Bridging Information Gaps with Comprehensive Answers: Improving the Diversity and Informativeness of Follow-Up Questions

Zhe Liu[1*] **Taekyu Kang**[1*] **Haoyu Wang**[1] **Seyed Hossein Alavi**[1,2] **Vered Shwartz**[1,2]
[1] University of British Columbia [2] Vector Insitute
{zheliu92, salavis, vshwartz}@cs.ubc.ca
{davidk15, macdude}@student.ubc.ca

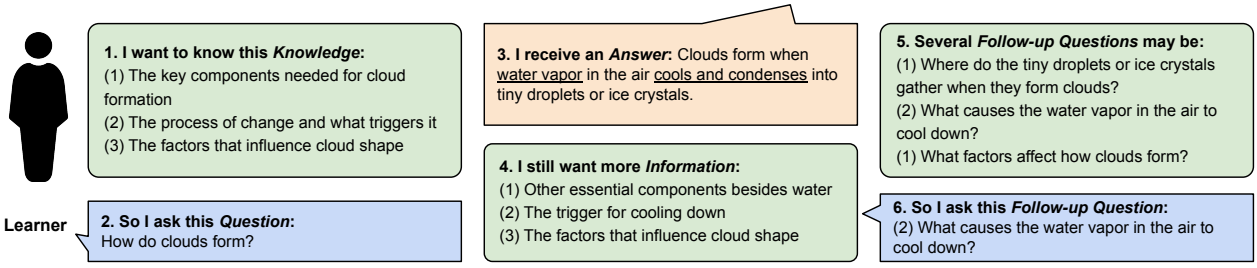

Figure 1: Illustration of a learner's cognitive process in generating follow-up questions. Green: Implicit cognitive stages; Callouts: Explicit (question, answer, follow-up) triplets collected for the dataset.

Abstract

Generating diverse follow-up questions that uncover missing information remains challenging for conversational agents, particularly when they run on small, locally hosted models. To address this, we develop an *information-gap-driven* knowledge distillation pipeline[1] in which a teacher LLM generates a comprehensive answer, contrasts it with the initial answer to identify information gaps, and formulates gap-bridging follow-up questions. Using this pipeline, we augment the existing FOLLOWUPQG dataset tenfold. We then fine-tune smaller student models on the augmented dataset to distill the teacher's knowledge. Experiments with selected teacher-student model pairs show that fine-tuned students achieve significantly higher informativeness and diversity than variations trained on the original dataset. These findings indicate that our pipeline, which mirrors the human cognitive process of information seeking, provides an efficient distillation channel from state-of-the-art LLMs to smaller models, enabling resource-constrained conversational systems to generate more diverse and informative follow-up questions.

1 Introduction

Asking questions is a fundamental mechanism for humans to acquire new information, particularly when existing information is incomplete. While large language models (LLMs) excel at passively answering questions from users, their ability to proactively guide conversations by identifying and addressing information gaps remains underdeveloped (Liu et al., 2025), with smaller models performing even worse. Therefore, the task of question generation (QG) has become a focal point in natural language processing (NLP) for its role in improving information-seeking dialogue systems (Chen et al., 2024)—including, making information seeking more accurate and efficient (Qi et al., 2020), resolving ambiguities (Li et al., 2017), and ultimately better understanding users' needs to provide suitable assistance across various domains (Laban et al., 2022; Arslan et al., 2024; Li et al., 2024).

While most existing QG tasks focus on generating questions directly answerable from a given context (Zhao et al., 2018; Pan et al., 2020; Ghanem et al., 2022)—a process that diverges from how humans infer and pursue missing information, Meng et al. (2023) propose the FOLLOWUPQG task, which requires models to generate *follow-up questions* that build on, but are not answerable by, the initial question-answer pair. They create the FOLLOWUPQG dataset and show that existing models often produce repetitive or context-bound questions that fail to target unexplored information (Meng et al., 2023). The core challenges of the FOLLOWUPQG task can be formulated into two dimensions: (1) identifying information gaps, the

[*] Denotes Equal Contribution.
[1] Code available at https://github.com/zheliu92/nlp_followupqg_public/

*Proceedings of the 14th Joint Conference on Lexical and Computational Semantics (*SEM 2025)*, pages 13–30
November 8-9, 2025

unanswered aspects of the initial question, and (2) generating diverse questions that target these gaps.

Building on traditional QG methods (Zhao et al., 2018; Pan et al., 2020; Ghanem et al., 2022), recent work attempts to generate information-seeking follow-up questions using preference optimization (Mazzaccara et al., 2024) and knowledge graphs (Liu et al., 2025), but still lack explicit mechanisms to model gaps or ensure diversity. To address these limitations, we propose an information-gap-driven teacher-student knowledge distillation pipeline. In our approach, a teacher LLM generates a hypothetical "complete" response to the initial question, contrasts it with the often incomplete initial answer to identify information gaps, and formulates gap-bridging follow-up questions. By generating multiple follow-up questions, each targeting some unanswered information, this pipeline ensures the diversity and informativeness of the follow-up questions. For example, in Figure 1, if the initial answer to "*how do clouds form?*" is "*clouds form when water vapor cools,*" a comprehensive answer might add "... *and condenses around dust particles,*" which explicitly exposes the gap through contrast and leads to an informative follow-up question such as "*What role do particles play in cloud formation?*"

Our pipeline can be applied across different teacher-student model pairs. In this work, we use GPT-4o (`2024-02-15-preview`) as the teacher model and BART-large as the student model to verify the pipeline. Specifically, we use GPT-4o to generate the comprehensive answers and follow-up questions. After verifying the quality of the follow-up questions via human evaluation, we then augmented the original FOLLOWUPQG training set tenfold and fine-tuned BART-large on both the original dataset and our augmented dataset. Leveraging GPT-4o to generate high-quality training data, and then distilling the teacher's knowledge into smaller models, our approach achieves strong performance at a significantly lower cost. The experimental results demonstrate significant improvements of the augmented dataset over the baselines, both in terms of quality (validity, relevance, informativeness, etc.) and diversity. Our contributions are as follows:

- We propose an *information-gap-driven* teacher-student knowledge distillation pipeline that generates follow-up questions through contrastive analysis of initial answers and generated comprehensive answers.
- We augment the FOLLOWUPQG training set with over 25,000 high-quality synthetic examples.
- Experimental results show that small models fine-tuned on our augmented dataset outperform peer small-model baselines and achieve near parity with representative LLM-based *Teacher* and *Chain-of-Thought* models.

2 Related Work

Question generation (QG) focuses on automatically generating semantically meaningful and well-structured questions based on a given text (Ali et al., 2010). While traditional QG techniques have made significant strides in domains such as machine comprehension (Du et al., 2017; Uto et al., 2023), e-commerce (Wang et al., 2021), and education (Luo et al., 2024), they primarily generate questions based on known answers. This approach contrasts sharply with human questioning behavior, which actively seeks new information from various perspectives. This limitation has led to the emergence of FOLLOWUPQG, a task whose goal is to generate questions that explore previously unanswered or underexplored aspects of a given text.

FOLLOWUPQG has evolved from simpler methods, such as template-based and retrieval-driven approaches (Kumar and Joshi, 2017; Soni and Roberts, 2019; B et al., 2020), to more advanced techniques that prioritize informativeness (Majumder et al., 2021; Mazzaccara et al., 2024). Knowledge-enhanced approaches, like those in Ge et al. (2023) and Gupta et al. (2022), leverage entity-relation pairs and knowledge graphs to improve the depth of the generated questions. Further advancing this, Liu et al. (2025) combined knowledge graphs with LLMs to increase question informativeness. Efforts to model human-like questioning behavior, such as InquisitiveQG (Ko et al., 2020), have relied on crowd-sourced follow-up questions written for news articles rather than those naturally generated by humans, leading to a lack of depth and cognitive diversity.

We follow the setting of the FOLLOWUPQG (Meng et al., 2023), which formalizes information-seeking follow-up question generation. Based on questions and answers from the ELI5 (explain like I'm 5) subreddit, follow-up questions in this dataset build upon—but are not answerable by—the initial question-answer pair, resembling real-world dialogues where follow-ups resolve ambiguities or deepen understanding.

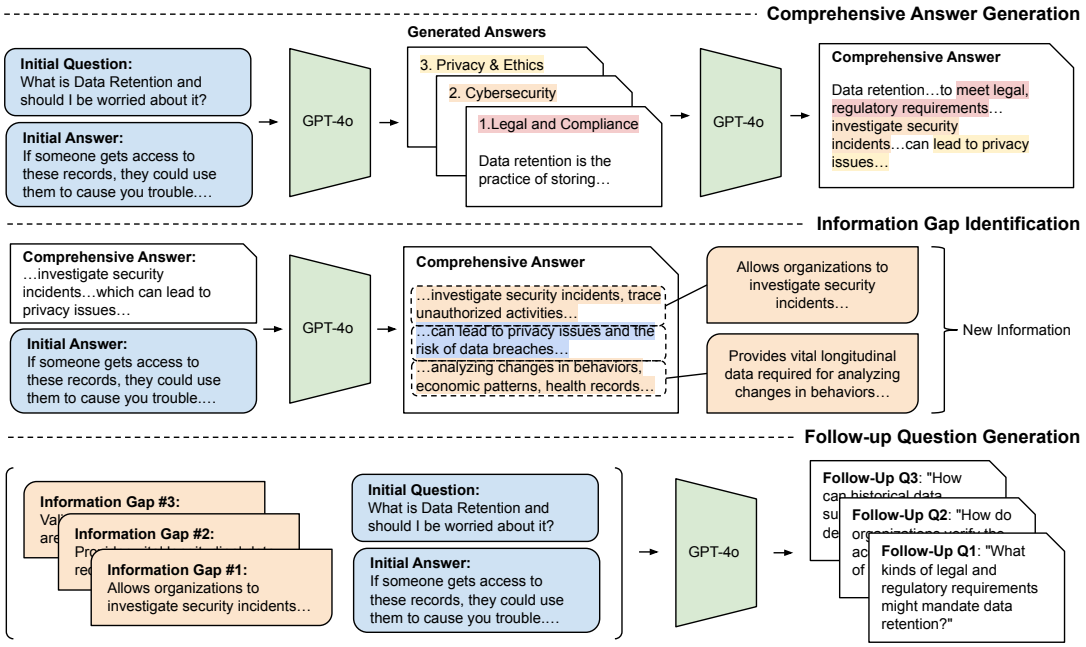

Figure 2: **Data augmentation pipeline.** For a Q&A pair, a comprehensive answer is first generated to the question. By comparing it with the initial answer, information gaps are identified. Finally, multiple follow-up questions are generated targeting those gaps.

Meng et al. (2023) found that models often produce questions that are either repetitive or fail to target unexplored information, thus lacking the cognitive diversity and variability seen in human questioning strategies (Sultan et al., 2020). While follow-up QG has made significant progress, existing approaches largely focus on generating questions directly, using various model architectures and knowledge enhancement techniques (Ge et al., 2023; Liu et al., 2025). Our work, however, takes a novel approach inspired by the human cognitive process that models information gaps and uses them to guide the follow-up question generation.

3 Data Augmentation

Effective FOLLOWUPQG requires models to infer and target gaps between the provided answer and the broader context of a conversation. Following the task definition by Meng et al. (2023): "to generate follow-up questions that seek new information given the initial question and answer", we denote the "initial question" as IQ, "initial answer" as IA, and the "follow-up question" as FQ. We identified critical limitations in the training dataset, including quality issues, which we addressed through dataset cleaning (§3.1). To overcome the small scale (2,790 instances) and low diversity of the dataset, we present a novel data augmentation pipeline (§3.2). Finally, we demonstrate that the augmented dataset retains high quality (§3.3).

3.1 Data Cleaning

The FOLLOWUPQG dataset is limited by its small scale, comprising 3,790 samples: 2,790 for training, 500 for validation, and 500 for testing. Within the 2,790 training instances, there are only 2,651 unique (IQ, IA, FQ) triplets, indicating duplication. Moreover, the dataset consists of 2,648 unique (IQ, IA) pairs, meaning that 99.8% of the (IQ, IA) pairs have only one reference FQ. Training models on this set could thus lead to poor follow-up question diversity. Our further analysis also uncovered data quality issues, likely stemming from automated data collection (see Appendix A). To improve the data quality, we did the following:

- **Deduplication**: We removed 139 duplicate (IQ, IA, FQ) triplets.
- **Reference quality check**: We manually filtered out 84 instances where the reference FQ diverged entirely from the initial question.
- **Sensitive content removal**: We excluded 24 instances involving topics like self-harm or crime, which LLMs are likely to refuse to answer.

The cleaned dataset (2,543 instances) retained broad topic coverage, containing 2,533 unique (IQ, IA) pairs.

3.2 Augmentation Pipeline

As discussed in §3.1, the limited scale of the dataset and the lack of follow-up question diversity hinder the coverage of diverse questioning strategies, restricting model generalization. To address this, we design a GPT-4o-based pipeline that augments the original dataset by generating additional follow-up questions. Our pipeline simulates human reasoning through three interconnected stages: comprehensive answer generation, information gap identification, and follow-up question generation.[2]

Comprehensive answer generation. To identify gaps in the IA, we generate a comprehensive answer (CA) that represents a complete and thorough response to the IQ. As shown in Figure 2, we prompt GPT-4o iteratively to generate answers to IQ that target different perspectives, and synthesize a unified CA. More specifically, GPT-4o was prompted to generate a combination of new and different answers that do not overlap with the other answers, where each answer focuses on a unique aspect not covered in the other generated answers.

Information gap identification. The next step is to identify key concepts or details discussed in the comprehensive answer (CA) but not covered in the initial answer (IA). This is done by prompting GPT-4o. As shown in Figure 2, the initial answer covers the topic of privacy issues but does not cover areas of cyber security (i.e. an information gap). To confirm the validity and reasonableness of the comprehensive answers and identified information gaps, we manually evaluated a random sample of comprehensive answers and ensured that they were accurate and reasonable. Examples of comprehensive answers can be seen in Table 7 and Table 23.

Follow-up question generation. Using the identified information gaps, we prompt GPT-4o to generate follow-up questions that address those gaps while maintaining contextual relevance to the IQ and IA. The generated questions must meet three criteria: be (1) answerable by the CA, (2) unanswerable by the IA, and (3) grounded in terminology and context from the IQ.

After augmentation, each (IQ, IA) pair now includes an average of 10.95 FQs. To preserve the original FOLLOWUPQG format, we automatically remove artifacts such as bullets or numbering from the generated FQs and merge them with the cleaned human-written examples. The resulting dataset comprises 27,874 samples—about 10× the original size—and better reflects the open-ended nature of human questioning, providing models with diverse, explicit signals for addressing information gaps.

3.3 Augmented Data Validation

To assess the quality of the generated follow-up questions, we conducted a human evaluation study on Cloud Connect, using Meng et al. (2023)'s survey. To ensure high-quality annotations, we restricted participation to native English-speaking annotators with a minimum of 1,000 completed annotation tasks and an approval rating exceeding 90%. A randomly sampled subset of 100 (IQ, IA, FQ) triplets was evaluated based on three key criteria: (1) whether the FQ was a valid question,[3] (2) whether any component of the triplet contained sensitive information, and (3) the degree of relatedness between the FQ and the (IQ, IA) pair. The full survey format, including example annotations, is provided in Appendix C. The results show that 94% of the FQs are labeled as valid, 92% as not sensitive, and 91% are related to the original (IQ, IA) pair. Inter-annotator agreement was moderate, with a Cohen's Kappa score of $\kappa = 0.73$ (Cohen, 1960).

4 Experiment Setup

Model Variants. To assess our proposed pipeline and augmented dataset, we fine-tuned BART-large (Lewis et al., 2020) (24 layers, 16 attention heads, hidden size = 1024) on several versions of the FOLLOWUPQG data (Meng et al., 2023), producing three model variants. BART-large is a seq2seq model that conditions on the concatenated IQ and IA to generate an FQ. We chose it as our base model because of its strong performance reported by Meng et al. (2023). As their implementation is not public, we reproduced their training setup (batch = 8, epochs = 10, Adam (Kinga et al., 2015)) and found that the original learning rate of 5e-5 caused instability, so we reduced it to 2e-5; all other hyperparameters remain unchanged.[4]

[2] Please refer to Appendix B for the LLM prompts used for the following stages.

[3] Following Meng et al. (2023), a valid question must be in a question format and ask for meaningful information, including Wh-questions (what/why/where/etc.), open-ended questions, probing questions, etc.

[4] Full hyperparameter details and reproduction results are provided in Appendix D.

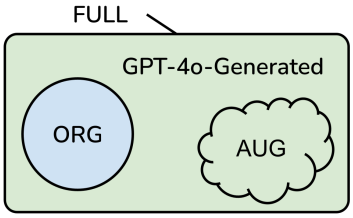

Figure 3: **The ORG/AUG/FULL Dataset.** ORG is the original dataset of ELI5 from the FollowupQG paper. We augment ORG using GPT-4o. FULL combines ORG with all GPT-4o-generated data (about 10× larger). AUG is a random sample of only GPT-4o-generated instances, equal in size to ORG and excluding its data.

We report three variants trained on the FOLLOWUPQG training set as shown in Figure 3: the *ORG* model is trained on the 2,790 original instances from Meng et al. (2023) and serves as our small-model baseline; the *AUG* model, trained on a size-matched random sample of 2,790 GPT-generated questions from our augmentation pipeline (Sec. 3.2) to isolate the effect of data quality; and the *FULL* model, trained on the entire 27,874-instance augmented dataset. All variants share identical hyperparameters and are evaluated on the original FOLLOWUPQG validation and test splits.

Decoding. To generate diverse but contextually relevant follow-up questions, we input the initial question and answer into the model in the following format: `IQ <SEP> IA`, and generate 10 follow-up questions by applying beam search with a beam width of 20, selecting the top 10 candidates. We added a diversity penalty of 10 to encourage unique outputs across the groups and set the temperature to $t = 1.0$ to maintain a balance between diversity and coherence. The maximum length for each generation is set to 1024 tokens. Duplicate generations are removed.

LLM Baselines. Beyond the three BART variants, we evaluated two GPT-4o baselines that represent current large-model performance. The *Teacher* setting executes the full information-gap pipeline from §3.2 on each test instance, returning GPT-4o's comprehensive answer, the identified gaps, and the resulting follow-up questions. The *CoT* setting applies the same chain-of-thought prompt but suppresses all intermediate reasoning, outputting only the follow-up questions. Duplicate generations are removed as in the BART variants.

Model	Total	Ungrammatical	Filtered (%)
ORG	2349	781	33.25
AUG	1895	68	**3.58**
FULL	2061	130	6.31

Table 1: Percentage of filtered-out ungrammatical FQs.

	Diversity			Length (in token)			
Model	Distinct-1 (%)	Distinct-2 (%)	Clusters per FQ	Avg.	Shortest	Longest	Std. Dev.
ORG	66.06	91.12	0.651	14.25	3	111	10.13
AUG	**77.36**	94.41	0.857	13.13	4	**24**	**2.98**
FULL	77.09	**94.85**	**0.866**	13.17	4	73	3.77
Teacher	77.00	95.07	0.869	11.96	3	25	3.19
CoT	80.65	96.34	0.878	16.36	5	30	4.02

Table 2: Automatic evaluation of follow-up question generation without human reference.

5 Results

To thoroughly assess the quality of the generated follow-up questions, we employ both automatic evaluation (§5.1) and human evaluation (§5.2). As a first step for both evaluations, we automatically identify and remove ungrammatical questions based on syntactic parsing (see Appendix E for a complete description of the filtering process). Table 1 shows the percentage of ungrammatical questions that were filtered out for each model. *AUG* (3.58%) and *FULL* (6.31%) produce far fewer ungrammatical FQs compared to *ORG* (33.25%), demonstrating their ability to generate more well-formed outputs. We focus the rest of our evaluation on the grammatical questions retained after the filtering.

5.1 Automatic Evaluation

Diversity. We assess the diversity of each set of FQs at the (IQ, IA) level and average the scores across the dataset. First, we report Distinct-n (Li et al., 2016), which measures the average distinct n-gram in the FQs associated with each (IQ, IA) pair. Table 2 shows that *AUG* and *FULL* achieve comparable Distinct-1/2 scores, both exceeding *ORG*. Moreover, the *AUG* Distinct-1/2 scores are comparable to those of the GPT-4o *Teacher* baseline and only slightly lower than those of the advanced *CoT* model.

We also compute a sentence-level diversity score. We embed the FQs using `all-mpnet-base-v2` (Reimers and Gurevych, 2019) and apply agglomerative clustering at a distance threshold of 1.0, normalizing the number of clusters by the num-

Model	BERT	Sent. Sim.	B1	B2	B3	B4	METEOR	ROUGE
ORG	86.28	76.74	40.34	8.49	2.54	1.15	17.57	19.09
AUG	85.72	71.91	32.54	4.02	0.69	0.17	13.84	11.07
FULL	85.74	72.42	32.95	4.19	0.85	0.25	14.16	11.79

Table 3: Automatic evaluation of follow-up question generation with human reference. *ORG* (baseline) performs slightly better.

ber of generated follow-up questions. A score of 1 denotes maximum diversity whereas lower values indicate that questions collapse into the same cluster. Again, Table 2 confirms the trend that our augmentation substantially improves the diversity. Moreover, for both metrics, *AUG* is statistically indistinguishable from the GPT-4o *Teacher* and *CoT* baselines, showing that our pipeline elevates small models to LLM-level diversity on the FOLLOWUPQG task.

Average question length. We report the average question length in terms of the number of tokens. We hypothesize that shorter questions are generally more readable. Table 2 lists the average length, shortest and longest follow-ups, and standard deviation (SD). The *ORG* model shows the greatest variation in question length (SD = 10.13). Notably, its longest follow-up (111 tokens) far exceeds *FULL* (73) and *AUG* (24). In contrast, *AUG* is the most consistent (SD = 2.98; max = 24), with *FULL* close behind (SD = 3.77)

Qualitatively examining the generated follow-up questions, we find that *AUG* and *FULL* generally produce concise, well-formed queries, while *ORG* sometimes generates very short, vague prompts (e.g., "So it's cultural?"). Meanwhile, the longer questions from *ORG* and *FULL* often include extraneous conversational filler. Overall, *AUG* maintains structured, concise outputs for follow-up questions, whereas *FULL* and *ORG* exhibit greater variability, occasionally producing overly long or conversational phrasing. More examples are provided in §6.1 and Appendix G.

Similarity to the references. To compare our results with those obtained by Meng et al. (2023), we perform identical automatic evaluations. We measure lexical overlap with BLEU-1-4 (Papineni et al., 2002), METEOR (Lavie and Agarwal, 2007), and ROUGE-L (Lin, 2004), and semantic similarity with BERTScore (Zhang* et al., 2020) and an embedding-based cosine score computed with `all-mpnet-base-v2` (Reimers and Gurevych, 2019), following Meng et al. (2023). For each metric, we compute the highest score across all generated follow-ups with the human reference and report the average for the entire dataset. Table 3 shows a consistent advantage for *ORG*. This is expected, as both its training data and the test set come from the original FOLLOWUPQG distribution. The lower BLEU scores for *AUG* and *FULL* reflect a common issue in open-ended QG: lexically diverse yet valid questions are under-rewarded by n-gram metrics (Pan et al., 2021). In contrast, the gap between *FULL* and *AUG* is much smaller on BERTScore and embedding similarity, which focus on semantic alignment and discount stylistic differences. Consequently, we turn to human evaluation to capture diversity and nuanced informativeness that automatic metrics may overlook.

5.2 Human Evaluation

We conducted a human evaluation to assess the quality of generated FQs on four criteria: validity, complexity (the level of reasoning required), relevance, and informativeness (Details listed in Table 4). We randomly sampled 30 (IQ, IA) pairs from the FOLLOWUPQG testing set and evaluated the generated FQs from all five models.

Human evaluation was conducted on Cloud Connect. To ensure high-quality annotations, we restricted participation to native English-speaking annotators with a minimum of 1,000 completed annotation tasks and an approval rating exceeding 90%. Annotators interacted with a structured evaluation interface (see Appendix H). Each task presented an initial question, its corresponding answer, and a generated FQ candidate. Annotators first assessed whether the FQ was valid. If deemed invalid, they proceeded directly to the next task. Oth-

	Question	Numeric Scale
Validity	Is the FQ question a valid question?	yes (1) / no (0)
	Does the FQ contain any of the following errors?	contains errors: • redundant • repetitive • wrong semantic collocation (1) / no errors (0)
Complexity	Does generating the FQ require reasoning?	complex (3) / moderate (2) / minimal (1) / no (0)
Relevance	How relevant is FQ to the initial question and answer?	strongly (3) / relevant (2) / slightly (1) / not (0)
Informativeness	Does the FQ elicit new information?	a lot (3) / some (2) / little (1) / no (0)

Table 4: The aspects evaluated in the human evaluation with respect to the follow-up question (FQ).

	ORG		*AUG*		*FULL*	
	Mean	**Variance**	**Mean**	**Variance**	**Mean**	**Variance**
Validity	0.7324	0.1964	**0.9065***	0.0849	0.8743	0.1102
Complexity	0.9274	1.0129	**1.4798***	0.9441	1.4454	0.7025
Relevance	1.6236	1.4716	**2.0935***	1.0225	1.7377	1.0269
Informativeness	0.7755	0.9563	**1.4517***	1.1297	1.2951	0.8223

Table 5: Human evaluation scores for each aspect comparing *ORG*, *AUG* and *FULL*. Best results are in bold; only statistically significant results are marked with an asterisk.

erwise, they answered four additional evaluation questions, as detailed in Table 4 (See Appendix F for the complete annotation guidelines). Each task was annotated by 3 annotators, yielding substantial inter-annotator agreement, with an average Cohen's Kappa of $\kappa = 0.77$ (Cohen, 1960).

Table 5 reports the mean and variance of each evaluation criterion for *ORG*, *AUG*, and *FULL*. Overall, *AUG* achieved the best results across all criteria, with a statistically significant difference from the other models (tested with a one-way ANOVA). Over 90% of the FQs generated by *AUG* were considered valid, and these questions were rated as relevant, somewhat informative, and minimally to moderately complex. *FULL* closely follows across aspects, while *ORG* lags behind. The only aspect on which *ORG* closely follows *FULL* is relevance, aligning with the findings of Meng et al. (2023) that current models perform well in maintaining relevance. In sum, the results clearly prefer the questions generated by *AUG*, which excel in validity, complexity, relevance, and informativeness—qualities essential for meaningful follow-up questions.

To gauge how effectively our information-gap pipeline distills LLM knowledge into smaller models, we repeated human evaluation on the GPT-4o *Teacher* and *CoT* baselines described in §4. Table 6 shows that both LLM baselines post slightly higher mean scores than the distilled *AUG* model, yet the difference is statistically significant only for the RELEVANCE metric. On validity, complexity, and informativeness, *AUG*—a BART-large model trained on a random, size-matched subset of our augmented data—remains statistically indistinguishable from the much larger *Teacher* and *CoT* models. These results underscore the strength of our pipeline: by contrasting initial answers with LLM-generated comprehensive answers, it simulates the human information-seeking process and produces synthetic follow-up questions rich in diversity and informativeness. Fine-tuning on this augmented data enables small, locally deployable models to reach teacher-level quality at a fraction of the inference cost, thereby making high-performance FOLLOWUPQG feasible on resource-constrained hardware.

	Teacher		*AUG*		*CoT*	
	Mean	**Variance**	**Mean**	**Variance**	**Mean**	**Variance**
Validity	0.9457	0.0514	0.9065	0.0849	**0.9647**	0.0342
Complexity	**1.6227**	0.6604	1.4798	0.9441	1.5725	0.7575
Relevance	2.1240	0.7928	2.0935	1.0225	**2.4863***	0.5500
Informativeness	1.6744	0.6502	1.4517	1.1297	**1.6784**	0.8095

Table 6: Human evaluation scores for each aspect comparing (i) *AUG* with *Teacher* and (ii) *AUG* with *COT* GPT-4o. Best results are in bold; only statistically significant results are marked with an asterisk.

The comparative results across models reveal key insights into the role of data quality versus quantity in the task of FOLLOWUPQG. Notably, *AUG*, trained on the same number of instances as *ORG* but consisting solely of GPT-4o-generated, high-quality, reasoning-heavy questions, consistently outperforms both *ORG* and *FULL* across most metrics, yielding greater validity, complexity, relevance, and informativeness. This indicates that data quality is more critical than dataset size. Despite consisting of ten times more training data, *FULL* failed to surpass *AUG*, likely because remaining lower-quality or low-reasoning examples from the original dataset dilute learning. These findings challenge the assumption that larger datasets automatically improve performance and underscore the value of targeted augmentation with strict quality control. Future work should explore strategies to scale data while maintaining rigorous curation to further advance follow-up question generation.

6 Analysis

To further understand the strengths and limitations of our method, we present a qualitative comparison of follow-up questions generated by all models for the same (IQ, IA) pair (§6.1), as well as an analysis of the expected information gain from the generated follow-up questions (§6.2).

6.1 Qualitative Analysis

In Table 7, we compare follow-up questions generated by the *ORG*, *AUG*, and *FULL* models for a given (IQ, IA) pair. Questions from *ORG* are often redundant—e.g., "What is a heuristic?"—or

IQ	How would you explain the concept of a heuristic?
IA	It took me a long time to understand heuristics and *I still don't know if I can ELI5 after a handful of bachelor and graduate degrees.* A heuristic is like a clue or a key to help you solve a problem or remember something. That's the best I got.
CA	A heuristic is a practical approach or mental strategy used for problem-solving and decision-making, particularly when dealing with complex or incomplete information. It serves as a cognitive shortcut, allowing individuals to reach quick and satisfactory solutions without the need for exhaustive analysis or perfect precision. By focusing on the most relevant aspects of a problem and often relying on past experiences, common sense, or general rules of thumb, heuristics simplify the decision-making process and reduce cognitive load. While heuristic methods are not guaranteed to provide the best solution, they are typically sufficient for achieving an immediate goal and are useful in many everyday situations where speed and practicality are prioritized over exhaustive accuracy.
FQ by *ORG*	– *How do you know if you can ELI5?* – *What is a heuristic?* – Thanks for the answer. So, what I am trying to understand is, *what is ELI5?* – *That's the best I got. I'm not a psychologist, but my brain picks out heuristics when I'm stressed and it usually works. I don't know if I can ELI5 after a handful of bachelor and graduate degrees. But I'm still working on it.*
FQ by *AUG*	– What are some examples of cognitive strategies that rely on heuristics? – How does the concept of a heuristic relate to the process of problem-solving? – What role does the hippocampus play in heuristic formation?
FQ by *FULL*	– What are some examples of heuristics used in decision-making? – What role do heuristics play in problem-solving? – How do heuristics *help in segmenting and segmenting information* to specific tasks or objectives? – *I'm not sure I can ELI5 after a handful of bachelor and graduate degrees. I'm not even sure if I can understand ELI4. I know that it's a heuristic. But I don't know if I understand ELII5. Like, I know what a heuristics are. And I know how to use a heymn to solve problems. So I'm asking if you can ELII4?*

Table 7: Example of follow-up question generated by three model variants, with comprehensive answers (ID 3182).

tangential, such as "How do you know if you can ELI5?" to the original responder that mentioned they didn't know if they could explain it to a 5-year-old (ELI5), thus drifting away from the target concept of heuristics. While the *FULL* model yields a wider range of relevant questions and excels in diversity, it occasionally produces tangential or wordy phrasing, for instance, "How do heuristics help in segmenting and segmenting information for specific tasks?", which hurts clarity. By contrast, *AUG* strikes the best balance of informativeness and diversity, offering focused, insightful questions like "What are some examples of cognitive strategies that rely on heuristics?" and "How does the concept of a heuristic relate to the process of problem-solving?". Additional examples can be found in Appendix G.

6.2 Quantifying Information Gain

In §5.2 we asked annotators to rate the informativeness of each follow-up question. We now introduce an automated alternative that requires no human raters, leveraging the GPT-4o "comprehensive answers" (CA; see definition in §3.2). We treat each CA as a proxy for the full body of information relevant to its (IQ, IA) context. An FQ is informative if it (i) cannot be answered from the IA alone—otherwise it adds no new information—and (ii) can be answered from the CA—otherwise it is likely irrelevant. Guided by this rule, we prompt GPT-4o to judge the answerability of every model-generated FQ against both the IA and the corresponding CA.

Model	Human-INF	GPT-INF-All (%)	GPT-INF-Sel (%)
ORG	0.7755	25.17	23.29
AUG	**1.4517**	**36.19**	**35.91**
FULL	1.2951	34.90	32.20

Table 8: Comparison of human-annotated informativeness scores and GPT-evaluated informative percentage across models.

Table 8 corroborates the human evaluation of informativeness: *AUG* produces the largest share of informative questions (36 %), followed by *FULL* (35 %) and *ORG* (25 %). Comparing the GPT-4o labels with human-annotated informativeness scores (§5.1) further validates the automatic method: annotators assigned higher mean scores to FQs the model classified as informative (1.29) than to those it did not (1.07). A two-sample t-test ($p = 0.0011$) confirms the statistical significance, although the effect size is small (Cohen's $d = 0.215$) (Cohen, 2013).

7 Conclusion

In this work, we proposed a novel approach to enhance the diversity and informativeness of follow-up questions by explicitly modeling information gaps via an LLM-generated comprehensive answer. We augmented the original FOLLOWUPQG dataset with GPT-4o and distilled this data into a small, locally deployable BART-large model. Experiments show that our pipeline enables the small model to outperform peer small-model baselines and to perform comparably to GPT-4o baseline models in terms of validity, complexity, relevance, and informativeness—all at a fraction of the inference cost. These results suggest that targeted, high-quality augmentation can be more impactful than merely increasing dataset size. They also demonstrate that our method offers a practical approach for improving information-seeking dialogues—by reducing ambiguities and enhancing LLM responses—even on systems with limited computational resources.

Future work could explore ways to increase follow-up-question diversity while reducing redundancy, and to extend the pipeline to downstream tasks involving multi-turn dialogue. We also encourage research on stronger automated metrics for evaluating question quality, given the high cost of human annotation and the limitations of current automatic measures.

Limitations

We acknowledge several limitations in our work. First, while our CA-based pipeline is effective in knowledge-driven contexts, its applicability to non-knowledge-based conversations, such as opinion-based questions (e.g., "What would you do in such a scenario?"), remains unclear, as the subjective judgment required in these conversations can be difficult for a generated CA to capture. Additionally, although our pipeline prioritizes informativeness, follow-up questions do not always need to introduce new information (Kurkul and Corriveau, 2018)—for example, requests for simpler explanations (e.g., "Can you explain this in an easier-to-understand way?").

Our work also calls for several future works and expansions. For example, our pipeline can be tested and evaluated on languages besides English, including low-resource languages. Moreover, given the pragmatic applicability of this pipeline and its focus on resource-efficiency, it would be pertinent to evaluate the compute-cost tradeoffs to help users make informed decisions. Lastly, our pipeline's performance can be evaluated on different combinations of Teacher-Student models as well. In the future, we hope to extend this method to support various types of follow-up questions and integrate it into downstream dialogue-based applications.

Ethical Considerations

All annotators involved in the human evaluation for this research were fairly compensated, with payment rates exceeding the local minimum wage to ensure equitable remuneration for their time and effort. Prior to recruiting annotators, ethical approval was obtained from the research ethics board at the authors' institution, ensuring that the human evaluation process adhered to ethical guidelines and that no harm was caused to any individual involved. Additionally, the FOLLOWUPQG dataset used in this work is publicly available, and we also released the new data created in this work, including the augmented data and generated comprehensive answers, to promote transparency and reproducibility in future work.

Acknowledgments

This work was funded, in part, by the Vector Institute, Canada CIFAR AI Chairs program, Accelerate Foundation Models Research Program Award from Microsoft, and an NSERC discovery grant.

References

Husam Ali, Yllias Chali, and Sadid A. Hasan. 2010. Automatic question generation from sentences. In *Actes de la 17e conférence sur le Traitement Automatique des Langues Naturelles. Articles courts*, pages 213–218, Montréal, Canada. ATALA.

Banu Arslan, Gokhan Eyupoglu, Semih Korkut, Kenan Ahmet Turkdogan, and Ertugrul Altinbilek. 2024. The accuracy of ai-assisted chatbots on the annual assessment test for emergency medicine residents. *Journal of Medicine, Surgery, and Public Health*, 3:100070.

Pooja Rao S B, Manish Agnihotri, and Dinesh Babu Jayagopi. 2020. Automatic follow-up question generation for asynchronous interviews. In *Proceedings of the Workshop on Intelligent Information Processing and Natural Language Generation*, pages 10–20, Santiago de Compostela, Spain. Association for Computational Lingustics.

Yunmo Chen, Tongfei Chen, Harsh Jhamtani, Patrick Xia, Richard Shin, Jason Eisner, and Benjamin Van Durme. 2024. Learning to retrieve iteratively for in-context learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7156–7168, Miami, Florida, USA. Association for Computational Linguistics.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Jacob Cohen. 2013. *Statistical power analysis for the behavioral sciences*. routledge.

Xinya Du, Junru Shao, and Claire Cardie. 2017. Learning to ask: Neural question generation for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1342–1352, Vancouver, Canada. Association for Computational Linguistics.

Yubin Ge, Ziang Xiao, Jana Diesner, Heng Ji, Karrie Karahalios, and Hari Sundaram. 2023. What should I ask: A knowledge-driven approach for follow-up questions generation in conversational surveys. In *Proceedings of the 37th Pacific Asia Conference on Language, Information and Computation*, pages 113–124, Hong Kong, China. Association for Computational Linguistics.

Bilal Ghanem, Lauren Lutz Coleman, Julia Rivard Dexter, Spencer von der Ohe, and Alona Fyshe. 2022. Question generation for reading comprehension assessment by modeling how and what to ask. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2131–2146, Dublin, Ireland. Association for Computational Linguistics.

Shrey Gupta, Anmol Agarwal, Manas Gaur, Kaushik Roy, Vignesh Narayanan, Ponnurangam Kumaraguru, and Amit Sheth. 2022. Learning to automate follow-up question generation using process knowledge for depression triage on Reddit posts. In *Proceedings of the Eighth Workshop on Computational Linguistics and Clinical Psychology*, pages 137–147, Seattle, USA. Association for Computational Linguistics.

D Kinga, Jimmy Ba Adam, et al. 2015. A method for stochastic optimization. In *International conference on learning representations (ICLR)*, volume 5, page 6. San Diego, California;.

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

Vineet Kumar and Sachindra Joshi. 2017. Incomplete follow-up question resolution using retrieval-based sequence to sequence learning. In *Proceedings of the 40th International ACM Sigir Conference on Research and Development in Information Retrieval*, pages 705–714.

Katelyn E Kurkul and Kathleen H Corriveau. 2018. Question, explanation, follow-up: A mechanism for learning from others? *Child Development*, 89(1):280–294.

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. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 102–111, Seattle, United States. Association for Computational Linguistics.

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, Prague, Czech Republic. Association for Computational Linguistics.

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 *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119, San Diego, California. Association for Computational Linguistics.

Jiwei Li, Alexander H Miller, Sumit Chopra, Marc'Aurelio Ranzato, and Jason Weston. 2017. Learning through dialogue interactions by asking questions. In *5th International Conference on Learning Representations, ICLR 2017*.

Shuyue Stella Li, Vidhisha Balachandran, Shangbin Feng, Jonathan S Ilgen, Emma Pierson, Pang Wei Koh, and Yulia Tsvetkov. 2024. Mediq: Question-asking llms and a benchmark for reliable interactive clinical reasoning. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

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

Jianyu Liu, Yi Huang, Sheng Bi, Junlan Feng, and Guilin Qi. 2025. From superficial to deep: Integrating external knowledge for follow-up question generation using knowledge graph and LLM. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 828–840, Abu Dhabi, UAE. Association for Computational Linguistics.

Haohao Luo, Yang Deng, Ying Shen, See-Kiong Ng, and Tat-Seng Chua. 2024. Chain-of-exemplar: Enhancing distractor generation for multimodal educational question generation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7978–7993, Bangkok, Thailand. Association for Computational Linguistics.

Bodhisattwa Prasad Majumder, Sudha Rao, Michel Galley, and Julian McAuley. 2021. Ask what's missing and what's useful: Improving clarification question generation using global knowledge. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4300–4312, Online. Association for Computational Linguistics.

Davide Mazzaccara, Alberto Testoni, and Raffaella Bernardi. 2024. Learning to ask informative questions: Enhancing LLMs with preference optimization and expected information gain. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 5064–5074, Miami, Florida, USA. Association for Computational Linguistics.

Yan Meng, Liangming Pan, Yixin Cao, and Min-Yen Kan. 2023. FollowupQG: Towards information-seeking follow-up question generation. In *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, Nusa Dua, Bali. Association for Computational Linguistics.

Liangming Pan, Wenhu Chen, Wenhan Xiong, Min-Yen Kan, and William Yang Wang. 2021. Zero-shot fact verification by claim generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 476–483, Online. Association for Computational Linguistics.

Liangming Pan, Yuxi Xie, Yansong Feng, Tat-Seng Chua, and Min-Yen Kan. 2020. Semantic graphs for generating deep questions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1463–1475, Online. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Peng Qi, Yuhao Zhang, and Christopher D. Manning. 2020. Stay hungry, stay focused: Generating informative and specific questions in information-seeking conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 25–40, Online. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. 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)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Sarvesh Soni and Kirk Roberts. 2019. A paraphrase generation system for EHR question answering. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 20–29, Florence, Italy. Association for Computational Linguistics.

Md Arafat Sultan, Shubham Chandel, Ramón Fernandez Astudillo, and Vittorio Castelli. 2020. On the importance of diversity in question generation for QA. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5651–5656, Online. Association for Computational Linguistics.

Masaki Uto, Yuto Tomikawa, and Ayaka Suzuki. 2023. Difficulty-controllable neural question generation for reading comprehension using item response theory. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 119–129, Toronto, Canada. Association for Computational Linguistics.

Yongzhen Wang, Kaisong Song, Lidong Bing, and Xiaozhong Liu. 2021. Harvest shopping advice: Neural question generation from multiple information sources in e-commerce. *Neurocomputing*, 433:252–262.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Yao Zhao, Xiaochuan Ni, Yuanyuan Ding, and Qifa Ke. 2018. Paragraph-level neural question generation with maxout pointer and gated self-attention networks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3901–3910, Brussels, Belgium. Association for Computational Linguistics.

A Problematic and Unrelated Instance Examples

Tables 9 and 10 show examples of a problematic instance and an unrelated instance from FOLLOWUPQG training set, respectively.

Initial Question	How exactly do you die if you touch electric wires? Do you burn? Suffocate? Heart stop?
Initial Answer	If the current going thru you is around 1 amp, the shock may give you a heart attack. So it's like the hospital shock paddles, only it kills you. If the current is like 10 amps, it probably won't give you a heart attack, and you may survive if your organs aren't fully cooked. Which they definitely could be if the current isn't immediately stopped. And cooked kidneys don't function that good. So to summarize, either heart attack or cooked organs/burned, depending on the current. Edit: To clarify, it takes less than 1 amp to stop your heart, and the current also has to go thru your heart. I just used 1 amp and 10 amps as sample values.

Table 9: Problematic Instance 1696: Self-Harm-Related Content.

Initial Question	How its possible for the light to travel constantly with out stop slowing down from the most old galaxies?
Initial Answer	Okay, an NFT is a NON-FUNGIBLE TOKEN. Let's take the Mona Lisa for example. [...] So what you've done is bought the off-brand version (online) instead of the real thing. Also, the guy you paid doesn't really own the Mona Lisa. The TLDR is that you're buying an online version of a product and it says you own it, but you don't own the real thing.
Follow-up Question	Thank you for this! My question is, what's the appeal? Is it just rich-guy-bragging-rights? [...] With, say, a baseball card you can display it and know that it's a rare item, and someday resell it at (you hope) a profit. Are some of these people just speculating and hoping the 'value' goes up?

Table 10: Unrelated Instance 641.

B LLM Prompts

	Prompts
Initial Answer	"Generate an answer focused on a single perspective only, without any conversational fillers. Do not repeat the question in the answer."
Next Answer	"Please provide a new answer focused on a different perspective, ensuring no overlap with previous answers. Focus on unique aspects or insights not covered earlier, and provide the answer only without any conversational fillers. Do not repeat the question in the answer."
Comprehensive Answer	"Synthesize the following answers into a single, comprehensive response. Integrate the key points and insights from each answer, ensuring a cohesive and well-rounded explanation. The final answer should be thorough and address multiple aspects of the question without unnecessary repetition."

Table 11: Comprehensive Answer Generation Prompts: GPT-4o first generates an answer from a single perspective, then iteratively provides non-overlapping answers from different perspectives, which are finally synthesized into a unified response.

Information Gap Identification & Follow-up Question Generation
"Generate all possible follow-up questions as candidates. These follow-up questions must be related to the original question, but must not be rephrases of the original question. These follow-up questions should be answerable by the complete answer. These follow-up questions should not be answered, covered, or detailed by the original answer, but must target terminologies mentioned in the original answer. Separate each follow-up question with '<sep>'."

Table 12: Follow-up Question Generation Prompt.

C Augmented Data - Human Annotation Guideline

Table 13 presents the job description and annotation questions for our human annotation task.

Job Description
Welcome, and thank you for participating in this text evaluation task! In this job, you'll be helping us verify the quality of follow-up questions generated by GPT. For each task, we will provide you with a pair consisting of a question and answer collected from Reddit's "Explain Like I'm Five" (ELI5) forum. You will be asked to evaluate the quality of the follow-up question generated by GPT. These questions and answers aim to provide layperson-friendly explanations for real-life queries. Here is an example of one task sample: Each task may contain noise, such as invalid follow-up questions, sensitive information, or questions unrelated to the original question or answer. Your role is to help us identify these noisy samples. For each task, you will be shown one triple (question, answer, follow-up question). Carefully review each component and answer the following questions based on your judgment:
Q1: Do you think the follow-up question is a valid question? **A.** Yes **B.** No
Q2: Does the initial question, answer, or follow-up question contain 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

Table 13: Task description and evaluation questions used for human annotation of augmented data.

C.1 Valid/Invalid Question 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 a question format and ask meaningful information, including Wh-questions (what/why/where/etc.), open-ended questions, probing questions and etc. Invalid questions like "10000 meters? really?", are often used in conversational speech to express feelings instead of asking for new information. Table 14 contains examples of valid and invalid follow-up questions.

Initial Question: Why is the sea calm in the mornings?	
Initial 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.	
Valid Follow-up	**Invalid Follow-up**
Why are winds always weak in the morning and very strong during the day?	Isn't it common sense that the sea is calmer in the morning?
Reason	**Reason**
The follow-up question is a "Why" question, asking specific reasons about the change of the winds. Therefore, it is a valid question.	This is a rhetorical question because it does not genuinely seek new information. It implies that the answer is obvious and does not contribute to the discussion.

Table 14: Examples of valid and invalid follow-up questions. For the given initial question and answer, the left column presents a valid follow-up question, while the right column features an invalid one, each accompanied by corresponding reasons below.

C.2 Inappropriate Question Guideline

Examples of racist comments include: "It's credit to your race," "Black people will not understand." Examples of hate speech include: "He should go back to where he comes from," "All Mexicans are rapists." Examples of offensive or rude comments include: "Women are not suitable for working in the IT field," "Gay will never understand." Table 15 contains an example of an inappropriate follow-up question.

Initial Question: Why do people develop eating disorders?	
Initial Answer: Eating disorders are complex mental health conditions influenced by a combination of genetic, psychological, environmental, and social factors. While societal beauty standards and pressures can contribute, eating disorders are not simply about wanting to be thin. Conditions like anorexia, bulimia, and binge-eating disorder involve intricate relationships between self-image, emotional regulation, and biological predispositions. Many individuals with eating disorders struggle with anxiety, depression, or trauma, which can further complicate their relationship with food.	
Inappropriate Follow-up	**Reason**
Why don't people with eating disorders just stop starving themselves and eat normally like everyone else?	This question is dismissive. The phrasing is insensitive and could be harmful to individuals struggling with these conditions.

Table 15: Example of an inappropriate follow-up question for the given initial question and answer, accompanied by corresponding reasons below.

C.3 Relevance Question Guideline

- **Strongly Related**: The follow-up question asks for specific definitions, particular reasons, or meanings directly from the original question and answer.
- **Related**: The follow-up question primarily seeks information from the original question or answer but also brings in additional, new information.
- **Slightly Related**: The follow-up question mainly addresses other cases but has some relevance to the original question or answer.
- **Not Related**: The follow-up question does not relate to the original question or answer.

Table 16 contains follow-up questions with various levels of relevance.

Initial Question: Why do airplanes leave white trails in the sky?	
Initial Answer: Those white trails are called contrails, short for condensation trails. They form when hot exhaust from the airplane's engines mixes with the cold air in the upper atmosphere. The water vapor in the exhaust condenses and freezes into tiny ice crystals, creating the white streaks you see in the sky. The persistence of these trails depends on humidity levels; if the air is dry, the contrail dissipates quickly, but if the air is humid, the contrail can linger for a long time.	
Strongly Related Question Example	**Related Follow-up Question Example**
Why do some contrails last longer than others?	Do contrails have any impact on the environment?
Reason	**Reason**
The follow-up question directly builds on the information provided in the answer, specifically regarding the persistence of contrails. Since the answer already mentions humidity as a factor, this question seeks further clarification, making it strongly related.	This follow-up question extends the topic of contrails by asking about their environmental impact. While the original answer does not discuss environmental effects, the question is still relevant because it builds on the phenomenon explained. Thus, it is considered related.
Slightly Related Question Example	**Not Related Follow-up Question Example**
Why do some airplanes make more noise than others?	What causes volcanoes to erupt?
Reason	**Reason**
The follow-up question is about airplanes, which is the general topic of the original question, but it shifts the focus from contrails to noise. While both topics are related to aviation, the connection between them is weak, making the question only slightly related.	The follow-up question introduces a completely unrelated topic (volcanoes) that has no connection to airplanes, contrails, or atmospheric conditions. Since it does not build on the original question or answer in any way, it is considered not related.

Table 16: Examples of follow-up questions' relevance for the given initial question and answer, accompanied by corresponding reasons below.

D Baseline Reproduce

To establish a baseline, we attempted to reproduce the results of Meng et al. (2023) using the reported parameters, as the original implementation was unavailable. We use BART-large, consisting of 24 layers, 16 attention heads, and a hidden dimension of 1024. The initial learning rate (5e-5) led to training instability, which we mitigated by adjusting it to 2e-5 while keeping other hyperparameters unchanged (batch size: 8, epochs: 10, optimizer: Adam (Kinga et al., 2015)). The training was conducted on an NVIDIA Tesla V100 GPU with 32GB of memory, taking approximately 6 hours per run. We set the random seed to 42. After multiple runs, our reproduced model yielded similar overall performance but with some variation-certain metrics improved while others slightly declined (see Table 17). This model served as the baseline for subsequent experiments.

E Filtering Ungrammatical Follow-Up Questions: Examples and Pseudocode

```
def is_valid_question(FQ, IQ, IA):
    return (
        # confirm that FQ ends with '?'
        contains_question_mark(FQ) and
        # use dependency parsing to
            check for
        # WH-Questions, Yes/No Questions
        # and Rhetorical Questions,
        is_question_dependency_parsing(
            FQ) and
        # checks for meaningless/invalid
             words
        # e.g. <QUS>, <EQT>
        not contains_invalid_word(FQ)
            and
        # checks for n-gram substring
            that is
        # duplicated in both FQ and (IQ/
            IA)
        not contains_duplicate_words(IQ,
             IA, FQ)
    )
```

	B1	B2	B3	B4	METEOR	ROUGE	BERT	Sentence Similarity
Reported ORG	17.22	7.11	3.89	2.61	8.00	13.35	-	-
Reproduced ORG	21.88	3.27	1.62	1.13	14.76	12.39	84.41	70.73

Table 17: Comparison of Results from Paper Report and Our Reproduction.

Error Type	**Example**
Improper delimiter	*> Constant reflux can cause cancer of the sphincter.<QUS> Is this true for people with GERD?*
Non-question statement	*I think that's the key. I'm not sure what you mean by "3 components". The rotation is defined by the third component. The fourth component is used to determine the degree of rotation. So, if you have a 3d vector, you can rotate it in any direction. But if you want to rotate it out to a 4d vector you need to use a fourth component.*
Excessive context duplication	*Now try doing it for tomorrow's price chart. When you look at the historical chart of a company's price you'll be able to pick good buying and selling points 100% of the time. Now try to do it for today's price* Is that impossible?

Table 18: Follow-up Question Error Types and Examples.

F Model Evaluation - Human Annotation Guideline

Table 19 presents the job description and annotation questions for our human annotation task.

Job Description
In this job, you'll be helping us evaluate the quality of follow-up questions generated by a language model called BART. For each task, we will provide you with a pair consisting of a question and answer collected from Reddit's "Explain Like I'm Five" (ELI5) forum. You will be asked to evaluate the quality of the follow-up question generated by BART. These questions and answers aim to provide layperson-friendly explanations for real-life queries. Our data may contain noise, such as invalid follow-up questions, errors, lack of reasoning, or follow-up questions unrelated to the original question or answer. Your role is to help us identify these noisy samples. For each task, you will be shown one triple (question, answer, follow-up question). Carefully review each component and answer the following questions based on your judgment:
Q1: Do you think the follow-up question is a valid question? **A.** Yes **B.** No
Q2: How relevant is the follow-up question to the original question and answer? **A.** Strongly Related **B.** Related **C.** Slightly Related **D.** Not Related
Q3: Does the follow-up question contain any of the following errors? **A.** No Errors **B.** Redundant **C.** Repetitive **D.** Wrong Semantic Collocation **E.** Other Errors
Q4: Does generating this follow-up question require reasoning? **A.** Requires complex amount of reasoning **B.** Requires moderate amount of reasoning **C.** Requires minimal amount of reasoning **D.** Does not require any reasoning
Q5: Does the follow-up question contain new information for the audience? **A.** Introduces a lot of new information **B.** Introduces some new information **C.** Introduces little new information **D.** Does not introduce any new information

Table 19: Task description and evaluation questions used for BART follow-up question evaluation.

F.1 Error Question Guideline

Does the follow-up question contain any of the following errors?

Identify any language issues in the follow-up question.

- **No Errors** – The follow-up question is appropriate and adds value.
- **Redundant** – The follow-up does not introduce any new information.
- **Repetitive** – The follow-up question closely mirrors the original question.
- **Wrong Semantic Collocation** – The question contains unnatural or incorrect phrasing.
- **Other Errors** – Any issues that do not fit the categories above.

Table 20 contains examples of follow-up questions with various error status.

Initial Question: How do vaccines work?	
Initial Answer: Vaccines work by training your immune system to recognize and fight specific germs. They contain harmless parts of the germ (or something similar) so that your body can learn to defend against it. This way, if you ever encounter the actual germ, your immune system can respond quickly and prevent illness.	
No Errors Example	**Redundant Example**
How does a vaccine train the immune system?	Are vaccines used to help the immune system recognize germs?
Reason	**Reason**
The follow-up question is well-formed, relevant, and adds value by diving deeper into a key concept from the original answer. It does not repeat information unnecessarily or contain any language errors.	The follow-up question is redundant because it merely restates information already provided in the initial answer without adding depth or prompting new discussion.
Repetitive Example	**Wrong Semantic Collocation Example**
What do vaccines do?	Do vaccines memorize diseases?
Reason	**Reason**
This follow-up question is nearly identical to the original question, simply reworded. Since it does not introduce new angles or expand on any details, it is considered repetitive.	The phrase"vaccines memorize diseases" is unnatural and incorrect in this context. A better way to phrase the question would be: "Do vaccines help the immune system remember diseases?"

Table 20: Examples of follow-up questions' error status for the given initial question and answer, accompanied by corresponding reasonings below.

F.2 Reasoning Question Guideline

Evaluate the level of reasoning needed to generate the follow-up question.

- **Complex reasoning** involves synthesizing multiple ideas or deeply analyzing information.
- **Moderate reasoning** requires interpreting the given content or slightly extending the discussion.
- **Minimal reasoning** involves simple comprehension or directly rephrasing information.
- **No reasoning** applies to questions that are direct repetitions or restatements without any thought process.

Table 21 contains examples of follow-up questions with various reasoning complexity.

Initial Question: How does sleep affect brain function?	
Initial Answer: Sleep is essential for brain function because it helps with memory consolidation, cognitive processing, and emotional regulation. During sleep, the brain strengthens neural connections, removes toxins, and allows different areas to reset for the next day.	
Complex Amount of Reasoning Example	**Moderate Amount of Reasoning Example**
What are the long-term cognitive effects of chronic sleep deprivation compared to occasional sleep loss?	How does sleep remove toxins from the brain?
Reason	**Reason**
This follow-up question requires complex reasoning because it involves comparing two different scenarios (chronic vs. occasional sleep deprivation) and analyzing their distinct long-term effects on cognition, requiring deeper thought and synthesis of information.	This follow-up question requires moderate reasoning because it builds on a specific detail from the original answer (toxin removal) and asks for an explanation of the biological process involved.
Minimal Amount of Reasoning Example	**Does Not Require Any Reasoning Example**
What are the benefits of sleep for memory?	Does sleep help with memory?
Reason	**Reason**
This follow-up question requires minimal reasoning as it only asks for elaboration on a topic already stated in the original answer (memory consolidation), without introducing any new angle.	This follow-up question does not require any reasoning since it directly repeats a fact already stated in the original answer, making it redundant.

Table 21: Examples of follow-up questions' reasoning complexity for the given initial question and answer, accompanied by corresponding reasons below.

F.3 Informativeness Question Guideline

Evaluate whether the follow-up question enriches the topic by providing or eliciting new information.

- **A Lot of New Information** indicates a significant amount of new knowledge is introduced.
- **Some New Information** suggests moderate enrichment.
- **Little New Information** implies minimal addition.
- **No New Information** means no new information is provided to the audience.

Table 22 contains examples of follow-up questions with various informativeness levels.

Initial Question: How do vaccines work?	
Initial Answer: Vaccines train the immune system to recognize and fight specific germs by introducing harmless parts of the germ or something similar. This prepares the body to respond quickly if exposed to the actual germ in the future.	
A Lot of New Information Example	**Some New Information Example**
What are the differences between traditional vaccines and mRNA vaccines?	How long does it take for a vaccine to provide immunity?
Reason	**Reason**
This follow-up question introduces a significantly new dimension by asking about different types of vaccines, which were not mentioned in the original answer, expanding the discussion substantially.	The follow-up question adds moderately new information by focusing on the timeline of immunity development, a relevant but additional detail not covered in the initial answer.
Little New Information Example	**Does Not Introduce Any New Information Example**
Do vaccines help prevent disease outbreaks?	Do vaccines help the immune system recognize germs?
Reason	**Reason**
The follow-up question slightly expands the discussion by addressing disease outbreaks, but it is already implied in the original answer, as vaccines train the immune system to fight germs.	This follow-up question does not add any new information as it directly restates a key point from the original answer in slightly different words.

Table 22: Examples of follow-up questions' informativeness for the given initial question and answer, accompanied by corresponding reasons below.

G Additional Examples

See Tables 23

H Interface Examples

See Figures 4 and 5

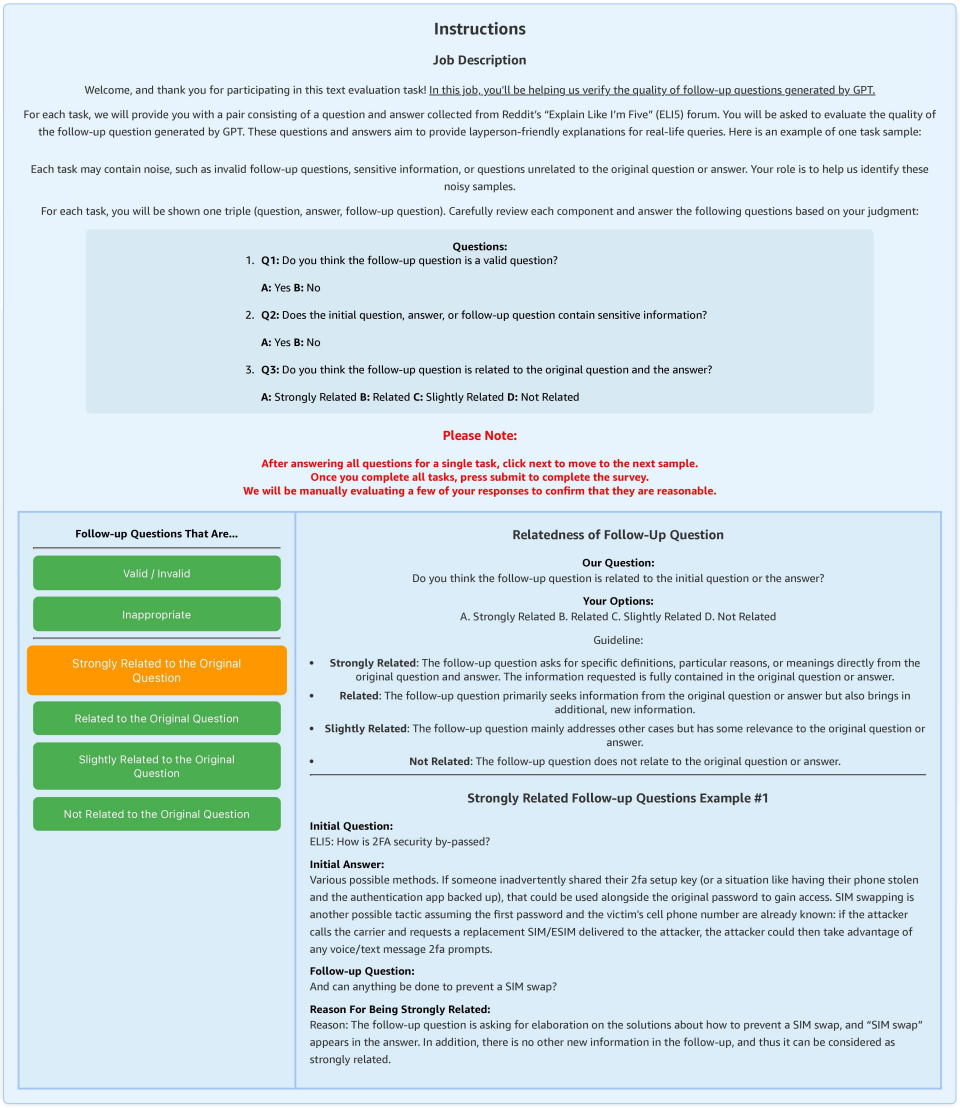

Figure 4: Human Evaluation Interface - Task Instructions and Examples.

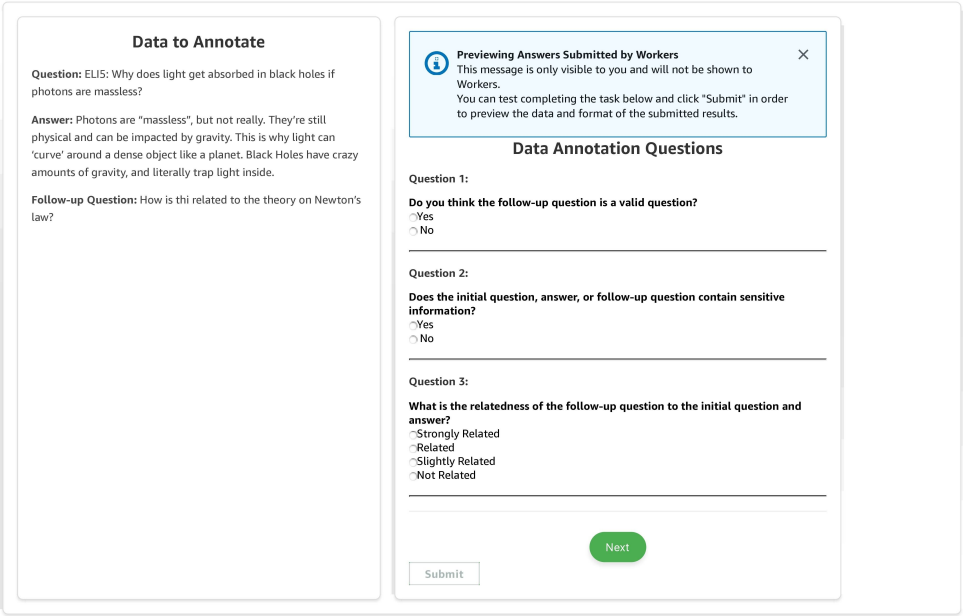

Figure 5: Human Evaluation Interface - Annotation.

IQ	What causes occupational domination?
IA	*In IT it is cultural. Women in the U.S. drop out of IT classes because they do not like the male-dominated culture.* Whereas India, Vietnam, and other countries where we outsource our IT have a much higher percentage of female programmers. *It is not something about women.* It is something about western IT classes and tech culture.
CA	Occupation domination occurs due to a combination of historical, social, economic, and geographic factors. *Historically*, certain groups have had greater access to education, resources, and opportunities, leading them to be more prevalent in specific jobs or industries. This is often reinforced by network effects, where individuals within a profession prefer to work with or hire people who share similar backgrounds, cultural norms, or experiences, thus perpetuating the concentration of certain demographics in those fields. *Legal and institutional* barriers can also play a significant role. Discriminatory policies or practices can systematically exclude certain groups from specific professions, allowing other groups, who do not face such barriers, to dominate those fields. *Geographic and environmental* factors further influence occupation domination. The presence of certain natural resources or particular climate conditions in an area can dictate the types of industries that develop there. Local populations often specialize in these industries, passing down specialized skills and knowledge through generations, which solidifies their dominance in those fields over time.
FQ by *ORG*	– But why do women drop out of IT classes? – *In IT it is cultural. Women in the U.S. drop out of IT classes because they do not like the male-dominated culture.* This is a myth. *It is not something about women.* It's a myth that women drop out because of the male-dominant culture. – *So it's cultural?* – What about women who want to be in tech but are not in the workforce? – Is there a cultural component of IT that makes it more difficult for women to break into it?
FQ by *AUG*	– What are some examples of how outsourcing has impacted the gender imbalance in IT? – How do cultural expectations influence the recruitment of female programmers? – Can you explain the concept of 'cultural fit' in the context of occupational domination?
FQ by *FULL*	– How do traditional gender roles in professions like medicine and engineering contribute to occupation domination? – Why might women feel more comfortable pursuing professions outside of traditional professions? – Can you explain the concept of 'perceived value' in the context of occupation domination?

Table 23: Example of follow-up question generated by three model variants, with comprehensive answers (ID 3168).