# A Novel RAG Framework with Knowledge-Enhancement for Biomedical Question Answering

Yongping Du*
*College of Computer Science*
*Beijing University of Technology*
Beijing, China
ypdu@bjut.edu.cn

Zikai Wang
*College of Computer Science*
*Beijing University of Technology*
Beijing, China
wangzikai@emails.bjut.edu.cn

Binrui Wang
*College of Computer Science*
*Beijing University of Technology*
Beijing, China
wbiinrw@emails.bjut.edu.cn

Xingnan Jin
*College of Computer Science*
*Beijing University of Technology*
Beijing, China
jinxingnan@outlook.com

Yu Pei
*College of Computer Science*
*Beijing University of Technology*
Beijing, China
peiyu@emails.bjut.edu.cn

2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) | 979-8-3503-8622-6/24/$31.00 ©2024 IEEE | DOI: 10.1109/BIBM62325.2024.10822837

***Abstract*—The biomedical question-answering system usually provide accurate and real-time responses, which is crucial for clinical decision-making and scientific research. Although large language models achieve remarkable results in general question-answering tasks, they still face challenges in specialized fields. This paper proposes a novel framework called RAG-Chain, which aims to enhance the performance of general-domain large models on special biomedical reasoning and question-answering tasks. The RAG-Chain framework improves the knowledge retrieval and generation abilities of general models by a multi-stage processing of external knowledge and automatic construction of chain-of-thought templates combined with self-consistency validation process of choice shuffling. The experimental results show that RAG-Chain improves the accuracy of the baseline model by an average of 6.9% on the MedQA dataset without the need for pre-training or fine-tuning in biomedical fields, verifying its strong adaptability and effectiveness in different large language models.**

***Keywords*—biomedical question-answering, large language models, RAG-Chain, choice shuffling**

## I. Introduction

The extensive and fast-changing nature of information in the biomedical field prevent researchers and medical professionals from accessing and understanding professional information accurately and quickly. With the development of artificial intelligence, especially the advancement of Natural Language Processing (NLP) technology, biomedical information retrieval and question-answering systems have become a hot research topic. These systems aim to simulate the decision-making process of human experts and provide professional answers to questions and information recommendations by analyzing and understanding natural language text.

In recent years, deep learning-based biomedical information processing models have made significant progress. For example, models based on the BERT architecture, such as BioBERT [3], improves the performance of biomedical question-answering task by pre-training and fine-tuning on specialized datasets in the biomedical domain. In addition, models such as BioLinkBERT [4] enhances the model's ability to learn inter-textual dependencies and knowledge by introducing cross-document links and knowledge graphs. These models demonstrate excellent performance on domain-specific tasks.

Despite these achievements, existing biomedical information processing models still have some limitations. Firstly, these models require a large amount of domain-specific data for pre-training and fine-tuning, which leads to a huge cost of computing resources and the limitations in model generality and deployability [5]. Secondly, due to the rapid updating of knowledge in the biomedical field, the usefulness and effectiveness of the model decreases with outdated training data. In addition, these existing models lack sufficient interpretability and stability when dealing with complex reasoning tasks, which is particularly prominent in scenarios such as clinical decision-making, where accuracy is extremely important [6].

In view of the above limitations, it is necessary to develop a biomedical information retrieval and question-answering system that does not require domain-specific pre-training and has high generality and stability [7]. The system needs to be adapted to new domain knowledge quickly while providing accurate answers and sufficient explanations in complex reasoning tasks. This paper provides in-depth research from multiple perspectives and the main contributions. We propose a novel framework that combines document chunking, retrieval rearrangement, document repackaging and multi-prompt template chain of thought to improve the performance of general large language models significantly on biomedical question-answering tasks

## II. Method

We propose the RAG-Chain framework that does not require domain-specific knowledge pre-training and fine-tuning of large language models, achieving comparable performance to general-domain large language models. As shown in Fig. 1, the RAG-Chain framework consists of three stages: (1) Comprehensive Chunking, Retrieval and Restructuring (C2R2), (2) Automatic Construction of Demonstrations (ACD) and (3) Choice Shuffling and Self-Consistency Validation (CSSCV).

### *A. Comprehensive Chunking, Retrieval and Restructuring*

**Chunking and Embedding.** The chunk size of a document affects performance significantly. Larger chunks provide more contextual information and enhance the LLMs' understanding of the question, but they also increase the inference time of LLMs. We select sentence-level chunking to achieve a balance between retaining the semantic meaning of the text and the efficiency of chunking.

Given a document $D = \{S_1, S_2, \ldots, S_n\}$, where $S_i$ represents the sentences appeared in $D$. A sentence segmentation algorithm used for document chunking divide

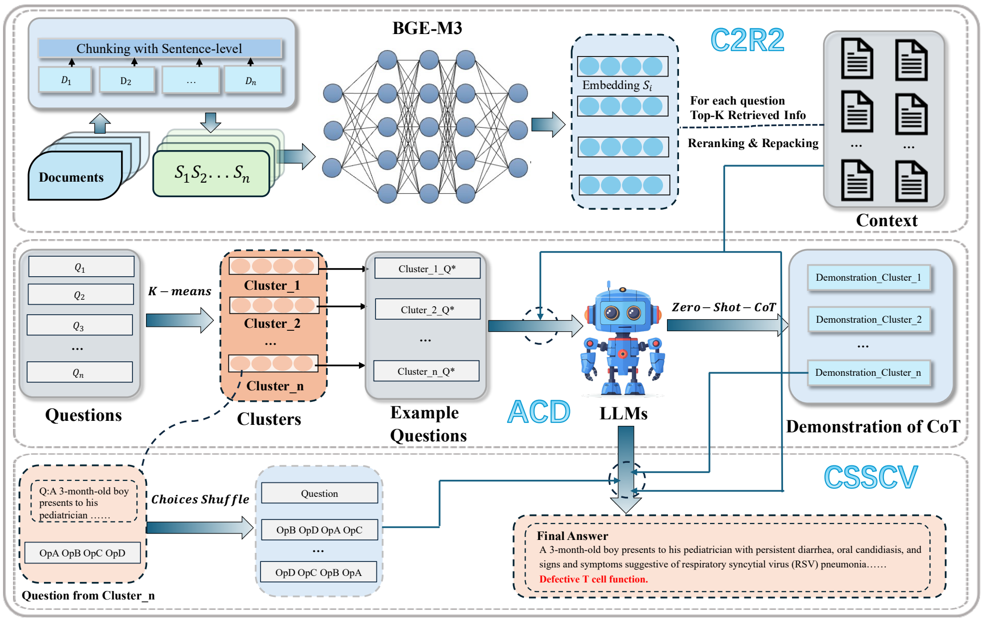

Fig. 1. The RAG-Chain framework, including the comprehensive chunking, retrieval and restricting, the automatic construction of demonstrations and choice shuffling and self-consistency validation.

continuous text into independent sentences. The algorithm uses sentence end-of-sentence markers $T = \{. \ , \ ? \ , \ \ !\}$ and other linguistic features to identify sentence boundaries, ensuring that each sentence remains complete and coherent semantically.

For each sentence $S_i$ in document $D$, the BGE-M3 model generates its embedding representation $\boldsymbol{e}_i$ as shown in (1).

$$\boldsymbol{e}_i = \text{BGE-M3}(S_i), \qquad \boldsymbol{e}_i \in \mathbb{R}^d \tag{1}$$

Where BGE-M3 [8] represents the encoding process of the model, and $\boldsymbol{e}_i \in \mathbb{R}^d$ is the d-dimensional embedding vector representation of sentence $S_i$.

The embedding vectors of all sentences form the embedding matrix $\boldsymbol{E}$ of document $D$, which consists of $n$ d-dimensional vectors as shown in (2).

$$\mathbf{E} = \begin{bmatrix} \mathbf{e}_1 \\ \mathbf{e}_2 \\ \vdots \\ \mathbf{e}_N \end{bmatrix} \in \mathbb{R}^{n\times d} \tag{2}$$

Therefore, the long document is converted into sentence-level embeddings, which are used for the following retrieval and generation.

**Retrieval and Reranking.** For a given question, the question is combined with the options to expand the original query, thereby improving the relevance of the retrieved result to the question. Let $q$ be the original query and $q'$ be the expanded query.

Based on the existing document embeddings, the similarity shown in (3) is calculated between the query $q'$ and each sentence $e_i$ in the document to obtain the Top-K most relevant sentences as the initial retrieval results.

$$sim(q', e_i) = \frac{q' \cdot e_i}{\|q'\| \cdot \|e_i\|} \tag{3}$$

After retrieving the Top-K related sentences, we assign different weights considering the positions of the sentences in the document. The overall score of each sentence is computed by (4) and the sentences are re-ranked. Here α and β represent different weights.

$$Score(e_i) = \alpha \cdot sim(q', e_i) + \beta \cdot Importance(e_i) \tag{4}$$

**Repacking.** In this study, we leverage research findings from the "lost in middle" [9] effect, which demonstrate that models perform better when relevant information is placed at the beginning or the end of the input sequence. During the document repackaging process, the document is recorded according to the importance and relevance of the sentence. The most critical sentence is placed at the beginning of the input sequence to capture the model's attention immediately, while secondary key sentence is placed at the end of the sequence to provide additional contextual support. The repacking approach guides the model to quickly capture the core information of the question and generate accurate answers by providing key information at both the beginning and end of the sequence.

### *B. Automatic Construction of Demonstrations*

Due to the instability of the Zero-Shot-CoT method, inspired by Zhang et al. [10], we propose an automatic construction method for Chain-of-Thoughts (CoT) explanations that include questions and reasoning chains based on demonstrations. Firstly, the BGE-M3 model is used to compute the vector representation $q_i \in \mathbb{R}^d$ for each question, where $i$ represents the question index and $d$ represents the vector dimension. Further, we use the k-means algorithm to cluster the question vectors, which aims to minimize intra-cluster error, specifically minimizing the following objective function as shown in (5).

$$J = \sum_{k=1}^{K} \sum_{\mathbf{q}_i \in C_k} \| \mathbf{q}_i - \mathbf{c}_k \|_2^2 \tag{5}$$

Here, $K$ is the number of clusters, $C_k$ is the set of vectors in the $k$-th cluster, and $\boldsymbol{c}_k$ is the centroid of the cluster which is calculated by averaging all vectors within the cluster as shown in (6).

$$\mathbf{c}_k = \frac{1}{|C_k|} \sum_{\mathbf{q}_i \in C_k} \mathbf{q}_i \tag{6}$$

Here, $| C_k |$ is the number of vectors in cluster $k$.

Therefore, the vectors are ranked within each cluster based on their distance to the cluster centroid, calculated using the Euclidean distance as shown in (7).

$$dist(\boldsymbol{q}_i, \boldsymbol{c}_k) = \| \mathbf{q}_i - \mathbf{c}_k \|_2 \tag{7}$$

The questions are sorted in ascending order based on the above distance. For each cluster $C_k$, we select the question vector $\boldsymbol{q}_i$ closest to the centroid $c_k$ as the representative question. Using large language models combined with the Zero-Shot-CoT technique, the representative question $Q_i^*$ and the corresponding retrieved content $C_i$ are taken as the input, generating the corresponding reasoning chain $R_i$ and answer $A_i$. The process is shown in (8).

$$(R_i, A_i) = \text{LLMs}(Q_i^*, C_i, \text{"Let's think step by step"}) \tag{8}$$

Specially, the generated reasoning chain $R_i$ and answer $A_i$ are combined with the representative question to form the demonstrations $D_i$ . This process is repeated for all $k$ clusters, generating $k$ demonstrations $D_1, D_2, \ldots, D_k$. These demonstrations are concatenated with the question to construct new prompts, which are taken as the input of the large model to obtain the answers.

### C. Choice Shuffling and Self-Consistency Validation

Large language models operate as black boxes with millions to billions of parameters, featuring complex, nonlinear decision-making processes. Research by Miyoung [11] shows that these models may favor certain options in multiple-choice questions. To mitigate this bias, we propose a choice shuffling self-consistency validation strategy.

First, before the model $M$ generates each reasoning path based on the chain-of-thought samples, we randomly shuffle the given options to reduce the model $M'$s potential bias toward the option order. This step can be represented as permuting the option set $O = \{o_A, o_B, o_C, o_D\}$ randomly to obtain a new option sequence $O'$.

Further, self-consistency validation of the answers is performed for different permutations of the multiple-choice options, aiming to identify the answer that is least sensitive to choice shuffling.

By randomizing the option order, the model's dependence on a specific option sequence is reduced and it encourages the model to focus more on the content of the options themselves. Additionally, the most stable answer is selected as the final result by self-consistency validation strategy, improving the reliability of the model's responses.

## III. Experiment And Analysis

We use Qwen2-7B-Instruct [12] as the base model and apply the proposed RAG-Chain framework for improvement. The experiments are conducted on the MedQA[13] and MedMCQA [14] datasets for biomedical question-answering tasks, and comparative experiments are performed with a series of models based on decoder architectures such as ChatGLM, Llama and Palm. The experimental results show that the proposed method achieves significant performance over the base model and outperforms previous related works.

### A. Experiment Settings

*1) Models:* The models used in this study include ChatGLM2 [15], ChatGLM3 [16], Qwen-chat [17], Qwen2-Instruct [12], Llama2-chat [18] and Llama3-Instruct [19]. These models cover different preferred languages.

*2) Datasets:*

**MedQA.** The MedQA datasets contains 1273 multiple choice questions used to evaluate medical specialist competency in the United States, Mainland China and Taiwan.

**MedMCQA.** The MedMCQA datasets presents exam questions. The "dev" subset of the dataset, upon which we report benchmark results (consistent with prior studies), contains 4183 questions, each with four multiple choice answers.

### B. Experiment Results

*1)* ***Performance Comparison Experiment:*** As shown in Table I, using Qwen2-7B-Instruct as the base model, the proposed RAG-Chain strategy performs best compared with the mainstream large language models subjected to few-shot learning, such as Llama2, Llama3-Instruct and Palm [20], which have the same scale of parameters. Additionally, it surpasses pre-trained models specifically designed for the biomedical field, such as PMC-Llama [21] and MEDITRON [22], and even performs better than the pre-trained models with 6.2 billion parameters or more that have undergone few-shot learning. Therefore, the proposed RAG-Chain strategy improve the accuracy of question-answering tasks effectively in the biomedical field without the need for domain-specific fine-tuning of the model.

TABLE I. Performance Comparison Results (Acc.)

| Model | MedQA | MedMCQA | Parameters |
|---|---|---|---|
| **MPT**(3-shot)[22] | 27.6 | 32.1 | 7B |
| **Falcon**(3-shot) [22] | 19.6 | 27.3 | 7B |
| **Llama-2**(3-shot) [22] | 35.4 | 37.9 | 7B |
| **PMC-Llama**(3-shot) [22] | 27.8 | 27.4 | 7B |
| **MEDITRON**(3-shot) [22] | 37.4 | 36.3 | 7B |
| **Llama3-Instruct** | 40.8 | 42.2 | 8B |
| **Flan-PaLM**(5-shot)[23] | 35.4 | 34.5 | 8B |
| **PaLM**(5-shot) [23] | 25.7 | 26.7 | 8B |
| **PaLM**(5-shot) [23] | 40.9 | 43.4 | 62B |
| **Flan-PaLM**(5-shot) [23] | 46.1 | 58.9 | 62B |
| **MEDITRON**(5-shot) [22] | 59.8 | 53.3 | 70B |
| **Llama-2**(5-shot) [22] | 58.4 | 52.4 | 70B |
| **Clinical-Camel** [22] | 56.8 | 46.7 | 70B |
| **PaLM**(5-shot) [23] | 58.9 | 54.5 | 540B |
| **Flan-PaLM**(5-shot) [23] | 60.3 | 56.5 | 540B |
| **Qwen2-Instruct+RAG-Chain (Ours)** | **60.6** | **63.7** | 7B |

*2)* ***Ablation Experiment:*** The RAG-Chain framework based on the Qwen2-7B-Instruct model, outperforming multiple pre-trained models in the biomedical field and general large language models. To have a comprehensive understanding of the contributions of each module in the framework, a series of ablation experiments are conducted on models with different language preferences, as shown in Table II. Here, C2R2 represents Comprehensive Chunking, Retrieval and Restructuring; ACD represents Automatic Construction of Demonstrations; CSSCV represents Choice Shuffling and Self-Consistency Validation.

TABLE II. ABLATION EXPERIMENT RESULTS

| Base Model | Methods | MedQA | | MedMCQA | |
|---|---|---|---|---|---|
| | | Acc. | ΔAcc. | Acc. | ΔAcc. |
| Qwen2-7B-Instruct | **Original** | 51.1 | - | 53.3 | - |
| | **+C2R2** | 56.0 | +4.9 | 59.1 | +5.8 |
| | **+C2R2 (CoT)** | 56.4 | +5.3 | 59.8 | +6.5 |
| | **+C2R2+CSSCV (CoT)** | 57.9 | +6.8 | 60.7 | +7.4 |
| | **+C2R2+ACD** | 59.3 | +8.2 | 61.9 | +8.6 |
| | **+C2R2+ACD+CSSCV** | **60.6** | **+9.5** | **63.7** | **+10.4** |
| Chatglm 2-6B | **Original** | 28.4 | - | 28.2 | - |
| | **+C2R2** | 31.8 | +3.4 | 33.3 | +5.1 |
| | **+C2R2 (CoT)** | 32.5 | +4.1 | 33.9 | +5.7 |
| | **+C2R2+CSSCV (CoT)** | 32.9 | +4.5 | 33.8 | +5.6 |
| | **+C2R2+ACD** | 33.3 | +4.9 | 35.1 | +6.9 |
| | **+C2R2+ACD+CSSCV** | **34.9** | **+6.5** | **35.6** | **+7.4** |
| Llama3-8B-Instruct | **Original** | 40.8 | - | 42.2 | - |
| | **+C2R2** | 44.9 | +4.1 | 46.7 | +4.5 |
| | **+C2R2 (CoT)** | 45.2 | +4.4 | 47.6 | +5.4 |
| | **+C2R2+CSSCV (CoT)** | 45.4 | +4.6 | 48.1 | +5.9 |
| | **+C2R2+ACD** | 46.1 | +5.3 | 49.6 | +7.4 |
| | **+C2R2+ACD+CSSCV** | **47.2** | **+6.4** | **50.1** | **+7.9** |

## IV. CONCLUSION

We propose the RAG-Chain framework, which demonstrates significant performance improvements in biomedical question-answering tasks. The framework improves the utilization of external knowledge and has the strong adaptability and effectiveness in different large language models. Specially, two kinds of strategies adopted in the framework, including the automatic construction of chain-of-thought templates and the choice shuffling self-consistency validation process, improve the performance of general large language models in biomedical reasoning and question-answering tasks effectively. The experimental results on biomedical datasets show that RAG-Chain improves the accuracy of the model in question-answering tasks significantly without the need for domain-specific pretraining or fine-tuning. It brings positive gains across different models and demonstrates its versatility. Furthermore, ablation experiments further verify the contributions of each component in RAG-Chain, verifying the framework's adaptability and effectiveness in multi-language scenarios. In the future, the RAG-Chain framework will be optimized to enhance the efficiency of external knowledge retrieval and explore its application potential in other types of tasks.

## ACKNOWLEDGMENT

This work is supported by the National Key R&D Program of China under grant NO.2023YFB3308004 and National Natural Science Foundation under Grant No.92267107.

## REFERENCES

[1] W. X. Zhao et al., “A Survey of Large Language Models,” Nov. 24, 2023, arXiv: arXiv:2303.18223. doi: 10.48550/arXiv.2303.18223.

[2] Y. Chang et al., “A Survey on Evaluation of Large Language Models,” ACM Trans. Intell. Syst. Technol., vol. 15, no. 3, pp. 1–45, Jun. 2024, doi: 10.1145/3641289.

[3] J. Lee et al., “BioBERT: a pre-trained biomedical language representation model for biomedical text mining,” Bioinformatics, vol. 36, no. 4, pp. 1234–1240, Feb. 2020, doi: 10.1093/bioinformatics/btz682.

[4] M. Yasunaga, J. Leskovec, and P. Liang, “LinkBERT: Pretraining Language Models with Document Links,” in Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), S. Muresan, P. Nakov, and A. Villavicencio, Eds., Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 8003–8016. doi: 10.18653/v1/2022.acl-long.551.

[5] K. Singhal et al., “Towards Expert-Level Medical Question Answering with Large Language Models,” May 16, 2023, arXiv: arXiv:2305.09617. doi: 10.48550/arXiv.2305.09617.

[6] E. Lehman, E. Hernandez, D. Mahajan, et al., “Do we still need clinical language models?,” in Conference on Health, Inference, and Learning, PMLR, 2023, pp. 578-597.

[7] P. Lewis, E. Perez, A. Piktus, et al., “Retrieval-augmented generation for knowledge-intensive NLP tasks,” Advances in Neural Information Processing Systems, vol. 33, pp. 9459-9474, 2020.

[8] Anonymous, “M3-Embedding: Multi-Linguality, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation,” Feb. 2024. [Online]. Available: https://openreview.net/forum?id=QvB5DoTxacN

[9] N. F. Liu et al., “Lost in the Middle: How Language Models Use Long Contexts,” Transactions of the Association for Computational Linguistics, vol. 12, pp. 157–173, Feb. 2024, doi: 10.1162/tacl_a_00638.

[10] Z. Zhang, A. Zhang, M. Li, and A. Smola, “Automatic Chain of Thought Prompting in Large Language Models,” Oct. 07, 2022, arXiv: arXiv:2210.03493. doi: 10.48550/arXiv.2210.03493.

[11] M. Ko, J. Lee, H. Kim, et al., “Look at the first sentence: Position bias in question answering,” in Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Association for Computational Linguistics (ACL), 2020, pp. 1109-1121.

[12] A. Yang et al., “Qwen2 Technical Report,” Jul. 17, 2024, arXiv: arXiv:2407.10671. doi: 10.48550/arXiv.2407.10671.

[13] X. Zhang, J. Wu, Z. He, X. Liu, and Y. Su, “Medical Exam Question Answering with Large-scale Reading Comprehension,” Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32, no. 1, Art. no. 1, Apr. 2018, doi: 10.1609/aaai.v32i1.11970.

[14] A. Pal, L. K. Umapathi, and M. Sankarasubbu, "Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering," in Conference on Health, Inference, and Learning, PMLR, 2022, pp. 248-260.

[15] A. Zeng et al., “GLM-130B: An Open Bilingual Pre-trained Model,” Oct. 25, 2023, arXiv: arXiv:2210.02414. doi: 10.48550/arXiv.2210.02414.

[16] T. GLM et al., “ChatGLM: A Family of Large Language Models from GLM-130B to GLM-4 All Tools,” Jul. 29, 2024, arXiv: arXiv:2406.12793. doi: 10.48550/arXiv.2406.12793.

[17] J. Bai et al., “Qwen-VL: A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond,” Oct. 12, 2023, arXiv: arXiv:2308.12966. doi: 10.48550/arXiv.2308.12966.

[18] H. Touvron et al., “Llama 2: Open Foundation and Fine-Tuned Chat Models,” Jul. 19, 2023, arXiv: arXiv:2307.09288. doi: 10.48550/arXiv.2307.09288.

[19] A. Dubey et al., “The Llama 3 Herd of Models,” Jul. 31, 2024, arXiv: arXiv:2407.21783. doi: 10.48550/arXiv.2407.21783.

[20] A. Chowdhery, S. Narang, J. Devlin, et al., “Palm: Scaling language modeling with pathways,” Journal of Machine Learning Research, vol. 24, no. 240, pp. 1-113, 2023.

[21] C. Wu, W. Lin, X. Zhang, Y. Zhang, W. Xie, and Y. Wang, “PMC-LLaMA: toward building open-source language models for medicine,” Journal of the American Medical Informatics Association, p. ocae045, Apr. 2024, doi: 10.1093/jamia/ocae045.

[22] Z. Chen et al., “MEDITRON-70B: Scaling Medical Pretraining for Large Language Models,” Nov. 27, 2023, arXiv: arXiv:2311.16079. doi: 10.48550/arXiv.2311.16079.

[23] K. Singhal, S. Azizi, T. Tu, et al., “Large language models encode clinical knowledge,” Nature, vol. 620, no. 7972, pp. 172-180, 2023.