ClimaQA: An Automated Evaluation Framework for Climate Question Answering Models

Veeramakali Vignesh Manivannan[1], Yasaman Jafari[1], Srikar Eranky[1], Spencer Ho[1], Taylor Berg-Kirkpatrick[1], Duncan Watson-Parris[1], Yian Ma[1], Leon Bergen[1], Rose Yu[1]

[1] University of California San Diego, San Diego, CA

vmanivannan@ucsd.edu,roseyu@ucsd.edu

UC San Diego

BERG LAB

ICLR 2025

Problem

How to evaluate foundation models effectively on domain specific scientific question-answers, as well as develop robust benchmark datasets for evaluation?

Our Methodology

*We propose **ClimaQA**: an adaptive, domain specific evaluation framework for climate science questions that utilizes a novel method to generate question-answer pairs as well as novel metrics to evaluate foundation models on these questions. To achieve this, we:*

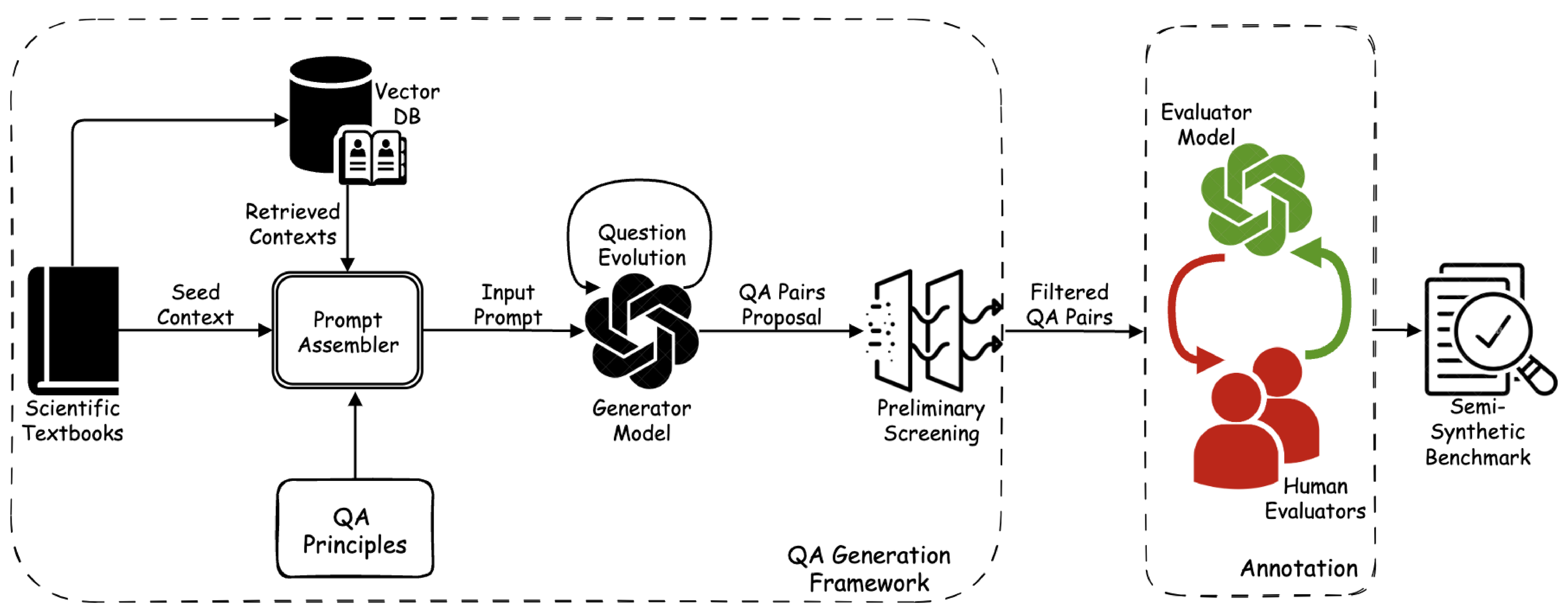
- Train a generator LLM to create base level scientific QA pairs using textbooks as context;
- Increase question complexity through prompt engineering and in-context learning;
- Allow domain experts to validate the question answer pairs that were generated by the LLM;

Comparison of Scientific Benchmark Datasets

Dataset	Domain	Source	Size	Automated	Validated	Multi-Task	Multi-Level
ScienceQA	Science	Hi-Scl Text	21000	✗	✓	✗	✗
Pira2	Ocean	Research	2250	✗	✓	✓	✗
SciQA	Comp Sci	ORKG	2500	✓	✓	✗	✗
Climate Crisis	Climate	None	20000	✓	✗	✗	✗
SciQAG-24D	Science	Research	8531	✓	✗	✗	✗
ClimaQA-Gold	Climate	Grad Text	566	✓	✓	✓	✓
ClimaQA-Silver	Climate	Grad Text	3000	✓	✗	✓	✓

Comparison of scientific benchmark datasets. Our **ClimaQA-Gold** dataset, about 566 pairs, are multi-task, multi-level, and validated by domain experts. Existing benchamrks either rely on manual expert annotation or fully on synthetic generation, which is inaccurate.

ClimaQA: Automated Question Generation Framework



- The generator LLM creates base level questions from the textbook contexts based on QA generation principles.
- The base questions are evolved by adding complexities
- These questions are then validated by domain experts
- The evaluator model adaptively learns to automatically validate the generated questions from the expert-labeled examples during the annotation phase

ClimaQA Benchmark Dataset

Dataset	Task	Base	Reasoning	Hypothetical	Total
ClimaQA-Gold	MCQ	126	72	47	245
	Freeform	54	52	55	161
	Cloze	-	-	-	160
ClimaQA-Silver	MCQ	501	264	235	1000
	Freeform	507	241	252	1000
	Cloze	-	-	-	1000

Contents of the ClimaQA dataset. Both ClimaQA-Gold and ClimaQA-Silver include 3 task-forms with varying levels of complexity for MCQ and Freeform.

```
Base

Question - What is a crucial factor to ensure when collecting data for calibration purposes?
Options -
   a) Using different solution sources for each data set.
   b) Consistency in equipment setup and data collection procedures.
   c) Changing the calibration locations frequently to avoid bias.
   d) Varying the nebuliser type for each calibration date.
Answer - b

------------------------------------------------------------------------------------------
Reasoning

Question - Why is consistency in equipment setup and data collection procedures considered a
crucial factor for collecting data for calibration purposes?
Options -
   a) It ensures that the calibration process is completed faster.
   b) It helps in minimizing errors and maintaining reliable and accurate measurements.
   c) It helps in identifying outliers in the data sets more effectively.
   d) It allows for easy integration of new equipment without affecting the calibration
      results.
Answer - b

------------------------------------------------------------------------------------------
Hypothetical Scenario

Question - How might the calibration accuracy be affected if the driers and DMA were changed
between different calibration sessions?
Options -
   a) The calibration accuracy would deteriorate due to inconsistent conditions.
   b) The calibration accuracy would fluctuate depending on the type of nebuliser used.
   c) The calibration accuracy would remain unaffected by the change in equipment.
   d) The calibration accuracy would improve due to the variability introduced.
Answer - a
```

Figure: Includes examples of different complexity levels for MCQ questions generated by our framework

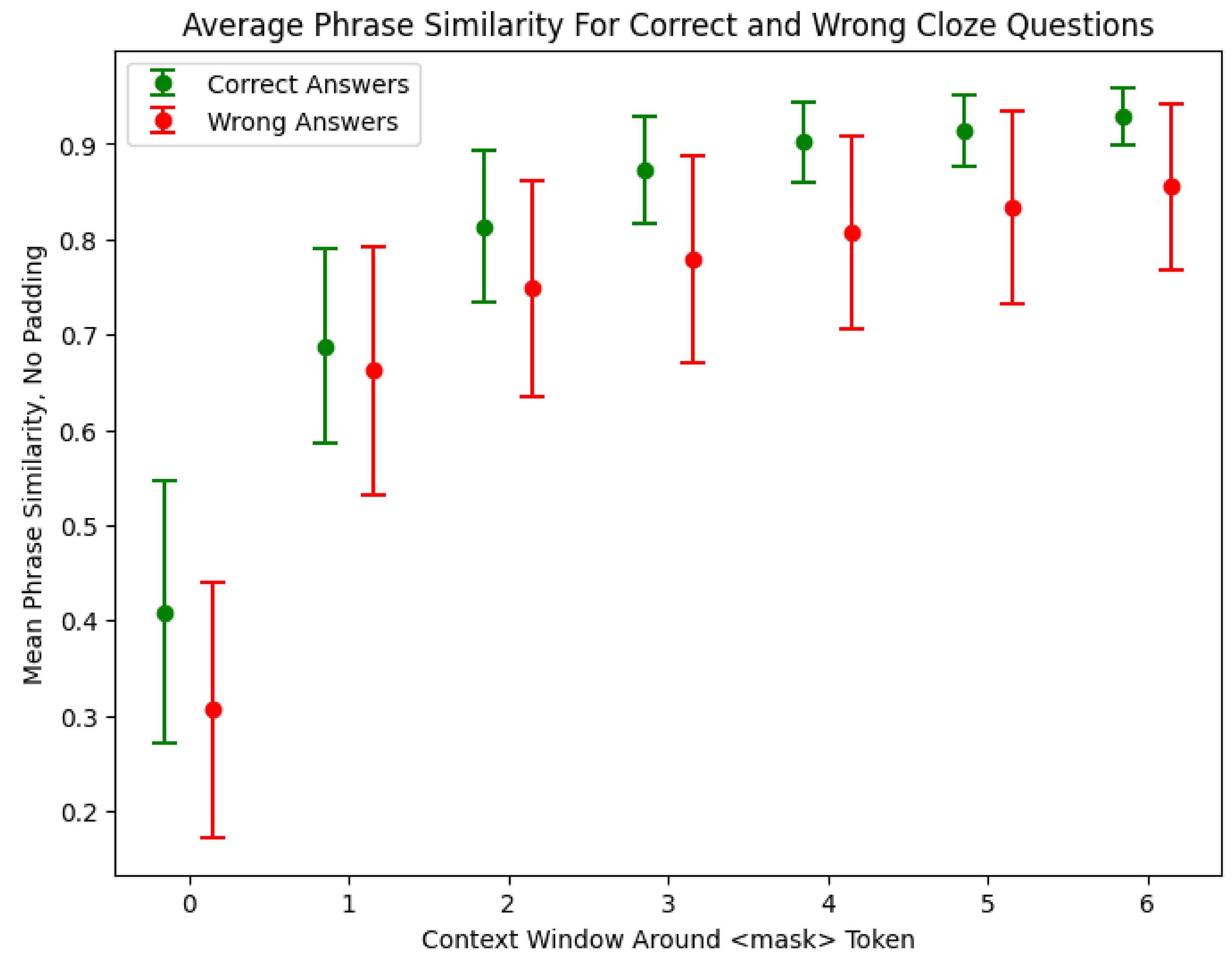
Link to our Arxiv paper

Evaluation Metrics

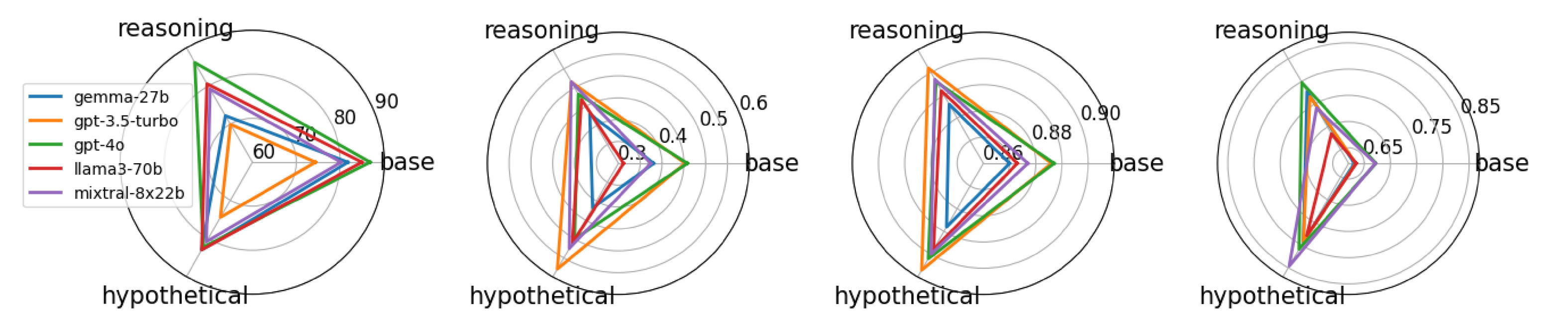
Models were evaluated on 3 types of questions - MCQ, Freeform, and Cloze.

- **MCQ** - Direct accuracy metric
- **Freeform** - BERTScore, BLEUScore, **Factual Accuracy** - used LLM as classifier to evaluate whether ground truth statement was SUPPORTED or REFUTED by the LLM output, and then used model logit scores as numerical metric.
- **Cloze** - Exact match, **Phrase Similarity** - Select a context window around the blank and report metric as the cosine similarity between the reference-filled and answer-filled phrases

You are a climate expert who annotates whether a given claim either SUPPORTS or REFUTES the presented evidence. You will be provided with the following input:
Evidence: ⟨*evidence*⟩
Claim: ⟨*claim*⟩
Respond with only one word: SUPPORTS if the claim supports the evidence and REFUTES otherwise.



Our phrase similarity metric is shown to be robust - on average most correct answers have a higher phrase similarity, whereas wrong answers have lower phrase similarity. A context window of 4 proves to be the most different.



(a) Accuracy (b) BLEU (c) BERTScore (d) Factual Accuracy

- Models struggle with reasoning in MCQ but not so in Freeform
- RAG outperforms all knowledge enhancement methods
- BLEU and BERTScore favour the generator model while Factual Accuracy does not
- Overall, GPT-4o generalizes well and dominates in all the tasks