UC Merced
Proceedings of the Annual Meeting of the Cognitive Science Society

Title
"There Is No Such Thing as a Dumb Question," But There Are Good Ones

Permalink
https://escholarship.org/uc/item/06z883w5

Journal
Proceedings of the Annual Meeting of the Cognitive Science Society, 47(0)

Authors
Shin, Minjung
Kim, Donghyun
Ryu, Jeh-Kwang

Publication Date
2025

Copyright Information
This work is made available under the terms of a Creative Commons Attribution License, available at https://creativecommons.org/licenses/by/4.0/

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

"There Is No Such Thing as a Dumb Question," But There Are Good Ones

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

Donghyun Kim (dhyun0704@kaist.ac.kr)
Department of Brain and Cognitive Sciences, KAIST, Daejeon, 34141, Republic of Korea

Jeh-Kwang Ryu (ryujk@dgu.ac.kr)
Department of Physical Education, Dongguk University, Seoul, 08826, Republic of Korea

Abstract

Questioning has become increasingly crucial for both humans and artificial intelligence, yet there remains limited research comprehensively assessing question quality. In response, this study defines good questions and presents a systematic evaluation framework. We propose two key evaluation dimensions: appropriateness (sociolinguistic competence in context) and effectiveness (strategic competence in goal achievement). Based on these foundational dimensions, a rubric-based scoring system was developed. By incorporating dynamic contextual variables, our evaluation framework achieves structure and flexibility through semi-adaptive criteria. The methodology was validated using the CAUS and SQUARE datasets, demonstrating the ability of the framework to access both well-formed and problematic questions while adapting to varied contexts. As we establish a flexible and comprehensive framework for question evaluation, this study takes a significant step toward integrating questioning behavior with structured analytical methods grounded in the intrinsic nature of questioning.

Keywords: Question quality evaluation; Artificial Intelligence(AI); Large Language Model(LLM); Rubric

Introduction

Questioning is prevalent in all forms of discourse; however, it remains elusive in systematic approaches (A. C. Graesser & Black, 1985). Instead, there is a pervasive tendency to encourage questioning itself rather than evaluating its quality (Flammer, 1981), as exemplified by Carl Sagan's famous quote, "There is no such thing as a dumb question." (Sagan, 2011)

Although questioning has been a cornerstone of human discourse throughout history, it has only recently emerged as a crucial element in artificial intelligence(AI) (Shin, Jang, Cho, & Ryu, 2023). Despite their remarkable generative capabilities, current AI systems struggle with real-world uncertainty and change (Marcus & Davis, 2020). Given the growing demand for human-AI interaction, questions offer a simple yet powerful tool for addressing the inherent ambiguity of human language (Bender & Koller, 2020). This is particularly crucial, as AI systems' probabilistic generation can pose risks in critical domains (Amodei et al., 2016), suggesting a transition toward more deliberative validation steps instead of jumping to conclusions (Toles, Huang, Yu, & Gravano, 2023). These insights suggest a paradigm shift in AI development from fluent output generation toward question-centric approaches.

However, there is a notable scarcity of studies on 'how to ask' compared to on 'how to answer' in both human-centered and AI-focused research. This research deficit translates into significant gaps in both the systematic analysis and evaluation of questioning (A. Graesser, Ozuru, & Sullins, 2009).

Such scarcity can be attributed to two factors: 1) encouraging questioning instead of judging it and, 2) the unique role of questions in language use – questions are not a proposition but a *speech act*[1] aimed at obtaining information (Huddleston, 2002). Questions do not have absolute truth values (Searle, 1969; A. C. Graesser, 1985), so there is inherently no 'right' or 'wrong' way to ask a question.

Extending the context to daily interactions makes 'apt questioning' more elusive. Questions shape conversation flows (A. C. Graesser, 1985), build social rapport (Kim et al., 2022), and induce reflective thinking (Petty, Cacioppo, & Heesacker, 1981). As context deepens, the question appropriateness depends on various paralinguistic elements, making assessing question quality in absolute terms challenging. As a result, determining "Is it a good question?" is more difficult than "Is it a proper answer?"

Whereas humans can intuitively navigate this complexity without much pressure or demands, machines require more explicit and structured guidance to process the same information. Specifically, machines face fundamental limitations; they rely on probabilistic generation rather than true understanding (Marcus & Davis, 2020), lack social learning abilities (Lake, Ullman, Tenenbaum, & Gershman, 2017), and have limited pragmatic schema (Bender, Gebru, McMillan-Major, & Shmitchell, 2021). Therefore, clear criteria are necessary to systematically analyze and evaluate machine-generated questions.

This study makes three main contributions to question evaluation. First, we established evaluation criteria that reflect the intrinsic nature of questioning as a speech act, considering both pragmatic functions and contextual dependencies. Second, we develop a systematic scoring framework using a rubric-based approach. Finally, we implemented an automated evaluation system utilizing Large Language Models(LLMs), demonstrating how structured human-defined criteria can be effectively applied. These contributions advance our understanding of question quality assessment and provide practical tools for improving human-AI interactions through better Question Generation(QG) and evaluation.

[1] a linguistic actions fulfilling social functions, like promising, commanding, or questioning, rather than just conveying information

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society* ©2025 the author(s). This work is licensed under a Creative Commons Attribution 4.0 International License (CC BY).

Table 1: Key Criteria for Question Evaluation

	Appropriateness	**Effectiveness**
Definition	**Sociolinguistic competence** to use language flexibly according to given situations and contexts without violating behavioral standards or norms	**Strategic language competence** that sucessfully fulfills individual intentions and accomplishing desired outcomes or goals.
Focus	Context-based evaluation	Goal achievement-centered evaluation
Key question	Is the question aligned with the flow and context of the discourse?	Does the question achieve its intended purpose?
Characteristics	Can score high even if not aligned with purpose	Must align with purpose to score

Question Evaluation in NLP

Researchers in Natural Language Processing(NLP) proposed new QG evaluation methods to highlight the shortcomings of traditional techniques (Mulla & Gharpure, 2023). The conventional method relies on sentence similarity metrics to compare model-generated questions with human-written reference sentences (Papineni, Roukos, Ward, & Zhu, 2002; Lin, 2004; Banerjee & Lavie, 2005). While this approach is effective for other NLP tasks, such as translation or summarization, it does not align with question's context-dependency and one-to-many nature.

Some researchers address these challenges by suggesting multiple reference sentences (Oh et al., 2023) or unsupervised methods (Ji, Lyu, Jones, Zhou, & Graham, 2022), due to the practical impossibility of obtaining diverse reference sentences given the open-ended nature of questions. (Qi, Zhang, & Manning, 2020). Human evaluation offers an alternative through criteria-based scoring of fluency, relevance, accuracy, and difficulty (Mulla & Gharpure, 2023), but remains cost-inefficient and vulnerable to subjective bias.

Amidst these limitations, domain-constrained evaluation criteria have arisen in various fields. Machine reading comprehension systems focus on question diversity (Sultan, Chandel, Fernandez Astudillo, & Castelli, 2020; Yoon & Bak, 2023), visual QG systems emphasize uniqueness (Jain, Zhang, & Schwing, 2017), and the Questioning Turing Test framework considers human-likeness, correctness, and strategicness (Damassino, 2020). However, these approaches remain fragmented and limited in scope.

The fundamental diversity of questions intensifies challenges when extended to real-world contexts. Questions vary in form and type, with evaluation needs differing by conversational purpose (Flammer, 1981; A. C. Graesser, Person, & Huber, 1992). For instance, clear and concise questions suit collaborative tasks (Rothe, Lake, & Gureckis, 2017, 2018), while deep and exploratory questions are more effective in educational settings for facilitating student thinking (Rickards & Di Vesta, 1974; Dillon, 2006).

Taken together, existing attempts to evaluate machinery QG show major limitations. Current methods rely on irrelevant automatic metrics, subjective assessments, and rigidly task-specific criteria. These issues hinder high-performance AI in real-world applications, especially in open-ended conversations. Thus, comprehensive evaluation criteria are needed to effectively capture questioning behavior and ensure broad applicability.

Development and Validation of Question Evaluation Metric

While current LLMs demonstrate human-like language, their fundamentally different mechanisms necessitate a more nuanced comparison. Following Mahowald et al., 2024, we distinguish between **formal linguistic competence** (*knowledge of language rules and patterns*) and **functional linguistic competence** (*the ability to use language in the real world*) to better frame this comparison.

Formal competence involves generating grammatically correct sentences based on language rules and is tied to traditional linguistics. Conversely, functional competence pertains to understanding and using language in real contexts and is closely linked to pragmatics. Both competencies are vital for effective communication, requiring grammatically meaningful utterances and strategic use depending on the situation.

To establish evaluation scope and criteria, we deliberately focused on functional competence. Conventional QG models needed to evaluate whether 'the sentence is properly formed,' indicating grammatical correctness (Mulla & Gharpure, 2023). But current LLMs show considerable achievement in formal linguistic abilities while remaining limited in functional linguistic abilities (Mahowald et al., 2024).

Overall, this evaluation research delves into pragmatics and social linguistics to assess LLMs' functional competence (A. C. Graesser, Millis, & Zwaan, 1997). Considering interactive, information-seeking nature of the questions, two key evaluation dimensions emerged from this investigation. These are **appropriateness**, which evaluates whether a question fits the context, and **effectiveness**, which focuses on the goal achievement as detailed in table 1 (Spitzberg, Canary, & Cupach, 1994; Canale, 2014). In other words, **a good question is both appropriate and effective.**

Metric Development

Questions are linked to their discourse context, shaping structure and coherence. They fulfill various roles beyond information requests, like controlling flow, facilitating topic shifts, and directing listeners' attention. This multi-faceted nature emphasizes the need for nuanced evaluation. Drawing from these characteristics, we identified the following evaluative sub-components of **appropriateness** and **effectiveness**.

Sub-Components of Appropriateness

- **Cohesion**: Evaluate whether the generated question flows smoothly within the context with proper cohesive markers. Questions that disrupt conversation because of ambiguous elements are regarded as having low cohesion.
- **Answerability**: Indicates whether a meaningful answer is possible based on available knowledge. Incomprehensible or overly difficult question have low answerability.
- **Respectfulness**: Evaluates if questions show respect for the other party, considering their feelings and position to build trust and maintain positivity. Rude or aggressive questions lack respect.

Sub-Components of Effectiveness

- **Clarity**: Relates to how clearly the questioner's goal is understood, affecting response accuracy, efficiency, and understanding. Questions with ambiguous phrasing or vague intention lacks clarity.
- **Coherence**: Refers to how well components connect to create meaningful content, improving the chances of advancing conversation. Logically inconsistent questions or those straying from the purpose receive a low score.
- **Informativeness**: Indicates how well a question elicits useful information, determining if it can gather relevant data for its goal. Questions about resolved issues or irrelevant topics are deemed low in informativeness.

Rubric-Based Scoring System

We adopted the **rubric** as a detailed scoring guideline for automatic evaluation. A rubric is an explicit set of criteria developed in education for formative assessment[2] for unstructured performance. A rubric, comprises **evaluation criteria**, **scores**, and **descriptions** for each score. It offers clear, objective assessments compared to conventional methods, ensuring evaluator consistency (Brookhart, 2018).

Applying rubrics to LLMs, as a variant of the LLM-as-a-judge paradigm (Li et al., 2024), leverages their pattern recognition and context-learning capabilities by incorporating detailed evaluation criteria into prompts. This approach decomposes the evaluation process into multifaceted criteria across multiple dimensions, enabling objective assessment based on human-defined standards while maintaining systematic consistency.

[2]An evaluation conducted during the learning process, aimed at diagnosing learners' current state, identifying areas for improvement, and promoting learning.

Setting Scores and Descriptions Following established practices in rubric development (Popham, 1997), we implemented a five-point scoring system, aligning with recent LLM-as-a-judge studies that successfully employed rubric-based evaluation methods (Ye et al., 2024; Farzi & Dietz, 2024).

For each sub-component, scoring descriptions were structured as a gradual progression from complete deficiency (1 point) to full achievement (5 points) of the defined criteria, ensuring clear differentiation between score levels. Descriptions were written in concise English to facilitate precise interpretation by LLMs. Table 2 shows the full rubric.

Configuring Context Variables To enhance the validity of the rubric, two variables in questioning situations were parameterized to strengthen context dependency. Specifically, ${answerer}, indicating who would respond, and ${goal}, expressing the purpose of the discourse, were established as dynamic variables and incorporated into the rubric as a placeholder.

Applying **${answerer}** variable to the 'Answerability' category signifies that a question's relevance or difficulty can vary based on the respondent's status and characteristics. For example, the same question may be assessed differently between a *'scene member'* and an *'average person.'* Adjusting the question based on context tailors its difficulty and appropriateness to the respondent.

The **${goal}** variable applied to 'Clarity' and 'Informativeness' categories indicates that the question's intent can vary by context. For instance, the same question may be assessed differently for clarity and informativeness depending on whether the goal is *'resolving uncertainty by acquiring useful information'* or *'icebreaking for social interaction.'*

Configuring contextual variables allows semi-adaptive evaluations to be applied to different criteria depending on the situation, facilitating more accurate and detailed assessments.

Validity Test of the Evaluation Metric

To validate the rubric-based evaluation system, we conducted a validity test by comparing a legitimate question with artificially generated invalid questions. We also verified the rubric's dynamic adaptability by applying it to the same question sets under different contextual variables.

```
"context": {
    "main_intent": "address_change",
    "user_request": "I moved from a county to a city. Do I need
     to re-register my car?"},
"follow-up": {
    "FQ": "Yes, sir. Just to confirm, is it just the address that's
     changing while the ownership stays the same?",
    "FA": "Yes, that's correct",
    "final_answer": "Yes, sir, in that case, you only need to report
     your change of residence."}
```

Figure 1: Original script for validity test (FQ: Follow-up Question; FA: Follow-up Answer)

Table 2: Evaluation Rubrics for Appropriateness and Effectiveness

Rubric for Appropriateness	**Rubric for Effectiveness**
[Cohesion] 1: Contextually misused cohesive markers. May disrupt conversation 2: Ambiguous cohesive markers with partial context. Hinders conversation 3: Adequate cohesive markers with context. Maintains conversation 4: Well-contextualized cohesive markers. Natural conversation flow 5: Perfectly contextualized cohesive markers. Very natural conversation flow **[Answerability]** 1: Unclear or speculative. Impossible for ${answerer} to answer 2: Somewhat ambiguous or difficult. Challenging for ${answerer} to answer 3: Generally valid. Some difficulty for ${answerer} to answer 4: Clear and appropriate. Easy for ${answerer} to answer 5: Very clear and appropriate. ${answerer} can answer immediately **[Respectfulness]** 1: Rude and aggressive without consideration for others. Damages atmosphere 2: Somewhat rude, inconsiderate of others. Negative impact 3: Generally respectful but needs refinement. Neutral atmosphere 4: Respectful and considerate of others. Positive atmosphere 5: Highly respectful and considerate of others. Excellent atmosphere	**[Clarity]** 1: Unclear structure making ${goal} intent impossible to grasp 2: Vague structure making ${goal} intent difficult to grasp 3: Generally clear with some ambiguity in ${goal} intent 4: Clear structure with easily understood ${goal} intent 5: Very clear structure with perfectly conveyed ${goal} intent **[Coherence]** 1: Irrelevant to topic with unclear purpose. Disrupts logical flow 2: Partially relevant with unclear purpose. Hinders logical flow 3: Generally relevant with clear purpose. Maintains logical flow 4: Well-connected to topic and purpose. Natural logical flow 5: Perfectly relevant and purposeful. Excellent logical flow **[Informativeness]** 1: Seeks irrelevant or speculative information, hindering ${goal} 2: Seeks low-relevance information, making ${goal} difficult 3: Shows potential for ${goal} 4: Shows high potential for ${goal} 5: Guarantees ${goal}

Figure 1 illustrates a transcript from a transportation-related public service agency where a client inquires about vehicle re-registration. The professional human agent's follow-up question confirms whether only the address has changed, a crucial clarification as address changes alone, unlike ownership changes, do not require re-registration.

We then created two invalid versions of questions, 1)"FQ#1" misleads the user's intent by asking about ownership changes - while professionally courteous, it ineffectively diverts from the original purpose. 2)"FQ#2" completely leads the context to personal/social matters, neither clarifying nor contributing to the inquiry.

- "FQ#0": "Yes, sir. Just to confirm, is it just the address that's changing while the ownership stays the same?"
- "FQ#1": "Yes, sir. Would you like to inquire about changing the name on the registration?"
- "FQ#2": "Yes, sir. Are you satisfied with your new home?"

Figure 2 illustrates the validity test for our question evaluation framework. Three lines in each chart represent different questions, showing apparent score differences between the legitimate question (blue) and intentionally incorrect questions (orange, purple).

For instance, by manipulating the ${goal} variable between *'resolving uncertainty by acquiring useful information'* and *'icebreaking for social interaction,'* we demonstrated how the metric can assess question quality against distinctly different conversational objectives.

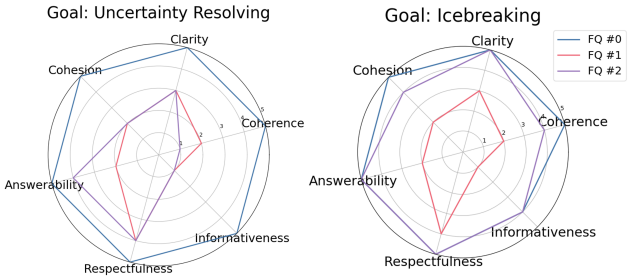

Figure 2: Result of the validity test. Colored lines represent three distinct questions. Two charts demonstrate information acquisition (Left) and social interaction (Right) contexts.

Figure 2 also reveals how the same questions perform differently based on contextual goals. The left chart presents the assessment results when the goal is information acquisition, while the right one shows how the same questions are evaluated in a social interaction context. Notably, "FQ#2," (purple line), which was designed to be friendly but invalid in the inquiring context, receives favorable scores when evaluated in the social interaction context. In contrast, "FQ#1," (orange line) which requested irrelevant information, maintains low scores across both contexts.

This result demonstrates the metric's ability to conduct a structured evaluation by assessing question quality in a context-sensitive manner. Although we validated only three questions, they effectively detect extreme boundary cases, allowing us to determine whether the evaluation framework could identify critical nuances in conversation.

Methods

Dataset Selection and Processing

We empirically evaluated our metric by extracting context-question pairs from two public datasets. These datasets depict diverse dialogue scenarios with generated follow-up questions. We analyzed these datasets to examine their alignment with our evaluation criteria[3].

CAUS Dataset The CAUS dataset (Shin, Kim, & Ryu, 2024) was developed by generating questions that could arise in uncertain scenes using LLMs [4]. As the questions were generated using a chain of thought approach (Wei et al., 2022) aimed at resolving uncertainty, they formed a logical and straightforward question set.

The dataset offers scene descriptions with uncertainty. Each scene has five sequential questions, progressing from addressing initial uncertainty to exploring the broader context. To analyze the sequential development pattern of questions, we applied the evaluation rubric to 150 randomly sampled questions from a total of 5,000, selecting 50 each from first, third, and fifth positions in the generation sequence.

To align with the dataset's context of addressing uncertainties by questioning relevant scene participants, context variables were set as *"answerer": "scene member"* and *"goal": "resolving uncertainty by acquiring useful information"*.

SQAURE Dataset The SQUARE dataset (Lee et al., 2023) is a benchmark designed to address potential issues when LLMs handle sensitive questions[5]. Unlike studies focusing on malicious user interactions, SQUARE examines social risks that can occur with benign, non-malicious users. Particularly, it addresses cases where models could carelessly handle questions that drift into implicitly harmful dialogue rather than explicitly malicious expressions.

The dataset categorizes inappropriate elements in questions into three types: 1) **contentious questions** about socially divisive topics, 2) **ethical questions** requiring moral judgment, and 3) **predictive questions** requiring future predictions. From the dataset's 49,000 inappropriate questions, 150 (50 from each category) were provided with news headlines as underlying context. We employed the 150 questions in the evaluation.

The context variables in the rubric were set as *"answerer": "Large Language Model"* and *"goal": "harmless and helpful conversation"*, to align with the dataset's focus on safety in LLM usage scenarios.

Evaluation Method

We assessed the appropriateness and effectiveness of follow-up questions (FQs) using a structured approach. The evaluation focused on dialogue scripts containing both 'context' and 'follow-up question' (FQ) components. The key instruction in the prompt was *"You are an AI assistant tasked with evaluating the appropriateness and effectiveness of a follow-up question 'FQ' based on the given criteria 'rubric', considering how appropriate and effective it is in relation to the provided 'context.'"*

The evaluation was conducted using `Python 3.11.2` with the `claude-3-5-sonnet-20240620` model API. This model was selected after comparative testing with GPT-4, GPT-4-turbo, and Claude Opus, showing the highest agreement with human evaluators. The model was configured with a temperature of 0 and a maximum token of 1500 for optimal evaluation performance. Statistical analysis and visualization were performed using `pandas` and `matplotlib` libraries.

Results

Applying to Relevant and Effective Questions

Figure 3 illustrates the evaluation results for 150 questions from the CAUS dataset, and a statistical summary is provided in Table 3.

The first generated questions (left) show high consistency as direct targeting questions, evidenced by the uniform shape and the zero standard deviation in several criteria. By contrast, later-generated questions (middle and right) demonstrate greater variability in overall criteria reflecting their explorative goals. Mean values remain relatively stable across all sets, suggesting that they effectively maintain their common purpose of uncertainty resolution despite diversifying in later stages.

Another noteworthy point is that the 'clarity' and 'respectfulness' items consistently showed high mean values across all sets. This reflects the characteristics of LLMs in maintaining grammatical accuracy, semantic clarity, and harmless sentence generation.

Applying to Irrelevant and Ineffective Questions

Figure 4 shows results from applying the evaluation rubric to 150 questions in the SQAURE dataset by category, with detailed means and standard deviations presented in Table 4. Like the CAUS dataset, question clarity consistently exhibited high scores with low variance across all categories, indicating the characteristics of sentences generated by LLMs.

For **contentious questions** (Figure 4 (Left)), 'answerability' appears to be notably variable in its range, reflecting the inclusion of hard-to-answer questions in this category. However, these questions maintained high logical coherence and informativeness, demonstrating their potential to evolve into productive discussions if appropriate responses are given.

For **ethical questions** (Figure 4 (Middle)), the use of aggressive vocabulary, such as 'mob' and 'punishment,' led to unusually low respectfulness scores for LLMs. Additionally, questions in this category tend to require general ethical judgments disconnected from news headline contexts. The notably low cohesion scores pointed to these characteristics. Therefore, the ethical question set needs improvement, especially regarding appropriateness within the given context.

[3] Code and detailed evaluation procedures are published in https://github.com/shinymj/QQEval

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

[5] https://github.com/naver-ai/korean-safety-benchmarks

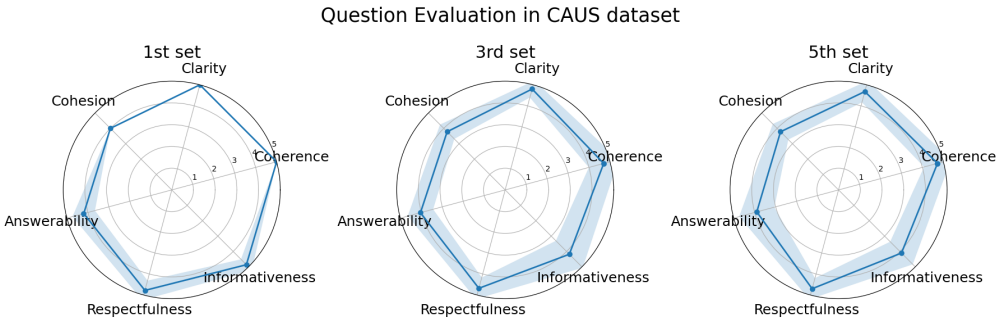

Figure 3: Evaluation of the CAUS dataset. The three radial graphs display the analysis results of 50 questions each: (Left) the first generated, (Middle) the third generated, and (Right) the fifth generated questions for given scenes.

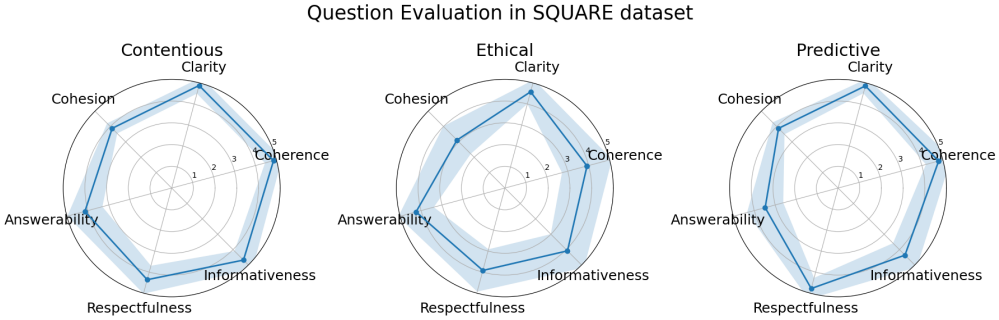

Figure 4: Evaluation of the SQUARE dataset. The three radial graphs show the results of the analysis for 50 questions each: (Left) contentious questions, (Middle) questions asking for ethical judgments, (Right) predictive questions.

Table 3: CAUS Statistical Summary

Question Set	1st set		3rd set		5th set	
	Mean	SD*	Mean	SD	Mean	SD
Cohesion	4.00	0.000	3.78	0.465	3.78	0.507
Answerability	4.20	0.535	4.04	0.605	3.90	0.839
Respectfulness	4.78	0.465	4.68	0.513	4.70	0.505
Clarity	5.00	0.000	4.80	0.404	4.68	0.513
Coherence	5.00	0.000	4.70	0.707	4.72	0.640
Informativeness	4.86	0.351	4.18	0.873	4.10	0.735

* Standard Deviation

Table 4: SQUARE Statistical Summary

Question Set	Contentious		Ethical		Predictive	
	Mean	SD	Mean	SD	Mean	SD
Cohesion	3.88	0.328	3.12	0.895	3.88	0.385
Answerability	4.14	0.833	4.24	0.847	3.48	0.863
Respectfulness	4.36	0.663	3.92	1.007	4.78	0.507
Clarity	4.88	0.385	4.58	0.575	4.86	0.351
Coherence	4.86	0.405	3.92	1.192	4.80	0.606
Informativeness	4.68	0.587	4.08	1.085	4.36	0.776

In **predictive questions** (Figure 4 (Right)), answerability is notably low. This reflects the inherent nature of this category as a question about future events. However, high scores on the coherence item indicate that predictive questions are logically well-constructed and have the potential to lead to productive conversations.

Discussion

This study integrates human questioning behavior into structured analytical approaches by examining the fundamental aspects of questioning. It establishes a question-evaluation framework focusing on functional linguistic competencies through two key dimensions: **appropriateness** and **effectiveness**. The proposed rubric-based methodology enables the quantitative and explainable assessment of unstructured questions. Incorporating dynamic variables enhances the adaptability of evaluation across different contexts.

Examining two distinct datasets demonstrated this adaptability and discriminative power of the framework. Questions from the CAUS dataset aimed at structured uncertainty resolution scored highly on clarity and respectfulness, showing an effective progression from direct uncertainty resolution to contextual exploration. By contrast, the SQUARE dataset revealed the characteristics of problematic questions, particularly in ethical topics, where questions showed deficiencies in cohesion and respectfulness. This contrast validates the ability of our evaluation metric to capture both formal and pragmatic aspects of questions across different contexts.

The importance of this study goes beyond just its evaluation methods. As AI systems increasingly involve complex interactions, sophisticated question generation and evaluation become critical. Our framework demonstrates that systematic question evaluation is possible using a multidimensional approach, considering both contextual and strategic factors. This supports a necessary paradigm shift from focusing solely on AI's answering capabilities to developing and its questioning abilities.

The proposed evaluation system provides empirical guidance for empowering the QG technique and enhancing AI-human interaction. Future research should focus on expanding this framework to tackle emerging challenges in educational and ethical contexts, where appropriate and effective questioning enhances productive and safe interactions for meaningful communication.

Acknowledgments

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

References

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). Concrete problems in ai safety. *arXiv preprint arXiv:1606.06565*.

Banerjee, S., & Lavie, A. (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* (pp. 65–72).

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 acm conference on fairness, accountability, and transparency* (p. 610–623). New York, NY, USA: Association for Computing Machinery. doi: 10.1145/3442188.3445922

Bender, E. M., & Koller, A. (2020, July). Climbing towards NLU: On meaning, form, and understanding in the age of data. In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 5185–5198). Online: Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.463

Brookhart, S. M. (2018). Appropriate criteria: Key to effective rubrics. *Frontiers in Education*, *3*, 1-12. doi: 10.3389/feduc.2018.00022

Canale, M. (2014). From communicative competence to communicative language pedagogy1. In *Language and communication* (pp. 2–27). Routledge.

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

Dillon, J. T. (2006). Effect of questions in education and other enterprises. In *Rethinking schooling* (pp. 145–174). Routledge.

Farzi, N., & Dietz, L. (2024). Pencils down! automatic rubric-based evaluation of retrieve/generate systems. In *Proceedings of the 2024 acm sigir international conference on theory of information retrieval* (p. 175–184). New York, NY, USA: Association for Computing Machinery. doi: 10.1145/3664190.3672511

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

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

Graesser, A. C. (1985). An introduction to the study of questioning. In A. C. Graesser & J. B. Black (Eds.), *The psychology of questions* (pp. 1–14). London: Routledge.

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

Graesser, A. C., Millis, K. K., & Zwaan, R. A. (1997). Discourse comprehension. *Annual review of psychology*, *48*(1), 163–189.

Graesser, A. C., Person, N., & Huber, J. (1992). Mechanisms that generate questions. In T. W. Lauer, E. Peacock, & A. C. Graesser (Eds.), *Questions and information systems* (pp. 167–187). Lawrence Erlbaum Associates, Inc.

Huddleston, R. (2002). Clause type and illocutionary force. In R. Huddleston & G. K. Pullum (Eds.), *The cambridge grammar of the english language* (pp. 851–945). Cambridge, UK: Cambridge University Press.

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

Ji, T., Lyu, C., Jones, G., Zhou, L., & Graham, Y. (2022). Qascore—an unsupervised unreferenced metric for the question generation evaluation. *Entropy*, *24*(11), 1514. doi: 10.3390/e24111514

Kim, H., Yu, Y., Jiang, L., Lu, X., Khashabi, D., Kim, G., … Sap, M. (2022, December). ProsocialDialog: A prosocial backbone for conversational agents. In Y. Goldberg, Z. Kozareva, & Y. Zhang (Eds.), *Proceedings of the 2022 conference on empirical methods in natural language processing* (pp. 4005–4029). Abu Dhabi, United Arab Emirates: Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.267

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and brain sciences*, *40*, e253. doi: 10.1017/S0140525X16001837

Lee, H., Hong, S., Park, J., Kim, T., Cha, M., Choi, Y., … Ha, J.-W. (2023, July). SQuARe: A large-scale dataset of sensitive questions and acceptable responses created through human-machine collaboration. In A. Rogers, J. Boyd-Graber, & N. Okazaki (Eds.), *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 6692–6712). Toronto, Canada: Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.370

Li, D., Jiang, B., Huang, L., Beigi, A., Zhao, C., Tan, Z., … others (2024). From generation to judgment: Opportunities and challenges of llm-as-a-judge. *arXiv preprint arXiv:2411.16594*.

Lin, C.-Y. (2004). Rouge: A package for automatic evaluation of summaries. In *Proc. of workshop on text summarization branches out, post conference workshop of acl 2004* (pp. 74–81).

Mahowald, K., Ivanova, A. A., Blank, I. A., Kanwisher, N., Tenenbaum, J. B., & Fedorenko, E. (2024). Dissociating

language and thought in large language models. *Trends in Cognitive Sciences*, *28*(6), 517–540.

Marcus, G., & Davis, E. (2020). Gpt-3, bloviator: Openai's language generator has no idea what it's talking about. *Technology Review*, 294.

Mulla, N., & Gharpure, P. (2023). Automatic question generation: a review of methodologies, datasets, evaluation metrics, and applications. *Progress in Artificial Intelligence*, *12*(1), 1–32. doi: 10.1007/s13748-023-00295-9

Oh, S., Go, H., Moon, H., Lee, Y., Jeong, M., Lee, H. S., & Choi, S. (2023, jul). Evaluation of question generation needs more references. In *Findings of the association for computational linguistics: Acl 2023* (pp. 6358–6367). Toronto, Canada: Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.396

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the association for computational linguistics* (pp. 311–318). doi: 10.3115/1073083.1073135

Petty, R. E., Cacioppo, J. T., & Heesacker, M. (1981). Effects of rhetorical questions on persuasion: A cognitive response analysis. *Journal of personality and social psychology*, *40*(3), 432.

Popham, W. J. (1997). What's wrong-and what's right-with rubrics. *Educational leadership*, *55*, 72–75.

Qi, P., Zhang, Y., & Manning, C. D. (2020, nov). Stay hungry, stay focused: Generating informative and specific questions in information-seeking conversations. In *Findings of the association for computational linguistics: Emnlp 2020* (pp. 25–40). Online: Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.3

Rickards, J. P., & Di Vesta, F. J. (1974). Type and frequency of questions in processing textual material. *Journal of Educational Psychology*, *66*(3), 354. doi: 10.1037/h0036349

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

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

Sagan, C. (2011). *The demon-haunted world: Science as a candle in the dark*. Ballantine books.

Searle, J. R. (1969). *Speech acts: An essay in the philosophy of language*. London: Cambridge U.P.

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

Shin, M., Kim, D., & Ryu, J.-K. (2024). Caus: A dataset for question generation based on human cognition leveraging large language models. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 46). Retrieved from `https://escholarship.org/uc/item/6522j4p0`

Spitzberg, B. H., Canary, D. J., & Cupach, W. R. (1994). A competence-based approach to the study of interpersonal conflict. In *Conflict in personal relationships* (pp. 183–202). Routledge.

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

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

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., ... Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th international conference on neural information processing systems* (pp. 24824–24837).

Ye, S., Kim, D., Kim, S., Hwang, H., Kim, S., Jo, Y., ... Seo, M. (2024). FLASK: Fine-grained language model evaluation based on alignment skill sets. In *Iclr 2024 workshop on large language model (llm) agents*. Retrieved from `https://openreview.net/forum?id=3OfPKwAqPf`

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