS FROM PERFORMANCE RESULTS

Figure 5:   Ocean task-level results. **Left**: Automatic evaluation. **Right**: Human evaluation. Compared to baselines, **OCEANGPT** performs better than *llama2-chat-7b*, *vicuna-1.5-7b* and *chatglm2-6b* in both two settings. The instance-level result is in Table 11 (Appendix A.2).

OCEANGPT can obtain better performance than previous open-sourced LLMs.   In Figure 5, we compare the performance of **OCEANGPT** with the three baseline models across 15 subtasks at the task-level in the ocean domain. We utilize both automatic and human evaluators, then compute the *win rate (%)* with baseline models. Compared to the baselines (*llama2-chat-7b*, *vicuna-1.5-7b*, *chatglm2-6b*)[2], **OCEANGPT** outperforms in the majority of tasks, which demonstrates the effectiveness of the proposed approach.

OCEANGPT excels in a range of ocean science tasks.   As shown in Figure 6, we present detailed automatic evaluation experimental results in the **OCEANBENCH**. It can be clearly seen that our model is superior to baseline language models in the vast majority of tasks. Note that previous open-sourced LLMs even fail to handle several expertise ocean tasks (e.g., Editing). While our designed multi-agent data generation framework can effectively act as experts in various subfields of the ocean domain, which indicates that **OCEANGPT** is a better expert in various ocean domains.

---

[2]We have trained OceanGPT-7B, thus we only compare open-sourced LLMs around 7B. We are running resources to train OceanGPT-30B, and will compare with more LLMs in the future.
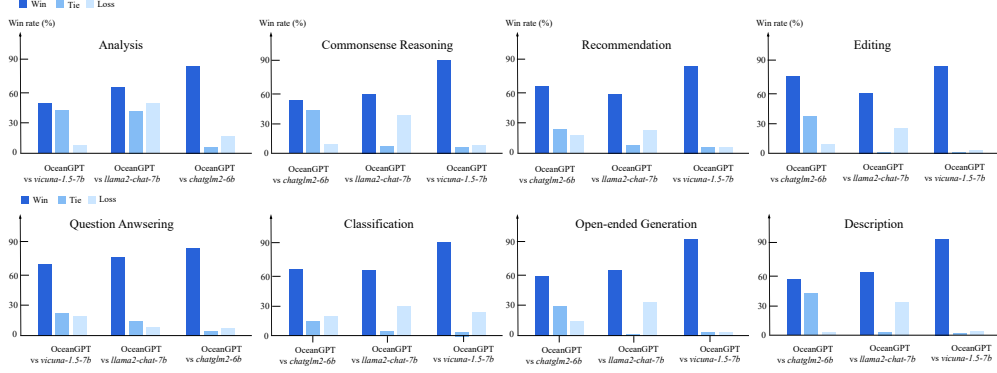
Figure 6: Evaluation results of **OCEANGPT** in the ocean science tasks in **OCEANBENCH**. The complete experimental results are shown in the Table 10 (Appendix A.2).
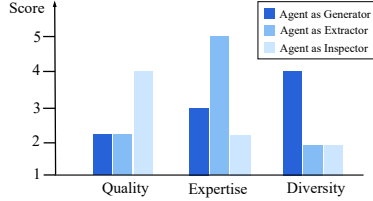

Figure 7: Performance analysis for agents.

DOINSTRUCT are the effective ocean data generators by multi-agent collaboration. As shown in Figure 7, we design three indicators to measure the data generation effect of our proposed method from the perspectives of knowledge quality, expertise and diversity. We use manual evaluation to calculate the scores of the three indicators from 1 to 5. The higher the score, the better the effect of the testing model. It can be seen that the evolving generator agent can effectively enhance the richness of ocean data. When the extraction agent is at work, the expertise of the content is greatly improved. At the same time, the inspector agent plays a significant role in enhancing the quality of the generated data. The results show that multi-agent collaboration is effective for ocean instruction generation.

5.2 EXPLORING THE POTENTIAL OF OCEANGPT

In this section, we explore the potential of **OCEANGPT** from the perspectives of ocean science and ocean engineering. For ocean science (Section 5.2.1), we focus on the key scientific issues of nuclear pollution in the ocean environment. For ocean engineering (Section 5.2.2), we use Gazebo [3] as the simulator (Manhães et al., 2016) to test **OCEANGPT**'s ability to control underwater robots.

5.2.1 OCEANGPT FOR OCEAN SCIENCE

In Figure 8, given the question *"How to conduct research on interfacial chemistry and toxicological effects of key radioactive nuclides"*, we compare the outputs of our model and *vicuna-1.5-7b*. The result shows that **OCEANGPT** demonstrates a higher level of knowledge expertise when describing the content of radioactive nuclide research. Its textual content is not only clear in structure and well-organized, but also covers various aspects of radioactive nuclide research, from experimental design to data analysis, and then to risk assessment and disposal guidelines. In contrast, although *vicuna-1.5-7b* has clear expression and logicality, it lacks depth and specific content related to radioactive nuclides. Overall, **OCEANGPT** has advantages in terms of knowledge expertise, quality, and richness. The complete outputs of both models are shown in the Table 5.

5.2.2 OCEANGPT FOR OCEAN ENGINEERING

Ocean engineering focuses on the design, development, and management of structures and systems within the ocean environment. It plays an indispensable role in harnessing the vast potential of the oceans while ensuring sustainable and secure maritime operations. To facilitate interaction between **OCEANGPT** and the external world, we integrate machine code instructions into its training data.

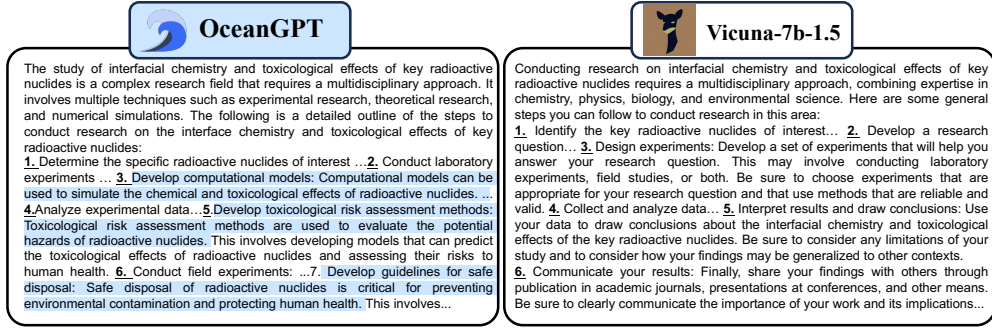[3] https://github.com/uuvsimulator/uuv_simulator

Figure 8: Case analysis on ocean science task. We use `blue` font to represent the different texts by **OCEANGPT**. The instruction is: `How to conduct research on interfacial chemistry and toxicological effects of key radioactive nuclides?`
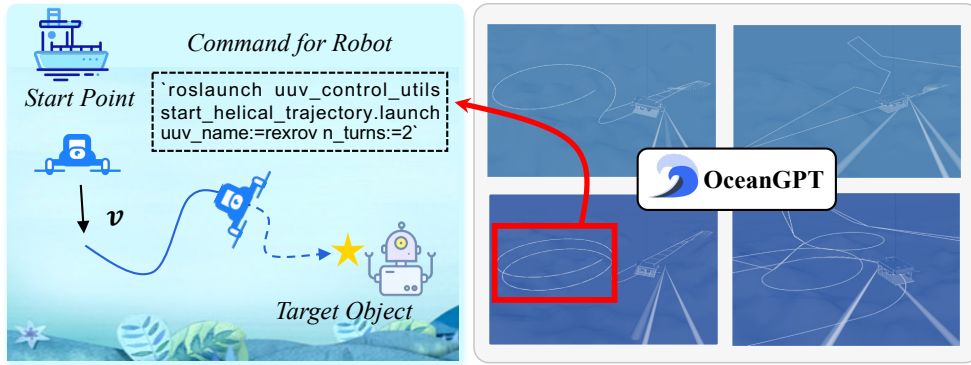

Figure 9: Our model can be instructed for underwater robot control in the simulation platform of Gazebo which shows **OCEANGPT** gains preliminary embodied intelligence capabilities.

As depicted in Figure 9, **OCEANGPT** can instruct underwater robots via code or console commands, allowing them to execute basic path-finding operations. In this example, by using natural language as a prompt, the **OCEANGPT** can automatically generate code (the robot generate a double helix path) for underwater robot to operate based on human instructions. In fact, the experimental result suggests that **OCEANGPT** has the potential to acquire embodied intelligence. Though we make preliminary attempts for ocean robot interaction, it paves the way for future advanced oceanic models to undertake intricate robotic control and planning tasks.

6 CONCLUSION

In this paper, we introduce **OCEANGPT**, the first-ever oceanographic pre-trained language model, which is expert in various ocean science tasks. To alleviate the difficulties for obtaining ocean data, we propose an domain construction framework called DOINSTRUCT, which constructs the ocean instruction dataset by multi-agent collaboration. Each agent in our designed framework is considered as an expert in a specific topic and is responsible for generating the corresponding data. Our generated dataset consists of diverse instructions to align the desired behaviors in ocean science issues. Additionally, we establish the first oceanography benchmark, **OCEANBENCH**, to evaluate the capabilities of LLMs in ocean domain. Though comprehensive analysis, we observe that **OCEANGPT** not only demonstrates a higher level of knowledge expertise for oceans science tasks but also gains preliminary embodied intelligence capabilities in ocean engineering. We will continue to improve **OCEANGPT** by training on larger corpus with larger models (e.g., 30B, 70B) and maintain **OCEANBENCH** by adding new data and tasks.

ETHICAL CONSIDERATIONS

The datasets and model in this paper are designed for exploratory analysis. Within the domain of LLMs, the distribution of both pre-training data and instruction data might exhibit biases, potentially influencing the model's outputs. Throughout experiments, **we have carefully examined the sampled data, eliminating any content that could be deemed toxic or offensive**. Yet, it's possible that **OCEANGPT** might occasionally yield incorrect or imprecise results. Hence, users leveraging the insights provided by this model should exercise their discretion and bear the risk.

REPRODUCIBILITY STATEMENT

Pre-training and instruction-tuning: `https://github.com/zjunlp/KnowLM`.

Instruction processing: `https://github.com/zjunlp/EasyInstruct` (The code is still under organization and will be integrated into EasyInstruct soon.).

Model will be released at HuggingFace soon.

REFERENCES

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL `https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html`.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, March 2023. URL `https://lmsys.org/blog/2023-03-30-vicuna/`.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. Palm: Scaling language modeling with pathways. *CoRR*, abs/2204.02311, 2022. doi: 10.48550/arXiv.2204.02311. URL `https://doi.org/10.48550/arXiv.2204.02311`.

Cheng Deng, Tianhang Zhang, Zhongmou He, Qiyuan Chen, Yuanyuan Shi, Le Zhou, Luoyi Fu, Weinan Zhang, Xinbing Wang, Chenghu Zhou, Zhouhan Lin, and Junxian He. Learning A foundation language model for geoscience knowledge understanding and utilization. *CoRR*, abs/2306.05064, 2023. doi: 10.48550/arXiv.2306.05064. URL `https://doi.org/10.48550/arXiv.2306.05064`.

Wayne E Esaias, Mark R Abbott, Ian Barton, Otis B Brown, Janet W Campbell, Kendall L Carder, Dennis K Clark, Robert H Evans, Frank E Hoge, Howard R Gordon, et al. An overview of

modis capabilities for ocean science observations. *IEEE Transactions on Geoscience and Remote Sensing*, 36(4):1250–1265, 1998.

Paul Falkowski. Ocean science: the power of plankton. *Nature*, 483(7387):S17–S20, 2012.

Yin Fang, Qiang Zhang, Ningyu Zhang, Zhuo Chen, Xiang Zhuang, Xin Shao, Xiaohui Fan, and Huajun Chen. Knowledge graph-enhanced molecular contrastive learning with functional prompt. *Nature Machine Intelligence*, 5:1–12, 05 2023. doi: 10.1038/s42256-023-00654-0.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021.

Jinhao Jiang, Kun Zhou, Zican Dong, Keming Ye, Wayne Xin Zhao, and Ji-Rong Wen. Structgpt: A general framework for large language model to reason over structured data. *CoRR*, abs/2305.09645, 2023. doi: 10.48550/arXiv.2305.09645. URL https://doi.org/10.48550/arXiv.2305.09645.

Xuchen Jin, Xianqiang He, Difeng Wang, Jianyun Ying, Fang Gong, Qiankun Zhu, Chenghu Zhou, and Delu Pan. Impact of rain effects on l-band passive microwave satellite observations over the ocean. *IEEE Trans. Geosci. Remote. Sens.*, 61:1–16, 2023. doi: 10.1109/TGRS.2022.3232402. URL https://doi.org/10.1109/TGRS.2022.3232402.

Zeljko Kraljevic, Anthony Shek, Daniel Bean, Rebecca Bendayan, James T. Teo, and Richard J. B. Dobson. Medgpt: Medical concept prediction from clinical narratives. *CoRR*, abs/2107.03134, 2021. URL https://arxiv.org/abs/2107.03134.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Salvatore Candido, and Alexander Rives. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023. doi: 10.1126/science.ade2574. URL https://www.science.org/doi/abs/10.1126/science.ade2574.

Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. Biogpt: generative pre-trained transformer for biomedical text generation and mining. *Briefings Bioinform.*, 23(6), 2022. doi: 10.1093/bib/bbac409. URL https://doi.org/10.1093/bib/bbac409.

Musa Morena Marcusso Manhães, Sebastian A. Scherer, Martin Voss, Luiz Ricardo Douat, and Thomas Rauschenbach. UUV simulator: A gazebo-based package for underwater intervention and multi-robot simulation. In *OCEANS 2016 MTS/IEEE Monterey*. IEEE, sep 2016. doi: 10.1109/oceans.2016.7761080. URL https://doi.org/10.1109%2Foceans.2016.7761080.

Michael Moor, Oishi Banerjee, Zahra Shakeri, Harlan Krumholz, Jure Leskovec, Eric Topol, and Pranav Rajpurkar. Foundation models for generalist medical artificial intelligence. *Nature*, 616: 259–265, 04 2023. doi: 10.1038/s41586-023-05881-4.

OpenAI. Gpt-4 technical report, 2023.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *NeurIPS*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/b1efde53be364a73914f58805a001731-Abstract-Conference.html.

Shuofei Qiao, Honghao Gui, Huajun Chen, and Ningyu Zhang. Making language models better tool learners with execution feedback. *CoRR*, abs/2305.13068, 2023a. doi: 10.48550/arXiv.2305.13068. URL https://doi.org/10.48550/arXiv.2305.13068.

Shuofei Qiao, Yixin Ou, Ningyu Zhang, Xiang Chen, Yunzhi Yao, Shumin Deng, Chuanqi Tan, Fei Huang, and Huajun Chen. Reasoning with language model prompting: A survey. In Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pp. 5368–5393. Association for Computational Linguistics, 2023b. doi: 10.18653/v1/2023.acl-long.294. URL https://doi.org/10.18653/v1/2023.acl-long.294.

Jack W. Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, H. Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, Eliza Rutherford, Tom Hennigan, Jacob Menick, Albin Cassirer, Richard Powell, George van den Driessche, Lisa Anne Hendricks, Maribeth Rauh, Po-Sen Huang, Amelia Glaese, Johannes Welbl, Sumanth Dathathri, Saffron Huang, Jonathan Uesato, John Mellor, Irina Higgins, Antonia Creswell, Nat McAleese, Amy Wu, Erich Elsen, Siddhant M. Jayakumar, Elena Buchatskaya, David Budden, Esme Sutherland, Karen Simonyan, Michela Paganini, Laurent Sifre, Lena Martens, Xiang Lorraine Li, Adhiguna Kuncoro, Aida Nematzadeh, Elena Gribovskaya, Domenic Donato, Angeliki Lazaridou, Arthur Mensch, Jean-Baptiste Lespiau, Maria Tsimpoukelli, Nikolai Grigorev, Doug Fritz, Thibault Sottiaux, Mantas Pajarskas, Toby Pohlen, Zhitao Gong, Daniel Toyama, Cyprien de Masson d'Autume, Yujia Li, Tayfun Terzi, Vladimir Mikulik, Igor Babuschkin, Aidan Clark, Diego de Las Casas, Aurelia Guy, Chris Jones, James Bradbury, Matthew J. Johnson, Blake A. Hechtman, Laura Weidinger, Iason Gabriel, William Isaac, Edward Lockhart, Simon Osindero, Laura Rimell, Chris Dyer, Oriol Vinyals, Kareem Ayoub, Jeff Stanway, Lorrayne Bennett, Demis Hassabis, Koray Kavukcuoglu, and Geoffrey Irving. Scaling language models: Methods, analysis & insights from training gopher. *CoRR*, abs/2112.11446, 2021. URL https://arxiv.org/abs/2112.11446.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilic, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurençon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris Emezue, Christopher Klamm, Colin Leong, Daniel van Strien, David Ifeoluwa Adelani, and et al. BLOOM: A 176b-parameter open-access multilingual language model. *CoRR*, abs/2211.05100, 2022. doi: 10.48550/arXiv.2211.05100. URL https://doi.org/10.48550/arXiv.2211.05100.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Kumar Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Nathaneal Schärli, Aakanksha Chowdhery, Philip Andrew Mansfield, Blaise Agüera y Arcas, Dale R. Webster, Gregory S. Corrado, Yossi Matias, Katherine Chou, Juraj Gottweis, Nenad Tomasev, Yun Liu, Alvin Rajkomar, Joelle K. Barral, Christopher Semturs, Alan Karthikesalingam, and Vivek Natarajan. Large language models encode clinical knowledge. *CoRR*, abs/2212.13138, 2022. doi: 10.48550/arXiv.2212.13138. URL https://doi.org/10.48550/arXiv.2212.13138.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca, 2023.

Christina Theodoris, Ling Xiao, Anant Chopra, Mark Chaffin, Zeina Sayed, Matthew Hill, Helene Mantineo, Elizabeth Brydon, Zexian Zeng, Shirley Liu, and Patrick Ellinor. Transfer learning enables predictions in network biology. *Nature*, 618:1–9, 05 2023. doi: 10.1038/s41586-023-06139-9.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, YaGuang Li, Hongrae Lee, Huaixiu Steven Zheng, Amin Ghafouri, Marcelo Menegali, Yanping Huang, Maxim Krikun, Dmitry Lepikhin, James Qin, Dehao Chen, Yuanzhong Xu, Zhifeng Chen, Adam Roberts, Maarten Bosma, Yanqi

Zhou, Chung-Ching Chang, Igor Krivokon, Will Rusch, Marc Pickett, Kathleen S. Meier-Hellstern, Meredith Ringel Morris, Tulsee Doshi, Renelito Delos Santos, Toju Duke, Johnny Soraker, Ben Zevenbergen, Vinodkumar Prabhakaran, Mark Diaz, Ben Hutchinson, Kristen Olson, Alejandra Molina, Erin Hoffman-John, Josh Lee, Lora Aroyo, Ravi Rajakumar, Alena Butryna, Matthew Lamm, Viktoriya Kuzmina, Joe Fenton, Aaron Cohen, Rachel Bernstein, Ray Kurzweil, Blaise Agüera y Arcas, Claire Cui, Marian Croak, Ed H. Chi, and Quoc Le. Lamda: Language models for dialog applications. *CoRR*, abs/2201.08239, 2022. URL https://arxiv.org/abs/2201.08239.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971, 2023a. doi: 10.48550/arXiv.2302.13971. URL https://doi.org/10.48550/arXiv.2302.13971.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288, 2023b. doi: 10.48550/arXiv.2307.09288. URL https://doi.org/10.48550/arXiv.2307.09288.

Martin Visbeck. Ocean science research is key for a sustainable future. *Nature communications*, 9 (1):690, 2018.

Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, Wayne Xin Zhao, Zhewei Wei, and Ji-Rong Wen. A survey on large language model based autonomous agents. *CoRR*, abs/2308.11432, 2023a. doi: 10.48550/arXiv.2308.11432. URL https://doi.org/10.48550/arXiv.2308.11432.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Qi Liu, Tianyu Liu, and Zhifang Sui. Large language models are not fair evaluators, 2023b.

Xingyao Wang, Zihan Wang, Jiateng Liu, Yangyi Chen, Lifan Yuan, Hao Peng, and Heng Ji. Mint: Evaluating llms in multi-turn interaction with tools and language feedback, 2023c.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-instruct: Aligning language models with self-generated instructions. In Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pp. 13484–13508. Association for Computational Linguistics, 2023d. doi: 10.18653/v1/2023.acl-long.754. URL https://doi.org/10.18653/v1/2023.acl-long.754.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. Finetuned language models are zero-shot learners. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. URL https://openreview.net/forum?id=gEZrGCozdqR.

Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, Rui Zheng, Xiaoran Fan, Xiao Wang, Limao Xiong, Yuhao Zhou, Weiran Wang, Changhao Jiang, Yicheng Zou, Xiangyang Liu, Zhangyue Yin, Shihan Dou, Rongxiang Weng, Wensen Cheng, Qi Zhang, Wenjuan Qin, Yongyan Zheng, Xipeng

Qiu, Xuanjing Huan, and Tao Gui. The rise and potential of large language model based agents: A survey. *CoRR*, abs/2309.07864, 2023. doi: 10.48550/arXiv.2309.07864. URL https://doi.org/10.48550/arXiv.2309.07864.

Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, and Daxin Jiang. Wizardlm: Empowering large language models to follow complex instructions. *CoRR*, abs/2304.12244, 2023. doi: 10.48550/arXiv.2304.12244. URL https://doi.org/10.48550/arXiv.2304.12244.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models. *CoRR*, abs/2306.13549, 2023. doi: 10.48550/arXiv.2306.13549. URL https://doi.org/10.48550/arXiv.2306.13549.

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, Weng Lam Tam, Zixuan Ma, Yufei Xue, Jidong Zhai, Wenguang Chen, Zhiyuan Liu, Peng Zhang, Yuxiao Dong, and Jie Tang. GLM-130B: an open bilingual pre-trained model. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/pdf?id=-Aw0rrrPUF.

Liangyu Zha, Junlin Zhou, Liyao Li, Rui Wang, Qingyi Huang, Saisai Yang, Jing Yuan, Changbao Su, Xiang Li, Aofeng Su, Tao Zhang, Chen Zhou, Kaizhe Shou, Miao Wang, Wufang Zhu, Guoshan Lu, Chao Ye, Yali Ye, Wentao Ye, Yiming Zhang, Xinglong Deng, Jie Xu, Haobo Wang, Gang Chen, and Junbo Zhao. Tablegpt: Towards unifying tables, nature language and commands into one GPT. *CoRR*, abs/2307.08674, 2023. doi: 10.48550/arXiv.2307.08674. URL https://doi.org/10.48550/arXiv.2307.08674.

Ningyu Zhang, Jintian Zhang, Xiaohan Wang, Honghao Gui, Kangwei Liu, Yinuo Jiang, Xiang Chen, Shengyu Mao, Shuofei Qiao, Yuqi Zhu, Zhen Bi, Jing Chen, Xiaozhuan Liang, Yixin Ou, Runnan Fang, Zekun Xi, Xin Xu, Lei Li, Peng Wang, Mengru Wang, Yunzhi Yao, Bozhong Tian, Yin Fang, Guozhou Zheng, and Huajun Chen. Knowlm technical report, 2023a. URL http://knowlm.zjukg.cn/.

Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Fei Wu, and Guoyin Wang. Instruction tuning for large language models: A survey. *CoRR*, abs/2308.10792, 2023b. doi: 10.48550/arXiv.2308.10792. URL https://doi.org/10.48550/arXiv.2308.10792.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona T. Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. OPT: open pre-trained transformer language models. *CoRR*, abs/2205.01068, 2022. doi: 10.48550/arXiv.2205.01068. URL https://doi.org/10.48550/arXiv.2205.01068.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. A survey of large language models. *CoRR*, abs/2303.18223, 2023. doi: 10.48550/arXiv.2303.18223. URL https://doi.org/10.48550/arXiv.2303.18223.

A   APPENDIX

A.1   LIMITATIONS AND ETHICAL CONCERNS

A.1.1   BIAS IN DATA DISTRIBUTION

In the realm of LLMs, the distribution of pre-training data and instruction data can be subject to substantial biases, which can shape the outputs of these models. Pre-training data for LLMs often comes from the internet, a vast and potentially biased source of information. The Internet content is inherently skewed, reflecting the biases of its contributors, and hence may not represent a balanced global perspective. Similarly, instruction data can also carry the biases of the humans who create these instructions. For instance, instruction developed by individuals with a particular cultural, socioeconomic, or educational background may inadvertently favor specific perspectives, languages, or communication styles and marginalize others. This bias in data distribution can result in models that reinforce existing prejudices, lack cultural sensitivity, or fail to accurately understand and generate content in underrepresented languages or dialects.

A.1.2   HALLUCINATION IN LARGE LANGUAGE MODELS

Although LLMs have shown tremendous success in general domains of NLP, there is a notable issue regarding their tendency to produce hallucinations. Hallucinations refer to instances where LLMs occasionally generate content that deviates from the user's input, contradicts previously generated context, or conflicts with established world knowledge. By developing strategies to address the issue of hallucination, LLMs can better align their outputs with user intent, preserve coherence within generated content, and enhance their overall utility in real-world applications.
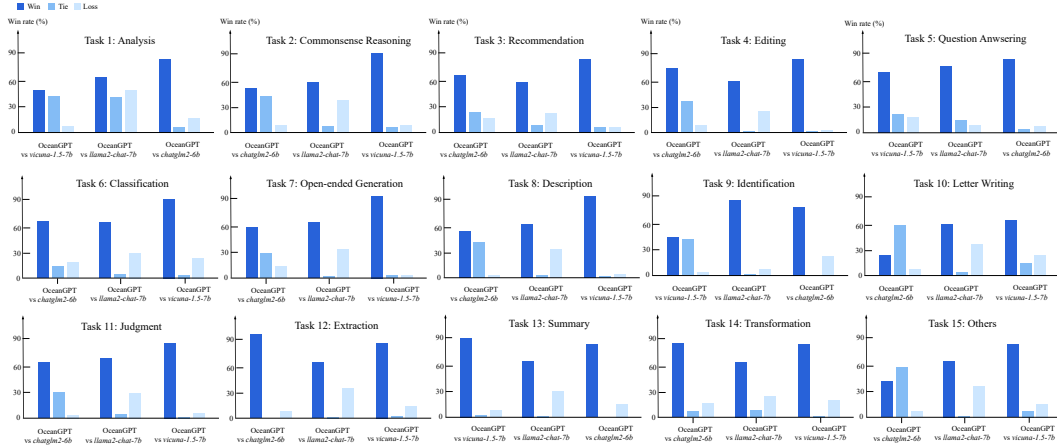
A.2   COMPLETE RESULTS IN OCEANBENCH

Figure 10:   Automatic evaluation results of **OCEANGPT** in all tasks in **OCEANBENCH**.

A.3   IMPLEMENTATION DETAILS

Journal	Num	Journal	Num
Oceanography	1,359	Ocean Science	1,168
Ocean Dynamics	2,825	Marine Biology	10,485
Aquatic Sciences	2,522	Fisheries Science	2,371
Maritime Studies	313	Marine Biotechnology	1,794
Ocean Science Journal	717	Physical Oceanography	455
Marine Geophysical Research	1,155	Journal of Oceanography	2,580
Journal of Physical Oceanography	3,588	Marine Ecology Progress Series	1,494
Fish Physiology and Biochemistry	3,251	Atmospheric and Oceanic Optics	1,156
Reviews in Fish Biology and Fisheries	1,158	Journal of Oceanology and Limnology	3,510
Journal of Atmospheric and Oceanic Technology	3,298	Journal of Ocean Engineering and Marine Energy	266
Journal of the Marine Biological Association of the United Kingdom	7,676	Journal of Offshore Mechanics and Arctic Engineering	2,015

Table 1: The statistics of the pre-training corpus. **Num** denotes the number of documents.

Task	Num	Task	Num	Task	Num
Analysis	674	Classification	895	Judgment	655
Commonsense Reasoning	1,024	Open-ended Generation	930	Extraction	1,078
Recommendation	1,089	Description	1,246	Summary	149
Editing	1,075	Identification	464	Transformation	401
Question Answering	1,230	Letter Writing	359	Others	157

Table 2: The statistics of **OCEANBENCH**. **Num** denotes the number of samples.
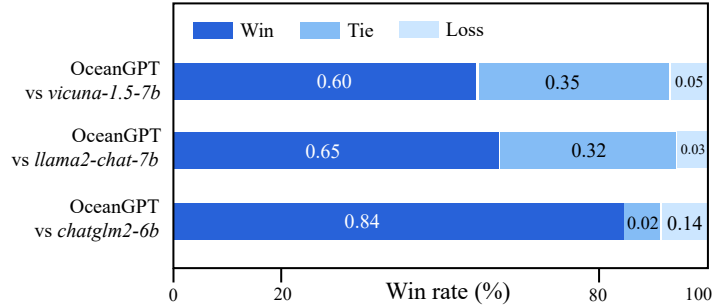
Figure 11: Instance-level results (automatic evaluation) in 15 tasks of **OCEANBENCH**.

Hyperparameter	
Fine-tuning method	LoRA
Batch Size	512
Device†	NVIDIA A800
GPU number	6
Learning Rate (LR)	$1e-4$
LoRA r	8
LoRA α	16
LoRA Dropout	0.05
Epoch	10

Table 3: Detailed experimental settings.

Figure 12: Distribution of our **OCEANBENCH**.

Prompt for "Fine-Tuned Agent as the Literature Extractor":
You are a helpful ocean assistant. You are to extract the question from each of the answer provided.
Answer: This is a seahorse, belonging to the family Syngnathidae. Seahorses are vertebrates commonly found in tropical and subtropical waters. They have unique morphology and biological characteristics and are important organisms in marine ecosystems.

Prompt for "Evolving Agent as the Generator":
Assuming you are an expert in marine engineering and resources, please keep the meaning of the following sentences unchanged and provide as much professional knowledge as possible.
Sentences:Please recommend some mineral resources found in the East China Sea.

Prompt for "Agent as the Inspector with Rule Constraints":
Assuming you are an inspector in marine science, please filter and judge the sentences in 'Sentences' based on the constraints provided below:
Constraints: Keyword Filter: Focus on literature that prominently mentions the terms 'coral reefs', 'ocean acidification', or 'deep-sea exploration'. Date Range: Only consider articles published between 2010 and 2022. Author Filter: Prioritize works by the Oceanic Research Institute. Type of Literature: Specifically look for 'experimental studies' and 'review articles'. Exclude 'conference papers'. Geographical Focus: Highlight research that pertains to the Pacific Ocean region. Language Constraint: Only select literature written in English. Abstract Inclusion: Ensure the abstract contains the phrase 'climate impact'. Abstract Exclusion: Exclude any literature whose abstract mentions 'laboratory simulation'.

Prompt for automatic evaluation using GPT4:
Please check if following sentences contain rich ocean related information. If so, output "related". Otherwise, output "unrelated".
Sentences: Dissolved organic carbon (DOC) represents the largest pool of reduced carbon in oceans and plays important roles in the ocean carbon cycle and food webs . DOC comprises nearly half of the riverine organic carbon flux into oceans. Riverine DOC is involved in numerous ecosystem functions, including key roles in chemical and biological processes. Refractory and labile DOC are, respectively, important for carbon sequestra-tion in the ocean and a vital food source for marine bacteria.

Table 4: The prompt example that we use in this work.

Instruction: How to conduct research on interfacial chemistry and toxicological effects of key radioactive nuclides?

OCEANGPT

The study of interfacial chemistry and toxicological effects of key radioactive nuclides is a complex research field that requires a multidisciplinary approach. It involves multiple techniques such as experimental research, theoretical research, and numerical simulations. The following is a detailed outline of the steps to conduct research on the interface chemistry and toxicological effects of key radioactive nuclides: 1. Determine the specific radioactive nuclides of interest: Different radioactive nuclides have different chemical properties and toxicological effects. Therefore, it is necessary to select specific radioactive nuclides of interest and establish a research plan. 2. Conduct laboratory experiments: Laboratory experiments are crucial for studying the chemical properties and toxicological effects of radioactive nuclides. The experiments involve measuring the chemical properties of radioactive nuclides, such as their reactivity, stability, and chemical reactivity. 3. Develop computational models: Computational models can be used to simulate the chemical and toxicological effects of radioactive nuclides. This involves developing mathematical models and algorithms to predict the behavior of radioactive nuclides under different conditions.
4. Analyze experimental data: Experimental data obtained from laboratory experiments need to be analyzed and interpreted. This involves studying the experimental results and comparing them with theoretical predictions.
5. Develop toxicological risk assessment methods: Toxicological risk assessment methods are used to evaluate the potential hazards of radioactive nuclides. This involves developing models that can predict the toxicological effects of radioactive nuclides and assessing their risks to human health.
6. Conduct field experiments: Field experiments are necessary to study the behavior of radioactive nuclides in real-world environments. This involves measuring the chemical and toxicological effects of radioactive nuclides in different ecosystems and studying their distribution and transport.
7. Develop guidelines for safe disposal: Safe disposal of radioactive nuclides is critical for preventing environmental contamination and protecting human health. This involves developing guidelines for the safe disposal of radioactive nuclides and their storage and transport.

---

vicuna-1.5-7b

Conducting research on interfacial chemistry and toxicological effects of key radioactive nuclides requires a multidisciplinary approach, combining expertise in chemistry, physics, biology, and environmental science. Here are some general steps you can follow to conduct research in this area: 1. Identify the key radioactive nuclides of interest: Start by identifying the specific radioactive nuclides that you want to study. This may include elements such as uranium, plutonium, and cesium, among others. 2. Develop a research question: Once you have identified the key radioactive nuclides, develop a research question that addresses an important aspect of their interfacial chemistry and toxicological effects. This question should be specific and focused, and should guide your research efforts. 3. Design experiments: Develop a set of experiments that will help you answer your research question. This may involve conducting laboratory experiments, field studies, or both. Be sure to choose experiments that are appropriate for your research question and that use methods that are reliable and valid. 4. Collect and analyze data: Once you have conducted your experiments, collect and analyze the data you have generated. Be sure to carefully document your methods and results, and to interpret your data in a clear and concise manner. 5. Interpret results and draw conclusions: Use your data to draw conclusions about the interfacial chemistry and toxicological effects of the key radioactive nuclides. Be sure to consider any limitations of your study and to consider how your findings may be generalized to other contexts. 6.Communicate your results: Finally, share your findings with others through publication in academic journals, presentations at conferences, and other means. Be sure to clearly communicate the importance of your work and its implications for interfacial chemistry and toxicology.

Table 5: Detailed case analysis on ocean science task. The input prompt instruction is How to conduct research on interfacial chemistry and toxicological effects of key radioactive nuclides?