FOLLOWUPQG: Towards Information-Seeking Follow-up Question Generation

Yan Meng*
University of Amsterdam
y.meng@uva.nl

Liangming Pan*
University of California, Santa Barbara
liangmingpan@ucsb.edu

Yixin Cao
Singapore Management University
caoyixin2011@gmail.com

Min-Yen Kan
National University of Singapore
kanmy@comp.nus.edu.sg

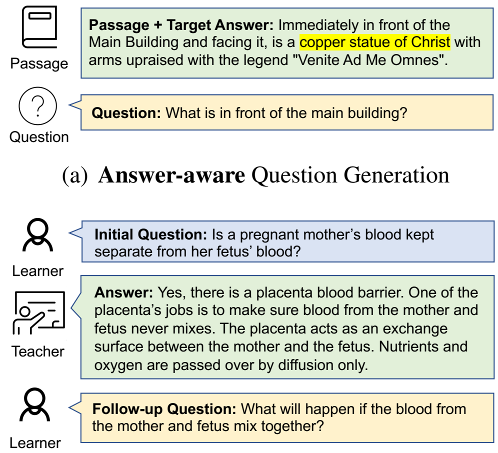
Abstract

Humans ask follow-up questions driven by curiosity, which reflects a creative human cognitive process. We introduce the task of *real-world information-seeking follow-up question generation (FQG)*, which aims to generate follow-up questions seeking a more in-depth understanding of an initial question and answer. We construct FOLLOWUPQG, a dataset[1] of over 3K real-world (initial question, answer, follow-up question) tuples collected from a Reddit forum providing layman-friendly explanations for open-ended questions.

In contrast to existing datasets, questions in FOLLOWUPQG use more diverse pragmatic strategies to seek information, and they also show higher-order cognitive skills (such as *applying* and *relating*). We evaluate current question generation models on their efficacy for generating follow-up questions, exploring how to generate specific types of follow-up questions based on *step-by-step* demonstrations. Our results validate FOLLOWUPQG as a challenging benchmark, as model-generated questions are adequate but far from human-raised questions in terms of informativeness and complexity.

1 Introduction

Question asking is considered a fundamental cognitive process. People typically ask concise and natural questions to seek information (Ram, 1991). *Question Generation* (QG) has recently gained much interest, targeting the study of how intelligent systems can generate relevant questions. This can evaluate the cognitive reasoning ability of models while benefiting many downstream tasks, such as generating assessments for course materials in education (Laban et al., 2022) and enriching training data for question answering (Pan et al., 2021a).



(a) **Answer-aware** Question Generation

(b) **Information-seeking** Question Generation

Figure 1: Examples of (a) *answer-aware QG* and (b) *information-seeking QG*.

Existing works (Duan et al., 2017; Zhao et al., 2018; Pan et al., 2020; Ghanem et al., 2022) focus on generating simple factoid questions, while few works to date target complex practical questions. The task of QG is often framed as generating questions from a source text and a specific target answer from reading comprehension datasets like SQuAD (Rajpurkar et al., 2016), as exemplified by Figure 1(a). Although useful in practical applications, such generated questions are quite different from actual human questions. First, they do not reflect the information-seeking nature of human question-asking, since the model already knows the answer beforehand. Second, they also do not reflect the creative human cognitive process in question-asking such as inferences and synthesis.

To bridge this gap, we propose the task of *real-world information-seeking follow-up question generation (FQG)*, which aims to generate *follow-up questions* that seek new information given the *initial question* and the *human-provided answer*. For example, the follow-up question in Figure 1 extends the provided answer to a reasonable counterfactual situation. Conventional follow-up question

*Authors contributed equally.
[1]Data available at https://github.com/vivian-my/FollowupQG

Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers), pages 252–271
November 1–4, 2023. ©2023 Association for Computational Linguistics

generation works focus on benefiting multi-hop reasoning QA systems (Malon and Bai, 2020) or generating multi-turn conversational questions (Reddy et al., 2019; Richardson et al., 2023). In contrast, our task is more practical and challenging, since it requires a higher level of cognition to know what one does not know (Miyake and Norman, 1979). First, it demands a deep comprehension of the teacher-provided answer, identifying the uncertainty or gaps in knowledge; and second, applying high cognitive skills such as analogy to generate a meaningful follow-up question.

In this paper, we construct a dataset, FOLLOWUPQG, containing 3,790 real-world (initial question, answer, follow-up question) tuples. We collect the data from the Reddit forum *Explain Like I'm Five*[2] which contains real-life questions and self-contained answers. The layperson-friendly nature of this forum makes the question and answer highly comprehensible, serving as a suitable context for follow-up question generation. We further ask crowd-workers to select relevant follow-up questions from replies to the answer as they are real curiosity-driven questions by humans. Our data analysis shows that FOLLOWUPQG captures a variety of high cognitive skills in question-asking, such as relating and causal inference.

We establish benchmarks on this data using GPT-Neo (Black et al., 2021), BART (Lewis et al., 2020) and T5 (Raffel et al., 2020). Automatic and human evaluation reveals that the best model can generate fluent follow-up questions. However, they still fall short of human questions in terms of semantic validity, complexity, and informativeness. Also, we find that ∼30% of the generated questions do not seek new information.

We note that one limitation of fine-tuning pretrained language models on QG task is to control *what to ask* and *how-to ask*. Inspired by recent prompting methods on large language models (Wei et al., 2022; Saha et al., 2023), we investigate on *chain-of-thought* prompt-based learning via GPT family[3], and observe that incorporating an intermediate reasoning chain can better control the models to ask specific types of questions compared to the standard promoting. However, there is still a large improvement in generating specific high-level questions. These observations make FOLLOWUPQG a challenging benchmark for advancing QG.

[2] https://www.reddit.com/r/explainlikeimfive/
[3] ChatGPT, GPT-3.5, GPT-4

2 Related Work

Question Generation (QG) aims to automatically generate questions from textual input. Existing QG studies (Du et al., 2017; Du and Cardie, 2018; Nema et al., 2019; Pan et al., 2020; Murakhovs'ka et al., 2021) are typically trained and evaluated on reading comprehension benchmarks such as SQuAD (Rajpurkar et al., 2016) and HotpotQA (Yang et al., 2018). Questions in those datasets are designed to test machine's reading comprehension ability, which fail to reflect the information-seeking nature of human question-asking. This gap has led to work on "answer-agnostic" question generation (Subramanian et al., 2018; Wang et al., 2019; Pan et al., 2019), in which the target answer is not given to the model as input. However, the sources of data are still reading comprehension datasets and the generated questions are still required to be answerable by the input source, which are quite different with the follow-up questions in our work, which aims to seek for unknown information from known knowledge.

To explore the generation of real human-like information-seeking questions, prior works have investigated generating clarification questions for forum posts in *StackExchange* (Rao and III, 2018; Kumar and Black, 2020), Amazon product reviews (Rao and III, 2019; Majumder et al., 2021), and online courses (Chen et al., 2018). However, clarification is only one of the pragmatic goals in asking follow-up questions. FOLLOWUPQG covers broader types of information-seeking behaviors beyond clarification, such as association, analogy, critical evaluation, and generalization. In addition, instead of focusing on restricted and highly-technical domains like StackExchange and Amazon products, we select *Explain Like I'm Five* as the underlying data source, which covers a boarder range of real-life topics (Fan et al., 2019).

The closest prior work is InquisitiveQG (Ko et al., 2020). They asked crowd-workers to write follow-up questions for news articles and trained models for follow-up question generation. However, our analysis reveals that crowd-sourced questions in InquisitiveQG are typically shallow in reasoning and biased towards monotonous cognitive skills, in contrast with our natural follow-up questions collected from the web. In addition, our work focuses on a scenario different with InquisitiveQG but common in real-life: asking follow-up questions based on the initial question and its answer.

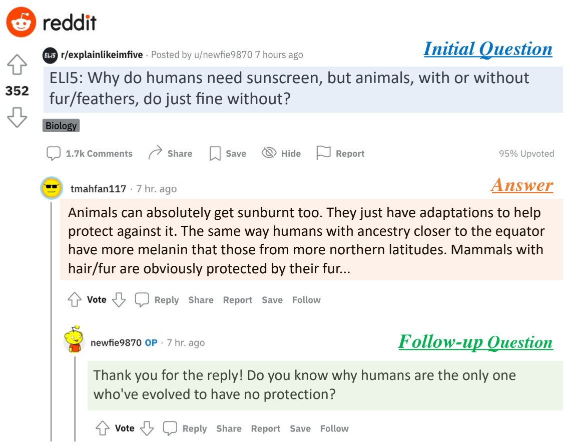


Figure 2: Sample *Explain like I'm five* (ELI5) forum.

3 The FOLLOWUPQG Dataset

We construct the FOLLOWUPQG dataset as follows. The follow-up questions and the source documents are collected from *Reddit* [4](§ 3.1). We first collect around 200,000 posts that contains question, answer, and replies to the answer with a site-specific web crawler. Then, we automatically select data samples that contain follow-up questions in § 3.2. Afterward, the selected 10,890 data samples are further validated by online workers from Amazon Mechanical Turk (AMT) (§ 3.3). The final dataset contains 3,790 high-quality samples.

3.1 Data Sources

To gather real-world information-seeking questions, we initially explored several websites which provide forums to ask open-ended questions, such as Quora, Khan Academy, as well as numerous Reddit forums (subreddits). After a careful comparison, we choose to focus on the subreddit *Explain Like I'm Five* (ELI5), where users are encouraged to provide answers which are comprehensible by a five year-old child. ELI5 is appealing because the questions are close to real-life and the answers are self-contained, and thus rely less on prior specialized knowledge. Their high comprehensibility makes the question and answer suitable to serve as the context of follow-up question generation.

3.2 Data Collection

A thread in the ELI5 forum (Figure 2) usually consists of: (1) a thread title, usually in question format and is considered as the *Initial Question*, (2) a vote that measures the quality of the thread, (3) top-level comments, most of which are detailed answers to the initial question, and (4) replies to the top-level comments, and many of them are asking follow-up questions to the answer.

Fan et al. (2019) have collected a large amount of (question, answer) pairs from ELI5 for question-answering. However, we could not reuse their corpus since they did not collect the follow-up questions. Therefore, we implement a site-specific web crawler to massively crawl data from the ELI5 forum. The crawler is built on *Pushshift API* and *Reddit API*, which give access to the post ID, body, vote, and the comments. We restrict the data collection size to 200,000 and only collect the first three levels of comments.

We then define rules based on regular expressions to automatically filter out the invalid samples in the crawled data. A thread is considered invalid if: 1) its thread title is not a question, 2) the answer is not self-contained (shorter than 30 characters[5]) or receive low votes, or 3) the replies to the answer do not contain any question. After applying this automatic filtration, 10,890 data samples remain.

3.3 Crowd-Sourced Data Validation

We find that the automatic-filtered data samples are still noisy. Especially, some replies in question-format are irrelevant to the initial question and answer or contain toxic or offensive contents. To ensure that our final corpus contain high quality follow-up questions, we design a crowd-sourcing task for data validation. We release 10,000 HITs (Human Intelligence Tasks) on the AMT platform, evenly divided into 10 batches. Each HIT presents the crowd-worker with one data sample of *(initial question, answer, follow-up question)*. To conduct human validation, we ask workers to answer three questions as follows:

- **Q1:** Is the follow-up question a complete question asking for new information?
- **Q2:** Does the data sample contain controversial topics, such as racism, hate speech, sexual topics, or offensive comments?
- **Q3:** What is the relatedness of the follow-up question to the initial question and the answer?, where workers use the 5-point classification set of "strongly related", "related", "slightly related", or "not related".

[4]License of usage: https://www.redditinc.com/policies/data-api-terms

[5]A pilot study is conducted to check the answers ranging from 10 to 50 characters, and results show that answers shorter than 30 characters are generally less informative.

Context	FollowupQG Examples	Category	Ratio
Initial question: Is a pregnant mother's blood kept separate from her fetus' blood? **Answer:** Yes, there is a placenta blood barrier. One of the placenta's jobs is to make sure blood from the mother and fetus never mixes. The placenta acts as an exchange surface between the mother and the fetus. Nutrients and oxygen are passed over by diffusion only.	What does the placenta exchange?	Definition	23%
	Why is it that nutrients and oxygen can only be passed over by diffusion?	Interpretation	38%
	What will happen if the blood mixes?	Counterfactual	19%
	Will the placenta still function if the woman is not pregnant?	Relating	6%
	Could someone give some suggestions on keeping the blood completely separate?	Creative	11%

Table 1: Question examples of different types of pragmatic functions in FollowupQG. The question-triggering text spans in the context are highlighted in different colors.

To select qualified workers, we restrict our task to workers who are located in five native English-speaking countries[6] and who maintain an approval rating of at least 90%. To ensure the annotations fulfil our guidelines, we give ample examples in our annotation interface with detailed explanations to help workers understand the requirements. The detailed annotation guidelines are in Appendix A. Each data sample is annotated by two different workers. We find substantial agreement between annotators, with an average Cohen's Kappa is 0.78, where the inter-annotator Kappa for Q1, Q2, and Q3 are 0.80, 0.61 and 0.92, respectively.

To evaluate the quality of annotation, we add 50 test samples to each batch of HITs. We get an average test accuracy of 0.73 for all 10 batches, indicating the high-quality of the data annotation. In the end, 112 workers participated in the task, with 96.35% average acceptance rate. The average completion time for one HIT is around 40 seconds, and we set payment at USD 1.00/HIT. To construct the final dataset, we retain only the samples that are annotated as high-quality[7] by both annotators, resulting in 3,790 instances. We randomly select 2,790 for training, 500 for validation, and 500 for testing.

4 Data Analysis

The pragmatic functions of human-raised follow-up questions and their required cognitive skill levels are crucial for understanding the mechanism of human question-asking. These factors should be studied for building efficient question generators. In § 4.1, we first characterize the pragmatic functions of questions in FollowupQG in accordance with the cognitive skills defined Bloom's Revised Taxonomy (Anderson et al., 2001). Then in § 4.2, by comparing against existing datasets, we will show that the questions in FollowupQG are of higher level and have richer pragmatic functions.

4.1 Categories of follow-up questions

We analyze 800 questions randomly sampled from our dataset and find that most follow-up questions fall into one of the following five categories that correspond to different cognitive levels in Bloom's Taxonomy. We show examples of each category in Table 1, where question-triggering text spans in the context are highlighted.

• **Definition**: 23.6% of questions seek clarifications for the definition or meaning of entities or facts in the context. Examples are: *What is the definition of ...?* We map these to the *Remembering* level in Bloom's taxonomy.

• **Interpretation**: 38.9% of questions seek interpretations for reasons, means, goals, or background information to gain a deeper understanding of the answer. Examples are: *Could you explain the reason ... ?* They correspond to the *Understanding* level in Bloom's taxonomy.

• **Counterfactual**: 18.7% of questions apply the learned knowledge in the answer to a reasonable counterfactual case. Examples are: *What will happen if ... ?* These mostly correspond to the *Applying* level in Bloom's taxonomy.

• **Relating**: 6.3% of questions ask patterns or relationships between existing examples in the context and other related cases, which belong to *Analysis* level in Bloom's taxonomy. Examples are like: *What is the relationship between ... ?*

• **Creative**: 11.1% of questions require the asker's creative thinking to invent new solutions or suggestions for learned facts in the context. They belong to *Creating* level in Bloom's taxonomy. Examples are: *Could ... be changed to improve ...?*

• **Others**: 1.3% are rhetorical questions, *e.g.*, expressing surprise by asking *Oh, really?*.

[6] Australia, Canada, Ireland, United Kingdom, USA.

[7] Choosing "Yes" answer for Q1, "No" answer for Q2, and choosing "strongly related" or "related" for Q3.

Dataset	Avg. of Words		Distribution of Cognitive Skills					Most Frequent Question Types		
	Ques.	Doc.	Rem.	Und.	App.	Anal.	Crea.	1st	2nd	3rd
FOLLOWUPQG	**43.6**	143.5	23	38	**19**	6	**11**	other	why	how
SQuAD (Rajpurkar et al., 2018)	9.9	134.8	**100**	0	0	0	0	what	how	when
LearningQ (Chen et al., 2018)	16.9	**1729.5**	18	**56**	13	**15**	3	why	other	what
InquisitiveQG (Ko et al., 2020)	7.1	150.4	46	49	5	0	0	what	why	how

Table 2: Descriptive features and statistics of FOLLOWUPQG and the datasets in comparison. We follow the Bloom's Taxonomy (Anderson et al., 2001) to define the cognitive skills of questions. **Rem.**: Remembering; **Und.**: Understanding; **App.**: Applying; **Anal.**: Analyzing; **Crea.**: Creating. For question types, we follow Liu et al. (2019) to categorize questions based on the interrogative word and define 9 question types: who, where, when, why, which, what, how, boolean, other.

In summary, 62.5% of human-raised follow-up questions are clarification questions asking for definitions and interpretation, while 36.1% of questions require higher-level cognitive thinking. This shows that FOLLOWUPQG has a relatively high proportion of questions that promote deep reasoning, considering the fact that asking deep questions is challenging for humans, revealed by prior studies (Graesser and Person, 1994; Dillon, 1988).

4.2 Comparison with existing datasets

We further compare FOLLOWUPQG with three existing QG datasets: SQuAD (Rajpurkar et al., 2016), the most widely-used dataset for answer-aware QG, and LearningQ (Chen et al., 2018) and InquisitiveQG (Ko et al., 2020), two similar datasets designed for information-seeking QG.

Table 2 shows the comparison on question and document length, question categories, and the leading question words. The question category is based on the level of cognitive skill defined in the Bloom's Taxonomy. We reuse the analytic results of Chen et al. (2018) for SQuAD and LearningQ. For InquisitiveQG, we analyze question categories by manually annotating 100 sampled questions.

Our findings are as follows. First, questions in FOLLOWUPQG are much longer than in other datasets. The reason is that natural follow-up questions usually contain additional context that is either a conditional clause to limit the scope of the question, or a summarization of the user's understanding of the context.Such additional context is often given before the actual questioning sentence to make the whole follow-up question more complete and clear. The include of additional context makes FOLLOWUPQG closer to real-world question-asking. Second, FOLLOWUPQG has a more balanced distribution of questions in terms of cognitive skills, and a high percentage of questions (36%) in high cognitive levels such as *applying* and *creating*. This makes FOLLOWUPQG significantly different with SQuAD, which is designed to test the reading comprehension ability on low cognitive skill level (*i.e. remembering*). Although InquisitiveQG also contain a high percentage of high-level questions, the key distinction is that their questions are written by crowd-workers instead of natural-occurring, which results in questions that are typically short and generic (*e.g.*, *Is there a particular example?*). LearningQ collect real questions from an online educational platform, therefore containing a large portion of clarification questions. Compared with FOLLOWUPQG, the source contexts of LearningQ (course materials and video captions) are much noisy and considerably longer, making it hard to model and evaluate the problem of FQG.

5 Follow-up Question Generation

In this section, we evaluate the ability of three pre-trained language models to generate follow-up questions via fine-tuning, while Section 6 explores large language models' ability via prompting. Through comprehensive evaluation, we discover the strengths and limitations of current models for follow-up question generation and identify areas ripe for future research.

5.1 Models

We choose three generation models that have shown state-of-the-art results on answer-aware QG: BART (Lewis et al., 2020), T5 (Raffel et al., 2020), and GPT-Neo (Black et al., 2021). We use *Huggingface* to implement BART-large and T5-base models, and fine-tune these two models on the training set of FOLLOWUPQG by predicting the follow-up question given the concatenation of the initial question and the answer as input[8]. We use

[8] `Initial Question <SEP> Answer`

Models	B1.	B2.	B3.	B4.	MET.	ROU$_L$.	FLU.	REL.	COM.	INF.
BART	**17.22**	**7.11**	**3.89**	**2.61**	**8.00**	**13.35**	4.54	**0.99**	1.36	**1.31**
T5	13.69	4.32	1.85	1.02	5.79	12.49	**4.89**	0.95	**1.51**	1.26
GPT-Neo	14.08	4.09	1.89	1.20	5.26	11.65	4.56	0.35	1.29	1.26

Table 3: Automatic and human evaluation performance for pre-trained language models on FOLLOWUPQG. **B1.**: BLEU1; **B2.**: BLEU2; **B3.**: BLEU3; **B4.**: BLEU4; **MET.**: METEOR; **ROU$_L$.**: ROUGE$_L$; **FLU.**: Fluency (1–5); **REL.**: Relevance (0–1); **COM.**: Complexity (1–3); **INF.**: Informativeness (1–3).

the *aitextgen*[9] library for implementing GPT-Neo, and the input sequence for fine-tuning this model is the concatenation of the initial question, answer and follow-up question[10] . In the testing time, only initial question and answer is given[11].

The batch size for BART, T5 and GPT-Neo is 8, 8 and 16, and we fine-tune for 10 epochs. We use Adam (Kingma and Ba, 2015) as the optimizer, with a learning rate of 5e-5 for all models. All the models are training on 1 RTX-4080 GPU. Table 4 shows the details of the models.

Model	Hidden Dimension	Layer	Head
BART-large	1024	24	26
T5-base	768	12	12
GPT-Neo	768	12	12

Table 4: Model details

5.2 Automatic Evaluation

We automatic evaluate the generated questions using BLEU1–4 (Papineni et al., 2002), METEOR (Lavie and Agarwal, 2007) and ROUGE-L (Lin, 2004). Results are shown in the top rows of Table 3 (Rows 1–3). In general, all models achieve much lower scores in automatic metrics, compared with their performance on answer-aware QG. For example, BART achieves a BLEU4 of 21.3 on SQuAD (Pan et al., 2021b), while on FOLLOWUPQG, it only achieves a BLEU4 of 2.61. Similar observations also hold for T5 and GPT-Neo. This is largely due to the open-ended nature of follow-up question generation. Compared with answer-aware QG, where the target answer is given and the questions are mostly factoid, follow-up questions are more open-ended, where the model may generate other plausible questions different from the human references, leading to low performance in n-gram based evaluation metrics. This open-ended nature of follow-up questions makes the automatic evaluation less informative.

[9]https://github.com/minimaxir/aitextgen

[10]Initial Question <SEP> Answer <QUS> Follow-up Question

[11]Initial Question <SEP> Answer <QUS>

5.3 Human Evaluation

To better evaluate the quality of generated questions, we conduct human evaluation on 100 randomly sampled pairs in the test set of FOLLOWUPQG. We ask four workers to rate the questions raised by humans and the questions generated by different models for these samples. Workers are blinded by the identity of the models in the annotation. For each question, we ask workers to give ratings on four criteria: *Relevance*, *Fluency Complexity*, and *Informativeness*. The detailed criteria are shown in our designed questionnaire in Appendix B. We average the scores from the workers on each question, reporting averaged performance.

We find that questions generated by BART and T5 achieve comparable scores with human questions in terms of *fluency* and *relevant*. However, the *complexity* and *informativeness* scores are much lower. This indicates that pre-trained models face challenges in solving the key issues of the FQG task, which aims at generating deep and informative questions. Furthermore, the Pearson correlation between automatic and human evaluation results is around 0.38, indicating a weak relationship. FOLLOWUPQG poses a new challenge for developing more faithful question evaluation metrics.

6 Controllable Follow-up QG

We see that the key difficulty in follow-up question generation is due to its open-ended nature. This increases the difficulty of controlling *what to ask* and *how to ask* for models. We now explore large language models' ability in tackling controllable follow-up question generation via in-context learning. Instead of relying on supervised fine-tuning methods, we adopt the idea of simply "prompting" the model with a few input–output exemplars to guide models to generate similar types of follow-up questions. Inspired by (Saha et al., 2023) works on

Creative Prompt

Initial Question: In a perfectly enclosed, all white room, why would the room go dark even though all light is reflected?

Answer: I think a better hypothetical here is an all mirrored room or the inside of a mirrored sphere. Light is a form of energy and this can dissipate over time…

Standard

Follow-up: Based on this, I want to raise a creative follow-up question: Can you think of any practical way to generate light inside the perfectly spherical reflector room?

Chain-of-Thought

Step 1: Based on this, I want to raise a creative follow-up question.
Step 2: I would like to know some suggestions for the light generation inside a reflector room.
Step 3: My follow-up question is: Can you think of any practical way to generate light inside the perfectly spherical reflector room?

Figure 3: Standard and chain-of-thought *creative* prompt examples. Chain-of-thoughts are highlighted.

utilizing chain-of-thought reasoning steps for the summarization task, we also give an intermediate reasoning step in a given prompt *(initial question, answer chain-of-thought, follow-up question)* to show its effectiveness for question generation.

6.1 Experimental Setting

We create standard and chain-of-thought prompts for each type of follow-up question in § 5.1, including *definition*, *interpretation*, *counterfactual*, *relating*, and *creative*. Figure 3 illustrates one example of a *creative* prompt for both the standard and chain-of-thought settings[12]. Specifically, chain-of-thought prompts aim to enhance the ability of large language models to accurately control the patterns of follow-up questions during the generation process with an intermediate reasoning step. To evaluate the results of controllability, we elicit language models to generate follow-up questions for 50 sampled (*initial question, answer*) pairs for different types of prompts. This results in 500 generated questions in total. To verify whether prompting an LLM in this way can bring controllability, we manually annotate the question types of the generated questions and calculate the question type accuracy, by comparing whether the types of the generated questions match the prompt type.

[12]We list the complete set of exemplars in Appendix C

6.2 Result Analysis

We evaluate chain-of-thought prompting on ChatGPT, GPT-3.5 (text-davinci-003), and GPT-4, respectively. Figure 4 shows the distribution of generated question types by utilizing standard and chain-of-thought prompts for the large language models. First, we observe that generating the same type of questions for the given prompts for low-level question types is relatively simpler for large language models, as compared to higher-order question types. For example, all three models have relatively higher accuracy ($\sim$60%) in generating *definition* and *interpretation* questions when given corresponding standard or chain-of-thought prompts. However, accuracies drop ($\sim$40%) when generating *relating* or *creative* questions.

Secondly, the evaluation of different language models indicates that GPT-4 outperformed other models in terms of controllable question generation tasks, particularly for high-order question generation. GPT-4 achieves around 66% accuracy in generating *creating* questions while ChatGPT and GPT-3.5 only reaches around 50% when using chain-of-thought prompts.

Third, our findings indicate that incorporating a chain-of-thought reasoning step results in an improvement in controllable accuracy when generating follow-up questions, particularly for higher-order question types. Notably, GPT-4 showcases an approximately 16% increase in accuracy in generating "creative" questions when compared to using standard prompts alone. However, ChatGPT and GPT-3.5 still exhibit relatively lower accuracy in controlling high-level questions, even with the utilization of chain-of-thought prompts. These results suggest potential future directions for further advancements in addressing the challenge of controlling and improving accuracy in generating high-level follow-up questions for these models.

6.3 Case Study

To give a clear understanding of the differences between follow-up questions raised by humans and large language models, we compare several model-generated questions with different prompts and human-raised questions in our dataset FOLLOWUPQG. Table 5 shows one *relating* human-raised follow-up question and six model-generated questions with *relating* standard and chain-of-thought prompts. More examples are shown in Appendix D In general, we observe that all models

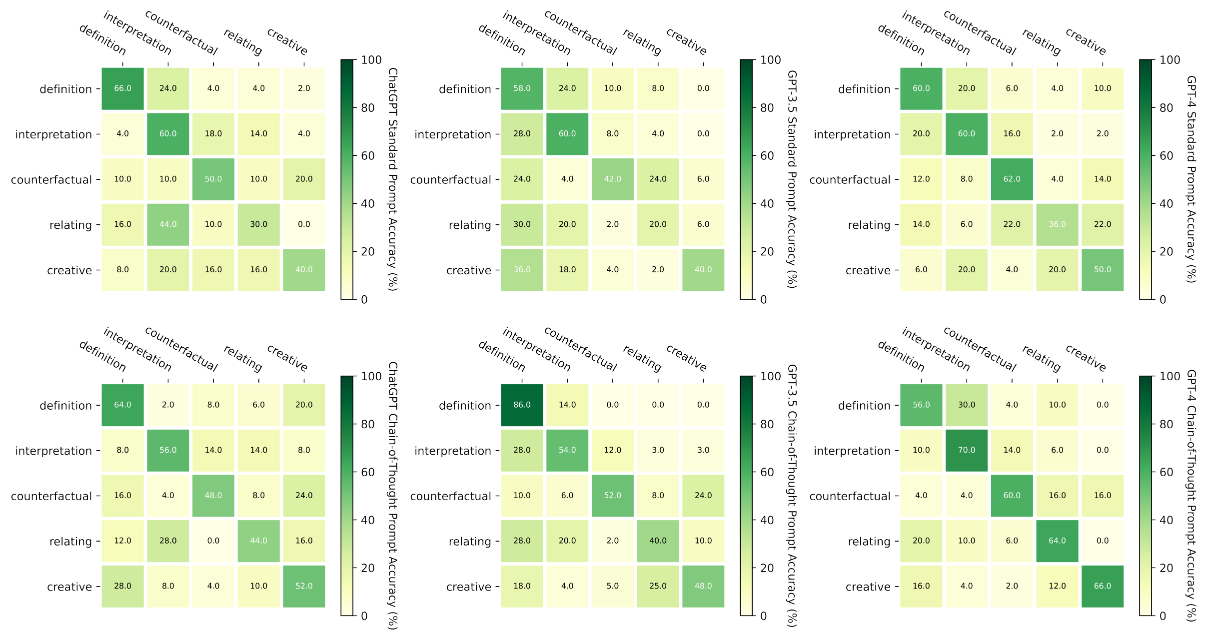


Figure 4: Distribution of generated question types by using different prompts on ChatGPT, GPT-3.5, GPT-4. Vertical labels represent prompts, horizontal labels represent question types.

Initial Question: Why are the size of insects small? Not larger as human beings? **Answer:** In the simplest terms, millions of years ago there were insects that were larger and over 7-8 feet long. It had to do with the higher percentage of oxygen in the atmosphere back then. But as the levels of gases in the atmosphere stabilized and Oxygen decreased, so did the size of these insects... **Human:** May I ask, why was there a higher percentage of oxygen? Was it because there were more plants and no human pollution? And why did it go down?	
ChatGPT (CoT): Are there any other examples of animals being affected by changes in atmospheric gases?	**ChatGPT (Standard):** Are there any mechanisms that insects have involved to counteract the lack of oxygen in the air?
GPT3.5 (CoT): What specific factors cause the decrease in oxygen levels that ultimately led to the decrease in insect size?	**GPT3.5 (Standard):** Why did the oxygen concentration decrease overtime?
GPT-4 (CoT): How did higher oxygen concentrations in the past contribute to the growth and survival of giant insects like the ones in prehistoric times?	**GPT4 (Standard):** Could insects theoretically be engineered to have more effective mechanisms, allowing them to grow larger in our current atmosphere?

Table 5: Examples of model-generated follow-up questions via standard and chain-of-thought relating prompts.

are able to generate fluent and logical follow-up questions based on the given context *(initial question, answer)*. However, they still have some limitations compared with human-written questions.

First, humans are capable of proposing *relating* questions by providing new and specific relevant examples. For instance, when presented with a context regarding a higher percentage of oxygen, humans can generate additional factors not mentioned in the given context, such as "increased plant population" or "human-induced pollution," as shown in Table 5. However, both ChatGPT and GPT-4 struggle to perform this task effectively. While GPT-4 may generate associations with new terms, the generated content appears to be more generic and lacks the specific and concrete examples that humans provide.

Second, LLM-generated questions tend to follow a formal and formulaic pattern, which contrasts with the characteristics of our dataset that are layer-person friendly and conversational in nature. This underscores another challenge of using this prompting approach with LLMs to control the style of generated questions.

7 Conclusion and Future Work

We explore the problem of information-seeking follow-up question generation by presenting FOLLOWUPQG, a dataset consisting of 3K *(initial question, answer, follow-up question)* tuples that represent real-life human question-asking, including rich pragmatic functions and diverse cognitive skills. We then propose question generation models on this data via fine-tuning and *chain-of-thought* prompting. Extensive evaluations demonstrate several difficult aspects of follow-up question generation, showing FOLLOWUPQG is a challenging dataset that deserves future investigation. Future works include how to promote higher-order deep questions, how to control the generation process, how to improve the evaluation metrics, and how to adapt follow-up QG in practical applications such as education.

Acknowledgement

This research has been supported by the WING research group of the National University of Singapore and Lee Kong Chian Fellowship awarded by Singapore Management University. We would like to thank Professor Kan for his support, and anonymous reviewers for their valuable feedback on how to improve the paper.

Limitations

We acknowledge several limitations in our work. First, follow-up questions are difficult to evaluate with current automatic evaluation metrics, especially judging whether the questions are seeking new information. Despite human evaluation is involved in this work, it is time-consuming and costly, and also it is difficult to reproduce and guarantee the evaluation consistency.

Second, since the data is collected from the online question-answering forum, the pragmatic functions we found in FOLLOWUPQG may not cover all types of follow-up questions in real-life. Although FOLLOWUPQG covers diverse types of follow-up questions from low to high cognitive levels, follow-up questions raised in other scenarios (*e.g.*, in the classroom, in paper review, in conversation) might be different and are worthwhile to explore. For example, *criticizing* questions rarely appear in our dataset, probably because in forum QA, the questioners are often less knowledgeable than the answerer in the domain they are asking for. However, in paper reviewing, *criticizing* questions may be more commonly seen.

Third, for the modeling part, we focus on revealing the limitations of state-of-the-art large language models in follow-up question generation. Although we design one method to improve the generation via *chain-of-thought* prompting, they are quite straightforward and only slightly contribute to generating deeper and more informative questions. More specialized model designs should be explored in the future to improve this, such as modeling the reasoning chain or discourse structure.

References

Lorin W Anderson, David R Krathwohl, Peter W Airasian, Kathleen A Cruikshank, Richard E Mayer, Paul R Pintrich, James Raths, and Merlin C Wittrock. 2001. A taxonomy for learning, teaching, and assessing: A revision of bloom's taxonomy of educational objectives, abridged edition. *White Plains, NY: Longman*.

Sid Black, Leo Gao, Phil Wang, Connor Leahy, and Stella Biderman. 2021. GPT-Neo: Large Scale Autoregressive Language Modeling with Mesh-Tensorflow.

Guanliang Chen, Jie Yang, Claudia Hauff, and Geert-Jan Houben. 2018. Learningq: A large-scale dataset for educational question generation. In *International Conference on Web and Social Media (ICWSM)*, pages 481–490.

JT Dillon. 1988. Questioning and teaching. a manual of practice.

Xinya Du and Claire Cardie. 2018. Harvesting paragraph-level question-answer pairs from wikipedia. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1907–1917.

Xinya Du, Junru Shao, and Claire Cardie. 2017. Learning to ask: Neural question generation for reading comprehension. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1342–1352.

Nan Duan, Duyu Tang, Peng Chen, and Ming Zhou. 2017. Question generation for question answering. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 866–874.

Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. 2019. ELI5: long form question answering. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 3558–3567.

Bilal Ghanem, Lauren Lutz Coleman, Julia Rivard Dexter, Spencer McIntosh von der Ohe, and Alona Fyshe.

2022. Question generation for reading comprehension assessment by modeling how and what to ask. In *Findings of Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 2131–2146.

Arthur C Graesser and Natalie K Person. 1994. Question asking during tutoring. *American educational research journal*, 31(1):104–137.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Wei-Jen Ko, Te-Yuan Chen, Yiyan Huang, Greg Durrett, and Junyi Jessy Li. 2020. Inquisitive question generation for high level text comprehension. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6544–6555.

Vaibhav Kumar and Alan W. Black. 2020. Clarq: A large-scale and diverse dataset for clarification question generation. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 7296–7301.

Philippe Laban, Chien-Sheng Wu, Lidiya Murakhovs'ka, Wenhao Liu, and Caiming Xiong. 2022. Quiz design task: Helping teachers create quizzes with automated question generation. *CoRR*, abs/2205.01730.

Alon Lavie and Abhaya Agarwal. 2007. METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 228–231.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 7871–7880.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81.

Bang Liu, Mingjun Zhao, Di Niu, Kunfeng Lai, Yancheng He, Haojie Wei, and Yu Xu. 2019. Learning to generate questions by learning what not to generate. In *International World Wide Web Conference (WWW)*, pages 1106–1118.

Bodhisattwa Prasad Majumder, Sudha Rao, Michel Galley, and Julian J. McAuley. 2021. Ask what's missing and what's useful: Improving clarification question generation using global knowledge. In *Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*, pages 4300–4312.

Christopher Malon and Bing Bai. 2020. Generating followup questions for interpretable multi-hop question answering. *ArXiv*, abs/2002.12344.

Naomi Miyake and Donald A Norman. 1979. To ask a question, one must know enough to know what is not known. *Journal of verbal learning and verbal behavior*, 18(3):357–364.

Lidiya Murakhovs'ka, Chien-Sheng Wu, Tong Niu, Wenhao Liu, and Caiming Xiong. 2021. Mixqg: Neural question generation with mixed answer types. *CoRR*, abs/2110.08175.

Preksha Nema, Akash Kumar Mohankumar, Mitesh M. Khapra, Balaji Vasan Srinivasan, and Balaraman Ravindran. 2019. Let's ask again: Refine network for automatic question generation. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3312–3321.

Liangming Pan, Wenhu Chen, Wenhan Xiong, Min-Yen Kan, and William Yang Wang. 2021a. Unsupervised multi-hop question answering by question generation. In *Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*, pages 5866–5880.

Liangming Pan, Wenhu Chen, Wenhan Xiong, Min-Yen Kan, and William Yang Wang. 2021b. Zero-shot fact verification by claim generation. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 476–483.

Liangming Pan, Wenqiang Lei, Tat-Seng Chua, and Min-Yen Kan. 2019. Recent advances in neural question generation.

Liangming Pan, Yuxi Xie, Yansong Feng, Tat-Seng Chua, and Min-Yen Kan. 2020. Semantic graphs for generating deep questions. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1463–1475.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 311–318. ACL.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21:140:1–140:67.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for squad. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 784–789.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100, 000+ questions for machine comprehension of text. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2383–2392.

Ashwin Ram. 1991. A theory of questions and question asking. *Journal of the Learning Sciences*, 1(3-4):273–318.

Sudha Rao and Hal Daumé III. 2018. Learning to ask good questions: Ranking clarification questions using neural expected value of perfect information. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 2737–2746.

Sudha Rao and Hal Daumé III. 2019. Answer-based adversarial training for generating clarification questions. In *Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*, pages 143–155.

Siva Reddy, Danqi Chen, and Christopher D. Manning. 2019. CoQA: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266.

Christopher Richardson, Sudipta Kar, Anjishnu Kumar, Anand Ramachandran, Zeynab Raeesy, Omar Khan, and Abhinav Sethy. 2023. Learning to retrieve engaging follow-up queries. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 2009–2016, Dubrovnik, Croatia. Association for Computational Linguistics.

Swarnadeep Saha, Shiyue Zhang, Peter Hase, and Mohit Bansal. 2023. Summarization programs: Interpretable abstractive summarization with neural modular trees.

Sandeep Subramanian, Tong Wang, Xingdi Yuan, Saizheng Zhang, Adam Trischler, and Yoshua Bengio. 2018. Neural models for key phrase extraction and question generation. In *Workshop on Machine Reading for Question Answering@ACL*, pages 78–88.

Siyuan Wang, Zhongyu Wei, Zhihao Fan, Yang Liu, and Xuanjing Huang. 2019. A multi-agent communication framework for question-worthy phrase extraction and question generation. In *AAAI Conference on Artificial Intelligence (AAAI)*, pages 7168–7175.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Huai hsin Chi, F. Xia, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. *ArXiv*, abs/2201.11903.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2369–2380.

Yao Zhao, Xiaochuan Ni, Yuanyuan Ding, and Qifa Ke. 2018. Paragraph-level neural question generation with maxout pointer and gated self-attention networks. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3901–3910.

A Human Annotation Guideline

A.1 Job Description

The aim of this data annotation task is to select high quality (Knowledge Goal Question, answer, follow-up question) tuples from online forums. The raw data was crawled from an online educational forum *Reddit* in the topic of *"explainlikeimfive"*, which aims to provide lay person-friendly explanations for real-life questions. We collected many (question, answer) pairs from the forum, together with the replies to the answer. A common type of reply to the answer is *asking follow-up questions*. We selected out those follow-up questions and format each data entry as a triple of (initial question, answer, follow-up question). An example is as follows:

Initial Question: When a disease infects you, most of them aim to kill you. But if they kill you, they die too. So why would they want to kill themselves?

Answer: They don't aim to kill you, they just survive and spread or not. Some diseases may be new to a particular environment, and the host may be unprepared for a new disease thus resulting in worse sickness and death for the host. Obviously , in general it does not suit the disease if the host dies too quickly and thus does not get a chance to pass it on but near in mind that the disease may have a natural reservoir elsewhere.

Follow-Up Question: If so why harm the host? If the disease could have peace with the body then it could multiply and spread.

However, there is much noise in the data we collected. For example, the follow-up question is not a valid question, the data entry contains sensitive information and the follow-up question may be not relevant to the question or the answer. **Your job is to help us identify and remove those noisy samples.**

For each HIT, we will show you one triple of (question, answer, follow-up question), you need to read the question, answer and follow-up question carefully and then answer the following three questions based on your judgement.

Q1: Do you think the follow-up question is a valid question?

A. Yes B. No

Q2: Do you think the (initial question, answer, follow-up question) contains sensitive information?

A. Yes B. No

Q3: Do you think the follow-up question is related to the original question and the answer?

A. Strongly related B. Related C. Slightly related D. Not related

When you answered the three questions, click submit to jump to the next sample.

A.2 Detailed Guidelines

A.2.1 Guideline for Q1

Q1: Do you think the follow-up question is a valid question?
A. Yes B. No

Guideline: The follow-up question might contain multiple sentences but it should consist of at least one valid question. A valid question must be in *question format* and *ask meaningful information*, including Wh-questions (what/why/where/etc.), open-ended questions, probing questions and etc. Examples for *invalid questions*: "10000 meters? really?", which are often used in conversational speech to express feelings instead of asking new information.

Examples that are considered INVALID question:

Initial Question: Eli5: Why do seahorses still existed?

Answer: That's not how evolution works. Evolution is survival of the fittest, not survival of the most useful. Seahorses will keep existing as long as they have food and reproduce faster than they are hunted.

Follow-up Question: Ohhh, I can't believe it. Seahorses will keep existing, really?

Reason: "Seahorses will keep existing, really?" is in question format but is invalid, because it is not asking information but just expressing suspicious feeling.

Initial Question: Why do tapes have screen tearing and splitting, but modern media doesn't?

Answer: Information on a VHS was an analog signal. To fit all of it on a tape in a linear form would end up with miles of tape. So they record the information in diagonal stripes across the tape, which is much wider than it would need to be if it just held a single track. This is then played back through a spinning head that is place at the same angle relative to the tape as the information on it. The tape moved by the head as the head spins, reading each stripe in a fluid motion, creating the smooth video you see on playback. Pausing it stops the tape moving, but not the spinning head. So the image you see is whatever that spinning head is reading repeated over and over. Hence the "between frames" effect.

Follow-up Question: That's right, I remember watching a Technology Connections video about how the head itself had to be skewed because of that. I remember as a kid looking inside the VCR and wondering "what is that thing, is it broken?" Now I know.

Reason: "what is that thing, is it broken?" is just the questioner's memory, not a question asking information in the original question or the answer.

Examples that are considered VALID question:

Initial Question: ELI5: Why is the sea calm in the mornings?

Answer: There are two types of waves which can turn a flat sea into a rougher one - swell waves and wind waves. Swell waves can arrive at any time of day, but because wind waves are generated by the wind, they only develop when the wind begins to blow steadily. Since wind speeds are often low at night, and increase during the daytime, wind waves often die out during the night, leading to a relatively flat sea (perhaps with swell waves) in the early morning. During the day, the wind waves increase in size as the wind speed increases, leading to a rougher, more choppy, sea surface during the afternoon and evening.

Follow-up Question: Now it's time for my question. Why are winds always weak in the morning and very strong during the day?

Reason: The follow-up question is a "Why" question, asking specific reasons about the change of the winds during the day. Therefore, it is a valid question.

A.2.2 Guideline for Q2

Q2: Do you think the (initial question, answer, follow-up question) contains racism, hate speech, sexual topics, offensive or rude comments?

A. Yes B. No

Guideline: Examples for racism comments: "It's credit to your race", "Black people will not understand". Examples for hate speech: "He should go back to where he comes from", " All Mexicans are rapists". Examples for offensive or rude comments: "Women are not suitable for working in IT field.", "Gay will never understand".

Examples that contain INAPPROPRIATE information:

Initial Question: ELI5: Do Asexual people not feel orgasms or pleasure?

Answer: Asexual people don't feel sexual attraction. That is, the thing that makes a sexual person look at someone else and think "I want to have sex with that person". Asexual people can still feel sexual desire or sexual arousal, but it usually won't be targeted towards someone in particular. I had an asexual roommate, and she did feel sexual arousal, but didn't feel sexual attraction. So, in the rare time that she masturbated, she would be able to become aroused by porn or sexual material, but her arousal wasn't centered in imagining herself performing the actions or having those actions performed on her.

Follow-up Question: Is it like, how? Do their brains work in a way that disallows them of sexual thinking? I don't mean anything bad, believe me, but I can't understand how something as primal and ingrained in our genes as reproducing is nullified by a factor of personality.

Reason: Sexual topic is involved in this data entry, which can be considered as the sensitive data.

A.2.3 Guideline for Q3

Q3: Do you think the follow-up question is related to the initial question or the answer?

A. Strongly Related B. Related C. Slightly Related D. Not Related

Guideline: "Strongly Related": The follow-up question asks specific definition, particular reasons or meanings in the original question and the answer. The information asked is all included in the original question or the answer. "Related": The follow-up question mainly asks information occurred in the original question or the answer but also involves other new information. "Slightly Related": The follow-up question mainly asks other cases but the cases are relevant to the original question or the answer. "Not Related": The follow-up question asks nothing relevant to the original question or the answer.

Examples that are STRONGLY RELATED to the original question:

Initial Question: ELI5: How is 2FA security by-passed?

Answer: Various possible methods. If someone inadvertently shared their 2fa setup key (or a situation like having their phone stolen and the authentication app backed up), that could be used alongside the original password to gain access. SIM swapping is another possible tactic assuming the first password and the victims cell phone number is already known: if the attacker calls the carrier and requests a replacement SIM/ESIM delivered to the attacker, the attacker could then take advantage of any voice/text message 2fa prompts.

Follow-up Question: And can anything be done to prevent a SIM swap?

Reason: The follow-up question is asking for elaboration on the solutions about how to prevent a SIM swap, and "SIM swap" appears in the answer. In addition, there is no other new information in the follow-up, and thus it can be considered as strongly related.

Examples that are RELATED to the given answer.

Initial Question:ELI5: What is an NFT?

Answer: They're kind of like digital trading cards. You buy a thing that represents a digital thing (an image or tweet or whatever else), and then you can turn around and sell that token for hopefully more money than you paid for it. Some important things to note: * Owning the token provides you with no practical rights to the thing (for example, you can't prevent others from using or displaying it) * The token is safe from counterfeiting because it's backed by the block-chain (no one else can steal your token or produce a fake one) * Because it's backed by the block-chain, every transaction that happens with your token uses a *lot* of electricity, which is a problem.

Follow-up Question: Is it possible to monetize the usage or display of the NFT? Like royalties for memes?

Reason: This follow-up question mainly asks information about NFT, but related to the case "royalties for memes". Therefore, it can be considered as the "Related" question.

Examples that are SLIGHTLY RELATED to the question or the given answer.

Initial Question: ELI5: Why turning something off, then on again actually fixes issues?

Answer: Turning something off brings it back to a known good state, and powering it up puts it through a known good set of steps. During the time it was on, errors and junk data and black magic might have occurred, giving bad data to good processes, resulting in poor performance.

Follow-up Question: Is it practically the same for software when uninstalling and then reinstalling it?

Reason: This follow-up question is asking whether the information in the answer is also suitable for other cases. The main goal for this question is to ask the principle of "software install" but also mentions the information in the answer. Therefore, it is slightly related to the given answer.

Examples that are NOT RELATED to the given answer.

Initial Question: ELI5: Why does light get absorbed in black holes if photons are massless?

Answer: Photons are "massless", but not really. They're still physical and can be impacted by gravity. This is why light can 'curve' around a dense object like a planet. Black Holes have crazy amount of gravity, and literally traps light inside.

Follow-up Question: Could you also explain the theory on Newton's law?

Reason: Newton's law is nothing related to the question or the answer. Therefore, it can be considered as "NOT RELATED" case.

A.3 Quality Check

We will check the quality of your data at the back-end in the Amazon-Turk. If 80% of your annotated data entries are invalid, we will probably reject all your annotated data samples. Instead, if your annotated data samples are all in line with our standards, we will give you a bonus as an extra reward. In detail, to control the annotation quality, the test samples will be inserted into each batch. According to the accuracy of the test samples, if the accuracy is lower than 60%, the annotation from this participant will be rejected.

B Human Evaluation Questionnaire

Table 6 shows the human evaluation questionnaire.

Questionnaire
Q1. Whether this follow-up question is relevant to the original question and the answer? ○ Yes ○ No
Q2. Whether this follow-up question contains the following errors: ☐ correct ☐ redundant ☐ invalid ☐ context copy ☐ wrong semantic collocation ☐ others
Q3. Whether this follow-up question requires reasoning to generate? ○ Yes, and very hard ○ Yes, but slightly simple ○ No
Q4. Whether this follow-up question reveals new information? ○ Yes, and a lot ○ Yes, but a few ○ No

Table 6: Questionnaire designed for human evaluation, where ○ and ☐ indicate single-item and multiple-item selection respectively.

C Chain-of-thought Prompt

Definition Prompt
Question: ELI5: How do you take a picture/screenshot of a bluescreen? You can write down the message or take a photo of the screen.
Answer: If it occurs within a virtual machine, then you can screenshot within the host OS...
Step 1: Based on this, I want to raise a definition question.
Step 2: I want the information 'virtual machine' to be further clarified.
Step 3: My follow-up question is: Thank you for your explanation. I have a few follow-up questions though: What exactly is a virtual machine?

Question: ELI5: How does the heart compensate when losing a lot of blood? For example a head injury or something more severe?
Answer: It beats faster to push what blood you do have to where your body still needs it...
Step 1: Based on this, I want to raise a definition question.
Step 2: I want the information 'put pressure on the wound' to be further clarified.
Step 3: My follow-up question is: Doesn't it also regulate pressure by constricting blood vessels to keep the flow at the same pressure?

Question: ELI5 - What exactly is the voltage measurement across your neuron?
Answer: Voltage is a measure of potential difference...
Step 1: Based on this, I want to raise a definition question.
Step 2: I want the information 'potential difference' to be further clarified.
Step 3: My follow-up question is: So what exactly is the potential difference? is it not the δV between two points of an electric field?

Question: ELI5: In Archeology, why do they have to dig to get to stuff? Was the ground plane lower then and higher now, or do things sink

over time? Pompeii was literally buried in volcanic ash.
Answer: Rome is a city built upon a city...
Step 1: Based on this, I want to raise a definition question.
Step 2: I want the information 'Rome' to be further clarified.
Step 3: My follow-up question is: Interesting point about Rome. What were the previous cities that existed before Rome?

Question: ELI5: How does a cell phone determine how much charge is left?
Answer: My understanding is that batteries output a constant voltage until ...
Step 1: Based on this, I want to raise a definition question.
Step 2: I want the information 'voltage regulator' to be further clarified.
Step 3: My follow-up question is: is this through using a voltage regulator that accepts variable input voltage? What is keeping the phone from running on less than 3v, as long as the amperage/resistance change accordingly to keep it in working range?

Interpretation Prompt
Question: Why do car insurance companies charge you more to retain your full right to sue someone?
Answer: Because if you delegate your right to sue for pain and suffering to them...
Step 1: Based on this, I want to raise an interpretation question.
Step 2: I want the information 'how to delegate the right and the insurance' to be explained.
Step 3: My follow-up question is: When selecting a policy, it gives two options of either limit the right or full right. Why can't we just sue someone individually and cut out the insurance as a middleman?

Question: Why don't dentists and eye doctors have to go to regular medical school?
Answer: The simplest explanation is that dentistry and optometrists are more about tools...
Step 1: Based on this, I want to raise an interpretation question.
Step 2: I want the information 'masses curve spacetime' to be explained.
Step 3: My follow-up question is: Then there is the flip side to my question as well, why are there not more specialized schools? A podiatrist doesn't need to know the same things an OB does but they both have the same gen ed in medical school.

Question: How can you convert liabilities to equity?
Answer: There are two primary ways for a company to raise money: taking on liabilities...
Step 1: Based on this, I want to raise an interpretation question.
Step 2: I want the information 'the relationship between liability and equity' to be explained.
Step 3: My follow-up question is: Oh ok I see, makes sense now. So essentially there will be more shareholders and thus more equity, but the liabilities would still have to be paid in the long run. Why is this conversion usually seen as a good thing?

Question: why are fish sometimes exempt from some vegetarian diets?
Answer: A 'vegetarian' who eats fish would likely be more aptly called a 'Pescatarian'...
Step 1: Based on this, I want to raise an interpretation question.
Step 2: I want the information 'fish and diets' to be explained.
Step 3: My follow-up question is: Why are fish sometimes the exception, health, morality, or otherwise?

Counterfactual Prompt
Question: Why can't we lower ocean levels manually?
Answer: You could, but the amount of difference...
Step 1: Based on this, I want to raise a counterfactual question.
Step 2: I want to know the counterfactual cases on 'digging for the fossil fuel'
Step 3: My follow-up question is: What if we were to hypothetically be capable to drill down into the Challenger Deep?

Question: ELI5: when we react to noises, why do we react faster when they are louder?
Answer: The startle reflex is a reaction in the brain that kicks the body into ...
Step 1: Based on this, I want to raise a counterfactual question.
Step 2: I want to know the counterfactual cases on 'the level of noises'.
Step 3: My follow-up question is: What makes the

louder noise likely to trigger the reflex if we look at the processes in the body?

Question: If soap and water clean the body, then what makes towel dirt after a bath?
Answer: When you dry yourself with a course...
Step 1: Based on this, I want to raise a counterfactual question.
Step 2: I want to know the counterfactual cases on 'alternative methods to avoid making towel dirt'.
Step 3: My follow-up question is: What if I just wrap a towel around my waist and walk around and air dry?

Question: How are gorillas and similar animals taught sign language?
Answer: They aren't. At least not in the..
Step 1: Based on this, I want to raise a counterfactual question.
Step 2: I want to know the counterfactual cases on 'language without repeating patterns'.
Step 3: My follow-up question is: What is language if not repeating patterns to communicate with another?

Question: What happens if you get sued for an amount of money significantly higher than you could ever actually pay?
Answer: It's why they sell umbrella coverage...
Step 1: Based on this, I want to raise a counterfactual question.
Step 2: I want to know the counterfactual cases on 'umbrella coverage'.
Step 3: My follow-up question is: What if you don't have umbrella coverage?

Relating Prompt
Question: How does a USB-C charger stay in the port?
Answer: It's round and smooth on the outer surface...
Step 1: Based on this, I want to raise a relating question.
Step 2: I want to know the relationship between the case of 'usb-c' and to relevant cases 'micro-b'.
Step 3: My follow-up question is: How come usb c is soo much better than micro b?

Question: What does this actually mean when people say private people tell you so little about themselves but you think you know a lot about them?
Answer: People tend to fill in the blanks...
Step 1: Based on this, I want to raise a relating question.
Step 2: I want to know the relationship between the case of a 'private person' and relevant cases of the new descriptions of being private.
Step 3: My follow-up question is: Is there a more specific word for this beyond being private?

Question: How is putting LED screens on grocery store cooler doors, instead of simple glass doors, a profitable choice?
Answer: The idea is that you save energy... *Step 1*: Based on this, I want to raise a relating question.
Step 2: I want to know the relationship between the case of 'putting LED screens on cooler doors' and the case of 'simple glass doors'.
Step 3: My follow-up question is: Except how is that better than a glass door you can see through?

Question: How can choosing pictures of certain objects (stop signs, cross walks, cars, etc.) prove that you are not a robot on websites?
Answer: Because how bots select images on ...
Step 1: Based on this, I want to raise a relating question.
Step 2: I want to know the relationship between the case of 'software made by the bot' and to new relevant cases of 'human-like bot'.
Step 3: My follow-up question is: But why haven't the bot makers made their software behave more like a human?

Question: Why do some countries' currencies go into million for such a small amount of value?
Answer: For example, in Thailand... *Step 1*: Based on this, I want to raise a relating question.
Step 2: I want to know the relationship between the case of 'domination' and new relevant cases.
Step 3: My follow-up question is: But why they don't do any re-domination to make things easier?

Creative Prompt
Question: In a perfectly enclosed, all-white room, why would the room go dark even though all light is reflected?
Answer: I think a better hypothetical ...
Step 1: Based on this, I want to raise a creative question.

Step 2: I want to know new solutions or suggestions for the case of ’light generation inside a reflector room’.
Step 3: My follow-up question is: Can you think of any practical way to generate light inside the perfectly spherical reflector room?

Question: During a live televised sports match, how do they get the replay footage edited in so quickly?
Answer: I design control rooms and their replays systems...
Step 1: Based on this, I want to raise a creative question.
Step 2: I want to know new solutions or suggestions for the ’career path’.
Step 3: My follow-up question is: What career path led you to that?

Question: How are PGP signatures used to verify users on forums?
Answer: You have to have access to the private ...
Step 1: Based on this, I want to raise a creative question.
Step 2: I want to know new solutions or suggestions for the ’private key’.
Step 3: My follow-up question is: How would they know you had the private key?

Question: How does your brain ’realize’ that you’re home?
Answer: I kinda feel the same..
Step 1: Based on this, I want to raise a creative question.
Step 2: I want to know new solutions or suggestions for the ’unexplained human instinct’.
Step 3: My follow-up question is: why would you say that it can’t be explained?

Question: How do Small Business Owners Pay Themselves?
Answer: You pay yourself as an employee and so you are taxed ...
Step 1: Based on this, I want to raise a creative question.
Step 2: I want to know new solutions or suggestions for ’how to define a phenomenon’.
Step 3: My follow-up question is: Any idea how they define “reasonable”?

D Case study

Examples of human-written follow-up questions and model-generated questions for standard and chain-of-thought prompts for definition (Table 7), interpretation (Table 10) counterfactual (Table 8) and creative (Table 9) types.

Initial Question: Why are PS4 and PS5 unable to read PS1 or PS2 discs? **Answer:** PS1 and some PS2 disks are based on the CD standard. The drive in a PS4 (and I assume PS5) doesn't read CDs. Among other things, it's only equipped with lasers for DVD (650nm) and Blu-ray (450nm). **Human:** A lot of PS2 games were on DVD though. Why can't the console read those?	
ChatGPT (CoT): What exactly is the CD standard? Is it the disc size and shape, or is it the way data is stored on the disc?	**ChatGPT (Standard):** What is the difference between CDs, DVDs, and Blu-ray discs in terms of their laser technology and storage capacity?
GPT3.5 (CoT): What is the difference between the CD standard and the DVD/Blu-ray standards that make the PS4 unable to read the PS1 or PS2 discs?	**GPT3.5 (Standard):** So does that mean that if the PS4 had the right lasers, it could read the PS1 and PS2 discs?
GPT-4 (CoT): Could you elaborate more on the CD standard? Also, are there any technical reasons why they didn't include the ability to read CDs on the PS4 and PS5 other than cost-saving measures?	**GPT4 (Standard):** So the reason for the inability to read PS1 or PS2 discs on PS4 and PS5 is due to the different disc formats and laser technology used, correct? Are there no possible software solutions to overcome this hardware limitation or enable backward compatibility?

Table 7: Examples of human written follow-up questions and model-generated questions via standard and chain-of-thought definition prompts.

Initial Question: What are scripts in terms of code? **Answer:** Typically a script refers to a small application written in a high-level programming language that isn't compiled into an executable file which typically automates a process rather than being an application that a user interacts with. **Human:** So like, say I coded a quick grade calculator in Python. That would be a script right? While an actual game in C++ that I could run would be a program?	
ChatGPT (CoT): Can you give an example of a script and how it is used for automation?	**ChatGPT (Standard):** Can you give some examples of common tasks that are automated by scripts in various fields like IT, finance, or web development?
GPT3.5 (CoT): So when you compile a script, does that mean it is converted into an executable file and can then be used as an application? What are the other benefits of compiling a script?	**GPT3.5 (Standard):** How do you compile a script?
GPT-4 (CoT): Can you give some examples of scripting languages and their common uses?	**GPT4 (Standard):** What are some examples of scripting languages and what kinds of processes can be automated with scripts?

Table 8: Examples of human written follow-up questions and model-generated questions via standard and chain-of-thought interpretation prompts.

Initial Question: What difference do the vocal cords of a person who can sing have in comparison to someone who can't? **Answer:** Absolutely nothing. The vocal cords are an instrument, the same as a trumpet or clarinet. Anyone is capable of singing well, the difference between someone who is good and who is bad is practice, technique, and for some, proper training. There have been studies proving that anyone can be taught to sing despite the fact that they have any practice in singing or not. **Human:** Are there any biological differences in someone's singing ability when it comes to how well they can hear the true pitch that they are putting out? Not vocal chord-related obviously, but the idea that anyone can kind of baffles me. I mean my friend. Asking for a friend.	
ChatGPT (CoT): What about the influence of genetics on singing ability? Are there any studies or evidence that suggest certain genetic traits may make it easier or harder for individuals to develop singing skills?	**ChatGPT (Standard):** If everyone can be taught to sing well, then why do some people seem to have a natural talent for singing?
GPT3.5 (CoT): If someone has great technique and practice but still has difficulty singing at certain pitches, is it possible that some people just have a lower ceiling of what they can do and others don't?	**GPT3.5 (Standard):** If everyone can learn techniques to sing well, then why do some people seem to be talented at singing?
GPT-4 (CoT): Are there any genetic factors that influence people's singing ability?	**GPT4 (Standard):** If there are no physical differences in vocal cords between good and bad singers, are there any other factors, such as genetics or anatomy, that contribute to someone's natural singing ability or vocal range?

Table 9: Examples of human written follow-up questions and model-generated questions via standard and chain-of-thought counter-factual prompts.

Initial Question: Why do we have an internal monologue and how does it work? **Answer:** From what I've read it's believed that some people have a connection between their frontal lobe where conscious thought is processed and their auditory processing center where speech is processed and constructed. Not everyone has this link however. About 25% of people don't and have no internal monologue. About 25% of people conduct this internal monologue all the time, and the rest only do part of the time. **Human:** This is such a strange concept to me. How do you think without words? My inner monologue never stops.	
ChatGPT (CoT): Could the absence of an internal monologue be beneficial in any way?	**ChatGPT (Standard):** Could you give me some examples of potential benefits of internal monologue?
GPT3.5 (CoT): What are the potential benefits of having an internal monologue?	**GPT3.5 (Standard):** Are there any specific benefits to having an internal monologue?
GPT-4 (CoT): Since there's a considerable difference in the presence of an internal monologue among people, are there any known advantages or disadvantages associated with having or not having an internal monologue?	**GPT4 (Standard):** Is there a correlation between having an internal monologue and certain personality traits or cognitive abilities, such as creativity or problem-solving skills?

Table 10: Examples of human written follow-up questions and model-generated questions via standard and chain-of-thought creative prompts.