How to Engage Your Readers?♦ Generating Guiding Questions to Promote Active Reading

Peng Cui[1], **Vilém Zouhar**[1], **Xiaoyu Zhang**[2], **Mrinmaya Sachan**[1]
ETH Zürich Department of Computer Science[1], ETH AI Center[2]
peng.cui@inf.ethz.ch

Abstract

Using questions in written text is an effective strategy to enhance readability. However, what makes an active reading question good, what the linguistic role of these questions is, and what is their impact on human reading remains understudied. We introduce GUIDINGQ, a dataset of 10K in-text questions from textbooks and scientific articles. By analyzing the dataset, we present a comprehensive understanding of the use, distribution, and linguistic characteristics of these questions. Then, we explore various approaches to generate such questions using language models. Our results highlight the importance of capturing inter-question relationships and the challenge of question position identification in generating these questions. Finally, we conduct a human study to understand the implication of such questions on reading comprehension. We find that the generated questions are of high quality and are almost as effective as human-written questions in terms of improving readers' memorization and comprehension.

github.com/eth-lre/engage-your-readers

1 Introduction

Questions play an important role in reading comprehension. Through actively raising questions and seeking answers from the content during reading, readers can deeply engage with the text and achieve better comprehension (Bharuthram, 2017; Syamsiah et al., 2018). However, asking good questions is challenging and requires complex skills.

How can we facilitate readers' active thinking and questioning during reading?♥[1] An effective approach could be presenting valuable questions *explicitly* in the text, which is a recognized strategy to enhance readability and engage readers (Haggan, 2004). An example is shown in Figure 1. The writer first uses a title question (Q1) to arouse the interest of potential readers. Then, a group of questions (Q2-3) surfaces in the beginning to introduce the central topics to be explored. In what follows, more questions (Q4-5) are raised and elaborated with the moving of discussion, holding the reader's attention throughout the reading. From a linguistic perspective, these questions not only build up a coherent discourse structure (Curry and Chambers, 2017), but also serve as a communicative device that constructs a virtual dialogue between the writer and potential readers, thereby making the text more engaging and interactive (Hyland, 2002).

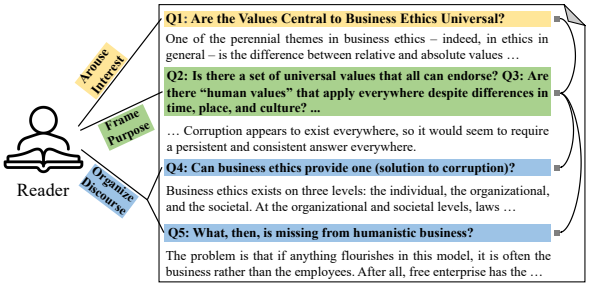

Figure 1: We generate interconnected questions with diverse rhetorical functions *during reading* to engage readers and improve comprehension.

In this paper, we refer to these in-text questions as *guiding questions*. Despite being widely used, there is little understanding of the effect of such questions on human reading. To fill this gap, we analyze how expert writers use guiding questions and explore how to model these questions with advanced language models. Further, we hypothesize that these author-posed questions can complement and encourage readers' spontaneous self-questioning, thereby fulfilling the goal of active reading. Based on these motivations, we address the following research questions:

- **RQ1**: What is the use, distribution, and role of guiding questions in formal writing?♣
- **RQ2**: How well can language models understand

[1]Superscript symbols indicate the roles of questions in this paper, detailed in Section 3.

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 11749–11765

and generate guiding questions?♣

- **RQ3**: What is the effect of these questions on human reading comprehension?♣

To answer these questions, we start by curating GUIDINGQ, a dataset of 10,577 guiding questions from research articles and textbooks. Since the two source texts are written by expert writers, we assume the questions are carefully designed to enhance readability and thus ideal for our research. Through qualitative and quantitative analysis on GUIDINGQ, we summarize the question roles based on their **discourse** and **interactional** effects, and present their usage and distribution across the domains (RQ1, Sections 3 and 4). Then, we explore various approaches to model these questions from three interrelated aspects (RQ2, Section 5): identifying question positions (where to ask), predicting question focus (what to ask), and finally generating the questions (how to ask). Our results highlight the importance of inter-question relationships and the challenge of question position identification in generating guiding questions.

Finally, to validate whether and how the generated guiding questions can facilitate active reading (RQ3, Section 6), we conduct a carefully designed human study where participants read articles with or without questions, complete a post-reading summarization test, and later evaluate the questions. The results demonstrate that the generated questions are not only of high quality but can help readers produce better summaries, indicating an improved memory retention and understanding of the high-level information of the article.

2 Related Work

Here we focus on discussing the linguistic role of questions. Questions can be used as a tool for discourse planning. The Question Under Discussion (QUD) framework uses questions to interpret the discourse relationship between textual units within a document (Van Kuppevelt, 1995; Roberts, 2012; Benz and Jasinskaja, 2017). For example, the relationship between *"S_a: A night of largely peaceful protests ended early Monday in a bloody"* and *"S_b: Hours earlier, Egypt's new interim leadership had narrowed in on a compromise candidate to serve as the next prime minister."* can be described by question "*What happened before the clash?*", where S_a, which elicits the question, is called anchor sentence and S_b is the answer sentence. Studying QUD with modern NLP techniques is a relatively new field where most efforts focus on data construction (De Kuthy et al., 2018; Westera et al., 2020; Ko et al., 2022; Wu et al., 2023).

While QUD provides a more flexible way to represent discourse connections compared to predefined fixed relationships (e.g., elaboration, condition) than other frameworks like Rhetorical Structured Theory (RST; (Mann and Thompson, 1988)), how to use these *implicit* questions to assist readers is not straightforward. In this paper, we instead focus on questions that are *explicitly* presented in a text by the author, which we call guiding questions. We argue such questions are more powerful. Besides their discourse effect, these questions directly interact with readers. Therefore, they have the potential to influence readers' behaviors (Curry and Chambers, 2017). For example, articles titled with questions are found to obtain more downloads (Jamali and Nikzad, 2011).

In terms of the goal, QUD strives to explore the dense space of questions to interpret exhaustive intra-document relationships, while we aim to distill sparse but crucial questions to engage readers without overwhelming them. We highlight the question position identification task as a preliminary step before generating questions. This dimension is neglected by previous QUD or generic Question Generation (QG) studies. However, when questions are used as a communicative tool to interact with readers *during* reading, when and where to raise them is arguably important. We hope this study sheds light on a greater understanding of the implication of questions in reading comprehension.

3 Taxonomy of Guiding Questions

This section describes our taxonomy of guiding questions. The taxonomy is built upon the discussion in Hyland (2002), which we adapt based on our dataset (Section 4). In particular, we classify questions into five different roles based on their different discourse or interactional effects. Their definitions and examples are provided below.

- **Arouse Interest.**♦ The first category refers to questions that appear in titles. Since the title is generally the reader's first encounter with a text, formulating it as an intuitive question can grab the reader's attention.

> **Questions in titles:**
> *How do Philosophers arrive at truth?*
> *Is there no quantum form of Einstein Gravity?*
> *Why do house-hunting ants recruit in both directions?*

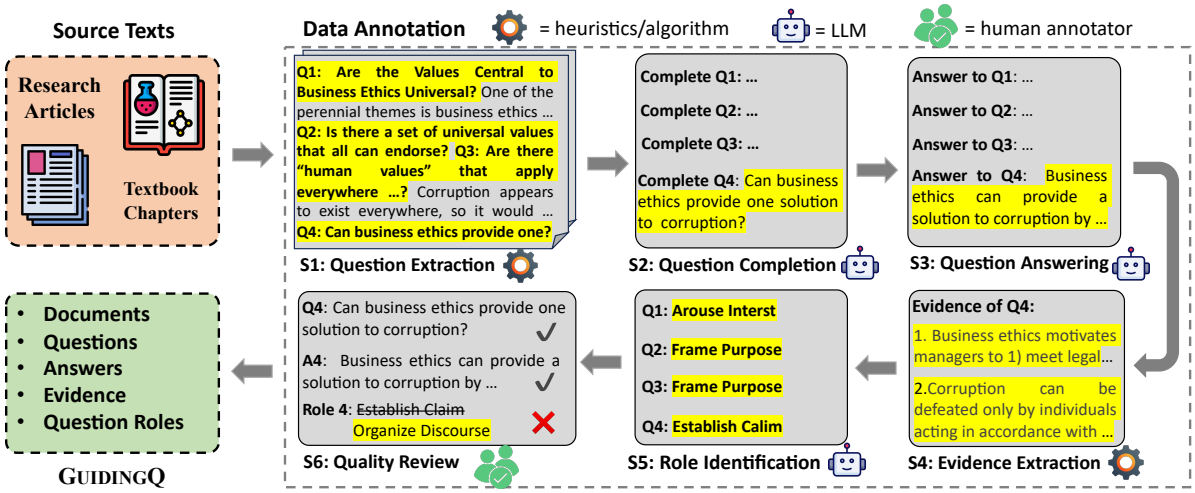

Figure 2: The construction pipeline of GUIDINGQ. Important outputs are highlighted

Note that although the title question usually reveals the main topic of an article, the breadth of the content is not always encapsulated by it

- **Frame Purpose.**♣ This type of question often surfaces and clusters in the beginning section to foreground the central topics to be explored.

> Several aspects of this theory need further investigation. *Is it possible to achieve predictable refractive changes? Can this be achieved through an intact epithelium?* ... This paper describes the use of a novel device ...

Writers pose these questions to provide an agenda for the article and then pick them up again in later sections to direct readers through the reading.

- **Organize Discourse.**♠ Questions can also serve as *subheadings* to structure the text, guiding readers by explicitly introducing shifts in information and identifying what will be discussed in the ensuing section.

> *What are the advances of telecommuting?* The term telecommuting emerged in the 1970s to ... *What are the drawbacks of telecommuting?* In 2013, Yahoo's then-CEO, Marissa Mayer, ended ... *What are the ethical challenges of telecommuting?* ...

Noticeably, such questions usually appear multiple times throughout an article, collectively creating a sense of progression toward a greater understanding of the topic.

- **Establish Claim.**♥ Another use of questions is to introduce and emphasize the writer's arguments rather than to seek the reader's interaction or viewpoint.

> *What contributes to a corporation's positive image over the long term?* Many factors contribute, including a reputation for treating customers and employees fairly and for engaging in business honestly.

A distinct feature of such questions is that the writer often provides a clear answer (i.e., the argument), usually close to the question, thereby limiting the reader's alternative interpretations to the preferred one.

- **Provoke Thought.**★ Finally, there are some "genuine" questions that do not anticipate specific responses within the text. Therefore, they can facilitate the reader's active thinking to the greatest extent.

> If the technological resources of today's governments had been available to the East Germany Stasi and the Romanian Securitate, *would those repressive regimes have fallen? How much privacy and freedom should citizens sacrifice to feel safe?* `[END]`

It's worth noting that different roles are not mutually exclusive. For instance, an AROUSE INTEREST question may also provoke thoughts and vice versa. Nevertheless, we focus on understanding the main role of a question.

4 GUIDINGQ DATASET

In this section, we first discuss the choice and rationale of source texts (§ 4.1), followed by the construction pipeline (Figure 2) of the GUIDINGQ dataset (§ 4.2). Then, we present a series of distributional features of guiding questions (§ 4.3).

4.1 Source Texts

We select **scientific articles** and **textbooks** as the source texts to build the dataset. Our choice is based on two considerations. First, their writer-reader discourses have a clear communicative intent, either peer-to-peer or teacher-to-student, which can motivate the use of questions. Second, they are formal texts written by experts, ensuring

GuidingQ	Textbook	Research Articles
# documents	621	1,501
# avg. words /doc.	2,404	2,140
# questions	3,593	6,964
# avg. words / question	13.8	21.0
# avg. questions / doc.	5.79	4.63

Table 1: Statistics of the GUIDINGQ dataset

that questions presented in them are strategically used to enhance readability. Specifically, we use textbook chapters collected from an online free publisher OpenStax[2] (Singh et al., 2023) and research papers from the arXiv and PubMed datasets (Cohan et al., 2018).

4.2 Construction Pipeline

We describe the main steps of collecting and annotating GUIDINGQ below.

S1: Question Extraction. We start by extracting questions from source texts by detecting interrogative marks. We only keep documents with at least three questions, indicating the writer actively used questions in the writing.

S2: Question Completion. Since the extracted questions are a part of the source texts, they are not always semantically complete due to omissions or unclear pronouns, e.g., *What central point might constitute such a code?* Therefore, we first identify and complete such questions based on the context. We do this because it is the first step to understanding the meaning of such questions.

S3: Question Answering. Next, we generate the answer to each question. In particular, the answer should be detailed enough and solely based on the article. Therefore, we use the article's words as the answer whenever possible. If a question is not discussed in the article (e.g., PROVOKE THOUGHTS question), we label it as "NO ANSWER."

S4: Evidence Extraction. Given a produced answer, we automatically extract *supporting sentences* from the article as evidence. We do this by greedily searching a set of sentences that has the maximum Rouge score with the answer, which is the standard way to find Oracle sentences for extractive summarization systems (Nallapati et al., 2017). For questions without answers as per **S3**, we directly label them as "NO EVIDENCE."

S5: Question Role Identification. Finally, we identify the role of each question. We make this the last step as the information collected in previous steps (complete question, answer, evidence) can help in understanding the role of a question.

[2] openstax.org

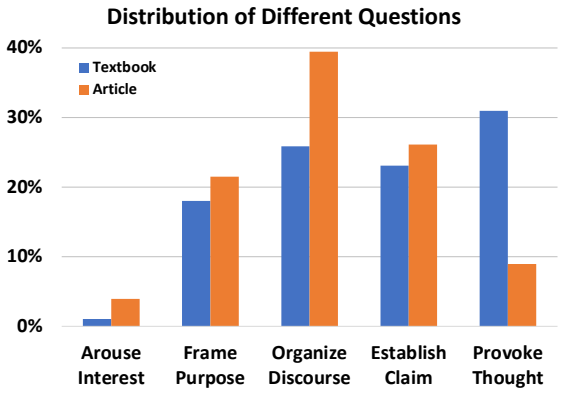

Figure 3: Distribution of different question roles.

Recent studies have shown that LLMs with carefully designed prompts are comparable to or even better than human annotators (He et al., 2023; Zhang et al., 2023; Törnberg, 2023). To reduce human effort and enable data scaling in the future, we use ChatGPT to complete the annotation steps **S2, S3,** and **S5**. See prompts in Tables 8, 9, and 10.

S6: Quality Control. We adopt a Human-LLM co-annotation paradigm (Li et al., 2023). Concretely, we ask the model to provide a confidence level for each of its outputs. Examples below a certain level are reviewed by human annotators, followed by an overall quality review, detailed in Appendix A.

4.3 GUIDINGQ Analysis

The statistics of the GUIDINGQ dataset is summarized in Table 1. In general, textbook chapters use questions slightly more frequently than research articles. This is possibly because teacher-to-student interactions are more inclined to involve questions than peer-to-peer ones.

What is the distribution of question roles?♠ In Figure 3, we compare the distributions of different questions on the two subsets. As can be seen, research articles contain more FRAME PURPOSE and ORGANIZE DISCOURSE questions, while textbooks favor PROVOKE THOUGHT questions possibly due to their educational purpose. Besides, AROUSE INTEREST (title) questions are more common in articles than textbooks, although both proportions are small. A possible reason is that research articles exist in a competitive environment where potential readers are confronted with a large number of papers, under which circumstances interrogative titles could help attract readers (Haggan, 2004; Jamali and Nikzad, 2011).

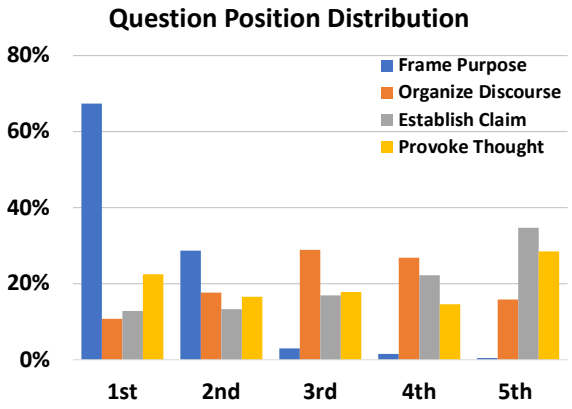

Figure 4: Position distribution of different questions. We omit AROUSE INTEREST questions as they are in titles by definition.

How diversely do human writers use guiding questions?♠ To understand this, we measure the number of question roles used in an article. The results are shown in Table 2. Since research articles contain fewer questions per article, their question roles are slightly less diverse than those in textbooks. Overall, the majority of articles (70%+) use less than three types of questions, suggesting room for improvement in human questioning strategies.

# Question Types	1	2	3	4	5
Textbooks	21.1	44.8	26.8	7.0	0.2
Articles	35.2	45.5	17.2	2.0	<0.1

Table 2: Distribution (%) of the number of unique question roles in one article.

Where are guiding questions asked?♠ We divide articles into five equally sized segments and calculate the percentage of questions that appear in each part. As in Figure 4, there is a strong positional bias among different questions. FRAME PURPOSE questions mostly appear in the first and second segments. ORGANIZE DISCOURSE questions are largely located in the middle segments, while PROVOKE THOUGHT questions tend to emerge in the last segment, consistent with their expected functions. ESTABLISH CLAIM questions are skewed towards the end of the text. A possible explanation is that writers make more and more conclusive arguments with the progression of discussion.

Where are guiding questions answered?♠ We measure the distance between a question and its farthest evidence sentence (if any) in Figure 5. This can be considered the scope a question acts on the discourse of the document, i.e., from when the question is raised to when it is closed. In this sense, PROVOKE THOUGHT questions go beyond the content as they do not have specific answers. We find the two subsets show a similar distribution. FRAME PURPOSE questions serve as the outline of a document, therefore having the largest range. ORGANIZE DISCOURSE questions are usually answered within one or two paragraphs, consistent with their subheading role. In contrast, ESTABLISH CLAIM questions have prompt answers as the claims.

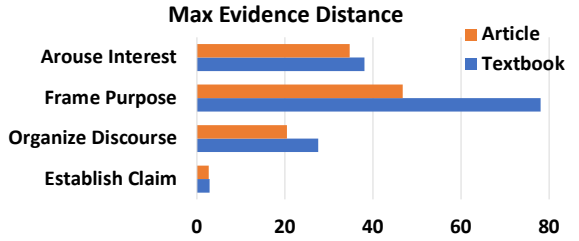

Figure 5: The average distance (in terms of sentence numbers) between a question and its farthest evidence.

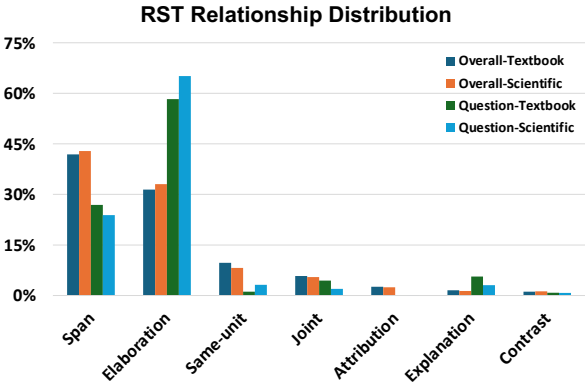

Figure 6: Distribution of RST discourse relationships with respect to questions and overall units. Relationships accounting for less than 1% are omitted.

How are questions related to other text units?♠ As guiding questions are integral to the article, we also analyze the discourse relationships between questions and other text units[3] within the same document using an RST (Mann and Thompson, 1988) parser[4]. In Figure 6, we compare the distribution of discourse relationships concerning questions with those of all text units. The analysis reveals that questions exhibit a higher proportion of concrete relationships, such as "elaboration" and "explanation," and a lower proportion of general relationships, such as "span" and "same-unit" compared to other units. This suggests incorporating ques-

[3]Following RST, we consider Elementary Discourse Unit (EDU) as the minimal unit.

[4]github.com/EducationalTestingService/rstfinder

tions and their answers in writing might add to the coherence of the discourse structure.

5 Guiding Question Generation

In this section, we first describe our methods to model guiding questions and then report experimental results.

5.1 Task Formulation

Given a document $\mathcal{D} = \{s_1, ..., s_n\}$ of n sentences, our goal is to learn a sequence of questions $\mathcal{Q} = \{q_1, ..., q_m\}$, their positions in the article $\mathcal{P} = \{p_1, ..., p_m\}$, and (optionally) their answer information $\mathcal{A} = \{a_1, ..., a_m\}$. In particular, $1 \leq p_i \leq n$ is the index of the sentence after which q_i should be asked. We call these sentences **anchor sentences** $\{s_{p_1}, ..., s_{p_m}\}$.

5.2 Data Preparation

To construct training examples, we remove questions from articles and reconstruct them based on the corrupted articles such that the model can learn how to use guiding questions as human writers. In concrete, given an article $\mathcal{D}$, we extract its questions and locate their positions to obtain $\mathcal{Q}$ and $\mathcal{P}$. For each question q_i, its answer information is a set of keywords $a_i = \{w_1, ..., w_{|a_i|}\}$ extracted from its evidence sentences obtained during the annotation (Section 4.2). We take this form to exclude redundant information in the full answer and reduce input (output) length for efficiency consideration.

Since deleting sentences would create incoherence and reduce learning $\mathcal{P}$ to identify where sentences are removed, we use `gpt-3.5-turbo-1106` to assess the coherence of missing positions and, if necessary, eliminate the incoherence by making small edits around the missing positions. Since there could still be nuanced differences in edited positions, we perform the same "delete and smooth" operation on 1% randomly selected non-question sentences as noise, detailed in Appendix B.1.

5.3 Modeling

In what follows, we describe three popular QG paradigms considered for our task: **Pipeline**, **Multitask**, and **Joint Generation** (Ushio et al., 2023). All approaches are unified as a text generation task and use Flan-T5 (Chung et al., 2022) as the backbone model, which we finetune on our dataset.

Pipeline. We decompose the task into three subtasks to learn $\{\mathcal{P}, \mathcal{A}, \mathcal{Q}\}$ with independent models.

First, a **Position Predictor** (PP) identifies the positions of questions. A naive way is to directly generate the position indices $\{p_1, ..., p_m\}$ conditioned on $\mathcal{D}$. However, this requires learning a mapping between sentences and numerical symbols. Instead, we opt to identify the *anchor sentences* by training PP to copy them from $\mathcal{D}$:

$$\tilde{C} = \underset{C}{\operatorname{argmax}}\, P_{\theta_{\mathsf{PP}}}(C|\mathcal{D}). \quad (1)$$

where the output $\tilde{C} = [s_{p_1}|...|s_{p_m}]$ is a concatenation of anchor sentences separated by "|". $\mathcal{P}$ can be obtained by relocating copied sentences in $\mathcal{D}$. When there is no exact match, we use BM25 to get the most similar sentence.

Then, we highlight the target position in $\mathcal{D}$ by inserting a special mark [Question] after the anchor sentence and use an **Answer Extractor** (AP) to generate answer keywords a_i.:

$$\mathcal{D}^{(i)} = [s_1, ..., s_{p_i}, [\text{Question}], ..., s_n], \quad (2)$$

$$\tilde{a}_i = \underset{a}{\operatorname{argmax}}\, P_{\theta_{\mathsf{AE}}}(a|\mathcal{D}^{(i)}), \quad (3)$$

where $\mathcal{D}^{(i)}$ is the document marked at position p_i.

Finally, a **Question Generator** (QG) generates questions based on predicted positions and extracted answers:

$$\tilde{q}_i = \underset{q}{\operatorname{argmax}}\, P_{\theta_{\mathsf{QG}}}(q|\mathcal{D}^{(i)}, a_i). \quad (4)$$

Note that the PP model predicts all positions in one pass while the other two generate output one by one.

MultiTask. The multitask model still consists of the three components described above. However, instead of independently training three models, we train a unified model for all tasks in a multitask learning manner. In practice, we mix the training examples of the three tasks and distinguish them by adding different task prefixes before the inputs.

Joint Model. As observed in Figure 1, guiding questions tend to be related to each other. Therefore, we consider jointly generating all questions at the same time. To be specific, we use a template function to convert $\{\mathcal{P}, \mathcal{A}, \mathcal{Q}\}$ into a flattened sequence $\mathcal{T}(\mathcal{P}, \mathcal{A}, \mathcal{Q}) = \{t(p_1, a_1, q_1)|t(p_2, a_2, q_2)...\}$ where $t(p, a, q) =$ "`Position:`s_p`#Answer:`a`#Question:`q". The

Models	Textbook					Scientific				
	# Q	**Rouge-L**	**Meteor**	**BertScore**	**Dist-1/2 (↓)**	**# Q**	**R-L**	**Meteor**	**BertScore**	**Dist-1/2 (↓)**
GPT-4 (0-shot)	7.58	15.7	18.9	84.3	64.4/48.9	5.57	**14.3**	**19.7**	82.8	68.4/49.1
Pipeline_{250M}	1.73	12.9	13.7	77.4	71.4/51.6	1.69	12.7	7.93	79.5	66.1/48.3
Multitask_{250M}	1.65	15.4	15.2	78.1	70.6/49.0	1.84	13.7	9.26	80.8	58.9/43.8
Joint_{250M}	3.41	16.9	19.3	82.8	66.8/47.5	2.08	12.2	9.89	81.0	78.9/51.3
$\text{Joint}^{\text{R}}_{250M}$	3.77	18.3	20.8	84.5	65.4/46.9	2.16	13.1	10.2	81.0	75.2/49.8
$\text{Joint}^{\text{R}}_{780M}$	3.81	30.5	28.1	87.7	65.6/47.1	2.40	12.4	9.38	82.5	71.3/48.9
$\text{Joint}^{\text{R}}_{3B}$	3.95	36.7	34.8	88.7	65.6/46.8	2.46	11.7	9.59	82.6	69.1/46.9
$\text{Joint}^{\text{R}}_{11B}$	3.95	**56.4**	**49.5**	**92.1**	64.1/46.5	2.55	13.5	10.3	**82.9**	68.5/47.2
Reference	5.46	100.	100.	100.	61.7/46.3	4.50	100.	100.	100.	66.7/48.9

Table 3: Results of finetuned and prompt-based models on the GUIDINGQ dataset. Main takeaways: 1) human guiding questions are interrelated (as per Dist-N); 2) Joint generation w/ question role performs the best among fine-tuned models; 3) Generated questions can capture the main message of human questions (as per BertScore).

model is trained to directly generate the whole target sequence $\mathcal{T}$ based on the article $\mathcal{D}$:

$$\tilde{\mathcal{T}} = \underset{\mathcal{T}}{\text{argmax}}\, P_{\theta_{\text{JT}}}(\mathcal{T}|\mathcal{D}). \quad (5)$$

In doing this, each question is also conditioned on previous ones, enabling the model to learn the inter-question connections.

For the joint model, we also introduce a variant that additionally generates question roles. This is done by inserting "`Role: question role`" between the position and answer field of each entry. We denote this model as Joint^{R}.

Since the above methods are all formalized as a text generation task, their training objectives take a similar form:

$$\mathcal{L} = -\sum_{i}^{|Y|} \log P_\theta(y_i|y_1, ..., y_{i-1}, X), \quad (6)$$

where X and Y are the input and output sequence of the corresponding task. See Appendix B.2 for training details.

5.4 Automatic Evaluation

The main results are presented in Table 3, where we also include zero-shot prompted GPT-4 (`gpt-4-1104-preview`) with the prompt in Table 13. We report the average number of generated questions **# Q**, **Rouge (L)** (Lin, 2004), **Meteor** (Banerjee and Lavie, 2005), and **BertScore** (Zhang et al., 2020). Besides, we use **Dist-N (1/2)** (Li et al., 2016), the percentage of distinct n-grams in generated questions, to measure the balance between *diversity* and *relevance* of guiding questions. Since there is no one-to-one map between generated and ground-truth questions, we concatenate all the questions as a whole sequence to compute reference-based metrics.

Overall, fine-tuned models generate fewer questions than references, while zero-shot GPT-4 generates more. We found that Flan-T5 tends to output short sequences; therefore, we attribute this to their different pre-training paradigms. Jointly generating question roles can boost performance, which is expected as they are indicative of a series of distributional features. The best fine-tuned models achieve remarkable BertScores. This suggests that the generated questions successfully replicate the main information of human questions.

For the textbook dataset, the joint model generally performs the best among fine-tuned approaches. In particular, it best resembles the Dist-1/2 results of human questions, while others tend to generate more independent questions (higher Dist-1/2). This suggests that the success of this model is possibly because it better learns the inter-question relationships. As for the scientific set, the multitask model achieves competitive results with the joint model. We conjecture the more complex content and sparser questions increase the difficulty of learning the question relationship, which could diminish the advantage of joint generation.

Finally, we scale up the parameter size of the best-performing model Joint^{R} up to 11B. We can see that scaling the model size results in significant performance gain on the textbook set but little on the scientific set. This demonstrates the challenge of understanding scientific language with LLMs.

Question Position. Since the generated questions are different from references, naively matching their positions (e.g., recall, precision) is not a suitable way. Instead, we evaluate a question's

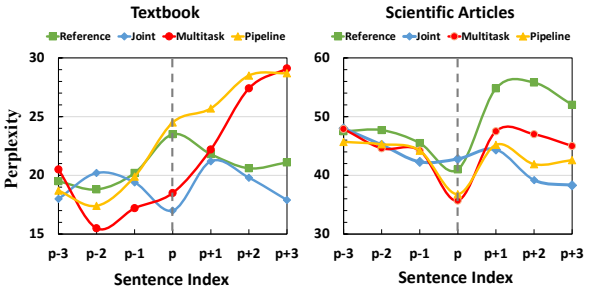

Figure 7: Average word perplexity of the question (index p) and its surrounding sentences. We omit questions by 0-shot GPT-4 as we find their positions are sensitive to prompts and temperatures.

position by measuring how well it fits the context using perplexity. Specifically, let p be the index of question q. We insert q in p and calculate the average word perplexity of sentences $[s_{p-3}, s_{p-2}, s_{p-1}, q, s_{p+1}, s_{p+2}, s_{p+3}]$, each conditioned on its previous three sentences. The results are presented in Figure 7. Interestingly, human questions in textbooks create a local peak of PPL in the context. We explain this phenomenon from the connection between *surprisal* and *salience*: text units with higher salience are usually less expected, making them stand out from the context and attract the reader's attention (Rácz, 2013; Zarcone et al., 2016; Blumenthal-Dramé et al., 2017). This is consistent with the interactional purpose of using questions. However, none of the models replicate this salience effect, either lowering the PPL (joint model) or increasing the PPL without a quick decline (pipeline/multitask).

The scientific articles show a different pattern where human questions are with lower PPL. We conjecture this is because scientific articles usually make sufficient discussions before proposing a question. As a consequence, their questions are highly predictable by prior sentences. Finetuned models manage to replicate the effect, possibly because research articles and their question usage are usually structured (e.g., proposing questions in the introduction) and thus easier to learn.

6 Human Study

Finally, we conduct a between-group human study to investigate the impact of guiding questions on reading comprehension. Participants are asked to read articles with (or without) questions, after which we gather their feedback and analyze their information retention and understanding to gain a holistic view of the effect of guiding questions.

		Reference	Generated
Quality	Relevance	4.0	4.2
	Position	4.2	4.3
	Importance	4.4	4.2
Usefulness	Engaging	3.9	4.0
	Understanding	4.2	4.1
	Overall	4.3	3.9

Table 4: Average scores by participants. "Quality" is evaluated on each single question, while "Usefulness" is evaluated on all questions of an article.

6.1 Experiment Design

Procedure. The human study runs on a crowdsourcing platform Prolific and consists of 4 stages.

1. **Demographic Questionnaire**: We collect participants' demographic information (Appendix C.1) and consent before the experiment.
2. **Article Reading**: Participants read an assigned article in 20 minutes with or without questions beside the article. Details of the reading interface are in Appendix C.2.
3. **Comprehension Test**: After reading, participants are asked to write a summary of at least 100 words *without* access to the article.
4. **Evaluation**: Finally, we ask participants to rate the usefulness and quality of questions (If any).

Participants. We select 45 participants with the criteria being at least C1 command of English. Each participant received 10 GBP/hour on average. We randomly assign them to one of three groups:

- **Control**: reading w/o questions.
- **Reference**: reading w/ expert-written questions.
- **Generated**: reading w/ generated questions.

Test Articles. We use the Joint^{R} model to generate questions for five textbook chapters[5] selected from different domains spanning *Business*, *Philosophy*, *Sociology*, *Political Science*, and *Psychology*. A few small adaptations are made to make them better fit the study (e.g., length reduction). To focus on question quality, we generate the same number of questions as reference[6]. Each article is assigned to 3 participants to average-out the effect of individual articles.

[5]We only consider textbook passages because scientific articles are challenging for general readers.

[6]We do this by truncating over-generated questions or disallowing the <eos> token until we get enough questions.

	Summary Quality (1-5)		
	Coherence	**Consistency**	**Informativeness**
Ctrl.	$2.38_{\pm 0.09}$	$2.52_{\pm 0.08}$	$2.31_{\pm 0.09}$
Gen.	$3.14_{\pm 0.08}$	$\mathbf{3.28}_{\pm 0.08}$	$2.98_{\pm 0.05}$
Ref.	$\mathbf{3.28}_{\pm 0.04}$	$3.23_{\pm 0.06}$	$\mathbf{2.99}_{\pm 0.08}$

Table 5: Quality scores ($\text{Mean}_{\text{std.}}$) of summaries from three groups over 5 runs.

Sum. \ Ans.	**Reference**	**Generated**
Reference	**55.2**	–
Generated	–	**56.3**
Control	52.8	52.5

Table 6: Entailment score between guiding questions summaries and answers with BertScore recall (×100).

6.2 Question Evaluation

Participants evaluate the intrinsic quality of each question from three aspects: *relevance*, *position*, and *importance* (detailed in Table 14). From a later survey, we found that participants may have different interpretations of these dimensions despite our instructions. Nevertheless, the results in the first section of Table 4 prove that the average quality of generated questions is on par with (if not better than) ground truth ones.

6.3 Effect on Reading Comprehension

User Perceived Usefulness. Results in the second part of Table 4 prove that the generated questions are as good as human questions in perceived usefulness. See the questionnaire in Table 14.

Improved Memorization and Comprehension. We use summary quality as a proxy to measure participants' memorization and comprehension. The evaluation consists of three dimensions: *Coherence*, *Consistency*, and *Informativeness*, and is based on GPT-4, which has shown a superior correlation with human annotations on summary evaluation (Liu et al., 2023), detailed in Appendix C.3. As shown in Table 5, the reference and generated groups achieve comparable scores, and both outperform the control group by a large margin. An additional between-group summary analysis, including summary time, length, and n-gram overlap, is shown in Table 15.

An intuitive explanation for the improved quality is that users memorize question-related information and incorporate them into the summary. To verify this, we measured the *entailment* score between summaries and answers to guiding questions using BertScore recall (BS_r):

$$\textsc{EntScore} = \frac{1}{|S|}\sum_{s\in S}\frac{1}{|Q_s|}\sum_{q\in Q_s}\text{BS}_r(s, a_q), \tag{7}$$

where s is a summary, Q_s is the set of guiding questions of the source article of s, and a_q is the answer to q.

The results in Table 6 show that the reference and generated groups incorporated more answer information compared to the control group, indicating improved memorization. However, the difference in recall scores is not as pronounced as the difference in summary quality (Table 5). This suggests that readers do not simply compile question-answer pairs into summaries. Therefore, we believe the improvement in summary quality also reflects a deeper understanding facilitated by the guiding questions.

Reading Time. The reading speeds of different groups are shown in Table 7. When questions are displayed next to the articles, the users spend a longer time reading. On the one hand, this is a signal of enhanced engagement. On the other hand, this potentially implies an extra cognitive load caused by guiding questions.

	Auth.	**Gen.**	**Contr.**
Reading Speed (w/s)	1.00	0.98	1.39

Table 7: Reading speeds of different user groups measured by words per second (w/s).

Participant Feedback. Finally, we gathered user feedback through a preliminary interview-based study, detailed in Appendix C.5. In summary, participants confirmed the benefits of guiding questions and discussed various aspects of these questions, highlighting both consistent preferences and nuanced complexities. We hope these insights will inspire future research in this area.

7 Conclusion

This paper studies the discourse and interactional role of guiding questions in textbooks and scientific articles. We explore various approaches for modeling these questions, providing insights into how to model this task and highlighting challenges to be solved. We validate our results with human studies, which demonstrate reading with guiding questions can improve the high-level memorization and understanding of human readers.

Broader Impacts

In this study, we analyzed the use of questions in academic and educational articles, demonstrating their benefits for reading comprehension. While questions can enhance engagement, they can also increase readers' cognitive load, as evidenced by longer reading times (Table 7). Additionally, questions may introduce unintended nuances for communication, such as creating unequal social relationships (Hyland, 2002). Therefore, it is important to be aware of these mixed effects when using guiding questions in writing.

Limitations

We summarize the limitations of this study into the following open questions.

Is our question role taxonomy generalizable to other domains?★ Our investigation of the role of guiding questions is initially focused on textbooks and scientific articles. However, different domains might use questions differently. Nevertheless, our analysis (Section 4.3) uncovers distributional features that are indicative of question functions, such as their positions and question-answer relationships. These findings offer insights that could be generalized to understand the roles of questions in broader contexts.

How to align guiding questions with individual preferences?★ Our model aims to replicate guiding questions crafted by human writers. However, these questions may not always resonate with individual readers, given their different reading goals and prior knowledge. We expect that personalized generation (Cui and Sachan, 2023), which takes into account user profiles, would yield more helpful questions.

How has the role of questions evolved?★ It is important to note that the use of questions could change over time. For instance, Ball (2009); Jiang and Hyland (2022) have analyzed the distribution shift of questions in titles over the past decades. In this study, we did not take the temporal dimension into account, and the conclusions are based on contemporary texts. Therefore, the findings of this paper may not remain consistent in the future.

References

Rafael Ball. 2009. Scholarly communication in transition: The use of question marks in the titles of scientific articles in medicine, life sciences and physics 1966–2005. *Scientometrics*, 79(3):667–679.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Anton Benz and Katja Jasinskaja. 2017. Questions under discussion: From sentence to discourse.

Sharita Bharuthram. 2017. Facilitating active reading through a self-questioning strategy: Student and tutor experiences and reflections of the strategy use. *Journal for Language Teaching*, 51(2):85–103.

Alice Blumenthal-Dramé, Adriana Hanulíková, and Bernd Kortmann. 2017. Perceptual linguistic salience: Modeling causes and consequences.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. 2018. A discourse-aware attention model for abstractive summarization of long documents. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 615–621, New Orleans, Louisiana. Association for Computational Linguistics.

Peng Cui and Mrinmaya Sachan. 2023. Adaptive and personalized exercise generation for online language learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10184–10198, Toronto, Canada. Association for Computational Linguistics.

Niall Curry and Angela Chambers. 2017. Questions in english and french research articles in linguistics: A corpus-based contrastive analysis. *Corpus Pragmatics*, 1:327–350.

Kordula De Kuthy, Nils Reiter, and Arndt Riester. 2018. QUD-based annotation of discourse structure and information structure: Tool and evaluation. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Madeline Haggan. 2004. Research paper titles in literature, linguistics and science: Dimensions of attraction. *Journal of Pragmatics*, 36(2):293–317.

Xingwei He, Zhenghao Lin, Yeyun Gong, Alex Jin, Hang Zhang, Chen Lin, Jian Jiao, Siu Ming Yiu, Nan Duan, Weizhu Chen, et al. 2023. Annollm: Making large language models to be better crowdsourced annotators. *arXiv preprint arXiv:2303.16854*.

Ken Hyland. 2002. What do they mean? questions in academic writing. *Text & Talk*, 22(4):529–557.

Hamid R Jamali and Mahsa Nikzad. 2011. Article title type and its relation with the number of downloads and citations. *Scientometrics*, 88(2):653–661.

Feng Kevin Jiang and Ken Hyland. 2022. Titles in research articles: Changes across time and discipline. *Learned Publishing*.

Wei-Jen Ko, Cutter Dalton, Mark Simmons, Eliza Fisher, Greg Durrett, and Junyi Jessy Li. 2022. Discourse comprehension: A question answering framework to represent sentence connections. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11752–11764, Abu Dhabi, United Arab Emirates. 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.

Minzhi Li, Taiwei Shi, Caleb Ziems, Min-Yen Kan, Nancy Chen, Zhengyuan Liu, and Diyi Yang. 2023. CoAnnotating: Uncertainty-Guided work allocation between human and large language models for data annotation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1487–1505, Singapore. Association for Computational Linguistics.

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.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

William C Mann and Sandra A Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text-interdisciplinary Journal for the Study of Discourse*, 8(3):243–281.

Ramesh Nallapati, Feifei Zhai, and Bowen Zhou. 2017. Summarunner: A recurrent neural network based sequence model for extractive summarization of documents. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31/1.

Péter Rácz. 2013. *Salience in sociolinguistics: A quantitative approach*, volume 84. Walter de Gruyter.

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. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Craige Roberts. 2012. Information structure: Towards an integrated formal theory of pragmatics. *Semantics and pragmatics*, 5:6–1.

Noam Shazeer and Mitchell Stern. 2018. Adafactor: Adaptive learning rates with sublinear memory cost. In *International Conference on Machine Learning*, pages 4596–4604. PMLR.

Janvijay Singh, Vilém Zouhar, and Mrinmaya Sachan. 2023. Enhancing textbooks with visuals from the web for improved learning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11931–11944, Singapore. Association for Computational Linguistics.

Nur Syamsiah, Zainal Rafli, and Sakura Ridwan. 2018. Self–questioning strategy on reading comprehension process. In *5th Asia Pasific Education Conference (AECON 2018)*, pages 120–129. Atlantis Press.

Petter Törnberg. 2023. Chatgpt-4 outperforms experts and crowd workers in annotating political twitter messages with zero-shot learning. *arXiv preprint arXiv:2304.06588*.

Asahi Ushio, Fernando Alva-Manchego, and Jose Camacho-Collados. 2023. An empirical comparison of LM-based question and answer generation methods. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 14262–14272, Toronto, Canada. Association for Computational Linguistics.

Jan Van Kuppevelt. 1995. Discourse structure, topicality and questioning. *Journal of linguistics*, 31(1):109–147.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.

Matthijs Westera, Laia Mayol, and Hannah Rohde. 2020. TED-Q: TED talks and the questions they evoke. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1118–1127, Marseille, France. European Language Resources Association.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Yating Wu, Ritika Mangla, Greg Durrett, and Junyi Jessy Li. 2023. QUDeval: The evaluation of questions under discussion discourse parsing. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5344–5363, Singapore. Association for Computational Linguistics.

Peiwen Yuan, Shaoxiong Feng, Yiwei Li, Xinglin Wang, Boyuan Pan, Heda Wang, and Kan Li. 2023. Batcheval: Towards human-like text evaluation. *arXiv preprint arXiv:2401.00437*.

Alessandra Zarcone, Marten Van Schijndel, Jorrig Vogels, and Vera Demberg. 2016. Salience and attention in surprisal-based accounts of language processing. *Frontiers in psychology*, 7:844.

Ruoyu Zhang, Yanzeng Li, Yongliang Ma, Ming Zhou, and Lei Zou. 2023. LLMaAA: Making large language models as active annotators. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13088–13103, Singapore. Association for Computational Linguistics.

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*.

A Annotation Details

We use `gpt-3.5-turbo-1106` with a temperature of 0.2 for data annotation. The prompts used for question completion (QC), question answering (QA), and question role identification (QRI) are listed in Tables 8, 9, and 10.

[Task Description]
You will be given some questions extracted from an article and their surrounding texts. Your task is to check whether these questions are self-contained. For example, "What constitutes such a code?" is not a self-contained question due to the unclear "code". If a question is not self-contained, complete it based on its context; otherwise, output it as it is.

[Input]
Question 1: {question 1}
Context 1: {context of question 1}
...

[Output] (Please strictly organize your output in the following format)
Question 1: <complete question 1>
...

Table 8: Prompt for question completion. We use the surrounding 10 sentences of a question as its context.

[Task Description]
You will be given an article and some questions. In particular, these questions are posed in the article and highlighted by a marker "[Question]" before them. Your task is to answer these questions based on the article. Start by reading the article carefully and locating the given questions. For each question, check whether it is answered in the article. If not, output "no answer"; otherwise, provide an answer that should be as detailed as possible and faithful to the article. Finally, provide a confidence level from 1 to 5 for each generated answer, where 1 is the lowest and 5 is the highest.

[Input]
Article: {article with questions marked}
Question 1: {question 1}
...

[Output] (Please strictly organize your output in the following format)
Answer 1: <answer to question 1>
Confidence 1: <confidence level, 1-5>
...

Table 9: Prompt for question answering.

We ask the model to provide confidence levels for outputs of QA and QRI as these two tasks are arguably more challenging. 437 questions have a confidence level of 1 on both tasks. An expert annotator (author of this work) manually validated these questions. Note that our goal is not to annotate as many examples as possible, but to check whether there are unknown question roles beyond our taxonomy and analyze common bad cases to improve prompts. We found that most cases with low confidence are because of their mixed roles. For example, when a question introduces an argument and meanwhile starts related discussions, it could be both ESTABLISH CLAIM and ORGANIZE DISCOURSE question. For such ambiguous cases, we recommend determining their main roles based on their question-answer relationships, as shown in Figure 5. For example, if the abovementioned question has an immediate answer, its main role should be ESTABLISH CLAIM . Note that our taxonomy is to provide a holistic understanding of question functions rather than a hard classification system.

[Task Description]
You will be given an article and some questions. In particular, these questions are presented in the article and play different roles. Your task is to identify their role. The definition and example of each question role are described below.
{Definitions and examples of question roles}
Make sure you understand the above instructions clearly. Start by reading the article carefully and locating the positions of the given questions. Check each question-answer pair and write down your analysis about its role based on the above definition. Finally, output its role and provide a confidence level from 1 to 5 for your judgment, where 1 is the lowest and 5 is the highest.

[Input]
Article: {article}
Question 1: {question 1}
Answer 1: {answer to question 1}
...

[Output]
Analysis 1: <analysis for the role of question 1>
Role 1: <role of question 1>
Confidence 1: <confience of output, 1-5>

Table 10: Prompt for question role identification.

Task	Criteria	Yes (%)
QC	The question is self-contained	90.5%
QA	The answer is overall acceptable	83.6%
QRI	The identified role is correct or	79.4%

Table 11: Quality review of automatic annotation.

Finally, we sample 10 research articles and 10 textbook chapters with 116 questions in total and

review the accuracy of annotation. As shown in Table 11, decent results are observed on all tasks. We plan to scale the dataset and update the annotation with more powerful LLMs available in the future.

B Experiment Details

B.1 Article Processing

To process an article into the training format, we delete questions from the article and eliminate incoherence by prompting `gpt-3.5-turbo-1104` with the instruction in Table 12. To reduce the cost, we build the paragraph using the 5 sentences before and after the deleted question, which is enough to assess or restore the coherence of the local context according to our qualitative inspection. When performing this operation on random sentences, we disallow sentences around (distance<10) any already deleted or question sentences to be selected in order to avoid severe incoherence.

[Task Description]
Given a paragraph where a sentence has been removed and replaced with "[MASK]," your task is to assess whether the paragraph remains coherent without the missing sentence. If yes, simply remove the [MASK] token. If not, please edit the text around [MASK] to restore its coherence. You can only make necessary and minimal edits, leaving the majority of the paragraph verbatim. You can not introduce new information or change or remove existing information.

[Input]
Input Paragraph: {paragraph with a missing sentence}

[Output]
Coherent paragraph: <coherent paragraph>

Table 12: Prompt for coherence maintenance.

B.2 Training setup

We split the dataset into 90% training and 10% test sets. Our implementations are based on the Transformers Library (Wolf et al., 2020). In concrete, for all approaches, we fine-tune the Flan-T5 for up to 10 epochs with a learning rate of $5e-5$ and batch size of 32. Following Raffel et al. (2020), we employ the AdaFactor (Shazeer and Stern, 2018) optimizer and do not use warm-up. An early stop strategy is applied when the loss on the validation set does not decrease in three continuous epochs. We use 4 Nvidia Tesla A100 cards with 40 GB GPU memory for training. One epoch takes around half an hour. At inference, we use beam search

[Task Description]
Given an article, your task is to incorporate several questions into the text to enhance its readability and make it more engaging.
For each question, first determine its position in the article by copying the sentence after which the question should be raised, then provide a set of keywords of the answer to the question. Finally, generate the target question.

[Input]
{article}

[Output]
Output 1:
Position: <sentence precedes the question>
Answer Keywords: <keywords separated by ",">
Question: <question 1>

Output 2:
...

Table 13: Prompt for GPT-4 question generation

decoding with a beam size of 4. All evaluations are conducted with the default parameters in their public implementations.

Since the articles in our dataset are relatively long, we truncate them and keep sentences whose indices fall within $[max(0, p_0 - 5), min(|D|, p_m + 5)]$, where $|\mathcal{D}|$ is the number of sentences in article $\mathcal{D}$, and p_0, p_m are the index of the first and last question respectively. In other words, we keep at least 10 sentences as the context for predicting each question. If the truncated document exceeds the model's context length, we split it into segments of roughly the same length and run each segment separately.

C Human Study Details

C.1 Demographic Information of Participants

The average age was 29 years, with 25 female and 20 male participants. The simplified ethnicity distribution is: 23 white, 15 black, and 7 Asian. All information is on a self-identification basis.

C.2 Reading Interface

Figure 8 shows a screenshot of the reading interface. To highlight guiding questions, we display them on the right side of the article. In order to measure paragraph-level reading time, participants need to click the "Finished" button at the end of each paragraph to reveal the next one.

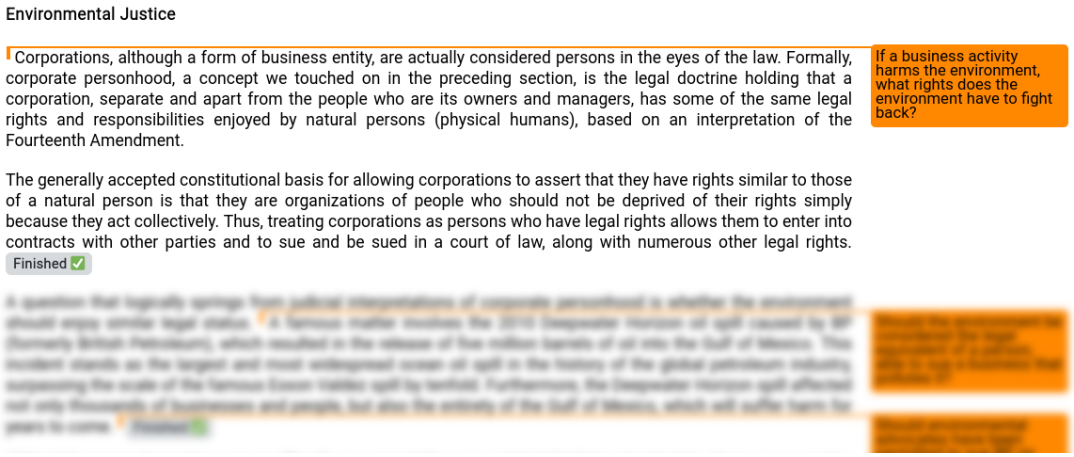

Figure 8: Main reading interface. We highlight reference or generated questions on the right side. In order to measure paragraph-level reading time, participants have to click *Finished* at the end of each paragraph in order to reveal the next one.

Engaging	The questions helped in keeping me **engaged** with the text.
Understanding	The questions improved my **understanding** of the structure and main ideas of the article.
Overall	**Overall**, I prefer to have such questions during the reading.
Relevant	Is the question is **relevant** to the context?
Position	Is the question raised at an **appropriate position** and not distracting?
Important	Is the question **important** to the central topic of the article?

Table 14: Evaluation questions used in the human study.

	Sum. Length	Sum. Time	N-gram Overlap (%)		
			uni-gram	bi-gram	tri-gram
Contr.	117	619	70.8	6.22	0.69
Auth.	123	627	74.1	6.45	0.71
Gen.	146	745	**76.0**	**8.84**	**1.44**

Table 15: Averaged summary length (words), summarization time (seconds), and n-gram overlap between summaries and articles.

C.3 Summary Evaluation

We prompt GPT-4 (`gpt4-1104-preview`) with a temperature of 0.7 to evaluate the quality of collected summaries. The prompt is shown in Table C.3, where we adopt the chain-of-thought (Wei et al., 2022) and batch evaluation (Yuan et al., 2023) prompting strategy.

C.4 Additional Summary Analysis

We provide additional analysis about collected summaries in Table 15, including summary time, length, and n-gram overlap between summaries and their sources. We can observe the reference and generated group produced longer summaries than the control group and spent longer time accordingly. The authentic group's summaries and the generated group's summaries are similar in quality, but the former is more concise. This can be attributed to the greater significance of reference questions, consistent with user ratings in Table 4. In addition, summaries from both the reference and specifically generated groups show a larger n-gram overlap, providing further evidence of improved memorization.

C.5 Observations from Preliminary Study

Before the reported human study, we conducted preliminary interviews and follow-up surveys with 15 participants, five from each group, as detailed in Sec. 6.1. The aim is to validate the experiment design and gather user preferences and feedback on presenting questions during their article reading experience. We refer to participants as `group_id` (R/C/G) + `user_id` (1-5), where R, G, and C stand for Reference, Generated, and Controlled.

Many participants expressed positive feedback on the generated questions, noting that "*the questions are easy to understand*" (G5). Notably, compared to the author-curated questions with relatively complex terms and sentences, our generated questions had simpler vocabulary and shorter sentences and facilitated quicker context comprehension. Some participants even found it helpful to use the questions to "*remember the content of the whole article*" and "*better understand it again*" (G1). We summarize the observations from the interview into three points.

Prompt Template

[Task Description]
Your will be given three summaries written for an article titled "{title}". Please act as an impartial judge and evaluate the quality of these summaries in terms of {metric}. Please make sure you read and understand the following instructions carefully. Please keep this document open while reviewing and refer to it as needed.

[Evaluation Criteria]
{metric description}

[Evaluation Steps]
Read the source article carefully and identify the main topic and key points.
Read each summary carefully. Check if the summary meets the above criteria and provide an explanation for your judgment.
Assign a {metric} score for each summary on a scale of 1 (lowest) to 5 (highest). Decimal scores are particularly encouraged.

[Input]
Article: {article}
Summary 1: {summary_1}
Summary 2: {summary_2}
Summary 3: {summary_3}

[Output] (Please strictly organize your output in the following format)
Explanations:
Analysis of summary 1: <your analysis>
Analysis of summary 2: <your analysis>
Analysis of summary 3: <your analysis>

Scores:
Score for summary 1: <score only, 1 - 5, preferably decimal>
Score for summary 2: <score only, 1 - 5, preferably decimal>
Score for summary 3: <score only, 1 - 5, preferably decimal>

{metric description}

Coherence (1-5) - the collective quality of all sentences. The summary is well-structured and well-organized. The summary should not just be a heap of related information, but should build from sentence to sentence to a coherent body of information about a topic.
Consistency (1-5) - the factual and conceptual alignment between the summary and the source article. The summary faithfully and precisely conveys the messages and ideas from the source material, without distortion or misinterpretation.
Informativeness (1-5): the extent to which the summary encapsulates the essential and relevant information. Informativeness is not merely about including various pieces of information, but selecting the most crucial elements that offer a comprehensive understanding of the topic.

Table 16: Prompt template and metric descriptions for summary evaluation.

Consistent Question Preference. While each participant has their unique criteria for defining a good question, we have observed a consistent preference for questions that are both intriguing and informative. Among the 10 participants exposed to supported questions, whether authentic or generated, seven conveyed a preference for questions that stimulate thought and reflection. These questions may offer insights (R5) or highlight specific details (G1), but they must be inherently challenging and motivate reading (G5). Conversely, participants tend to dislike questions that are too easy or have immediate answers within the article (R4, R5, G5). This observation provides valuable insights for refining our future question generation process by allowing us to better control the level of difficulty and align with participants' preferences.

Relevance, Distractibility, and Helpfulness. In the survey, we asked participants to assess the relevance, distractibility, and helpfulness of each question on a 5-point Likert scale. Half believe a question's relevance depends on whether the following sentences answer it. Meanwhile, shorter questions tend to be less distracting, with their distractibility rating inversely proportional to perceived helpfulness. Interestingly, conflicting viewpoints arose, with some participants considering certain questions neither distracting but useless (R1) or helpful yet irrelevant (G1, G4). This nuanced understanding of human complexity calls for more research in human-AI collaboration research to find adaptive question generation solutions.

Question Position and Scenario Matters. The two observations mentioned above may vary slightly depending on the question's position and the reading scenario. Notably, five participants experienced a "cold start" during the reading task, expressing difficulty in "*getting into the context of an article at first*" (R1). Therefore, their initial preference leans towards easier and more relevant questions at the article's beginning. As they familiarize themselves with the context after reading a few paragraphs, their inclination shifts towards more intriguing and divergent questions. Furthermore, participants' standards for a good question may fluctuate in different scenarios. Distinctions were made between first-time reading versus content reviewing (R2), learning scenarios versus examination scenarios (R4), and serious learning versus casual reading (G4). These considerations could serve as additional factors in our future iterations.