## I. Overview of Changes

The paper has been updated with a revised evaluation method, which now uses the **Claude model as a meta-reviewer** to combine the scores from other LLM judges. In response to feedback, we have also added significant detail throughout the paper to clarify our methodology, justify our design choices, and improve reproducibility. All experimental results were re-run using this new method, which means that **all figures and tables have been replaced with new ones**. The text has been edited to reflect these new results and to better explain the study's context and contributions.

## II. Section-by-Section Changes

### Abstract & Introduction

- The abstract was updated to include the new finding that the smaller Phi-2 model performed well, sometimes better than larger models.
- The introduction was revised to clarify the paper's contributions, including the creation of the CDQG dataset, the new evaluation process, and its validation.

### 2. Related Works

- This section was updated with more recent citations and now includes new subsections on "LLMs for Reasoning" and "LLMs for Evaluation".
- A discussion has been added to compare our framework with existing methods, such as "Ask to Learn," clarifying the distinction between our non-conversational, knowledge-driven context and prior conversational approaches.

### 3. CDQG Framework

- **Evaluation Process:** This section details a key change in the methodology.
    - Three LLMs (GPT-3.5 Turbo, Mistral 8x7b, and Gemini) now provide initial scores.
    - The **Claude model has been introduced as a "metareviewer"** to analyze these scores and produce a final score.
    - A justification for the evaluation metrics has been added, explaining how **relevance, coherence, and diversity** were chosen to collectively capture the essential dimensions of curiosity-driven inquiry (i.e., pertinence, logical depth, and breadth of exploration).
- **Dataset:**
    - The dataset size has been increased to 1,988 statements. The counts for each subject and difficulty level have been updated in Table 1.
    - Further details on the dataset construction process have been added, clarifying how statements were generated with GPT-4 **under the oversight of domain experts (PhD students)** who categorized and validated the difficulty

levels.
    - The process for validating the difficulty levels was updated, now reporting an average Cohen's Kappa of 0.639 based on feedback from three annotators.

4. Models & 5. Results

- The list of evaluated models was updated.
- **All figures and tables in the results section are new**, reflecting the updated evaluation method.
- The analysis of model performance was rewritten. Key new findings include:
    - The **Phi-2** model performs very well, often comparable to larger models.
    - **Gemini's** performance is now described as mixed.
- A new subsection on **Metric Correlations** was added, with a corresponding figure.

6. Ensuring the Validity of CDQG

- This validation section was updated.
- **Noise-Addition Study:** The results from this study were updated to show how adding noise to the questions affects the scores.
- **Human Evaluation:** The process was significantly updated to enhance its reliability.
    - It now involves **four graduate student annotators** (previously one).
    - The evaluation was conducted on a larger sample of about **1,000 data points**.
    - Agreement between human scores and the Claude model's scores was measured using Cohen's Kappa, with the new agreement values reported for each metric.

7. Discussion & 8. Conclusion

- The discussion section was expanded with new subsections on potential applications for the research.
- The conclusion was revised to summarize the paper's updated findings.

III. Figure and Table Updates

- **Figure 1 (Framework Diagram):** Updated to show Claude as the meta-reviewer.
- **Table 1 (Dataset Splits):** Updated with the new dataset counts.
- **All other figures and tables (2, 3, 4, etc.) are new** and present the results from the latest experiments. They provide new data on model performance, metric correlations, and the impact of the noise-addition study.

IV. Enhancements for Reproducibility

- To improve the reproducibility of our work, a new **Appendix E (Model Configuration Details)** has been added. This appendix provides specific details on the settings and configurations used for each of the evaluated LLMs.