QUDSELECT: Selective Decoding for Questions Under Discussion Parsing

Ashima Suvarna[♡*] **Xiao Liu**[◇*] **Tanmay Parekh**[♡] **Kai-Wei Chang**[♡] **Nanyun Peng**[♡]
[♡] Computer Science Department, University of California, Los Angeles
[◇] Wangxuan Institute of Computer Technology, Peking University
{asuvarna31,tparekh,kwchang,violetpeng}@cs.ucla.edu
lxlisa@pku.edu.cn

Abstract

Question Under Discussion (QUD) is a discourse framework that uses implicit questions to reveal discourse relationships between sentences. In QUD parsing, each sentence is viewed as an answer to a question triggered by an anchor sentence in prior context. The resulting QUD structure is required to conform to several *theoretical criteria* like answer compatibility (how well the question is answered), making QUD parsing a challenging task. Previous works construct QUD parsers in a pipelined manner (i.e. detect the trigger sentence in context and then generate the question). However, these parsers lack a holistic view of the task and can hardly satisfy all the criteria. In this work, we introduce **QUDSELECT**, a joint-training framework that selectively decodes the QUD dependency structures considering the QUD criteria. Using instruction-tuning, we train models to simultaneously predict the anchor sentence and generate the associated question. To explicitly incorporate the criteria, we adopt a selective decoding strategy of sampling multiple QUD candidates during inference, followed by selecting the best one with criteria scorers. Our method outperforms the state-of-the-art baseline models by 9% in human evaluation and 4% in automatic evaluation, demonstrating the effectiveness of our framework. Code and data are in https://github.com/asuvarna31/qudselect.

1 Introduction

Discourse structure describes the relationships between different sentences of an article or conversation. The ability to understand discourse structure is crucial for natural language processing tasks such as text summarization (Durrett et al., 2016), conditional generation (Narayan et al., 2023), and narrative understanding (Xu et al., 2024). Recent works have adapted the Question Under Discussion (QUD) framework to analyze discourse structures (Benz and Jasinskaja, 2017; Riester et al., 2021). In the QUD framework (Van Kuppevelt, 1995; Roberts, 2012), the relationships between sentences in an article are characterized by (implicit) free-form questions. Each question is evoked by an anchor sentence in prior context, and answered by an answer sentence in the subsequent content. For instance, in Figure 1, the relationship between sentence 3 (referred to as s_3) and the previous context is that s_3 answers the question "Which movie has the most Oscar nominations?" evoked by the anchor sentence s_1.

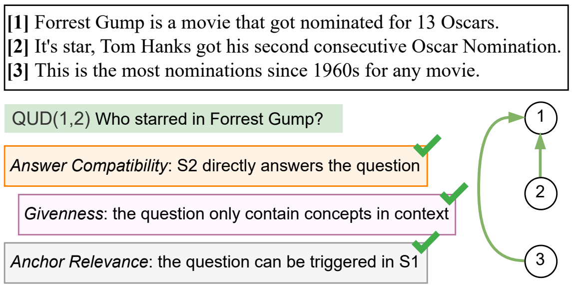


Figure 1: An article snippet along with the associated QUD dependency structure. Each edge from s_i to s_j with attribute q indicates sentence s_j anchors the question q, and sentence s_i answers the question q.

The QUD structures involve contextually-grounded questions that adhere to three theoretical criteria (De Kuthy et al., 2018; Wu et al., 2023; Riester et al., 2018): a) *answer compatibility*: the question must be answerable by the answer sentence in the discourse, like s_2 directly answers the question "Who starred in Forrest Gump?" in Figure 1; b) *givenness*: the question should only contain concepts that are accessible to the reader from prior context or common knowledge, like "Forrest Gump" in the question; and c) *anchor relevance*: the question should be relevant to the anchor sentence, e.g., the aforementioned question can be triggered in s_1.

[*]Equal contribution.

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 1288–1299
November 12-16, 2024 ©2024 Association for Computational Linguistics

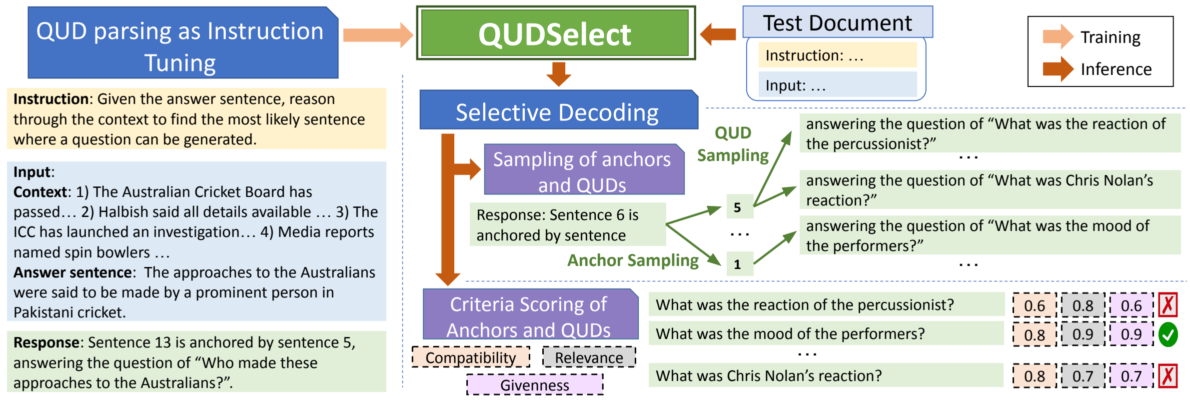


Figure 2: Overview of our QUDSELECT framework.

Previous works on QUD parsing break down the task into two steps: anchor selection and question generation. De Kuthy et al. (2020) develop a rule-based method for the question generation step, Ko et al. (2023) train task-specific models for each step, while Wu et al. (2023) prompt large language models (LLMs) in a stepwise manner. However, these approaches lack a holistic view of the task, causing the predicted QUDs to often fail to satisfy all the criteria. For instance, GPT-4 fails to generate questions that are fully grounded on the anchor sentence in 50% of the cases.[1]

To address these challenges, we propose QUDSELECT, a joint-training framework that selectively decodes QUD structures by incorporating the criteria, as shown in Figure 2. Specifically, we instruction-tune models to jointly predict the anchor sentence and the corresponding question given an answer sentence (e.g., s_{13}) and prior context (e.g., $s_1, \ldots, s_{12}$ of the article). We propose selective decoding where we sample multiple anchor and question pairs, score them using criteria scorers, and finally, select the best scored pair.

Experiments conducted on the DCQA (Ko et al., 2022) dataset show that QUDSELECT outperforms baselines by ~9% on average in human evaluation. To reduce resource and cost-intensive expert evaluation, we develop automatic evaluators trained on human annotations, and conduct a larger-scale automatic evaluation. The automatic evaluation results show that QUDSELECT achieves around a ~4% improvement over the selected baselines. Further analyses reveal that the performance could be further improved with more selected candidates.

[1]This is observed from the human annotations in the QUD evaluation dataset QUDEVAL (Wu et al., 2023).

2 Related Work

QUD is a linguistic framework that analyzes discourse and pragmatics by viewing each sentence as an answer to an implicit question triggered in prior context (Van Kuppevelt, 1995; Roberts, 2012; Benz and Jasinskaja, 2017). While theoretical discussions around QUDs relied on constructed examples, Riester (2019) introduced an annotation framework for reconstructing QUDs from data. Westera et al. (2020), Ko et al. (2022) and Hesse et al. (2020) annotated Ted-talk transcripts and news articles respectively in an expectation-driven manner, where questions are triggered while reading (i.e., unseen discourse progression) while De Kuthy et al. (2018) annotated two interview transcripts with full, hierarchical questions.

Recent works have begun adapting QUD for automatic discourse parsing (Ko et al., 2022, 2023; Wu et al., 2023), narrative graph construction (Xu et al., 2024) and decontextualization of scientific documents (Newman et al., 2023). QUD fits well for understanding the structure and coherence of texts that are intended to provide argumentation (Liu et al., 2024) and complex reasoning (Hu et al., 2022), and has potential applications to enhance document understanding in information extraction (Parekh et al., 2023, 2024a; Huang et al., 2024) with applications in wider domains like epidemiology (Parekh et al., 2024b) and biomedical science (Ma et al., 2023). Ko et al. (2023) introduced a QUD parser trained on DCQA (Ko et al., 2022) that consists of an anchor selection and a question generation pipeline. Wu et al. (2023) evaluated QUDs generated by LLMs by few-shot prompting in a two-step manner: question generation followed by anchor generation. Xu et al. (2024) followed

Model	Answer Compatibility			Givenness			Anchor Relevance			Avg. (↑)
	Dir. (↑)	Unfocus.	No Ans.(↓)	No New (↑)	Ans. leak. (↓)	Hall. (↓)	Fully G. (↑)	Partial. G.	No G. (↓)	
AUTOMATIC EVALUATION										
Pipeline	68.2	4.5	27.3	83.7	10.0	6.3	63.6	0.0	36.4	71.8
LLaMA2-7B	67.4	12.9	19.7	88.3	6.7	5.0	52.7	17.7	29.6	69.5
+ QUDSELECT	70.4	8.2	21.4	**91.8**	**6.0**	**2.2**	61.0	12.4	26.6	74.4
Mistral-7B	71.4	8.7	19.9	89.3	**6.0**	4.7	58.0	15.9	26.1	72.9
+ QUDSELECT	<u>74.1</u>	9.0	<u>16.9</u>	86.5	7.2	6.2	**68.3**	11.0	<u>20.7</u>	<u>76.3</u>
GPT-4	**92.7**	3.3	**4.0**	78.7	18.9	2.4	51.9	32.0	16.1	74.4
+ QUDSELECT	90.0	4.1	5.9	80.0	15.0	5.0	62.5	21.4	**16.0**	**77.5**
HUMAN EVALUATION										
Pipeline	52.5	15.0	32.5	53.8	28.7	17.5	50.0	32.5	17.5	52.1
Mistral-7B	67.0	15.4	17.6	60.3	23.6	16.1	58.6	29.0	12.4	62.0
+ QUDSELECT	**67.1**	20.0	**12.9**	**77.6**	**20.0**	**2.4**	**68.2**	24.7	**7.1**	**71.0**

Table 1: Automatic and human evaluation results. Numbers are in percentages (%). Best results are in bold, and the best results of open-source models (if not the best overall) are underlined. Avg. indicates the average ratio of ideal QUDs (the first option of each criterion). We abbreviate Direct Answer as Dir. Ans., Indirect Answer as Indir. Ans., Answer Leakeage as Ans. Leak., Hallucination as Hall., and Grounded as G.

a QUD style annotation for generating narrative graphs by incorporating retrospective questions triggered from succeeding context.

3 The QUDSELECT Framework

Task Formulation Given a document with n sentences $D = \{s_1, s_2, \ldots, s_n\}$, QUD parsing aims to build a QUD dependency tree. We formulate the QUD parsing task as edge-level prediction following previous works (De Kuthy et al., 2018; Ko et al., 2023): given an answer sentence $s_i \in \{s_2, \ldots, s_n\}$[2], models are asked to predict the anchor sentence $a_i \in \{s_1, \ldots, s_{i-1}\}$ and generate the question q_i.

Overview Figure 2 illustrates the structure of our QUDSELECT framework. We first instruction tune a joint QUD parser §3.1. Then, we propose selective decoding §3.2 to select the best candidate from sampled ⟨*anchor sentence, question*⟩ pairs.

3.1 QUD Parser Training

Unlike previous works that use separate models for anchor prediction and question generation, we exploit the instruction following ability of LLMs (Wang et al., 2022) to perform these two steps *jointly*, as demonstrated in Figure 2 (left). This joint inference provides the model with a holistic view of the task. Given the answer sentence s_i and context of sentences prior to s_i, models are instructed to output the anchor a_i and the question q_i. We provide the instruction-response template in Appendix A.

[2]The first sentence s_1 is the root of the QUD dependency tree, and does not anchor on any other sentence

3.2 Selective Decoding

To incorporate specific criteria during inference, we sample multiple ⟨*anchor sentence, question*⟩ candidates and select the best one by using simple criteria scorers.

To generate multiple QUD candidates for a context $\{s_1, \ldots, s_{i-1}\}$ and an answer sentence s_i, we sample multiple anchor sentences and question candidates by *selectively* utilizing beam-search with a wide beam while decoding. Following prior work (De Kuthy et al., 2018; Benz and Jasinskaja, 2017; Wu et al., 2023), we assume that every answer sentence has a corresponding question. First, for anchor prediction, we prompt the model with *sentence* s_i *is anchored by sentence* using a beam size k to generate k possible anchors. Post deduplication of anchor candidates, we again utilize beam-search with size k to generate k question candidates for each anchor sentence. This encourages diversity in both the prediction of anchor sentences and questions.

We apply m criteria $\mathcal{C} = \{c_1, \ldots, c_m\}$ to assess the quality of generated candidates from different aspects. Each criterion assigns a score $c_j(a, q) \in [0, 1]$ to a candidate $\langle a, q \rangle$, and the overall score is the summation of all criteria $\Sigma_{j=1}^{m}(c_j(a, q))$. The candidate with the highest overall score is selected as the final prediction.

Criteria Scorers. We consider the three key principles of QUD as our criteria: answer-compatibility, givenness, and anchor relevance. We implement *reference-free* and *training-free* scorers for each of them.

Answer Compatibility: This criterion indicates that the question q should be answerable by the answer sentence s_i. We regard this as a natural language inference (NLI) task, and use the probability that s_i entails q measured by an off-the-shelf NLI model (`bart-large-mnli`) as the compatibility score.

Givenness: This criterion evaluates if the question only consists of information from the context. An ideal question should be naturally invoked from the context, without concepts that appear out of thin air. We measure the givenness with content word overlap between q and the context $s_{1...i-1}$. We extract lemmas L_q and L_c of all content words (nouns, verbs, adjectives, and adverbs) in the question and the context, and compute the givenness score as $|L_q \cap L_c|/|L_q|$.

Anchor Relevance: This criterion measures if the question q is relevant to the anchor sentence a. Similar to the givenness score, we approximate it with content word overlap between a and the focus of q. We regard the maximum noun phrase of q as its focus f_q, and extract lemmas L_{fq} and L_a of all content words in f_q and a. The relevance score is computed as $|L_{fq} \cap L_a|/|L_{fq}|$.

4 Experimental Setup

Models and Datasets We utilize the DCQA dataset (Ko et al., 2022) for training and evaluating QUD parsers. The DCQA dataset consists of 22k English questions across 606 news articles. We use two instruction-tuned models LLaMA2-7B (Touvron et al., 2023) and Mistral-7B (Jiang et al., 2023) as base models of our framework. To explore the effectiveness of selective decoding on closed-source models, we also apply it to GPT-4 (Achiam et al., 2023). We sample $k = 10$ candidates for each answer sentence. Implementation details can be found in Appendix A.

Baselines We compare against two existing QUD parsers: the Pipeline approach (Ko et al., 2023) and GPT-4 prompting (Wu et al., 2023). We also provide ablation of not using selective decoding during inference, i.e., QUDSELECT with $k = 1$.

Human Evaluation We follow the annotation guidelines outlined in QUDEVAL (Wu et al., 2023) and evaluate the generated QUDs for answer compatibility, givenness, and anchor relevance. Detailed classification of the criteria is in Appendix B. We evaluate 100 questions across 8 articles from the DCQA test set. We recruit three annotators from Amazon's Mechanical Turk after extensive training and screening. We report the majority vote results and achieve an average inter-annotator agreement of 68.3% averaged across all evaluated dimensions. More details are in Appendix C.

Automatic Evaluation While human evaluation is more accurate for evaluating the efficacy of QUD parsing models, it is time-consuming and expensive to collect at scale. To this end, we apply supervised classifiers to judge the generated QUDs. Specifically, we train RoBERTa classifiers (Liu et al., 2019) on the expert annotated data in QUDEVAL for answer compatibility and anchor relevance, and Longformer (Beltagy et al., 2020) for givenness due to the longer context length. We achieve a macro F1 score of 0.48 for answer compatibility, 0.42 for givenness, and 0.53 for anchor relevance, outperforming or matching the best existing automatic evaluators. Detailed comparisons with other evaluators are in Appendix D. We conduct the automatic evaluation on on 400 questions per model across 22 articles from the entire DCQA test set.

5 Results and Analysis

5.1 Main Results

Automatic Evaluation Results. Table 1 (top) reports the automatic evaluation results. QUDSELECT (Mistral-7B) outperforms the previously established pipeline baseline on all the three criteria. And QUDSELECT improves the performance of instruction tuned Mistral-7B, LLaMA2-7B and GPT-4, leading to $\sim 4\%$ improvement over models without QUDSELECT.

Human Evaluation Results Table 1 (bottom) reports the human evaluation results. We compare the best open-source model from Table 1, QUDSELECT (Mistral-7B), with Pipeline and Mistral-7B. QUDSELECT (Mistral-7B) generates 67% directly answered questions, 78% questions with no unseen concepts, and 68% fully grounded questions. This highlights the effectiveness of our framework in generating QUDs that satisfy the desired criteria.

Error Analysis Our detailed classifications of the evaluation metrics (Appendix §B) allow us to categorize the various errors made by the models. We find from Table 1 that GPT-4 generates higher percentage of directly answered QUDs but these QUDs are more likely to have answer leakage errors. This indicates that GPT-4 tends to include

QUDSELECT (Mistral)	
Answer: s_3 Anchor: s_1 QUD: “Why is it important that U.S. exports of nuclear material cannot be adequately traced from country to country?”	✓Direct answer ✓No new concepts ✓Fully grounded
Answer: s_4 Anchor: s_2 QUD: “Who commissioned the report?”	✓Direct answer ✓No new concepts ✓Fully grounded
Pipeline (Ko et al. (2023))	
Answer: s_3 Anchor: s_2 QUD: “What does Glenn think is the future outlook on nuclear materials?”	✗Non answer ✗Answer leakage ✓Partially grounded
Answer: s_4 Anchor: s_2 QUD: “Who is the Sen. Glenn from?”	✗Nonsensical question

Table 2: Example QUDs generated by QUDSELECT (Mistral) and the pipeline method for a test article. The full article text can be found in Appendix Figure 5. s_i indicates the i-th sentence in the article.

aspects from the answer sentence in the question that increases the answer compatibility but reduces the givenness. We find that QUDSELECT improves GPT-4 performance by reducing the answer leakage error and improving the relevance of the anchor. Overall, we find that QUDSELECT improves the validity of the answers and increases the grounding of the questions in the anchor which leads to performance improvements for all models.

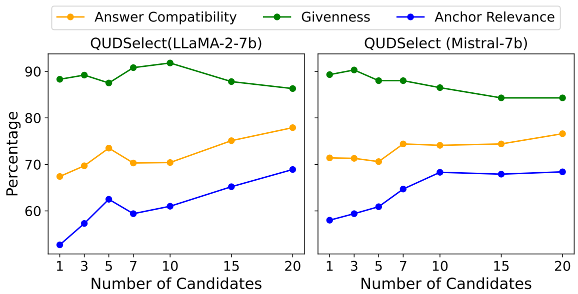


Figure 3: Hyperparameter analysis on the number of candidates. QUDSELECT shows improved performance with an increased number of candidates.

5.2 Hyperparameter Study

To study the performance sensitivity of QUDSELECT to the number of candidates k, we vary k from 1 to 20 for QUDSELECT (LLaMA2-7B) and QUDSELECT (Mistral-7B) and show the performance in Figure 3. The performance reveals an upward trend as k grows for Answer Compatibility and Anchor Relevance while Givenness is sacrificed by a small margin for better overall performance. With $k = 10$, QUDSELECT significantly outperforms the selected baselines without significant runtime overhead.

5.3 Case Study

In Table 2, we show the QUDs generated by QUDSELECT (Mistral-7B) and the Pipeline model for a news article (Appendix Figure 5) along with the human annotations for each question. Most QUDs generated by QUDSELECT (Mistral-7B) are explicitly answerable, include no unseen concepts, and are fully grounded in the anchor. In contrast, the Pipeline method generates incomplete questions or incompatible question-answer pairs for the given article. This demonstrates the overall effectiveness of QUDSELECT in generating high-quality QUDs.

6 Conclusion

In this work, we propose QUDSELECT, a joint framework for generating QUD structures by integrating key *theoretical criteria*. To achieve this, we reformulate the QUD parsing as an instruction tuning task and selectively decode the candidate questions and anchors. Furthermore, we develop automated evaluation methods trained on expert annotations to reduce the reliance on labor-intensive expert evaluations and facilitate model development for QUD parsing. Experiments demonstrate that QUDSELECT significantly outperforms baselines in both automatic and human evaluations.

Acknowledgements

We thank Hritik Bansal and Sidi Lu for their constructive comments. We thank the anonymous reviewers for their helpful discussions and suggestions. Our work was supported by Optum Labs, Amazon Alexa AI Research Award, an Amazon Research Award via UCLA Science Hub and the Amazon Fellowship (Tanmay Parekh) and we express our gratitude for their support.

Limitation

QUDSELECT generates the QUD structure as a dependency tree where each sentence is connected to a prior context via a question. This does not guarantee the generation of full, hierarchical QUDs where the answer of a QUD entails the answer of its descendants (Roberts, 2012). Furthermore, QUDSELECT generates each QUD edge independently and does not model the relationships between questions. Thus, we leave the exploration of such discourse level constraints to future work.

Sampling Cost. Although the time cost increases when sampling more candidates for QUDSELECT, the number of sampled unique anchors does not increase, due to the limited number of reasonable anchors in an article. The average number of unique anchors is less than 3 when $k = 20$. Therefore, the growth of sampling cost is approximately linear to k. We find that increasing the number of candidates leads to an increase in the model performance §5.2.

Ethical Consideration

Our framework relies on open-source and closed-source LLMs that may generate harmful and biased outputs. Therefore, it should be used with human supervision. For human evaluation, we recruit annotators from Amazon's Mechanical Turk, and all annotators are fairly paid more than $15 USD per hour (which varies depending on the time spent per HIT), which is higher than the national minimum wage where the annotators are recruited.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational linguistics*, 34(4):555–596.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

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

Kordula De Kuthy, Madeeswaran Kannan, Haemanth Santhi Ponnusamy, and Detmar Meurers. 2020. Towards automatically generating questions under discussion to link information and discourse structure. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5786–5798.

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)*.

Greg Durrett, Taylor Berg-Kirkpatrick, and Dan Klein. 2016. Learning-based single-document summarization with compression and anaphoricity constraints. *arXiv preprint arXiv:1603.08887*.

Christoph Hesse, Anton Benz, Maurice Langner, Felix Theodor, and Ralf Klabunde. 2020. Annotating quds for generating pragmatically rich texts. In *Proceedings of the Workshop on Discourse Theories for Text Planning*, pages 10–16.

Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2021. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Nan Hu, Zirui Wu, Yuxuan Lai, Xiao Liu, and Yansong Feng. 2022. Dual-channel evidence fusion for fact verification over texts and tables. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5232–5242.

Kuan-Hao Huang, I-Hung Hsu, Tanmay Parekh, Zhiyu Xie, Zixuan Zhang, Prem Natarajan, Kai-Wei Chang, Nanyun Peng, and Heng Ji. 2024. TextEE: Benchmark, reevaluation, reflections, and future challenges in event extraction. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 12804–12825, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

AQ Jiang, A Sablayrolles, A Mensch, C Bamford, DS Chaplot, D de las Casas, F Bressand, G Lengyel, G Lample, L Saulnier, et al. 2023. Mistral 7b (2023). *arXiv preprint arXiv:2310.06825*.

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.

Wei-Jen Ko, Yating Wu, Cutter Dalton, Dananjay Srinivas, Greg Durrett, and Junyi Jessy Li. 2023. Discourse analysis via questions and answers: Parsing dependency structures of questions under discussion. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 11181–11195, Toronto, Canada. Association for Computational Linguistics.

Xiao Liu, Yansong Feng, and Kai-Wei Chang. 2024. Casa: Causality-driven argument sufficiency assessment. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5282–5302.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Mingyu Derek Ma, Alexander Taylor, Wei Wang, and Nanyun Peng. 2023. DICE: Data-efficient clinical event extraction with generative models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15898–15917, Toronto, Canada. Association for Computational Linguistics.

Aman Madaan, Amrith Setlur, Tanmay Parekh, Barnabas Poczos, Graham Neubig, Yiming Yang, Ruslan Salakhutdinov, Alan W Black, and Shrimai Prabhumoye. 2020. Politeness transfer: A tag and generate approach. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1869–1881, Online. Association for Computational Linguistics.

Shashi Narayan, Joshua Maynez, Reinald Kim Amplayo, Kuzman Ganchev, Annie Louis, Fantine Huot, Anders Sandholm, Dipanjan Das, and Mirella Lapata. 2023. Conditional generation with a question-answering blueprint. *Transactions of the Association for Computational Linguistics*, 11:974–996.

Benjamin Newman, Luca Soldaini, Raymond Fok, Arman Cohan, and Kyle Lo. 2023. A question answering framework for decontextualizing user-facing snippets from scientific documents. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3194–3212.

Tanmay Parekh, I-Hung Hsu, Kuan-Hao Huang, Kai-Wei Chang, and Nanyun Peng. 2023. GENEVA: Benchmarking generalizability for event argument extraction with hundreds of event types and argument roles. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3664–3686, Toronto, Canada. Association for Computational Linguistics.

Tanmay Parekh, I-Hung Hsu, Kuan-Hao Huang, Kai-Wei Chang, and Nanyun Peng. 2024a. Contextual label projection for cross-lingual structured prediction. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5738–5757, Mexico City, Mexico. Association for Computational Linguistics.

Tanmay Parekh, Anh Mac, Jiarui Yu, Yuxuan Dong, Syed Shahriar, Bonnie Liu, Eric Yang, Kuan-Hao Huang, Wei Wang, Nanyun Peng, and Kai-Wei Chang. 2024b. Event detection from social media for epidemic prediction. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5758–5783, Mexico City, Mexico. Association for Computational Linguistics.

Arndt Riester. 2019. Constructing qud trees. In *Questions in discourse*, pages 164–193. Brill.

Arndt Riester, Lisa Brunetti, and Kordula De Kuthy. 2018. Annotation guidelines for questions under discussion and information structure. *Information structure in lesser-described languages: Studies in prosody and syntax*, pages 403–443.

Arndt Riester, Amalia Canes Nápoles, and Jet Hoek. 2021. Combined discourse representations: Coherence relations and questions under discussion. In *Proceedings of the First Workshop on Integrating Perspectives on Discourse Annotation*, pages 26–30.

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

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models.

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

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Anjana Arunkumar, Arjun Ashok, Arut Selvan Dhanasekaran, Atharva Naik, David Stap, et al. 2022. Super-naturalinstructions: Generalization via declarative instructions on 1600+ nlp tasks. *arXiv preprint arXiv:2204.07705*.

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.

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.

Liyan Xu, Jiangnan Li, Mo Yu, and Jie Zhou. 2024. Graph representation of narrative context: Coherence dependency via retrospective questions. *arXiv preprint arXiv:2402.13551*.

Da Yin, Xiao Liu, Fan Yin, Ming Zhong, Hritik Bansal, Jiawei Han, and Kai-Wei Chang. 2023. Dynosaur: A dynamic growth paradigm for instruction-tuning data curation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4031–4047.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

A QUDSELECT Implementation Details

We instruction-tune QUD parsers in the format of Figure 4. Similar to Yin et al. (2023), we apply LORA (low-rank adaptation, Hu et al. (2021)) with learning rate $2e-5$, $lora_{rank}=256$, and $lora_{alpha}=256$. Models are trained for 2 epochs with batch size 128. During inference, we sample QUD candidates with k beams and temperature 1. All the experiments are performed with 48GB NVIDIA A6000 GPUs.

Instruction: Given the answer sentence, reason through the context to find the most likely sentence where a question can be generated.

Input:
Context: {context}
Answer sentence: {Answer}

Response: Sentence {Answer ID} is anchored by sentence {Anchor ID}, answering the question of "{Question}".

Figure 4: Prompt format for instruction tuning QUD parsers.

B Evaluation Criteria Details

We follow the evaluation protocol outlined in (Wu et al., 2023) for our human and automatic evaluation.

- Answer Compatibility: This criterion indicates that the question q should be answerable by the answer sentence s_i. For evaluation, we classify each $q-s_i$ pair as a) *Direct and Explicit Answer (Dir.):* s_i answers the q explicitly, b) *Unfocused (Unfocus.):* some parts of s_i answer q indirectly, or c) *Not Answered:* s_i does not answer q.

- Givenness: This criterion evaluates if the question only consists of information from the context. An ideal question should be naturally evoked from the context, without concepts that are not accessible to the reader from common knowledge. This criterion has the following categories *a) No new concepts (No New):* q does not contain any concepts beyond the context or common knowledge, b) *Answer leakage (Ans. leak.):* q contains concepts that are not in the context but in s_i, c) *Hallucination (hall.):* q contains new concepts that are not answer-leakage.

- Anchor Relevance: This criterion measures if the question q is relevant to and naturally evoked from the anchor sentence a. This criterion has the following categories *a) Fully Grounded (Fully G.):* q contains concepts from anchor a, b) *Partially Grounded (Partial G.):* q contains some concepts from anchor a and is not directly addressing the focus of a, c) *Not grounded (No G.)*: q is completely irrelevant to a.

C Human Evaluation Details

We provide the annotation template and training materials in Figure 6 and 7. All annotators were recurited from Amazon's Mechanical Turk and fairly paid more than $15 USD per hour which varied depending on the time spend per HIT (more than the national minimum wage where the annotators are recruited). To ensure high quality annotations, the annotators were provided with extensive guidelines and training (Figure 7).

We measure inter-annotator agreement with Krippendorff's α. As shown in Table 3, annotators achieve "moderate" agreement across Answer Compatibility and Givenness (Artstein and Poesio, 2008). Since, relevance between two concepts (question and achor) is highly dependent on the annotators' comprehension of the article, we find that agreement score for Anchor Relevance is "fair" (Artstein and Poesio, 2008). We also note the pairwise agreement in Table 3. The agreements are comparable with those in QUDEVAL, and indicate a certain degree of subjectivity in QUD analysis.

	Comp.	**Givn.**	**Relv.**
Pair-Wise Agreement	70.0%	75.0%	60.0%
Krippendorff's α	0.68	0.64	0.43

Table 3: Inter-annotator agreement for human judges.

D Automatic Evaluator Details

We train automatic evaluators with the human annotations from QUDEVAL. Experienced human annotators assess the answer compatability, giveness, and anchor relevance of 2,040 machine-generated QUDs from 51 articles. We randomly split the articles into training/validation/test sets with the ratio of 60%/15%/25%.

We fine-tune classifiers for each criterion individually. Similar to Madaan et al. (2020), we use RoBERTa-large (Liu et al., 2019) as the backbone model of answer compatibility and anchor relevance, and Longformer-base (Beltagy et al., 2020) as the backbone model of givenness due to the longer context length. For answer compatibility, the input to the model is the question and the answer sentence, and the output is one of the three labels *Dir-Ans.*, *Unfocus.*, and *Not-Ans.* For givenness, the input is the context (sentences before the anchor sentence in the article) and the question, and the answer is one of the three labels *No-New.*, *Ans-leak.*, and *Hallu.* For anchor relevance, the input is the question and the anchor sentence, and the output is one of the three labels *Full.*, *Some.*, and *No-G.* Models are fine-tuned for 10 epochs with the learning rate $1e-5$ and batch size 32.

We report the F1 scores of our automatic evaluators in Table 4. For reference, we also provide the F1 scores of the random baseline, and the best reference-free and reference-based metrics from QUDEVAL (Wu et al., 2023). GPT-Scr (w/o ref) and GPT-Scr (w/ ref) indicate prompting GPT-4 to score without and with the human-annotated reference QUD. BERTScore means calculating the similarity between the candidate and reference QUD with BERTScore (Zhang et al., 2019). The rule-based method checks if all content words in the candidate question are presented in the context. Please refer to the QUDEVAL paper for more details. Note that the results of random and ours are conducted on our held-out test set, while the results of baseline evaluators are conducted on two held-out articles. Our evaluators are better than or comparable with the baselines, highlighting the credibility of using them in automatic evaluation.

Compatibility	Dir-Ans.	Unfocus.	Not-Ans.	Macro F1
Random	0.68	0.03	0.15	0.29
GPT-Scr (w/o ref)	0.70	0.05	0.36	0.37
BERTScore	0.51	0.14	0.43	0.36
Ours	**0.84**	**0.28**	**0.32**	**0.48**
Givenness	No-New.	Ans-leak.	Hallu.	Macro F1
Random	0.65	0.29	0.10	0.35
Rule-based	0.52	**0.40**	0.19	0.37
GPT-Scr (w/ ref)	0.65	0.35	0.1	0.37
Ours	**0.74**	0.23	**0.30**	**0.42**
Relevance	Full.	Some.	No-G.	Macro F1
Random	0.52	0.22	0.21	0.32
GPT-Scr (w/o ref)	0.73	**0.41**	**0.57**	**0.57**
GPT-Scr (w/ ref)	0.63	0.26	0.22	0.37
Ours	**0.79**	0.32	0.48	0.53

Table 4: Automatic evaluator assessment in F1.

E Evaluating the Correctness of the Selected Anchor

In §4 we focus on three criteria: answer compatibility, givenness and anchor relevance. We highlight that anchor relevance refers to the measure of relevance between the question and anchor (§B. Therefore, in our evaluation framework we evaluate the correctness of the selected anchor as how relevant it is to the question. An anchor that is incorrect or not relevant would be considered "not-grounded". From Table 1, we see that QUDSELECT reduces the percentage of not grounded questions generated by the model and therefore improves the overall quality of the QUDs generated. To further analyse the correctness of the anchor selection we report the agreement accuracy (Table 5) of the the selected anchor sentences with the human annotated anchors from the DCQA dataset. Note that this is a partial notion of accuracy and does not accurately represent the quality of a model, since it is natural for different questions to be triggered from different sentences (Ko et al., 2023).

Model	**Anchor Agreement**
Pipeline	47.9%
LLaMA2-7B	48.7%
+ QUDSELECT	45.7%

Table 5: Anchor agreement score between the selected anchor and the human-annotated anchors from the DCQA dataset.

F Article of Case Study

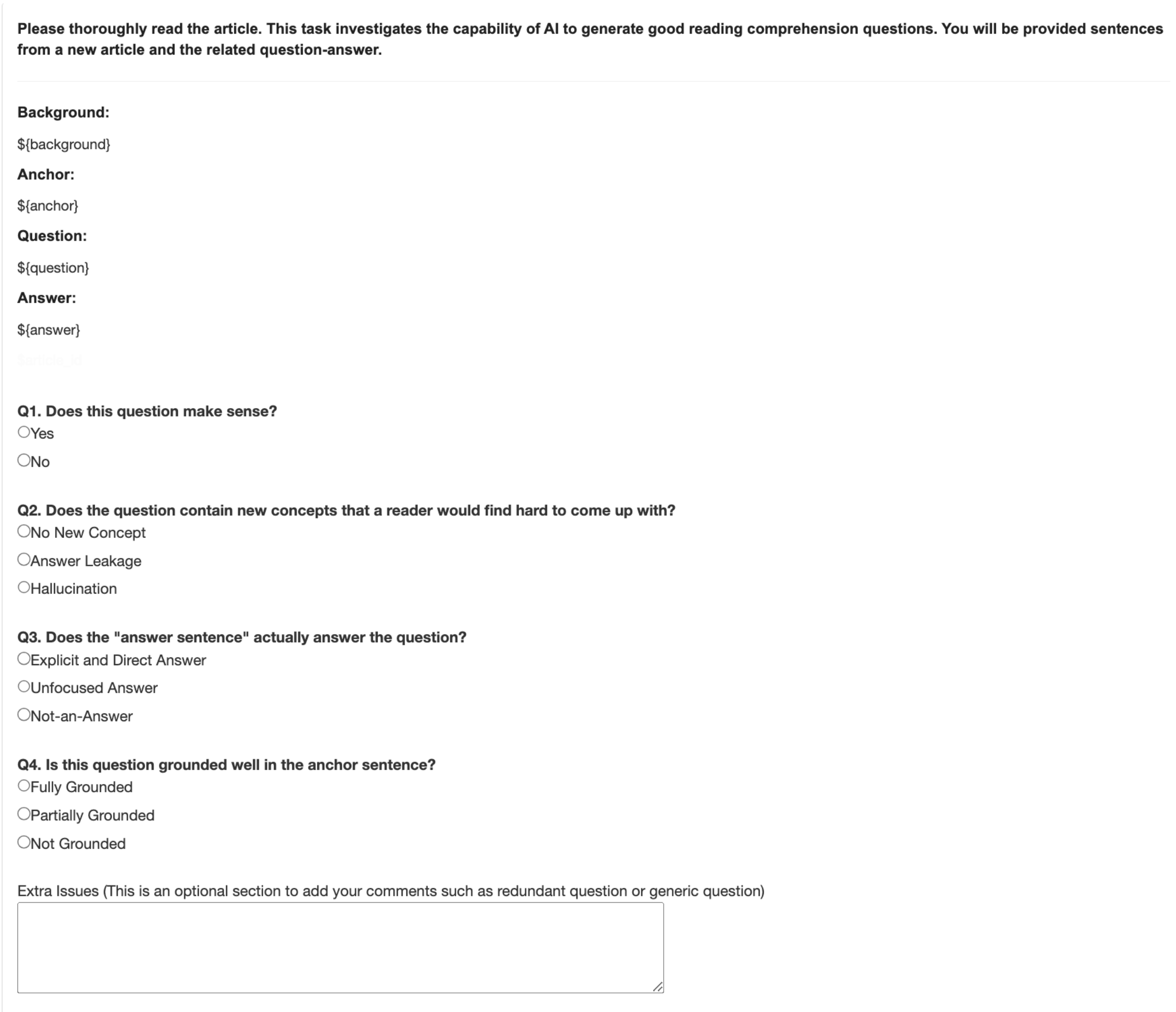
We provide the article snippet used in the case study in Figure 5. The article is from the DCQA dataset. We also provide questions generated by other models in Table 6.

1. U.S. exports of nuclear material cannot be adequately traced from country to country, according to a congressional report.
2. 'Scarcely a day goes by without a report of a new black market deal,' said Sen. John Glenn in a statement reacting to the report.
3. 'Given the staggering amount of nuclear materials we have exported, it could only be a matter of time before some of this deadly contraband proves to be of U.S. origin.'
4. As chairman of the Senate Committee on Governmental Affairs in the last Congress, Glenn commissioned the report from the General Accounting Office, which conducts investigations for legislators.
5. The report says hundreds of tons of plutonium and highly enriched uranium have accumulated worldwide, mostly from nuclear power generation.

Figure 5: Article snippet used in case study.

LLaMA2	
Answer: s_4 Anchor: s_3 QUD: “What is deadly contraband?”	✗Non answer ✓No new concepts ✗Partially grounded
Answer: s_3 Anchor: s_1 QUD: “Why is it difficult to trace nuclear material?"”	✗Non answer ✓No new concepts ✓Fully grounded
QUDSELECT (LLaMA2)	
Answer: s_4 Anchor: s_2 QUD: “Who requested the report?”	✓Direct answer ✓No new concepts ✓Fully grounded
Answer: s_3 Anchor: s_1 QUD: “What is the reason for the inability to trace nuclear material?”	✓Indirect Answer ✓No new concepts ✗Partially grounded
GPT4	
Answer: s_6 Anchor: s_6 QUD: “What does the congressional report reveal about the quantity of nuclear material that has accumulated globally?”	✗Generated the answer as the anchor and led to answer leakage
Answer: s_4 Anchor: s_2 QUD: “Who was responsible for commissioning the report on the traceability of U.S. nuclear material exports?”	✓No new concepts ✓Fully grounded

Table 6: Example QUDs generated by different models. The full article text can be found in Appendix Figure 5. s_i indicates the i-th sentence in the article.


Please thoroughly read the article. This task investigates the capability of AI to generate good reading comprehension questions. You will be provided sentences from a new article and the related question-answer.

Background:

${background}

Anchor:

${anchor}

Question:

${question}

Answer:

${answer}

Q1. Does this question make sense?

○Yes

○No

Q2. Does the question contain new concepts that a reader would find hard to come up with?

○No New Concept

○Answer Leakage

○Hallucination

Q3. Does the "answer sentence" actually answer the question?

○Explicit and Direct Answer

○Unfocused Answer

○Not-an-Answer

Q4. Is this question grounded well in the anchor sentence?

○Fully Grounded

○Partially Grounded

○Not Grounded

Extra Issues (This is an optional section to add your comments such as redundant question or generic question)

Figure 6: The annotation template for human evaluation. We ask annotators to classify the given QUD, anchor and answer for Givenness, Answer Compatibility, and Anchor Relevance.

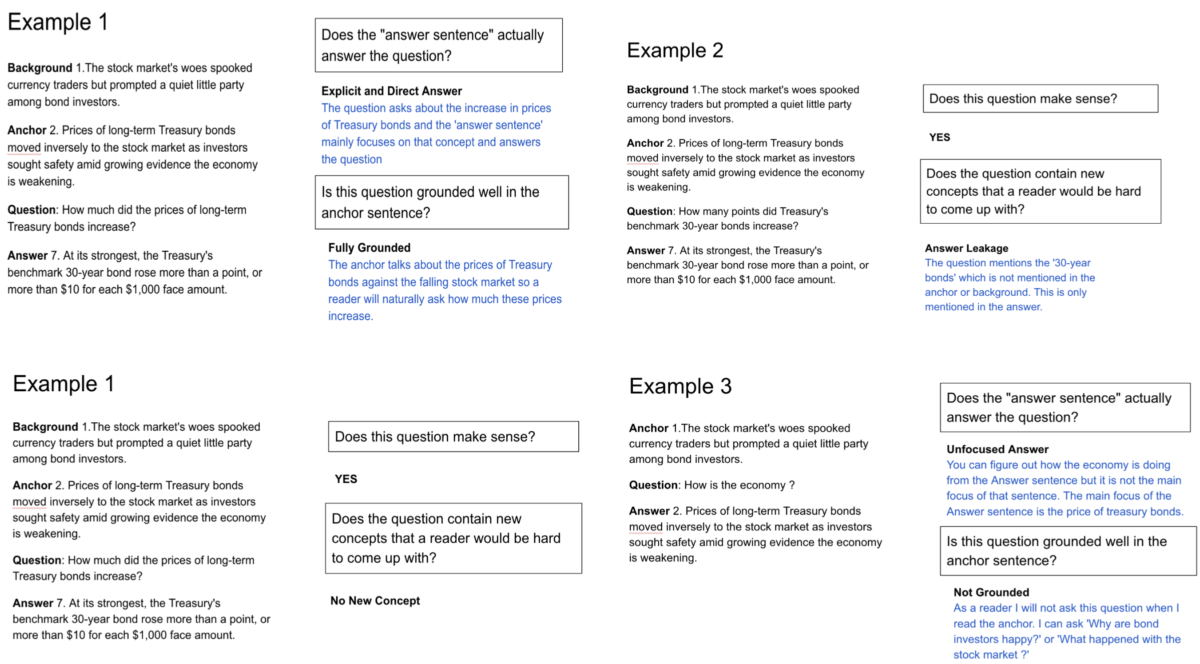


Figure 7: Additional training materials and instructions for human evaluation.