Chain of Questions: Guiding Multimodal Curiosity in Language Models

Nima Iji[1*], Peter Chapman[1] and Kia Dashtipour[1]

[1*]School of Computing, Edinburgh Napier University, Edinburgh, UK.

*Corresponding author(s). E-mail(s): 40656795@napier.ac.uk;
Contributing authors: p.chapman@napier.ac.uk;
k.dashtipour@napier.ac.uk;

Abstract

Reasoning capabilities in large language models (LLMs) have substantially advanced through methods such as chain-of-thought prompting and explicit step-by-step explanations. However, these improvements have not yet fully transitioned to multimodal contexts, where models must proactively decide which sensory modalities such as vision, audio, or spatial perception to engage when interacting with complex real-world environments. In this paper, we introduce the Chain of Questions (CoQ) framework, a curiosity-driven prompting approach that encourages multimodal language models to dynamically generate targeted questions regarding their surroundings. These generated questions guide the model to selectively activate relevant modalities, thereby gathering critical information necessary for accurate reasoning and response generation. We evaluate our framework on a novel multimodal benchmark dataset, assembled by integrating WebGPT, ScienceQA, AVSD, and ScanQA datasets. Experimental results demonstrate that our CoQ method significantly enhances a foundation model's ability to effectively identify and integrate pertinent sensory information. This supports modality alignment and provides a transparent reasoning path across diverse multimodal tasks, while establishing a foundation to evaluate accuracy and interpretability in future work.

Keywords: multimodal reasoning, large language models, chain of thought, curiosity-driven learning, question generation, sensor fusion

1 Introduction

Recent advancements in large language models (LLMs) have significantly enhanced their reasoning capabilities, primarily through techniques such as Chain-of-Thought (CoT)[1], which encourage models to explicitly generate intermediate reasoning steps before providing an answer. These methods have markedly improved the interpretability and accuracy of model outputs, particularly for textual reasoning tasks [1]. However, despite these advances, current models predominantly remain limited to unimodal, text-based interactions and often neglect the rich multimodal contexts present in real-world environments.

Human reasoning inherently integrates multiple sensory modalities—visual, auditory, spatial, and textual—to construct coherent interpretations of complex scenarios. For example, when navigating a bustling street, humans simultaneously interpret visual cues from traffic signs, auditory information from vehicle noises, spatial awareness from surrounding structures, and textual instructions from navigation apps. Such comprehensive multimodal reasoning allows humans not only to respond accurately but also to proactively seek missing information by directing attention to relevant sensory channels.

In contrast, existing multimodal language models (MLLMs) [2–7] typically treat modalities other than text as supplementary inputs, passively incorporating them into their reasoning processes. This passive modality integration constrains the models' ability to dynamically determine what additional sensory information is necessary for comprehending and addressing context-dependent tasks. Consequently, their applicability and effectiveness are significantly diminished in practical, dynamic, real-world scenarios requiring active sensory exploration.

To overcome these limitations, this paper proposes a novel approach designed explicitly to guide multimodal language models in proactively generating curiosity-driven questions that dynamically identify and engage relevant sensory modalities. This active questioning mechanism enables models to autonomously determine which modalities (vision, audio, spatial perception, etc.) should be activated to gather necessary information from their environment. The CoQ framework thereby represents a substantial advancement beyond passive multimodal integration approaches by promoting active, targeted sensory exploration, aligning model reasoning processes more closely with natural human cognition. Our approach introduces a new paradigm of "multimodal curiosity", enabling language models to systematically and selectively query their surroundings, enhancing both the interpretability and accuracy of their multimodal reasoning capabilities.

1.1 From Prompt to Sensors

To create more robust AI systems that better mirror human cognitive processes, it is essential to extend reasoning capabilities to actively include multimodal information. The CoQ designed explicitly to enhance multimodal reasoning in language models by guiding them to selectively query their environment through curiosity-driven questions. By dynamically generating these questions, the model identifies which sensory modalities are necessary to gather relevant information for solving a given task.

This process is implemented within the framework through four distinct conceptual stages:

$$\text{Prompt} \rightarrow \text{Question} \rightarrow \text{Task} \rightarrow \text{Sensor}$$

- **Prompt**: The initial textual input provided by the user.
- **Question**: Curiosity-driven inquiries that the model formulates to gather relevant multimodal data. A comprehensive list of possible questions is presented in Table 1.
- **Task**: Specific operations triggered by these questions, such as face recognition, speech-to-text (STT), or object detection.
- **Sensor**: Hardware or software-based modalities activated by tasks, including cameras, microphones, LiDAR sensors, etc.

Question	Task
What do I see?	Object Detection
Who am I looking at?	Captioning
What are they saying?	STT
What am I hearing?	Sound event detection
What is the sentiment?	Sentiment Analysis
What is the spatial location?	Spatial Detection
What is the pose?	Pose Estimation
What are they doing?	Action Recognition
Who is talking?	Speaker ID
What language?	Language ID

Table 1 Questions and corresponding tasks in the CoQ framework.

1.2 Example Workflow

An illustrative example demonstrating the CoQ framework is shown in Figure 1. Given a user prompt, the model formulates targeted questions about the environment, invokes corresponding tasks, and activates appropriate sensors. The sensor-derived observations are aggregated into a coherent multimodal context, enhancing the model's ability to accurately respond to complex, context-dependent prompts. This step-wise questioning and sensing procedure enables precise and contextually informed reasoning.

2 Background

Large Language Models (LLMs) have revolutionized natural language processing (NLP) by enabling systems to understand, generate, and reason with human language at unprecedented scales. Historically, translating complex human problems into formal programming languages was both challenging and resource-intensive [8, 9].

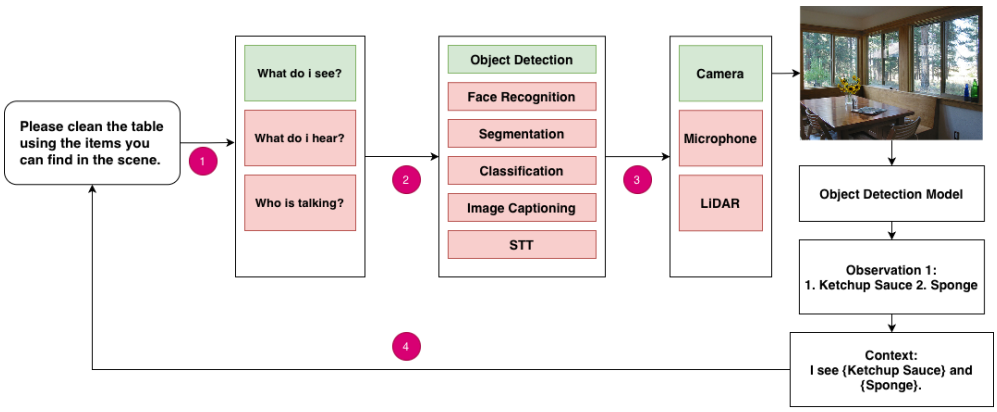

Fig. 1 Illustration of the CoQ framework for multimodal reasoning. Given a natural language prompt, the model generates a set of curiosity-driven questions, each mapped to specific perceptual tasks (e.g., object detection, speech-to-text). These tasks activate the corresponding sensors (e.g., camera, microphone) to gather environment-specific data. The collected observations are then aggregated into a coherent context, enabling the model to form a structured and grounded response. This process mirrors human-like inquiry and perception, enhancing reasoning through selective multimodal exploration.

NLP emerged as a response, developing algorithms for essential tasks such as text classification, summarization, sentiment analysis, and information extraction [10, 11].

Early NLP models, such as Recurrent Neural Networks (RNNs) [12] and Long Short-Term Memory networks (LSTMs) [13], were pivotal in sequential language processing but faced scalability and parallelization constraints. The introduction of Transformer architectures with self-attention mechanisms [14] represented a significant advancement, enabling efficient parallel training and the development of large-scale models like GPT [5], Gopher [15], and Chinchilla [16], which demonstrate remarkable linguistic fluency and reasoning capabilities.

Human communication inherently integrates multiple modalities such as text, visuals, audio, and spatial information. Computational multimodality involves effectively representing and processing these diverse data types to enrich understanding and generation tasks. Multimodal Language Models (MLLMs) extend traditional NLP models by integrating these multiple modalities into unified or coordinated representations, significantly improving contextual understanding [17, 18].

Prominent MLLMs include CLIP, aligning image and text embeddings for cross-modal retrieval and generation [5], ViLBERT, jointly processing visual and textual inputs for tasks such as visual question answering [4], and Flamingo, designed for few-shot multimodal interactions [6]. Fusion methods integrate modalities into shared latent representations [2, 3], while coordination and fission approaches manage separate yet interrelated modality-specific representations. Additionally, Socratic models leverage predictions from unimodal experts to guide textual reasoning in multimodal contexts without extensive retraining [19].

As language models increase in scale, their emergent reasoning abilities, including few-shot learning and generalization to novel problems, have substantially

improved [1]. However, larger model sizes alone do not guarantee consistent performance on complex reasoning tasks [15]. Consequently, researchers have introduced various prompting methods designed to enhance model reasoning abilities. Chain-of-Thought prompting explicitly guides models through intermediate reasoning steps, closely mimicking human cognitive processes [1]. Explanation-based prompting further strengthens model performance by requiring explicit explanations for generated responses, enhancing transparency and clarity [20]. Iterative prompting systematically decomposes complex problems into simpler sub-questions, facilitating structured problem-solving [21]. Moreover, self-consistency prompting improves reliability by generating multiple candidate solutions and selecting the most internally consistent response, thereby increasing confidence in the final output [22].

Building upon the challenges identified in multimodal reasoning, the CoQ framework aims to systematically incorporate diverse modalities into the reasoning process of language models. The central idea behind CoQ is to enable the language model to proactively generate targeted, curiosity-driven questions that explicitly map the required evidence in the user's prompt to corresponding modality-specific tasks. These tasks then activate relevant sensors either hardware-based (e.g., camera, microphone, LiDAR) or software-based (e.g., image recognition, speech-to-text) to collect the necessary observations.

3 Chain of Questions

In light of our discussion in the previous section, we talked about the challenges to bring multimodal inference to language models. In our proposed framework we are trying to add the information from other modalities in the reasoning process of the models. The main idea is to motivate the model to ask related questions that map the required evidences in the prompt to specific tasks. Then tasks would use different sensors (Hardwares or Software execution) to collect information about the outer environment of the model.

3.1 Framework Overview

CoQ framework operates through a structured pipeline comprising several interconnected stages. When the model receives a textual prompt (P), it first generates a series of modality-specific questions $Q = \{q_1, q_2, \ldots, q_k\}$ that clarify what additional multimodal information is required. Each question generated by the model corresponds to a specific task through a task selection function ($\mathcal{T}$), resulting in $T_i = \mathcal{T}(q_i)$. Subsequently, each identified task activates appropriate sensors via a sensor assignment function ($\mathcal{S}$), yielding $S_i = \mathcal{S}(T_i)$. The execution of each task T_i using sensor S_i results in an observation o_i.

Once all observations $O = \{o_1, o_2, \ldots, o_k\}$ are collected, they are aggregated to form a comprehensive multimodal context (C). The final answer (A) is inferred by integrating this context with the initial user prompt through a reasoning function F_a:

$$A = F_a\left(P,\ \text{Aggregate}\left(\{\text{Execute}(\mathcal{T}(q_i), \mathcal{S}(\mathcal{T}(q_i)))\}_{i=1}^{k}\right)\right)$$

3.2 Implementation Methodologies

CoQ framework can be implemented through two primary methodologies: few-shot learning and fine-tuning.

The few-shot learning approach requires minimal resources as it does not necessitate model retraining. Instead, the model is prompted with carefully designed examples that encourage curiosity-driven questioning to gather relevant multimodal information. This approach is beneficial in resource-constrained environments, allowing for quick and efficient deployment.

Alternatively, the fine-tuning approach involves training the foundation model explicitly to generate modality-specific questions during the reasoning process. While this method may incur higher computational costs, it potentially offers improved accuracy and consistency in question generation and multimodal reasoning outcomes. In our experimental evaluation, detailed in subsequent sections, we primarily adopted the few-shot learning method to demonstrate the framework's effectiveness.

3.3 System Configuration and Task Mapping

The responsibility for triggering these questions lies within the reasoning capabilities of the LLM itself. When utilizing the few-shot learning approach, we condition the LLM to recognize the set of valid questions by embedding them within the system prompt. This process simulates a form of "curiosity" regarding the external environment.

To implement the few-shot approach, we utilized the prompt structure detailed in Figure 2. This instructs the model to act as an intermediary agent that queries the environment before attempting to answer the user.

```
You are an AI assistant that generates clarifying questions about what
additional information is needed to answer an input.

TASK:
Determine whether the input requires information from another modality
(vision, audio, spatial location). If yes, output the corresponding
question(s). If no additional information is needed, output "None".

ALLOWED OUTPUTS:
- what do I see?
- what do I hear?
- what is the spatial location of the objects?
- None

IMPORTANT: You may combine outputs using " or ".
example: what do I see? or what do I hear?

EXAMPLES:
Input: What color is the car?
Output: what do I see?

Input: Who directed the movie Jaws?
Output: None

NOW ANSWER:
Input: {question}
Output:
```

Fig. 2 The few-shot prompt template used to elicit curiosity-driven questions from the LLM.

3.4 Context Integration

Upon processing this prompt, the LLM generates the relevant questions, thereby triggering the associated sensors and tasks. The results from these tasks are converted into a textual modality. For example, an object detection task might return a list of entities such as "flower, table, window."

To close the reasoning loop, these observations are injected as context into the prompt for the next iteration. This represents the most streamlined method for grounding the model's final response in the collected environmental data. Consequently, the final answer is generated by synthesizing the user's original inquiry with the newly acquired context, as shown below:

4 Dataset

Due to the novelty of multimodal curiosity-driven reasoning, existing datasets were insufficient for effectively evaluating the CoQ framework. Current datasets predominantly focus on enhancing language model capabilities within single modalities,

```
Contexts:
Context1 <- flower, table, window

Prompt:
I see {context1}

User is asking:
{question}
```

Fig. 3 Example of context integration into the final prompt.

primarily text-based reasoning. To properly evaluate multimodal curiosity and reasoning, we designed and constructed a comprehensive benchmark dataset by carefully integrating multiple existing datasets representing various modalities.

We combined several specialized datasets, including WebGPT [23], ScienceQA [24], AVSD [25], and ScanQA [26], each providing distinct modality contexts. This integration aimed to allow language models to dynamically determine whether additional multimodal information, such as visual, auditory, or spatial data, is required to answer a given prompt accurately.

The WebGPT [23] dataset primarily consists of textual modality, containing 19,578 human-generated prompts initially used to train GPT models. As the dataset exclusively comprises text-based prompts, we marked these instances as not requiring additional multimodal information.

In contrast, the ScienceQA [24] dataset features prompts with and without supplementary visual evidence. We explicitly divided ScienceQA prompts into two distinct categories: those accompanied by visual evidence (images) and those strictly textual. This classification allowed the language models to recognize prompts explicitly requiring visual input.

Additionally, we employed the AVSD [25] dataset to evaluate scenarios involving dialogues within video sequences. In such cases, models must integrate both auditory and visual modalities to comprehend the dialogue context adequately, necessitating the generation of appropriate curiosity-driven questions related to visual observation and auditory understanding.

Finally, the ScanQA [26] dataset, which includes 41,363 human-curated question-answer pairs based on 800 ScanNet 3D indoor scans, was incorporated to assess spatial modality comprehension. Prompts from ScanQA explicitly require the model to query spatial information about objects within a given environment.

After meticulous integration and categorization, our final multimodal benchmark dataset consists of 180,629 carefully labeled instances. This comprehensive dataset structure incorporates purely textual prompts, visually supported prompts, audiovisual dialogue prompts, and spatially oriented prompts. This categorization enables rigorous evaluation of a model's capability to effectively identify when additional multimodal information is necessary, thus thoroughly assessing multimodal reasoning performance within the proposed CoQ framework.

5 Environment Questioning Experiments

This section presents a detailed description of the experimental setup and results, thoroughly evaluating the efficacy of the CoQ framework to show curiosity in Language models. The primary goal of these experiments is to assess whether the CoQ method effectively prompts language models to generate appropriate curiosity-driven questions and thereby select suitable multimodal information. Additionally, we explore how this framework performs across models of varying sizes, highlighting its adaptability and robustness.

5.1 Implementation Details and Configuration

Our experiments involved multiple language models to comprehensively evaluate the effectiveness of the CoQ framework across different architectures and sizes. We selected four prominent models: FLAN T5 base (250 million parameters), FLAN T5 large (780 million parameters), FLAN T5 XL (3 billion parameters), and Llama 2 (7 billion parameters). These models represent varying levels of emergent reasoning capabilities, allowing us to examine how the CoQ method scales with model complexity.

Each model was configured with three different decoding strategies: (1) greedy decoding, selecting the token with the highest probability; (2) sampling, promoting diversity in token selection; and (3) beam search, which considers multiple potential outputs simultaneously, although with increased computational overhead. Sampling and beam search configurations particularly depend on the flexibility and robustness of the function mapping generated questions to corresponding multimodal tasks.

Model	**Parameters**	**Type**
FLAN T5 base	250 million	encoder/decoder
FLAN T5 large	780 million	encoder/decoder
FLAN T5 XL	3B	encoder/decoder
Llama 2	7B	decoder only

Table 2 Models used in our experiments with their parameter counts and architectural types.

Inference was performed using specialized GPU hardware P100 GPUs for FLAN T5 models and A100 GPUs for Llama 2 to ensure efficient computational performance. Given the dataset's substantial size, batching methods were employed during data loading and inference processes to optimize efficiency and resource usage.

To maintain computational efficiency and clarity in evaluating the CoQ framework, we primarily utilized a few-shot learning approach. Specifically, the models were prompted with explicit instructions and illustrative examples designed to elicit curiosity-driven multimodal questions relevant to each prompt.

The primary focus of these experiments was not on final answer generation, but rather on evaluating the accuracy and relevance of the questions generated by the models.

5.2 Experimental Results and Analysis

Our experimental results indicate substantial differences in model performance in generating relevant curiosity-driven multimodal questions, strongly correlated with model size and architectural design. The key results of our experiments are summarized in Tables 3 and 4.

Model	Match	Mismatch	Match %
FLAN T5 XL	137,701	42,928	76.2%
FLAN T5 Large	47,511	133,118	26.3%
FLAN T5 Base	31,861	148,768	17.6%
LLaMA 7B	79,355	101,274	43.9%

Table 3 Comparison of matched versus mismatched outputs across different model variants.

Table 3 illustrates how accurately each model, when using the CoQ framework, generated curiosity-driven questions that matched the expected modality label in the dataset. A match is defined as a generated question that correctly corresponds to the type of sensory evidence (for example, visual, auditory, or textual) required to answer the original question. For example, in the ScienceQA dataset, if a question requires visual information (such as an image) to be answered, the model is expected to generate a sub-question like "What do I see?", which forces the system to use a visual sensor (such as a camera) to acquire the relevant information. In this context, the model's ability to generate an appropriate, modality-specific sub-question demonstrates a deeper understanding of task requirements and sensor alignment.

In contrast, a mismatch occurs when the model either requests information from the wrong modality or fails to ask for any modality at all in cases where sensory data is necessary. For instance, if a visual input is required but the model generates a generic or text-only question like "What is happening?", without referencing the need to see or observe, it is considered a mismatch. Such cases indicate gaps in modality awareness or insufficient reasoning over the input structure.

The results show that FLAN T5 XL produced the highest number of correctly aligned questions (76.2%), while smaller variants like FLAN T5 Base (17.6%) and FLAN T5 Large (26.3%) struggled to match modality requirements. Despite its larger parameter count, LLaMA 7B underperformed (43.9%), likely due to its decoder-only architecture, which may prioritize generative diversity over precise alignment with task-specific input modalities.

Table 4 captures the curiosity rate of each model, measured by observing how frequently they generated modality-related questions in contexts where such questions were appropriate. Specifically, As shown in Table 1, we defined a fixed mapping between curiosity-driven questions and sensor tasks (e.g., "What do I see?" for object detection, "What am I hearing?" for sound event detection, "What is the sentiment?" for sentiment analysis). A response was counted as "asked" if the model generated one

Model	Asked	Did Not Ask	Asked (%)
FLAN T5 XL	144,547	36,082	80.0%
FLAN T5 Large	130,941	49,688	72.5%
FLAN T5 Base	60,773	119,856	33.6%
LLaMA 7B	74,036	106,593	41.0%

Table 4 Rate of modality-related questions generated by each model as a proxy for curiosity.

of these mapped questions in response to a prompt that required additional sensory input. If the model failed to do so in such cases, it was counted as "did not ask."

This measure serves as a proxy for the model's inherent curiosity or its responsiveness to the CoQ prompting strategy. A higher asking rate indicates a greater tendency to explore and engage with the environment. For instance, FLAN T5 XL asked appropriate modality-related questions in 80.0% of applicable cases, whereas LLaMA 7B and FLAN T5 Base did so only 41.0% and 33.6% of the time, respectively. These results highlight meaningful differences in how various models operationalize curiosity through language when prompted within the CoQ framework.

These findings emphasize the significant influence of both model architecture and parameter scale on the successful implementation of the CoQ framework. Additionally, they highlight the importance of refining prompting strategies and improving task-mapping functions to enhance both the relevance and precision of curiosity-driven questioning.

Overall, our experimental analysis confirms the effectiveness of the Chain of Questions framework in guiding multimodal curiosity and reasoning systematically. These results suggest promising avenues for further enhancements through advanced fine-tuning methods, optimized prompts, and deployment of larger, more sophisticated model architectures.

6 Evaluating CoQ against Static Multimodal

In this experiment, we evaluate the efficacy of the CoQ framework operating within an agentic loop compared to a static, single-step multimodal model. The primary objective is to assess whether an iterative, curiosity-driven approach yields superior performance over passive multimodal integration, particularly in tasks varying from simple object existence to complex text-based reasoning.

6.1 Experimental Setup

For the CoQ implementation, we utilized an agentic loop designed to simulate active reasoning. The process begins with a self-reflection step where the agent asks, "Do I have enough evidence to answer the user's prompt?" If the evidence is insufficient, the agent generates specific questions based on the system's capabilities defined in the prompt templates.

The architecture comprises FLAN-T5 as the core language model (processing only textual modality), integrated with expert tools: InstructBLIP[27] for image captioning

(answering "What do I see?") and Qwen2.5-VL-7B-Instruct[28] for Optical Character Recognition (OCR) (answering "What do I read?").

We compared this agentic approach against LLaVA-1.5 (7B) [29], a state-of-the-art multimodal model serving as the baseline. The evaluation employed two distinct datasets to assess performance across varying levels of complexity. First, we utilized the test split of the POPE[30] dataset (9,032 samples), which focuses on object existence verification (Yes/No) and represents lower-complexity perceptual tasks. Additionally, we assessed performance on the train split of the TextVQA[31] dataset (6,272 samples), a benchmark requiring complex OCR capabilities and reasoning that necessitates the agent to iterate and gather specific textual evidence from the scene.

6.2 Results and Analysis

The comparative results are presented in Table 5. Performance was measured using Accuracy (exact/soft match), F1 Score (token overlap), and Reasoning Steps (frequency of agent loops).

Dataset	Model	Accuracy (%)	F1 Score	Freq. Steps
POPE	CoQ Agent	**84.95**	**71.63**	3
(9,032 samples)	LLaVA 7B	84.91	3.86	1
TextVQA	CoQ Agent	**85.28**	**76.33**	3
(6,272 samples)	LLaVA 7B	56.73	16.94	1

Table 5 Performance comparison between the CoQ Agent and LLaVA Baseline on POPE and TextVQA datasets. The 'Freq. Steps' metric indicates the modal number of reasoning loops required.

The results highlight a significant divergence in performance based on task complexity. On the POPE dataset, both models achieved comparable accuracy (∼84.9%), indicating that for simple binary verification tasks, a static multimodal model is as effective as an agentic one in terms of correctness. However, the CoQ Agent demonstrated a vastly superior F1 Score (71.63 vs. 3.86). This discrepancy suggests that while LLaVA often answered correctly, its generative output likely contained extraneous tokens or conversational filler that lowered the overlap score, whereas the CoQ framework, guided by targeted questions, produced precise, ground-truth-aligned responses.

In contrast, the TextVQA dataset revealed the critical advantage of the CoQ framework's iterative approach. The CoQ Agent outperformed the LLaVA baseline by a substantial margin in accuracy (85.28% vs. 56.73%). Complex tasks such as reading text from images often require zooming, cropping, or re-evaluating specific regions—capabilities simulated by the CoQ agent's ability to ask clarifying questions (e.g., "What do I read?").

The analysis of reasoning steps underscores why CoQ provides a decisive advantage in complex multimodal reasoning. While the baseline model is constrained to a single zero-shot inference, the CoQ Agent operates through a dynamic multi-step cycle. This iterative process allows the agent to decompose intricate problems and actively retrieve

specific sensory data that a static single-pass model misses. Consequently, the CoQ framework demonstrates superior robustness where passive models fail, proving that an agentic, curiosity-driven loop is essential for outperforming standard multimodal architectures in tasks requiring high-level reasoning and detailed evidence gathering.

7 Conclusion

We introduced the CoQ framework, which guides a unimodal language model to produce curiosity-driven, modality-specific sub-questions that map to sensor tasks[cite: 8, 35]. The fundamental purpose of this framework is to leverage the reasoning capabilities of the language model, enabling it to autonomously decide what specific evidence is required from the environment to address a given query. We evaluated four models on a composite multimodal benchmark constructed from WebGPT, ScienceQA, AVSD, and ScanQA.

Our internal analysis used two descriptive metrics reported in the Results. First, the *Modality Match Rate* quantifies how often a generated sub-question corresponds to the required input modality for the item. Second, the *Asking Rate* quantifies how often a model produces any mapped modality question when additional sensory input is needed. Under these measures, FLAN T5 XL achieved the highest alignment and curiosity among the evaluated models, with a 76.2% match rate and an 80.0% asking rate. These results indicate that model architecture and scale are associated with differences in modality alignment and the propensity to ask sensor-triggering questions within CoQ.

Furthermore, we extended our evaluation to compare the CoQ agentic loop against a static multimodal baseline (LLaVA-1.5). On simple perceptual tasks (POPE), both models achieved comparable accuracy ($\sim$85%), though CoQ demonstrated superior precision with a significantly higher F1 score. Crucially, in complex reasoning scenarios (TextVQA), CoQ outperformed the baseline by a substantial margin, achieving 85.28% accuracy compared to 56.73%. This performance gap is attributed to the framework's iterative nature; while the baseline relies on a single zero-shot inference, CoQ's multi-step curiosity loop enables it to actively resolve ambiguity through targeted tool use.

Our findings highlight the substantial potential of curiosity-driven prompting strategies in multimodal contexts, proving that an active agentic loop is pivotal for outperforming standard models in tasks requiring high-level reasoning. Future research directions include refining the prompting methodology, enhancing task-mapping precision, and exploring fine-tuning approaches to achieve even more robust and accurate multimodal reasoning capabilities. Ultimately, the CoQ framework represents a significant step towards creating more sophisticated, contextually aware language models capable of effectively operating in real-world environments.

Acknowledgements. The authors would like to thank the School of Computing at Edinburgh Napier University for their support during this research.

Declarations

- **Funding**: Not applicable.

- **Conflict of interest/Competing interests**: The authors declare that they have no competing interests.
- **Ethics approval and consent to participate**: Not applicable.
- **Consent for publication**: Not applicable.
- **Data availability**: The integrated dataset (WebGPT, ScienceQA, AVSD, and ScanQA) is currently private and will be made available upon reasonable request.
- **Materials availability**: Not applicable.
- **Code availability**: The code for the CoQ framework is not publicly available at this time.
- **Author contribution**: Nima Iji: Conceptualization, Methodology, Experiments, Data Curation, Writing. Peter Chapman and Kia Dashtipour: Supervision, Review & Editing.

References

[1] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E.H., Le, Q.V., Zhou, D.: Chain-of-thought prompting elicits reasoning in large language models. In: NeurIPS, pp. 1800–1814 (2022)

[2] Poria, S., Cambria, E., Bajpai, R., Hussain, A.: A review of affective computing: From unimodal analysis to multimodal fusion. Information Fusion **37**, 98–125 (2017)

[3] Zhang, Y., Sidibé, D., Morel, O., Meriaudeau, F.: Deep multimodal fusion for semantic image segmentation: A survey. Image and Vision Computing **105** (2020)

[4] Lu, J., Batra, D., Parikh, D., Lee, S.: ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks (2019)

[5] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning, pp. 8748–8763 (2021)

[6] Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millicah, K., Reynolds, M., Ring, R., Rutherford, E., Cabi, S., Han, T., Gong, Z., Samangooei, S., Monteiro, M., Menick, J., Borgeaud, S., Brock, A., Nematzadeh, A., Sharifzadeh, S., Binkowski, M., Barreira, R., Vinyals, O., Zisserman, A., Simonyan, K.: Flamingo: A visual language model for few-shot learning. In: NeurIPS (2022)

[7] Shridhar, M., Manuelli, L., Fox, D.: Cliport: What and where pathways for robotic manipulation. In: Proceedings of the 5th Conference on Robot Learning (CoRL) (2021)

[8] Sadiku, M., Zhou, Y., Musa, S.: Natural language processing in healthcare. International Journal of Advanced Research in Computer Science and Software

Engineering **8**(5), 39 (2018)

[9] Zheng, H., Xu, K., Zhou, H., Wang, Y., Su, G.: Medication recommendation system based on natural language processing for patient emotion analysis. Academic Journal of Science and Technology **10**(1), 62–68 (2024)

[10] Jin, Z., Mihalcea, R.: Natural language processing for policymaking, pp. 141–162 (2023)

[11] Sgroi, G., Russo, G., Maglia, A., Catanuto, G., Barry, P., Karakatsanis, A., Rocco, N., Pappalardo, F., Group, E.W.: Evaluation of word embedding models to extract and predict surgical data in breast cancer. BMC Bioinformatics **22**(14), 631 (2022)

[12] Schmidt, R.: Recurrent neural networks (RNNs): A gentle introduction and overview (2019)

[13] Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Computation **9**(8), 1735–1780 (1997)

[14] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Polosukhin, I.: Attention is all you need. In: NeurIPS (2017)

[15] Rae, J., Borgeaud, S., Cai, T., Millican, K., Hoffmann, J., Song, F., Aslanides, J., Henderson, S., Ring, R., Young, S., Rutherford, E., Hennigan, T., Menick, J., Cassirer, A., Powell, R., Driessche, G., Hendricks, L., Rauh, M., Huang, P.-S., Irving, G.: Scaling language models: Methods, analysis & insights from training Gopher (2021)

[16] Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., Las Casas, D., Hendricks, L.A., Welbl, J., Clark, A., Hennigan, T., Noland, E., Millican, K., Driessche, G., Damoc, B., Guy, A., Osindero, S., Simonyan, K., Elsen, E., Vinyals, O., Rae, J.W., Sifre, L.: Training compute-optimal large language models. In: NeurIPS (2022)

[17] Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., Ng, A.: Multimodal deep learning. In: Proceedings of the 28th International Conference on Machine Learning, pp. 689–696 (2011)

[18] Arjmand, M., Dousti, M., Moradi, H.: TEASEL: A transformer-based speech-prefixed language model (2021)

[19] Zeng, A., Attarian, M., Ichter, B., Choromanski, K.M., Wong, A., Welker, S., Tombari, F., Purohit, A., Ryoo, M.S., Sindhwani, V., Lee, J., Vanhoucke, V., Florence, P.: Socratic models: Composing zero-shot multimodal reasoning with language. In: The Eleventh International Conference on Learning Representations (2023)

[20] Lampinen, A., Dasgupta, I., Chan, S., Mathewson, K., Tessler, M., Creswell, A., McClelland, J., Wang, J., Hill, F.: Can language models learn from explanations in context? In: Findings of the Association for Computational Linguistics: EMNLP 2022, pp. 537–563 (2022)

[21] Wang, B., Deng, X., Sun, H.: Iteratively prompt pre-trained language models for chain of thought. In: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pp. 2714–2730 (2022)

[22] Wang, X., Wei, J., Schuurmans, D., Le, Q.V., Chi, E.H., Narang, S., Chowdhery, A., Zhou, D.: Self-consistency improves chain of thought reasoning in language models. In: The Eleventh International Conference on Learning Representations (2023)

[23] Nakano, R., Hilton, J., Balaji, S., Wu, J., Long, O., Kim, C., Hesse, C., Jain, S., Kosaraju, V., Saunders, W., Jiang, X., Cobbe, K., Eloundou, T., Krueger, G., Button, K., Knight, M., Chess, B., Schulman, J.: WebGPT: Browser-assisted question-answering with human feedback (2021)

[24] Lu, P., Mishra, S., Xia, T., Qiu, L., Chang, K.-W., Zhu, S.-C., Tafjord, O., Clark, P., Kalyan, A.: Learn to explain: Multimodal reasoning via thought chains for science question answering. In: NeurIPS (2022)

[25] Alamri, H., Cartillier, V., Das, A., Wang, J., Cherian, A., Essa, I., Batra, D., Marks, T.K., Hori, C., Anderson, P., Lee, S., Parikh, D.: Audio-visual scene-aware dialog. In: CVPR (2019)

[26] Azuma, D., Miyanishi, T., Kurita, S., Kawanabe, M.: Scanqa: 3d question answering for spatial scene understanding. In: CVPR, pp. 19107–19117 (2022)

[27] Dai, W., Li, J., Li, D., Tiong, A.M.H., Zhao, J., Wang, W., Li, B., Fung, P., Hoi, S.: InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning (2023). https://arxiv.org/abs/2305.06500

[28] Team, Q.: Qwen2.5-VL (2025). https://qwenlm.github.io/blog/qwen2.5-vl/

[29] Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. In: NeurIPS (2023)

[30] Li, Y., Du, Y., Zhou, K., Wang, J., Zhao, W.X., Wen, J.-R.: Evaluating object hallucination in large vision-language models. arXiv preprint arXiv:2305.10355 (2023)

[31] Singh, A., Natarajan, V., Shah, M., Jiang, Y., Chen, X., Batra, D., Parikh, D., Rohrbach, M.: Towards vqa models that can read. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8317–8326 (2019)