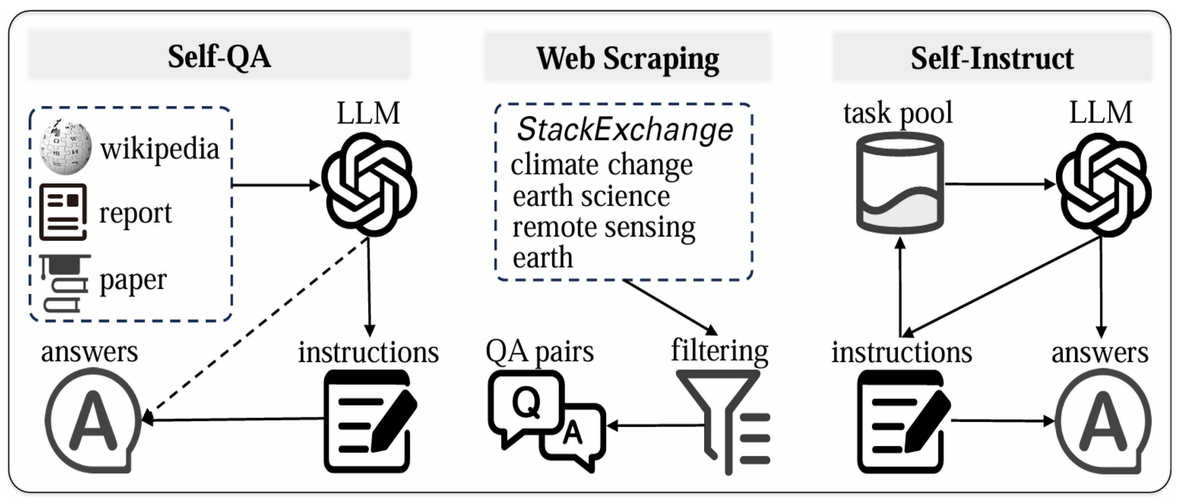
ClimateChat: Designing Data and Methods for Instruction Tuning LLMs to Answer Climate Change Queries

Zhou Chen, Xiao Wang, Yuanhong Liao, Ming Lin, Yuqi Bai
Tsinghua University, China

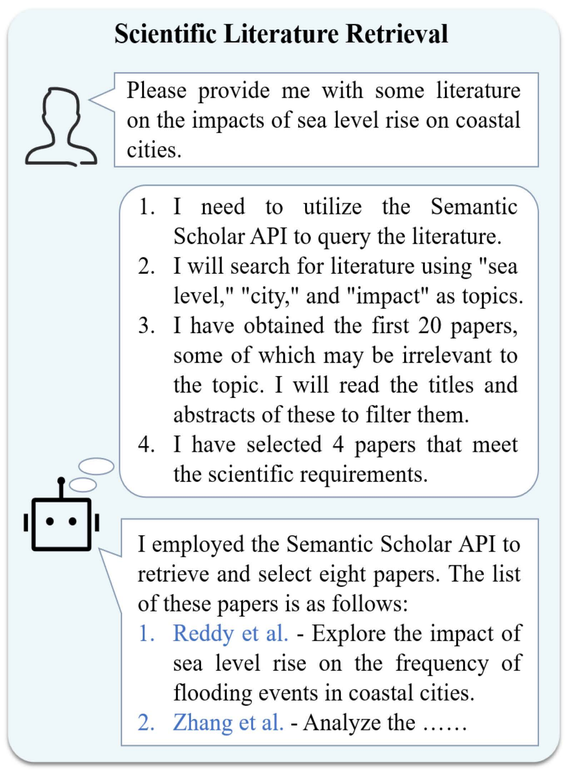
❖ Introduction

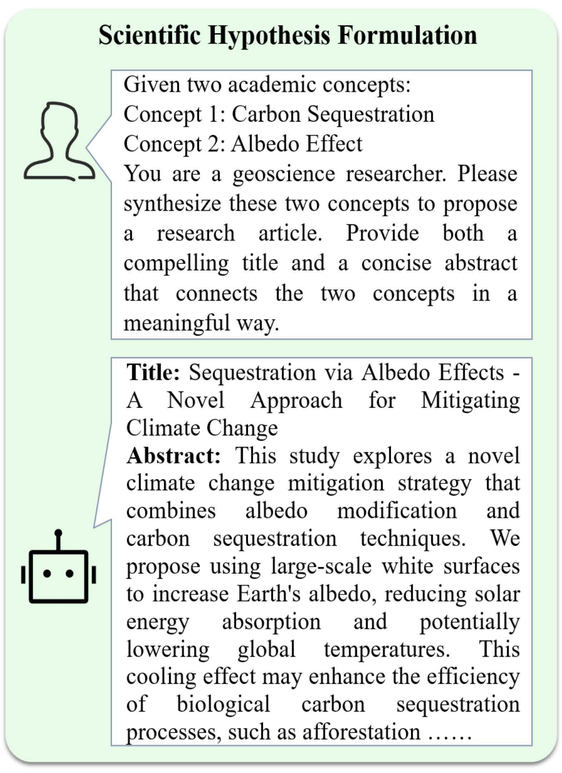
- As climate change intensifies, demand for climate science research grows. Large Language Models (LLMs) offer potential but lack high-quality, scalable instruction data for climate-specific tasks.
- We propose an automated method to generate diverse instruction data (ClimateChat-Corpus) using document-derived facts, web scraping, and seed instructions. This dataset fine-tunes open-source LLMs, yielding ClimateChat.
- ClimateChat outperforms baselines on climate Q&A tasks. Evaluations highlight the impact of base models and instruction data, underscoring the need for optimal model selection.
- This work advances climate-specific LLMs, providing a scalable data pipeline and empirical insights for future research.

❖ Method



❖ Application





❖ Experiments

Scores of baselines and ClimateChat on objective tasks

Model	Climate	Ecology	Environment	Health	Geography	Average
ClimateMistral	76.9±3.8	77.3±0.0	77.8±5.6	70.0±0.0	75.3±3.2	75.5±2.4
JiuZhouChat	81.9±2.3	89.5±2.7	90.6±6.7	81.5±6.0	87.9±3.2	86.0±2.0
ClimateChat	93.5±2.3	96.8±2.7	90.6±3.3	90.0±8.5	93.2±3.2	93.0±2.9

Scores of baselines and ClimateChat on subjective tasks

Model	Helpfulness	Depth	Accuracy	Relevance	Creativity	Detail	Average
ClimateMistral	124.5	111.0	134.0	134.0	76.0	119.0	116.4
JiuZhouChat	123.0	102.0	130.0	133.0	75.0	117.0	113.5
ClimateChat	131.0	129.0	134.0	134.0	97.0	132.0	126.2

Human evaluation results for subjective tasks of baselines and ClimateChat

Model	Significant Inaccuracy	Moderate Inaccuracy	Slight Inaccuracy	Accuracy
ClimateMistral	11	18	10	6
JiuZhouChat	3	3	10	29
ClimateChat	1	4	6	34

Performance comparison of different models with and without RAG

Model	w/o RAG			w/ RAG		
	Obj. Task	Subj. Task	Average	Obj. Task	Subj. Task	Average
JiuZhouChat	86.0±2.0	113.5	99.8±1.0	93.2±1.9	120.3	106.8±1.0
ClimateChat	93.0±2.9	126.2	109.6±1.4	93.8±1.7	129.6	111.7±0.9

❖ Experiments

- The LLM instruction-tuned with the ClimateChat-Corpus demonstrated significant improve ments in performance on climate change question-answering tasks
- Utilizing JiuZhou as the base model significantly boosted performance on climate change question-answering tasks
- While climate change instruction data can enhance LLM performance on related tasks, using an LLM without sufficient knowledge may lead to increased occurrences of hallucinations.
- Although the RAG method enhances LLM performance on objective tasks, it remains less effective than instruction tuning on subjective tasks.

Paper | Code (Train model) | SFT model | Base model

Related work:
JiuZhou: Open Foundation Language Models and Effective Pre-training Framework for Geoscience
PreparedLLM: Effective Pre-pretraining Framework for Domain-specific Large Language Models