CAUS: A Dataset for Question Generation based on Human Cognition Leveraging Large Language Models

Minjung Shin (mjshin77@snu.ac.kr)
Interdisciplinary Program in Cognitive Science, Seoul National University
Gwanak-ro, Gwanak-gu, Seoul, 08826, Republic of Korea

Donghyun Kim (kimdonghyun0704@gmail.com), Jeh-Kwang Ryu (ryujk@dgu.ac.kr)
Department of Physical Education, Dongguk University
Pildong-ro, Jung-gu, Seoul, 08826, Republic of Korea

Abstract

We introduce the **C**urious **A**bout **U**ncertain **S**cene (**CAUS**) dataset, designed to enable Large Language Models, specifically GPT-4, to emulate human cognitive processes for resolving uncertainties. Leveraging this dataset, we investigate the potential of LLMs to engage in questioning effectively. Our approach involves providing scene descriptions embedded with uncertainties to stimulate the generation of reasoning and queries. The queries are then classified according to multi-dimensional criteria. All procedures are facilitated by a collaborative system involving both LLMs and human researchers. Our results demonstrate that GPT-4 can effectively generate pertinent questions and grasp their nuances, particularly when given appropriate context and instructions. The study suggests that incorporating human-like questioning into AI models improves their ability to manage uncertainties, paving the way for future advancements in Artificial Intelligence (AI).

Keywords: question generation; uncertainty; curiosity; large language model

Introduction

The significance of questioning, rather than just answering, lies in the entity's latent capacity to seek information, regardless of whether the entity is a human or a machine. While questioning as a learning strategy comes naturally to humans, its implementation in machines, i.e., question generation (QG), is relatively recent (Duan, Tang, Chen, & Zhou, 2017; Chen, Yang, Hauff, & Houben, 2018; Zhou, Zhang, & Wu, 2019; Gong, Pan, & Hu, 2022). Although QG models have yielded positive results in active learning (Misra et al., 2018; Krishna, Lee, Fei-Fei, & Bernstein, 2022) and user engagement(Huang, Yeomans, Brooks, Minson, & Gino, 2017), this topic remains on the fringes rather than a mainstream of artificial intelligence (AI).

In recent years, AI models, especially Large Language Models (LLMs) have garnered significant attention for their proficient language generation. These models can even engage in zero-shot learning, generating text for scenarios not covered in their training data (Brown et al., 2020; Wei et al., 2022). However, closer examination reveals significant shortcomings in LLMs, particularly in planning tasks when subjected to systematic evaluations. While competent with surface-level structures, they struggle in deeper planning (Momennejad et al., 2023). Furthermore, while they outperform humans in generating text and images, their understanding performances are not as robust (West et al., 2023). These issues become evident when user instructions are ambiguous, leading to inconsistent or erratic responses, such as hallucinations (Gallegos et al., 2023) or vulnerabilities (Wang, Yue, & Sun, 2023). The main issue with LLMs is that they are often deployed to address inference tasks based on probabilistic contexts without engaging in follow-up questions, even in uncertain situations (Toles, Huang, Yu, & Gravano, 2023).

Then, are LLMs inherently incompetent at asking questions? Can't we improve AI models by implementing human questioning strategies to deal with uncertainty? With these questions in mind, we propose a text dataset named **CAUS** (**C**urious **A**bout **U**ncertain **S**cene) that emulates human cognitive processes for resolving uncertainty, with reasoning and asking questions. The main contributions of our hypothetical approach are as follows:

- Providing a diverse range of questions proper to specific scenarios.

- Classifying questions based on multi-dimensional criteria, considering their attributes, scope, and format.

- Establishing a collaborative system between LLMs and human researchers to enhance efficiency and relevance.

In our study, we focus on *sincere information-seeking questions*, which aim to resolve uncertainty in a given situation (Flammer, 1981; A. C. Graesser & Olde, 2003). By excluding social interaction elements like persuasive or request questions, we maintain a clear information-seeking intent. Our approach assumes a scenario where an agent asks an oracle to obtain specific information, thus resolving uncertainty.

We employ *Scene Description* texts as the starting point for question generations. For each scene, we produce *Reasoning*, which captures how humans clarify uncertain entities, and *Questioning*, which generates relevant questions to resolve uncertainties. Concurrently, we classify the *Type of Generated Question* as a basis for systematic question evaluation. The constructed dataset, along with the code and prompts, are distributed for further use [1].

Related Work

Research on Questioning

Research on asking questions is relatively scarce compared to answering them. This scarcity is due to two key issues: 1) difficulties with defining the scope of various questions

[1] https://github.com/lbaa2022/CAUS_v1

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.*

(e.g., rhetorical questions, questions given in plain text) and 2) various contexts in which they occur (e.g., requests, social coordination, expressing complaints) (A. C. Graesser & Black, 1985). Although research on questioning is limited, there is consensus that the cognitive process underlying questioning is *cognitive disequilibrium* and *the questioner's desire to resolve it* (Flammer, 1981; Otero & Graesser, 2001; A. C. Graesser & Olde, 2003; Loewenstein, 1994). The cognitive disequilibrium stems from knowledge gaps, anomalies, contradictions, discrepancies, unexpected outcomes, or goal-blocking obstacles (A. C. Graesser, Person, & Huber, 2013). The questioner's desire refers to humans' robust motivation for information exploration, i.e., curiosity (Golman & Loewenstein, 2018; Vazard & Audrin, 2022).

Empirical studies, especially in education, underscore the value of effective questioning in enhancing learning, advocating for the promotion of students' questioning skills (A. C. Graesser & McMahen, 1993; Rosenshine, Meister, & Chapman, 1996; Macagno, 2023). Outside of educational contexts, studies are primarily conducted in gaming, while quantifying uncertainty in open domains presents significant challenges. These studies indicate that adept questioners often demonstrate strong strategic thinking, but effective questioning is not an easy feat, even for humans (Rothe, Lake, & Gureckis, 2018). However, even if people usually ask inefficient questions, asking behavior itself plays an essential role in learning (Cervera, Wang, & Hayden, 2020) and desirable development (Chouinard, Harris, & Maratsos, 2007). To sum up, asking questions is crucial but challenging when formulating relevant ones.

Good Questioning

Most discussions on *what makes a good question* center around *learning*, reflecting the inquisitive nature of asking questions. In educational contexts, good questions encourage deep reasoning and active exploration. Inquiries, such as "why," "how," "what if," and "what if not," are valuable because they delve into causal, goal-oriented, and logical reasoning (A. C. Graesser & Olde, 2003; A. Graesser, Ozuru, & Sullins, 2009). Such questions are linked to understanding, problem-solving, reasoning, creativity, and other cognitive processes. They encourage learners to engage more profoundly with the material, enhancing learning and literacy (Macagno, 2023; Otero & Graesser, 2001).

Finding a solid and nominal criterion for good questions is challenging outside of pedagogy. Indeed, in the everyday life of humans, defining the exact and acute question is not always significant. However, defining what constitutes good or bad questions in the context of QG is essential. Thus, individual research efforts often define their own criteria to evaluate questions, reflecting the complex and diverse requirements of questioning tasks in different contexts.

In Battleship, a strategy guessing game, good questions target valuable information about the hidden configuration of the game board. The question should be specific to the context and aimed at resolving uncertainties about ship size and position (Rothe, Lake, & Gureckis, 2017). A proposal of the Questioning Turing Test emphasizes that questioning provides a more nuanced measure of AI than passive responding. The study suggests three evaluation criteria: *human-likeness*, *correctness* (i.e., whether the entity fulfills the inquiry), and *strategicness* (i.e., accessing goal by fewer questions) (Damassino, 2020).

Among various criteria, *question diversity* is underscored as a beneficial index of good questioning in improving QA results and user engagement. Diverse questions can cover broader topics, content, and difficulty levels, which is crucial for comprehensive understanding and assessment in learning contexts (Sultan, Chandel, Fernandez Astudillo, & Castelli, 2020). Studies have shown that when questions have varied types, syntax, and content, they require diverse answers, enhancing learning and evaluation processes (Yoon & Bak, 2023). Another study revealed the number of unique questions and novelty are positively related to the performance of a visual QG system regarding engagement and effectiveness (Jain, Zhang, & Schwing, 2017). However, evaluation criteria for existing research tend to be either excessively mechanical (e.g., automatic scoring based on similarity) or exceedingly subjective (e.g., ranked by human annotators).

Dataset Design

Based on the flow of the uncertainty resolution process in humans, as presented in Fig. 1, we suggest a dataset that aims to emulate epistemic curiosity. Briefly, when we encounter uncertainty, we first identify missing information and use a suitable thinking strategy, including making answerable questions.

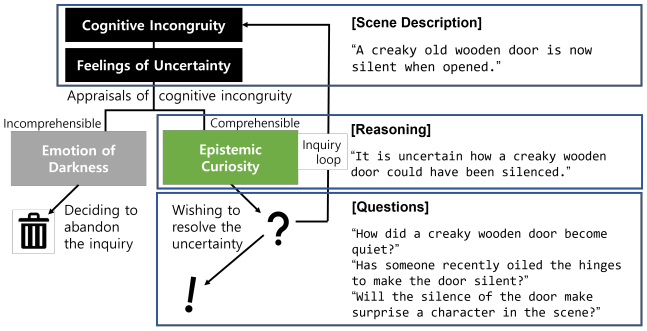

Figure 1: The main concept of the CAUS dataset aligned with human cognition. The uncertain `Scene Description` provides a context that causes "epistemic curiosity." The `Reasoning` sentences point out the uncertain point in a given scene. And the `Questions` represent efforts to resolve the uncertainty.

The CAUS dataset consists of 1K of *scene description* sentences, 1K of *reasoning* sentences, and 5K of *inquisitive question* sentences, which are all written in English. The data configuration is presented in Table 1, and the entire process of building the dataset is illustrated in Fig. 2.

Table 1: Example of configuration for each uncertainty class in the dataset. For detailed explanations of K-type and Q-type, refer to the **Question Classification** section, and (Shin et al., 2023)

Uncertainty Class	Contents		
Object	{**Scene Description**}: A hand mirror is seen on the kitchen counter.		
	{**Reasoning**}: It's unclear why a hand mirror would be found on the kitchen counter, a place typically reserved for cooking utensils and ingredients.		
	{**Questions**}:	[K-type]	[Q-type]
	Why is a hand mirror on the kitchen counter?	Causality	Intention disclosure
	Who left the mirror there?	Identity	Concept completion
	Was it used for a specific purpose in the kitchen?	Intention	Verification
	Is it common for this mirror to be misplaced?	Attributes	Verification
	What is the usual location of this mirror?	Spatial layout	Concept completion
Intention	{**Scene Description**}: A classmate sits alone during lunch breaks and avoids social interactions.		
	{**Reasoning**}: It's unclear why the classmate chooses to sit alone and avoid social interactions during lunch breaks.		
	{**Questions**}:	[K-type]	[Q-type]
	Why does the classmate prefer solitude during lunch breaks?	Intention	Intention disclosure
	What is the classmate's general attitude towards social interactions?	Internal state	Judging
	Are there any observable factors contributing to the classmate's isolation?	Causality	Cause elucidation
	Has the classmate always behaved this way or is it a recent change?	Temporal relation	Case specification
	How do other classmates react to this individual's behavior?	Internal state	Result account
Event	{**Scene Description**}: The bottom-ranked team now holds the championship trophy.		
	{**Reasoning**}: It's unclear how the bottom-ranked team, who were presumably underperforming, managed to secure the championship trophy.		
	{**Questions**}:	[K-type]	[Q-type]
	How did the bottom-ranked team manage to win the championship?	Procedure	Method explication
	What strategies did they employ to overcome their ranking?	Procedure	Method explication
	Who were the key players in their victory?	Identity	Concept completion
	What was the reaction of the top-ranked teams?	Internal state	Result account
	How will this victory impact the team's future performance?	Causality	Expectation

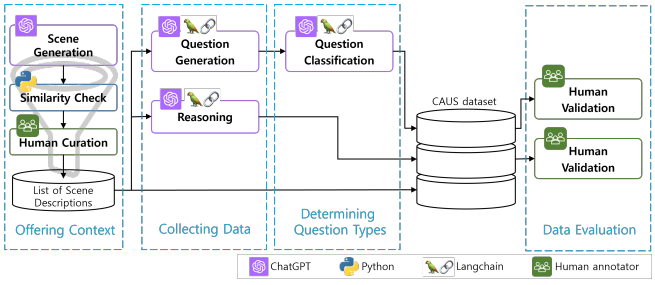

Figure 2: The pipeline for creating a dataset involves collecting, classifying, and evaluating potential questions evoked from scene descriptions. Symbols in the bottom box represent frameworks used in the pipeline.

Scene Description Generation

We crafted scene description texts containing intentional inconsistencies. These inconsistencies were designed to create objective information gaps and cognitive disequilibrium without involving any social dynamics. The texts provide a context that elicits epistemic curiosity and encourages exploration beyond the surface level.

Three uncertainty classes were established within scenes for the following reasons: 1) To avoid any bias towards limited features in the scene generation, and 2) To validate the capability to appropriately interpret and respond to nuances of different types of uncertainty.

We first carefully inspected the Situation Model (Zwaan, 1999) and the Event-Indexing Model (Zwaan, Langston, & Graesser, 1995) in order to refer to the well-formed structural framework for understanding narratives and examine the various elements they suggest (e.g., events, time, location, characters, objects, causality, motives, purposes, and plans). We then decided to focus on three distinct classes based on key elements of uncertainty: Object, Intention, and Event uncertainty. **Object uncertainty** refers to instances where objects in a scene are contextually out of place. **Intention uncertainty** addresses situations with unclear motivations behind a character's actions. **Event uncertainty** involves ambiguity in a specific stage of an event within the scene. This tripartite classification allows for a comprehensive elucidation of

Table 2: K-type category

#	Categories	#	Categories
K1	Identity	K7	Contents
K2	Class	K8	Procedure
K3	Attributes	K9	Causality
K4	Quantities	K10	Intention
K5	Spatial layout	K11	Internal state
K6	Temporal relation		

Table 3: Q-type category

#	Categories	#	Categories
Q1	Verification	Q8	Interpretation
Q2	Case specification	Q9	Cause elucidation
Q3	Concept completion	Q10	Intention disclosure
Q4	Feature specification	Q11	Result account
Q5	Quantification	Q12	Method explication
Q6	Definition	Q13	Expectation
Q7	Comparison	Q14	Judging

various uncertainties in scene interpretation.

We used the GPT-4 API released from OpenAI[2] to generate diverse scene descriptions. We provided the model with zero-shot instructions and set the temperature to 0.7 and 1, which is recommended for diverse outcomes (OpenAI Community, 2023). After gaining excessive sentences, we checked for similar sentences using the cosine similarity algorithm, and two researchers reviewed them to remove duplicates. The deduplicated list still required active human curation. Two researchers deleted or modified 1) inappropriate scenes involving biases on occupation, gender, etc., 2) unrealistic scenes that do not follow the laws of physics, 3) scenes in which uncertainty was diminished by subtext, and 4) scenes that lacked unexpectedness. The filtered list was then finalized by a third researcher to establish a list of scenes that contained uncertainty. After the meticulous inspection, we attained 1,000 scene description sentences, of which 328 were object-related, 364 were intention-related, and 308 were event-related uncertainties.

Reasoning and Query Generation

Reasoning Humans appraise uncertainty at a metacognitive level during the initial phase of resolution process (Fig. 1). We implemented the core function of uncertainty resolution by generating sentences that point out unclear aspects of the given scene through inferring. For the inference process, we adopted the GPT-4-0613 model, which demonstrated the highest test performance, with a temperature setting of 0 to produce deterministic outcomes. The reasoning process is carried out with the main instruction: `"Point out something unclear or uncertain from the scene in one statement using a relation pronoun (who, what, where, when, how, or why)"`.

Query Generation To implement the essential efforts for uncertainty resolution, asking questions, we presented the scene description to the GPT-4-0613 model and instructed it to generate questions addressing these uncertain aspects. As with the reasoning phase, we leveraged a model demonstrating the highest performance in tests. We also set the temperature to 0 to generate deterministic outcomes. To ensure a diverse range of questions, we instructed the model to sequentially create questions, starting from those *spotting the most uncertain feature* to those *exploring the situation*. In addition, no additional constraints were placed to allow us to observe the model's question-generating behavior. This approach aimed to capture a wide spectrum of inquiry types, reflecting the context of the scenes described. The questioning process is carried out with the main instruction: `"Create five different terse questions that can be derived from a scene, from directly targeting the uncertain aspect to gathering additional information from the situation."`

Question Classification

Alongside the query generation, we conducted 2-dimensional classifications of the generated questions referring to our prior work (Shin et al., 2023). The first dimension is knowledge type (K-type), which identifies missing information that can be the source or target of inquiry. Table 2 shows eleven different K-type categories that specify the potential class of missing information in interactive situations. These K-types range from simple and objective to complex and subjective as the number increases.

The second dimension of question classification is question type (Q-type), which represents the issue of how to express the inquiry. Table 3 displays fourteen different Q-type categories that classify inquiry expressions based on pragmatics. As the number grows, the questions become more profound and subjective, like the K-type categories.

In question classification, we utilized the GPT-4-0613 model and the Langchain library[3] by providing detailed instruction prompts for categorizing questions into different types. Although we fixed the model's temperature to 0 to achieve deterministic outcomes, slight variations were observed in the classification results with each iteration due to the inherent nature of LLMs. To mitigate this problem, we repeated the process three times and adopted the model's classification if at least two out of three repetitions agreed. Despite having more than ten category options for both K-type and Q-type, the model showed consistent classification performance. Across all three iterations, a majority of questions

[2]https://openai.com/

[3]An open-source Python package that offers a most standardized interface for various LLM applications compatible with experimenting with different ideas, prompts, and models. https://python.langchain.com/

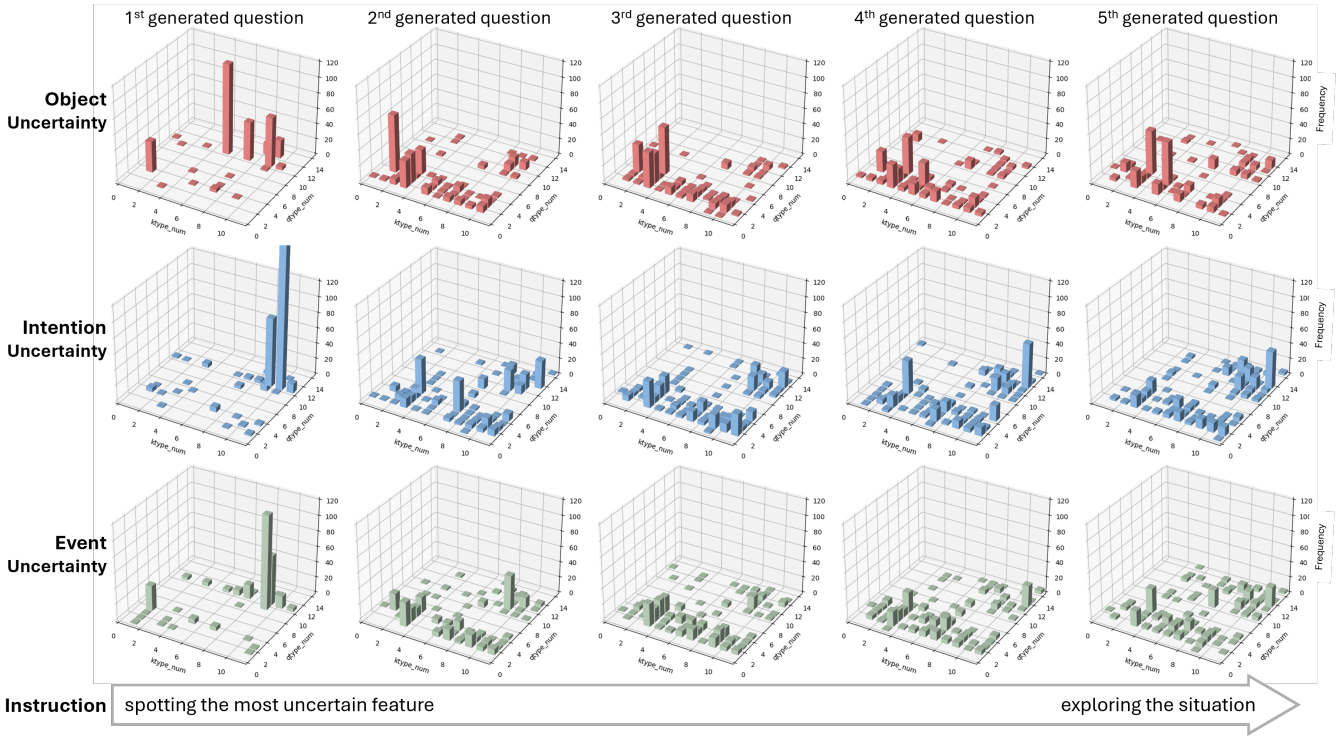

Figure 3: Question classification results. Each row indicates the scene class (i.e., (Top)*Object Uncertainty*. (Middle)*Intention Uncertainty*. (Bottom)Event Uncertainty.). Each column shows the generation order of the question. In each plot, the position of the bars corresponds to the question category, coordinated by k-type and q-type. The height of the bars indicates the frequency of the question for that type pair.

(90.6% for K-type, 91.4% for Q-type) were categorized identically, and most of the remaining questions (9.1% for K-type, 8.3% for Q-type) were made two same results out of three iterations. In very rare cases ($\sim$ 0.3%) where all three outcomes differed, we labeled the outcome '0, Undetermined.' The classification results collected from each condition were displayed in a question space (Fig. 3, further detailed in the **Experimental Results** section). The meaning of each position within the question space is summarized in Fig. 4.

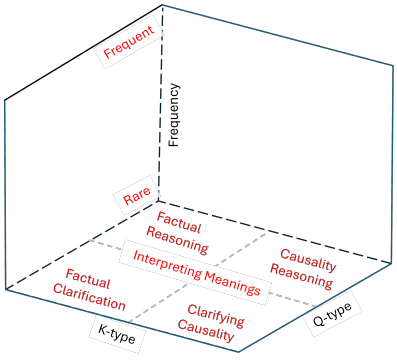

Figure 4: Denoting the question space for displaying question classification

Evaluation

Reasoning A random sample of 100 inferences, representing 10% of the dataset, was taken. Two researchers reviewed the generated inference sentences separately to determine whether they accurately pointed to the uncertainty in the scene. Then, the results of the reviews were consolidated and discussed to assess accuracy.

Question Classification We evaluated the classification results of 500 questions from 100 randomly selected scene descriptions leveraging the GPT model. The 500 questions represent 10% of the dataset. Two researchers created human-generated ground truth according to the definition of K-type and Q-type criteria. In creating the ground truth, researchers achieved an 85% inter-rater agreement, and discrepancies were resolved through discussion.

Experimental Results

Question Classification Results

Fig. 3 shows the distribution of queries, categorizing questions into K-type and Q-type. Five questions were generated for each scene description sentence, labeled with K-type and Q-type criteria, and plotted in the question spaces according to their labels. The plots were organized according to the generation order of each sentence.

The leftmost column shows classification result of the first generated questions, resulting in a highly clustered pattern. This concentration toward specific types reflects the instruction intended to spot the uncertain aspect directly. Most clusters fall under the *Causality Reasoning* area, focusing on inquiring about the antecedents or intentions behind the event. At the same time, a few pertain to *Factual Clarification*, focusing on identifying the subject of the action. However, in intention uncertainty (blue bars in the middle row), it is observed that almost all questions fall within the *Causality Reasoning* region, reflecting contexts where verifying the subject of the action is unnecessary.

On the other hand, those sets toward the right are more diverse and spread evenly throughout the question space, reflecting the instruction to gather additional information from the situation. In the object uncertainty (red bars in the top row), later questions predominantly belong to the *factual clarification* category, focusing on the attributes of objects. Conversely, for the intention uncertainty (blue bars in the middle row), the questions commonly aim to infer or clarify the causes or effects of actions. Regarding event uncertainty (green bars in the bottom row), a blend of object-related and behavior-related uncertainties, there is a tendency for the aforementioned patterns to intermingle, reflecting a combination of both attributes and causality.

Evaluation Results

Evaluation on Reasoning Sentences Upon evaluating a randomly sampled set of 100 inference sentences, it was determined that over 90% of the sentences were deemed appropriate. Sixty-one sentences were inferred within the scope of the words presented in the scene (e.g., "It's unclear why the colleague always keeps their office door closed."), while thirty sentences incorporated contextual cues for inference (e.g., "It's unclear why a renowned painter, *who typically exhibits in galleries or museums*, is selling his pieces on the street."). However, a minority of nine sentences constituted inappropriate inferences. These inappropriate inferences distorted the content by introducing irrelevant clues or departed from the laws of physics (e.g., "It's unclear how a refrigerator magnet is attached to a car door, *which is typically made of materials not receptive to magnets*.").

Evaluation on Question Classification We recorded the number of matches with the human-generated ground truth to evaluate the model-generated question classification. There was an 83.2% match (416 questions) between the model and human ground truth for the K-type questions. Similarly, for Q-type questions, an 83.6% match (418 questions) was observed. We noted that items yielding different outcomes in all three iterations were challenging even for human evaluators to categorize into a single type. Additionally, in cases where two out of three iterations produced the same result, the differing third outcome tended to be semantically adjacent to the other types.

Discussion

In this study, we designed a novel dataset based on human cognition, along with a pipeline generating human-like questions and question classification structures. Our approach leveraged cutting-edge tools such as the GPT-4 model, and we performed in-depth evaluations to assess its performance. We made the controlled uncertainty with scene descriptions that offer a text version of Out Of Distribution (OOD). The uncertainty was fine-tuned to remain predictable, though not as tightly structured as in a game context. The main focus was understanding the LLM's capability to identify and inquire about uncertain elements within a given context. We also explored whether LLM can capture the nuanced content and format of given questions by the question classification.

To wrap up, we revisit the two questions posed at the outset. Firstly, we asked: Are LLMs inherently incompetent at asking questions? Our findings challenge this prevailing idea, demonstrating that LLMs can generate appropriate questions with proper context and instruction. Moreover, LLM's high question classification performance also revealed that LLM is good at predicting the properties of questions, as if the model understands the questioner's purpose and motivation. In conclusion, LLMs seem to have the potential to ask and grasp the nuance.

Secondly, we considered: Can AI models be improved by implementing human questioning strategies when dealing with uncertainty? While this specific question was not directly addressed in our study, the findings offer a promising outlook on the potential applicability of such strategies. As motivated by the Vicuna model, which effectively enhanced the capabilities of a 13 billion parameter small LLM through fine-tuning for multi-round and long conversations (Chiang et al., 2023), it is evident that the strategic application of appropriate datasets can significantly bolster the utility of LLMs for specific purposes.

Acknowledging the limitations of our research, it is important to note that our work is highly sensitive to the specific prompting used. Our procedure, while optimized for the GPT-4 model, may not yield consistent results in the question classification phase with different models. Currently, this exploratory approach, aiming for optimal outcomes, is commonly seen in other studies involving LLMs (Reynolds & McDonell, 2021; Webson & Pavlick, 2022). However, this sensitivity underscores the necessity for developing more robust and model-agnostic prompting techniques. We also acknowledge an aspect in our hypothetical work that remains unaddressed, yet is crucial in human questioning: the role of desires and social interactions. This omission points to the need for further research integrating these human factors into the study of question generation and its application.

In summary, our research suggests that LLMs are not fundamentally limited in their ability to ask questions, and the application of human-like questioning strategies in AI models, particularly in dealing with uncertainties, holds substantial promise for future advancements.

Acknowledgments

We deeply thank the reviewers for providing kind and helpful comments. And this work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No. 20220-00951, Development of Uncertainty-Aware Agents Learning by Asking Questions)

References

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... others (2020). Language models are few-shot learners. *Advances in neural information processing systems*, *33*, 1877–1901.

Cervera, R. L., Wang, M. Z., & Hayden, B. Y. (2020). Systems neuroscience of curiosity. *Current Opinion in Behavioral Sciences*, *35*, 48–55. doi: 10.1016/j.cobeha.2020.06.011

Chen, G., Yang, J., Hauff, C., & Houben, G.-J. (2018). Learningq: a large-scale dataset for educational question generation. In *Proceedings of the international aaai conference on web and social media* (Vol. 12). doi: 10.1609/icwsm.v12i1.14987

Chiang, W.-L., Li, Z., Lin, Z., Sheng, Y., Wu, Z., Zhang, H., ... Xing, E. P. (2023, March). *Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.* Retrieved from https://lmsys.org/blog/2023-03-30-vicuna/ (Last accessed: 2024-05-09)

Chouinard, M. M., Harris, P. L., & Maratsos, M. P. (2007). Children's questions: A mechanism for cognitive development. *Monographs of the society for research in child development*, i–129.

Damassino, N. (2020). The questioning turing test. *Minds and Machines*, *30*(4), 563–587.

Duan, N., Tang, D., Chen, P., & Zhou, M. (2017). Question generation for question answering. In *Proceedings of the 2017 conference on empirical methods in natural language processing* (pp. 866–874). doi: 10.18653/v1/D17-1090

Flammer, A. (1981). Towards a theory of question asking. *Psychological Research*, *43*(4), 407–420.

Gallegos, I. O., Rossi, R. A., Barrow, J., Tanjim, M. M., Kim, S., Dernoncourt, F., ... Ahmed, N. K. (2023). Bias and fairness in large language models: A survey. *arXiv preprint arXiv:2309.00770*.

Golman, R., & Loewenstein, G. (2018). The desire for knowledge and wisdom. In G. Gordon (Ed.), *The new science of curiosity* (pp. 37–42). Hauppauge, NY: Nova Science Publishers, Inc.

Gong, H., Pan, L., & Hu, H. (2022). Khanq: A dataset for generating deep questions in education. In *Proceedings of the 29th international conference on computational linguistics* (pp. 5925–5938).

Graesser, A., Ozuru, Y., & Sullins, J. (2009). What is a good question? In M. G. McKeown & L. Kucan (Eds.), *Bringing reading research to life* (pp. 170–193). New York, NY: Guilford Press.

Graesser, A. C., & Black, J. B. (Eds.). (1985). *The psychology of questions*. Lawrence Erlbaum Associates, Inc.

Graesser, A. C., & McMahen, C. L. (1993). Anomalous information triggers questions when adults solve quantitative problems and comprehend stories. *Journal of Educational Psychology*, *85*(1), 136.

Graesser, A. C., & Olde, B. A. (2003). How does one know whether a person understands a device? the quality of the questions the person asks when the device breaks down. *Journal of Educational Psychology*, *95*(3), 524.

Graesser, A. C., Person, N., & Huber, J. (2013). Mechanisms that generate questions. In *Questions and information systems* (pp. 167–188). Psychology Press.

Huang, K., Yeomans, M., Brooks, A. W., Minson, J., & Gino, F. (2017). It doesn't hurt to ask: Question-asking increases liking. *Journal of personality and social psychology*, *113*(3), 430.

Jain, U., Zhang, Z., & Schwing, A. G. (2017, July). Creativity: Generating diverse questions using variational autoencoders. In *Proceedings of the ieee conference on computer vision and pattern recognition (cvpr)*.

Krishna, R., Lee, D., Fei-Fei, L., & Bernstein, M. S. (2022). Socially situated artificial intelligence enables learning from human interaction. *Proceedings of the National Academy of Sciences*, *119*(39), e2115730119. doi: 10.1073/pnas.2115730119

Loewenstein, G. (1994). The psychology of curiosity: A review and reinterpretation. *Psychological bulletin*, *116*(1), 75. doi: 10.1037/0033-2909.116.1.75

Macagno, F. (2023). Questions as dialogue games. the pragmatic dimensions of "authentic" questions. *Studies in Philosophy and Education*, *42*(5), 519–539.

Misra, I., Girshick, R., Fergus, R., Hebert, M., Gupta, A., & Van Der Maaten, L. (2018). Learning by asking questions. In *2018 ieee/cvf conference on computer vision and pattern recognition* (pp. 11–20). Salt Lake City, UT, USA: IEEE. doi: 10.1109/CVPR.2018.00009

Momennejad, I., Hasanbeig, H., Frujeri, F. V., Sharma, H., Ness, R. O., Jojic, N., ... Larson, J. (2023). Evaluating cognitive maps in large language models with cogeval: No emergent planning. *Advances in neural information processing systems*, *37*.

OpenAI Community. (2023, April). *Cheat sheet: Mastering temperature and top-p in chatgpt api.* Retrieved from https://community.openai.com/t/cheat-sheet-mastering-temperature-and-top-p-in-chatgpt-api/172683 (Last accessed: 2024-05-09)

Otero, J., & Graesser, A. C. (2001). Preg: Elements of a model of question asking. *Cognition and instruction*, *19*(2), 143–175. doi: doi.org/10.1207/S1532690XCI1902_01

Reynolds, L., & McDonell, K. (2021). Prompt programming for large language models: Beyond the few-shot paradigm. In *Extended abstracts of the 2021 chi confer-*

ence on human factors in computing systems. New York, NY, USA: Association for Computing Machinery. doi: 10.1145/3411763.3451760

Rosenshine, B., Meister, C., & Chapman, S. (1996). Teaching students to generate questions: A review of the intervention studies. *Review of educational research*, *66*(2), 181–221.

Rothe, A., Lake, B. M., & Gureckis, T. (2017). Question asking as program generation. In I. Guyon et al. (Eds.), *Advances in neural information processing systems* (Vol. 30). Curran Associates, Inc.

Rothe, A., Lake, B. M., & Gureckis, T. M. (2018). Do people ask good questions? *Computational Brain & Behavior*, *1*(1), 69–89. doi: 10.1007/s42113-018-0005-5

Shin, M., Jang, M., Cho, M., & Ryu, J.-K. (2023). Uncertainty-resolving questions for social robots. In *Companion of the 2023 acm/ieee international conference on human-robot interaction* (p. 226–230). New York, NY, USA: Association for Computing Machinery. doi: 10.1145/3568294.3580077

Sultan, M. A., Chandel, S., Fernandez Astudillo, R., & Castelli, V. (2020, July). On the importance of diversity in question generation for QA. In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 5651–5656). Online: Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.500

Toles, M., Huang, Y., Yu, Z., & Gravano, L. (2023). What is a good question? task-oriented asking with fact-level masking. *arXiv preprint arXiv:2310.11571*.

Vazard, J., & Audrin, C. (2022). The noetic feeling of confusion. *Philosophical Psychology*, *35*(5), 757-770. doi: 10.1080/09515089.2021.2016675

Wang, B., Yue, X., & Sun, H. (2023). Can chatGPT defend its belief in truth? evaluating LLM reasoning via debate. In *The 2023 conference on empirical methods in natural language processing.*

Webson, A., & Pavlick, E. (2022, July). Do prompt-based models really understand the meaning of their prompts? In M. Carpuat, M.-C. de Marneffe, & I. V. Meza Ruiz (Eds.), *Proceedings of the 2022 conference of the north american chapter of the association for computational linguistics: Human language technologies* (pp. 2300–2344). Seattle, United States: Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.167

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., ... Fedus, W. (2022). Emergent abilities of large language models. *Transactions on Machine Learning Research*. (Survey Certification)

West, P., Lu, X., Dziri, N., Brahman, F., Li, L., Hwang, J. D., ... others (2023). The generative ai paradox:"what it can create, it may not understand". In *The twelfth international conference on learning representations.*

Yoon, H., & Bak, J. (2023, December). Diversity enhanced narrative question generation for storybooks. In H. Bouamor, J. Pino, & K. Bali (Eds.), *Proceedings of the 2023 conference on empirical methods in natural language processing* (pp. 465–482). Singapore: Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.31

Zhou, W., Zhang, M., & Wu, Y. (2019). Question-type driven question generation. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (emnlp-ijcnlp)* (pp. 6032–6037). doi: 10.18653/v1/D19-1622

Zwaan, R. A. (1999). Situation models: The mental leap into imagined worlds. *Current directions in psychological science*, *8*(1), 15–18.

Zwaan, R. A., Langston, M. C., & Graesser, A. C. (1995). The construction of situation models in narrative comprehension: An event-indexing model. *Psychological Science*, *6*(5), 292-297. doi: 10.1111/j.1467-9280.1995.tb00513.x