AURA-QG: Automated Unsupervised Replicable Assessment for Question Generation

Rajshekar K, Harshad Khadilkar, Pushpak Bhattacharyya
Indian Institute of Technology Bombay
{rajshekark, harshadk, pb}@cse.iitb.ac.in

Abstract

Question Generation (QG) is central to information retrieval, education, and knowledge assessment, yet its progress is bottlenecked by unreliable and non-scalable evaluation practices. Traditional metrics fall short in structured settings like document-grounded QG, and human evaluation, while insightful, remains expensive, inconsistent, and difficult to replicate at scale. We introduce AURA-QG: an Automated, Unsupervised, Replicable Assessment pipeline that scores question sets using only the source document. It captures four orthogonal dimensions i.e., answerability, non-redundancy, coverage, and structural entropy, without needing reference questions or relative baselines. Our method is modular, efficient, and agnostic to the question generation strategy. Through extensive experiments across four domains i.e., car manuals, economic surveys, health brochures, and fiction, we demonstrate its robustness across input granularities and prompting paradigms. Chain-of-Thought prompting, which first extracts answer spans and then generates targeted questions, consistently yields higher answerability and coverage, validating the pipeline's fidelity. The metrics also exhibit strong agreement with human judgments, reinforcing their reliability for practical adoption. The complete implementation of our evaluation pipeline is publicly available.[1]

1 Introduction

Question Generation (QG) is a fundamental NLP task for many downstream applications (Jiang et al., 2023), including education, automated knowledge assessment, information retrieval, and conversational AI systems. In Knowledge Base (KB)-grounded settings, the ability to generate high-quality, relevant, and diverse questions is essential for building systems that can engage with structured information effectively. However, as research in QG continues to evolve, the challenge of evaluating generated questions remains a critical bottleneck (Zhang et al., 2021).

Human evaluation, although reliable, is time-consuming, expensive, and difficult to scale. It often involves manual annotation for factors such as relevance, fluency, answerability, and coverage, introducing inconsistencies and making large-scale benchmarking infeasible. Automatic evaluation metrics such as BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), or METEOR (Banerjee and Lavie, 2005), while commonly used, fail to capture the nuanced aspects of question quality, especially in structured or semantic contexts like KB-QG. These metrics typically rely on surface-level lexical overlap and require reference questions, limiting their applicability and generalizability.

To address these limitations, AURA-QG is designed to score question sets generated from plain-text documents such as PDF manuals, reports, fiction, or health brochures. Our current pipeline focuses exclusively on unstructured text and intentionally ignores tables, figures, and other structured layout elements. For any given input document and its corresponding question set, our system produces independent scores along four interpretable dimensions. Furthermore, our evaluation pipeline is designed for factoid or informational questions and does not extend to assessment-style questions requiring inference, abstraction, or synthesis. It primarily targets the base levels of Bloom's taxonomy (remembering and understanding) as elaborated in Appendix A. These dimensions aim to capture core qualities of effective question sets and are defined as follows:

- **Answerability:** (Nema and Khapra, 2018) Whether each question can be answered using information present in the source document.
- **Redundancy:** (Mai and Carson-Berndsen, 2023) The degree of semantic overlap be-

[1] https://github.com/rk620/AURA-QG

Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 2979–2992
December 20-24, 2025

tween the retrieved answer units of different questions in the set.

- **Coverage:** How well the collective set of questions captures the breadth and diversity of the input content (Laban et al., 2022).
- **Structural Entropy:** A proxy for the diversity of question templates used (Fabbri et al., 2020), capturing variation in syntactic patterns and structural forms across generated questions, indicating the use of question templates over repetitive phrasing.

These metrics reflect essential dimensions of question set quality such as semantic relevance, informational breadth, non-redundancy, and linguistic diversity, making the evaluation both principled and aligned with human expectations. While this captures their intuitive motivation, we formally define and operationalize each metric in the subsequent sections of the paper. Our contributions are:

- **AURA-QG Evaluation Pipeline:** We introduce a reference-free, interpretable, and fully automated evaluation framework for assessing document-grounded question sets across four axes-Answerability, Redundancy, Coverage, and Structural Entropy.
- **Multi-domain Applicability:** The pipeline supports question sets generated from diverse textual domains including manuals, health brochures, and narrative fiction, without relying on reference questions or gold answers.
- **Human Alignment Validation:** We conduct a large-scale human evaluation with up to 7 annotators per example and show that the pipeline's preferred question set agrees with the human majority in 75% of cases, with an 80% individual-level agreement when aligned (Section 6.5).
- **Metric-wise Agreement Analysis:** We present per-metric agreement statistics and directional consistency analysis, demonstrating that individual metrics frequently favor human-preferred sets (Section 6.1).
- **Open Evaluation Recipe:** We release a replicable, modular evaluation methodology, designed to scale across new datasets and QG systems, offering the community a practical alternative to manual evaluation (Appendix C).

2 Related Work

In this paper, we are concerned with the quality of a question set based on a given corpus of text. Note that this contrasts with related but different problems in reading comprehension (Deutsch et al., 2021) or conversational systems (Griol et al., 2013). Our objective (see Section 3 for details) is to measure the appropriateness of a set of questions *as a collection*, given a corpus of text, in order to anticipate possible queries raised about the content.

Previously, the topic of question generation has been considered from the point of view of educational material and assessment (Wang et al., 2022; Gorgun and Bulut, 2024). The focus of these studies is to comprehensively cover the material (which we call *coverage*), and to generate questions with different levels of difficulty. The latter is not relevant to this study, because our document-question sets are related to Frequently Asked Questions (FAQ) type of settings. Historically (Heilman and Smith, 2010; Kurdi et al., 2020; Mulla and Gharpure, 2023) automated question evaluation has tended to focus on semantic matching between the question-document pair or between generated and gold standard questions. While the former approach misses out on a strong comparison across the set of questions (for example, redundancy), the latter approach is difficult in practice because of the need to have access to gold standard questions.

An automated pipeline called QGEval (Fu et al., 2024) attempted to address the FAQ setting using seven proposed metrics in literature, including some metrics related to the ones we use in Section 4. However, they found that several of the metrics (either by definition or by method of computation) do not align well with human evaluation, which is considered gold standard. In this work, we show that our proposed methodology does have good agreement with human evaluation. Similarly, Nema and Khapra (2018) have shown that standard n-gram based metrics such as BLEU score also do not have good overlap with human evaluation. An older study (before the release of modern LLMs) also emphasizes the need for an independent evaluation mechanism (Kumar et al., 2018).

We believe that there is a need to define comprehensive QG evaluation metrics that can be automatically computed, and be aligned with human evaluation. In the rest of this paper, we define the problem formally and propose AURA-QG with accompanying experiments and ablation studies.

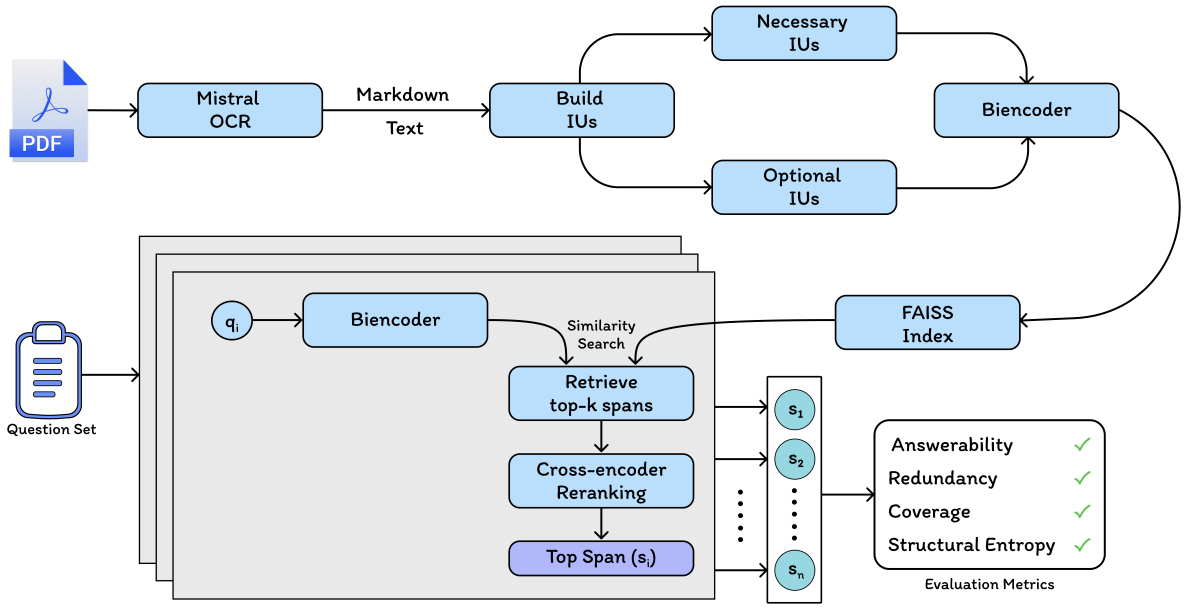

Figure 1: Overview of the evaluation pipeline. Given a document, Mistral OCR extracts Markdown text, from which necessary and optional Information Units (IUs) are constructed and indexed. For each question q_i in the question set (represented by overlapping rectangles), we retrieve and rerank top spans using semantic similarity, yielding a best-matching IU for metric computation (Answerability, Redundancy, Coverage, Structural Entropy).

3 Problem Description

Consider a raw document $\mathcal{A}$, on which questions are to be generated. The document contains artifacts which are not relevant for question generation, such as headers, image references, and other content. Therefore we consider a clean corpus $\mathcal{B} \subset \mathcal{A}$ which is relevant to question generation (QG). We also assume that a question set $\mathcal{Q}$ is provided through external means (we compare different approaches for generating $\mathcal{Q}$ in Section 5).

Then the automatic evaluation problem is to define and compute a set of n metrics $M : \mathcal{B} \times \mathcal{Q} \rightarrow \mathbb{R}^n$ that are aligned with human evaluation. Formally, this implies that some preference scoring function $S(M) : \mathbb{R}^n \rightarrow \mathbb{R}$ exists such that,

$$P\Big(S(M(\mathcal{Q}_1)) > S(M(\mathcal{Q}_2)) \mid H(\mathcal{Q}_1) > H(\mathcal{Q}_2)\Big) > 0.5, \quad (1)$$

where $H(\mathcal{Q}_i)$ is an aggregated human annotated score for question set $\mathcal{Q}_i$. We do not explicitly use the corpus $\mathcal{B}$ in the notation above, but it is understood that all scores (automatic and human) are on the same corpus.

The intuition behind (1) is that if the aggregated human-annotated score for one set of questions $\mathcal{Q}_1$ on $\mathcal{B}$ is higher than for an alternative set $\mathcal{Q}_2$, then it is likely that the automated score S based on the metrics M follows the same ranking order. In the next section, we propose the metrics M and the scoring function S that align with this objective.

4 Methodology

We present a modular pipeline designed to evaluate the alignment between a document and a set of questions generated from it. Our approach combines semantic indexing, answer span retrieval, structural refinement, and metric-based evaluation to comprehensively measure the quality of generated questions. The methodology is content-grounded and does not rely on external gold answers, making it adaptable across domains.

4.1 Information Unit Construction

To ground evaluation in the source text, we preprocess the document into Information Units (IUs), which serve as candidate answer spans. These include necessary units (core sentences and list blocks) and optional units (auxiliary spans such as sliding windows or decomposed list items) that enhance fine-grained retrieval. Optional units are semantically linked to their parent necessary units through similarity matching, ensuring hierarchical coverage and supporting more precise redundancy and coverage computation.

For list-type spans, we incorporate a refinement mechanism that adaptively narrows down multi-

item content into smaller, more relevant subsets whenever entire lists are too broad to be aligned with specific questions. While the full procedure relies on model confidence scores to decide when and how this refinement is triggered, we defer these technical details to Appendix B.

The complete methodology for IU construction, optional-unit similarity linkage, and dynamic list refinement is described in Appendix B.

4.2 Semantic Indexing and Span Retrieval

To assess answerability, we identify which IUs are most semantically aligned with it. All IUs $\mathcal{I}$ are encoded using a bi-encoder (`multi-qa-MiniLM-L6-cos-v1`), and the resulting vectors are indexed using FAISS (Facebook AI Similarity Search) (Johnson et al., 2019). For each question $q \in \mathcal{Q}$, we perform the matching process,

1. **Top-k retrieval** is performed using FAISS similarity search in the embedding space:

$$\text{Retrieve}(q) = \{i_1, i_2, ..., i_k\} \subset \mathcal{I}$$

2. **Reranking with cross-encoder** (`ms-marco-MiniLM-L-6-v2`) refines the ranking by assessing the contextual alignment between each candidate span and the question. A relevance score is assigned:

$$s_{q,i_j} = \text{Score}(q, i_j)$$

3. **Top span selection** identifies the most probable answer span:

$$i_q^* = \arg\max_{i_j} s_{q,i_j}$$

This two-stage approach, semantic retrieval followed by precise reranking, ensures that question–answer mappings are both efficient and semantically meaningful.

4.3 Evaluation Metrics

Our evaluation framework comprises four complementary metrics, each targeting a different facet of quality in the question set. These metrics are designed to be maximizable, offering a consistent interpretation of higher values as better performance.

The metrics are: Answerability – proportion of questions that can be confidently grounded in the source document.

Coverage – how comprehensively the question set spans the document's core content.

Non-Redundancy Score (NRS) – degree to which distinct questions map to distinct information units, rewarding diversity.

Structural Entropy – variation in question templates (e.g., what/why/how), capturing structural diversity.

4.3.1 Answerability

Answerability quantifies the proportion of questions that can be confidently grounded in the source document. A question is marked as 'answerable' if its top-ranked IU achieves a relevance score above a set threshold $\delta = 2.5$ (See Section 6.3 for experiments on optimal threshold consideration). Formally,

$$\text{Answerability} = \frac{|\{q \in \mathcal{Q} \mid s_{q,i_q^*} > \delta\}|}{|\mathcal{Q}|}$$

This metric reflects the reliability of the question generation process with respect to content.

4.3.2 Coverage

Coverage measures how comprehensively the generated questions span the document's core content. It focuses only on necessary IUs, evaluating what fraction of them are addressed by the question set. A necessary IU is considered covered either through direct retrieval or via indirect linkage through optional IUs, as detailed in Section 4.4. Let $\mathcal{N}_{covered} \subseteq \mathcal{N}$ be the set of covered necessary IUs. Then:

$$\text{Coverage} = \frac{|\mathcal{N}_{\text{covered}}|}{|\mathcal{N}|}$$

This metric captures the breadth of semantic alignment between the document and the question set.

4.3.3 Non-Redundancy Score (NRS)

Redundancy in question generation often manifests as multiple questions pointing to the same answer span. To reward diversity, we compute redundancy as the proportion of such repeated mappings and invert it to create a maximizable Non-Redundancy Score (NRS). Let $\mathcal{R} \subset \mathcal{Q}$ be the set of redundant questions, then:

$$\text{Redundancy} = \frac{|R|}{|\mathcal{Q}|}$$

$$\text{Non-Redundancy Score (NRS)} = 1 - \text{Redundancy}$$

The NRS encourages distinct questions to correspond to distinct pieces of information, improving the utility and informativeness of the question set.

4.3.4 Structural Entropy

Each WH-template targets a distinct informational need: *what* and *where* seek factoid answers, *why* prompts causal explanations, and *how* elicits procedural responses. A question set with high structural entropy thus reflects varied reasoning demands and serves as a proxy for evaluating the informativeness and functional breadth of a question set.

To assess the diversity of question templates, we extract the *question-template type*) from each question. Let T be the empirical distribution of question templates in the set. We first compute the standard entropy:

$$\text{SE} = -\sum_{t \in \mathcal{T}} P(t) \log_2 P(t)$$

Since the number of generated questions varies across passages and prompts, we normalize entropy using a length-aware adjustment. Let $|\mathcal{C}|$ denote the number of possible question template classes (e.g., what, why, how, etc.). We define the balanced entropy as:

$$\text{Balanced SE} = \frac{\text{SE} \cdot (|\mathcal{Q}|/|\mathcal{C}|)}{\text{SE} \cdot (|\mathcal{Q}|/|\mathcal{C}|) + |\mathcal{C}|}$$

Unless otherwise stated, all mentions of "Structural Entropy" or "Entropy" throughout the paper refer to this balanced form.

4.4 IU Coverage Propagation via Metadata

Determining whether a necessary IU is "covered" requires a nuanced treatment of retrieval paths. We extend coverage beyond direct retrievals by leveraging IU metadata that encodes parent-child relationships and semantic linkages.

A necessary IU $n \in \mathcal{N}$ is marked as *covered*, if it satisfies any of the following:

- **Direct match:** n is selected as the top span for some question.
- **Linkage via optional window:** An optional IU $o \in \mathcal{O}$ is selected, and $n \in \text{Link}(o)$ where $\text{Link}(o)$ denote the set of necessary IUs $n \in \mathcal{N}$ such that n is semantically linked to optional IU o via subtype-specific similarity matching as defined in Section B.1. This linkage is reflected in the *parent_chunk_id* metadata of o.
- **List-item propagation:** A bullet-type optional IU is selected, and its parent_chunk_id refers to a necessary list IU n.

These coverage rules are resolved by traversing the metadata structure generated during IU construction. This ensures that our coverage metric accurately reflects meaningful content engagement, even when the retrieval paths are indirect or span multiple IU levels.

In summary, the proposed methodology introduces a pipeline to evaluate question sets based on semantic alignment and structural diversity. By leveraging a hierarchy of information units, subtype-aware retrieval, and maximizable metrics grounded in coverage and linguistic variation, the framework ensures both depth and breadth. It moves beyond surface-level matching to assess the true informativeness of generated questions.

5 Experiments

We begin by validating our evaluation pipeline through human preference judgments to ensure that the automatic metrics align with human expectations. The results of this agreement analysis are presented in Table 1. Once established, we analyze metric behavior across domains, passage granularities, and prompting strategies. The main experimental results are summarized in Table 2.

5.1 Human Evaluation Setup

We evaluate the alignment of our automatic metrics with human judgment on 24 (passage, $\mathcal{Q}_1$, $\mathcal{Q}_2$) triplets, where $\mathcal{Q}_1$ and $\mathcal{Q}_2$ denote two alternative question sets generated for the same passage. In total, this setup results in roughly 1200 human-evaluated questions across all pairs. Two groups of 5-7 annotators (instructions given in Appendix C) independently select the better question set per triplet, based on answerability, coverage, redundancy, and structural entropy. Ties are resolved by a separate tie-breaker annotator, and the resulting preferences serve as ground truth for alignment analysis in Section 6. Throughout this paper, the term triplet refers to a combination of a passage and its two corresponding question sets, i.e., (passage, $\mathcal{Q}_1$, $\mathcal{Q}_2$).

5.2 Domains and Data

To evaluate generalizability across content styles, we select four document types with distinct structure and semantics:

- **Car Manuals** (technical, procedural),
- **Economic Surveys** (expository, policy-heavy),

- **Fictions** (narrative, open-ended),
- **Health Brochures** (concise, instructional).

These span instructional, analytical, narrative, and public-health genres ensuring that our metrics do not overfit to any single domain.

We segment the corpus into four levels of context: paragraph, page, chapter, and full-document, yielding 200 paragraphs, 200 pages, 40 chapters, and 4 full-documents in total. This experiment tests whether our metrics remain stable across varying passage sizes, and whether they appropriately reflect the contextual tradeoffs in question quality. For example, full-document inputs yield lower coverage due to token limits but higher structural entropy, while paragraph-level inputs show the opposite trend due to the limited scope. Validating the pipeline across granularity helps ensure it can scale across real-world use cases where document size varies significantly.

5.3 Prompting Strategies

We compare two prompting paradigms for question generation:

Zero-shot prompting, which directly asks the model to generate questions,

Chain-of-Thought (CoT) prompting, which encourages intermediate reasoning before question formulation by first extracting answer spans and then generating questions for each.

We used these prompting techniques to generate questions using `gemini-2.0-flash` (DeepMind, 2024), a model with built-in CoT reasoning capabilities. This axis allows us to evaluate whether our metrics can reflect established intuition that CoT-based prompting yields better questions. If our pipeline is sound, CoT-generated sets should consistently score higher in answerability and coverage, with non-redundancy and structural entropy being desirable but optional.

5.4 Cross-factor Validation

All prompting strategies are applied across all passage granularities and all document domains, yielding a dense evaluation matrix. This setup ensures that the observed trends are not artifacts of a specific prompt, length, or domain, but rather reflect the consistent behavior of the evaluation framework under diverse conditions.

Sign Agreement	Agreement Count	Proportion
Answerability	17	0.71
Coverage	19	0.79
Redundancy	12	0.50
Entropy	14	0.58
Majority Vote	18	**0.75**

Table 1: Sign agreement of individual metrics and majority-vote-based agreement with human preference across 24 triplets.

6 Results and Analysis

We begin by assessing the alignment between our automatic metrics and human judgments through a preference-based evaluation. Once validated, we analyze question generation quality across domains, passage granularities, and prompting strategies. The analysis highlights clear performance trends and architectural tradeoffs observed across the four metrics.

6.1 Metric Agreement with Human Scoring

We analyze the alignment between each automatic metric and human preferences over the 24 evaluated triplets. For each metric, we compute how often it assigns a higher score to the human-preferred question set, reporting both the count and the corresponding fraction. We additionally report how often the aggregate score across all four metrics favors the human choice. Full details of this agreement procedure are provided in Appendix C, and results are summarized in Table 1.

This initial agreement analysis provides strong empirical support for the alignment claim in (1), demonstrating that our automatic metrics consistently reflect human preferences, unlike prior approaches which lacked such validation.

6.2 Individual-Level Human Agreement

To further quantify how well the pipeline reflects individual human judgments, we compute the conditional probability of a randomly chosen human agreeing with the pipeline's decision, conditioned on whether it matched the set with the highest score.

When the pipeline's prediction aligns with the human majority, individual annotators **agree with it in approximately 80% of cases**, indicating a strong and consistent correspondence between automatic and human evaluation. Even when it diverges

Domain	Passage Level	Zero-Shot Prompting					CoT Prompting				
		Qs	Ans	Cov	NRS	Ent	Qs	Ans	Cov	NRS	Ent
Car Manual	Paragraph	8	86.5	92.9	**64.2**	42.6	15	**88.0**	**94.4**	56.2	**44.6**
	Page	16	94.4	85.1	**72.9**	55.1	29	**94.5**	**87.3**	72.0	**58.2**
	Chapter	58	92.6	53.2	**88.8**	68.6	110	**93.9**	**74.6**	76.3	**79.7**
	Full PDF	191	**95.8**	6.2	**95.1**	**41.3**	463	92.5	**15.1**	88.5	40.0
Economic Survey	Paragraph	12	89.2	87.8	71.8	**56.0**	18	**93.8**	**89.0**	**73.9**	52.6
	Page	15	90.4	77.8	**77.6**	56.0	31	**91.4**	**83.6**	73.2	**57.5**
	Chapter	56	93.1	56.1	**88.2**	72.8	133	**93.4**	**82.9**	84.3	**73.9**
	Full PDF	212	86.0	8.9	**94.8**	**83.7**	396	**92.2**	**10.0**	90.7	55.6
Narrative Fiction	Paragraph	8	40.8	57.9	**77.3**	32.5	12	**54.7**	**68.7**	71.7	**38.9**
	Page	18	47.8	26.7	**85.5**	48.2	35	**57.1**	**49.7**	81.5	**56.7**
	Chapter	45	56.6	7.9	**81.1**	60.6	257	**63.3**	**38.8**	77.8	**74.9**
	Full PDF	220	45.9	1.6	**91.1**	**73.0**	513	**51.3**	**3.4**	82.1	71.0
Health Brochure	Paragraph	7	69.4	88.3	72.7	**35.2**	9	**71.6**	**91.4**	**78.6**	29.8
	Page	14	72.6	75.7	**77.2**	42.7	20	**79.3**	**83.1**	70.8	**52.6**
	Chapter	48	**83.4**	42.7	**85.1**	68.0	106	79.7	**68.0**	84.5	**70.3**
	Full PDF	358	60.9	18.0	**96.3**	**67.5**	526	**75.7**	**24.3**	87.2	51.6

Table 2: Evaluation scores across domains, passage granularities, and prompting strategies. Qs: Average number of questions in that level, Ans: Answerability, Cov: Coverage, NRS: Non-Redundancy Score, Ent: Structural Entropy.

from the majority, **around 43% of annotators still concur with its choice** - showing that the pipeline's alternative selections are not arbitrary, but often represent legitimate minority perspectives within natural human variability. This sustained level of individual agreement across both consensus and disagreement cases underscores the robustness of the proposed evaluation framework and its reliability as a proxy for human judgment. A detailed explanation of this analysis is provided in Section 6.5 and further contrasted with standard inter-annotator agreement techniques in Section 6.6.

6.3 Mean Shift Analysis w.r.t. Threshold

We analyze the sensitivity of the evaluation pipeline to the relevance threshold (δ) used in answer span retrieval. Six thresholds ($\delta \in 0, 1.5, 2.5, 3.5, 5, 7.5$) were tested on the dataset. For each threshold, the four evaluation metrics were computed for both question sets ($\mathcal{Q}_1$, $\mathcal{Q}_2$), averaged within each passage, and then across all passages to obtain the mean metric value per threshold. This *mean shift analysis* captures how the average metric behavior changes as the model becomes more selective in accepting retrieved spans.

Unlike standard deviation analysis, which is affected by varying sample sizes, mean shift provides a clearer and more stable view of systematic trends.

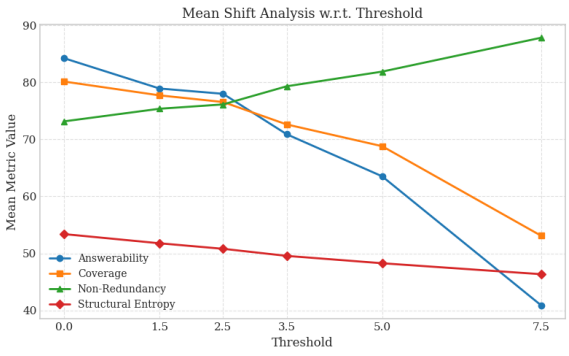

Figure 2: Mean metric variation across thresholds showing optimal balance at $\delta = 2.5$.

As shown in Figure 2, **Answerability** and **Coverage** drop with increasing threshold, since stricter filtering removes lower-confidence questions whose spans fall below the cutoff. In contrast, **NRS** increases steadily, as many overlapping or low-confidence questions are dropped at higher thresholds, naturally reducing repetition across the retained set. **Structural Entropy** remains mostly stable, indicating that question diversity is unaf-

fected by retrieval strictness.

A threshold of $\delta = 2.5$ offers the best trade-off. At this point, the system retains strong Answerability and Coverage while achieving high Non-Redundancy. Beyond this, further NRS gains come at the expense of sharp coverage losses. **Human evaluation agreement remains stable**, as the comparative setup ensures consistent relative rankings across thresholds. Hence, we adopt $\delta = 2.5$ as the default threshold for subsequent experiments.

6.4 Quantitative Analysis

Higher Non-Redundancy in Zero-Shot Prompting: Zero-Shot Prompting consistently exhibits higher non-redundancy scores (NRS) across almost all domains and passage granularities. This is because Zero-Shot strategies tend to generate fewer questions (almost half the number of questions generated by CoT prompting method), naturally reducing the chances of semantic overlap or repetitive patterns. The lower question volume results in a concise and less redundant question set, reinforcing the trend across domains.

High Coverage in CoT Prompting: CoT Prompting significantly outperforms Zero-Shot in Coverage (Cov) scores across nearly all passage levels and domains. This is an outcome of the two-step strategy in the CoT prompt: first, extracting all answer spans and then generating specific questions for each. This leads to thorough coverage of the passage. This validates the strength of CoT prompting for maximizing coverage.

Coverage–Redundancy Tradeoff in CoT: This increase in coverage from CoT prompting comes with a tradeoff. Non-redundancy scores drop as the model generates several closely related questions when multiple answer spans are extracted from the same sentence or block. During reranking, these often map back to the same blocks, reducing uniqueness. Thus, CoT's aggressive coverage yields denser but more overlapping question sets.

Low PDF Coverage Due to Token Limits: Coverage drops significantly at the Full PDF level, particularly under CoT, due to the token limit of the question-generating model. When presented with entire PDFs (often exceeding 250 pages), the model fails to process content holistically. This causes the model to either truncate inputs/outputs or focus on selective segments, resulting in undercoverage.

CoT Yields Higher Answerability: Answerability (Ans) is higher under CoT prompting in nearly 88% of test cases. The span-based prompting results in clearer, grounded and focused questions.

Weak Performance on Fiction Domain: The scores across all metrics for Fiction domain is notably lower. This is because, dialogues, narrative implicit context, and unstructured nature of narratives complicate the extraction of factual content.

Structural Diversity Reflected in Entropy: We also observe that Entropy generally increases from Paragraph to Chapter level across all domains and prompting strategies. This is expected, as larger content windows provide more freedom to frame structurally diverse questions. An exception is the Full PDF level, where the model's token constraints limit expressiveness, suppressing potential entropy gains.

Overall, Chain-of-Thought prompting shows strong gains in answerability and coverage, while Zero-Shot performs better in minimizing redundancy. Domain-specific behaviors and structural challenges further influence the quality of generated questions.

6.5 Conditional Probability of Human–Pipeline Agreement

As a proxy for measuring human–pipeline agreement, we compute the conditional probability that a randomly selected human annotator agrees with the pipeline's decision, conditioned on whether the pipeline's prediction aligns with the human majority. This formulation is particularly relevant in our majority-vote evaluation setup, as it captures how closely the pipeline's choices resonate with individual human judgments rather than only aggregate consensus.

Let:

- H_i: the number of human annotators for example i (typically 5 or 7),
- Y_H: the question set chosen by a randomly sampled human annotator,
- Y_P: the question set chosen by the pipeline,
- AGREE: event that Y_P matches the majority human vote,
- DISAGREE: event that Y_P does not match the majority human vote.

We compute:

$$\mathbb{P}(Y_H = Y_P \mid \text{AGREE}) = 0.80,$$
$$\mathbb{P}(Y_H = Y_P \mid \text{DISAGREE}) = 0.43.$$

Condition	Agreement Rate	Interpretation on alignment
Agreement	0.80	Strong alignment
Disagreement	0.43	Mild divergence

Table 3: Conditional human-pipeline agreement rates across 24 triplets with $H_i \in \{5, 7\}$.

The conditional probabilities reported in Table 3 provide a deeper view into how the pipeline's decisions relate to individual human judgments. A high value of $\mathbb{P}(Y_H = Y_P \mid \text{AGREE}) = 0.80$ suggests that when the pipeline's preference coincides with the majority human vote, this consensus is not superficial-most individual annotators also tend to choose the same question set. This reinforces that the automatic metrics not only approximate collective preferences but also capture consistent, human-interpretable patterns at the level of individual decision-making.

In contrast, $\mathbb{P}(Y_H = Y_P \mid \text{DISAGREE}) = 0.43$ reveals a complementary insight. Even when the pipeline's overall decision differs from the majority vote, nearly half of the human evaluators still side with it. This behavior indicates that the pipeline's disagreements are not random deviations but rather align with legitimate minority opinions that emerge in inherently subjective tasks like question evaluation. Such outcomes are common in linguistic assessments, where multiple interpretations can coexist and human consensus may not always represent the only valid perspective.

Together, these two probabilities suggest that the pipeline exhibits a stable, interpretable alignment with human reasoning patterns across both consensus and contested cases. The relatively high conditional agreement even under disagreement scenarios demonstrates that the evaluation pipeline captures underlying semantic and structural qualities that are meaningful to humans, rather than overfitting to surface patterns or majority biases. This further substantiates its reliability as an objective, replicable proxy for large-scale human evaluation, especially in domains where subjective variability is expected.

6.6 Comparison with Inter-Annotator Agreement Techniques

Traditional inter-annotator agreement (IAA) metrics such as Cohen's κ or Fleiss' κ quantify consistency *among* human annotators when assigning categorical labels. However, in our evaluation setup, annotators express *relative preferences* between two question sets rather than absolute labels, rendering such measures less informative. Our conditional probability formulation instead measures alignment *between* the pipeline and the distribution of human opinions, directly estimating the likelihood that a randomly chosen annotator agrees with the pipeline's decision under both consensus and disagreement scenarios. This approach thus provides a more interpretable and contextually appropriate proxy for human agreement in pairwise comparison settings, reflecting the pipeline's fidelity to human reasoning rather than mere annotator consistency.

7 Summary and Conclusion

In this work, we present a novel, fully automated, deterministic, LLM-free and reference-independent evaluation pipeline for document-grounded question generation, assigning interpretable scores across four axes: Answerability, Non-Redundancy, Coverage, and Structural Entropy. It evaluates questions with respect to the source document alone, making it highly scalable and adaptable to various domains and granularities. The pipeline is modular, transparent, and designed for integration into automated benchmarking workflows.

We validate our method across four diverse domains and four levels of input granularity, comparing Zero-Shot prompting with a Chain-of-Thought (CoT) strategy that first extracts all potential answer spans before question generation. Results show CoT significantly improves answerability and coverage, while revealing redundancy tradeoffs. We also observe that broader contexts yield higher entropy but may reduce coverage due to token limits.

Finally, strong alignment with human preferences confirms the reliability of our metrics, establishing this pipeline as a practical and insightful tool for benchmarking, model development, and deployment in real-world applications for scalable QG evaluation.

Limitations

Our evaluation pipeline is primarily tailored for lower-order cognitive tasks in Bloom's taxonomy, specifically, factoid or FAQ-style question sets that involve remembering and understanding discrete information. It does not support assessment-oriented questions that require higher-order reasoning such as application, analysis, or evaluation (e.g., inference-based MCQs or comprehension-style questions). Although our metrics are reference-free and interpretable, they do not account for conceptual breadth, while structural entropy captures WH-template diversity, it does not measure whether the questions span varied underlying concepts or information units. Additionally, our current pipeline does not handle scenarios where answering a question requires referring to structured elements such as tables, charts, or figures embedded in the document. Lastly, while we use Gemini-2.0-Flash for generation, our evaluation pipeline is not integrated with generation models and does not perform joint optimization, leaving room for co-adaptive design in future work.

References

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 MT and/or Summarization*, pages 65–72.

Google DeepMind. 2024. Gemini 1.5 technical report. https://deepmind.google/discover/blog/gemini-15.

Daniel Deutsch, Tania Bedrax-Weiss, and Dan Roth. 2021. Towards question-answering as an automatic metric for evaluating the content quality of a summary. *Transactions of the Association for Computational Linguistics*, 9:774–789.

Alexander Fabbri, Patrick Ng, Zhiguo Wang, Ramesh Nallapati, and Bing Xiang. 2020. Template-based question generation from retrieved sentences for improved unsupervised question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4508–4513. Association for Computational Linguistics.

Weiping Fu, Bifan Wei, Jianxiang Hu, Zhongmin Cai, and Jun Liu. 2024. Qgeval: Benchmarking multi-dimensional evaluation for question generation. *arXiv preprint arXiv:2406.05707*.

Guher Gorgun and Okan Bulut. 2024. Exploring quality criteria and evaluation methods in automated question generation: A comprehensive survey. *Education and Information Technologies*, 29(18):24111–24142.

David Griol, Javier Carbó, and José M Molina. 2013. An automatic dialog simulation technique to develop and evaluate interactive conversational agents. *Applied Artificial Intelligence*, 27(9):759–780.

Michael Heilman and Noah A Smith. 2010. Good question! statistical ranking for question generation. In *Human language technologies: The 2010 annual conference of the North American Chapter of the Association for Computational Linguistics*, pages 609–617.

Hang Jiang, Junnan Liu, Jing Liu, Yue Zhang, and Graham Neubig. 2023. Closed-book question generation via contrastive learning. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, 7(3):535–547.

Vishwajeet Kumar, Ganesh Ramakrishnan, and Yuan-Fang Li. 2018. Putting the horse before the cart: A generator-evaluator framework for question generation from text. *arXiv preprint arXiv:1808.04961*.

Ghader Kurdi, Jared Leo, Bijan Parsia, Uli Sattler, and Salam Al-Emari. 2020. A systematic review of automatic question generation for educational purposes. *International journal of artificial intelligence in education*, 30(1):121–204.

Antoine Laban, Yahong Sennrich, and Andy Way. 2022. Discord questions: A computational approach to diversity analysis in news. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (Workshop)*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, pages 74–81.

Long Mai and Julie Carson-Berndsen. 2023. I already said that! degenerating redundant questions in open-domain dialogue systems. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Student Research Workshop)*, pages 226–236. Association for Computational Linguistics.

Nikahat Mulla and Prachi Gharpure. 2023. Automatic question generation: a review of methodologies, datasets, evaluation metrics, and applications. *Progress in Artificial Intelligence*, 12(1):1–32.

Preksha Nema and Mitesh M. Khapra. 2018. Towards a better metric for evaluating question generation systems. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3950–3959. Association for Computational Linguistics.

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

Zichao Wang, Jakob Valdez, Debshila Basu Mallick, and Richard G Baraniuk. 2022. Towards human-like educational question generation with large language models. In *International conference on artificial intelligence in education*, pages 153–166. Springer.

Ruqing Zhang, Jiafeng Guo, Lu Chen, Yixing Fan, and Xueqi Cheng. 2021. A review on question generation from natural language text. *ACM Transactions on Information Systems (TOIS)*, 40(1):1–43.

A Bloom's Taxonomy and Evaluation Scope

Bloom's Taxonomy is a hierarchical classification of cognitive skills used to assess learning outcomes, ranging from basic knowledge recall to advanced creative tasks. The taxonomy is structured in six ascending levels of complexity:

- **Remember**: Recall of facts and basic concepts (e.g., define, list, memorize)
- **Understand**: Explain ideas or concepts (e.g., describe, identify, classify)
- **Apply**: Use information in new situations (e.g., implement, solve, demonstrate)
- **Analyze**: Draw connections among ideas (e.g., differentiate, compare, test)
- **Evaluate**: Justify a decision or position (e.g., argue, critique, judge)
- **Create**: Produce new or original work (e.g., design, write, formulate)

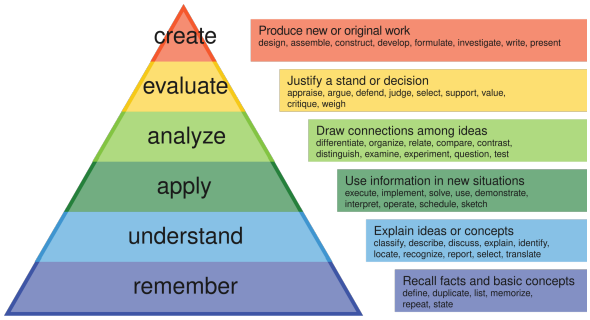

Figure 3: Bloom's Revised Taxonomy pyramid with verbs and descriptions for each cognitive level.
Image source: Wikimedia Commons

Our evaluation framework is designed for factoid or FAQ-style question sets that correspond to the two foundational levels: Remember and Understand. These include questions that require recalling factual information or explaining basic concepts. Higher-order cognitive tasks such as applying, analyzing, or synthesizing information are beyond the current scope of our automatic evaluation pipeline. Hence, this work does not evaluate assessment-type questions that aim to test deeper reasoning, judgment, or creativity.

B Information Units (IU) Construction

The pipeline begins by processing the source document, typically provided as a PDF file. We employ an OCR-based Markdown extractor to convert each page into text while preserving some structural cues like headings and bullet points. However, many of these visual artifacts (e.g., headings, image references, captions, footnotes, and tables) are not meaningful for content understanding and are filtered out during preprocessing.

The resulting clean text is split into a set of content blocks, denoted as $\mathcal{B} = \{b_1, b_2, ..., b_n\}$. These blocks are further categorized based on structural features: *paragraph-type blocks* consist of continuous prose and are segmented into individual sentences, while *list-type blocks* are preserved as full bulleted or numbered lists without splitting into individual items. After cleaning, paragraph sentences undergo filtering to remove short or content-poor units, whereas list blocks are retained as composite spans. Each selected sentence or complete list block is treated as a candidate answer span and referred to as an **Information Unit (IU)**. These IUs are grouped into:

- $\mathcal{N}$: the set of **necessary IUs**, consisting of individual sentences from cleaned paragraphs and full list blocks that represent the core document content.
- $\mathcal{O}$: the set of **optional IUs**, which includes sliding sentence windows extracted from paragraph-type blocks and decomposed list items used for auxiliary coverage.

This distinction between $\mathcal{N}$ and $\mathcal{O}$ is intentional and grounded in their functional roles during evaluation. While $\mathcal{N}$ anchors the evaluation to core content that a high-quality question set should collectively cover, $\mathcal{O}$ provides finer-grained and overlapping units that support flexible retrieval for per-question scoring. Using smaller windows or decomposed items in $\mathcal{O}$ enables more precise detection of answerability and redundancy, without compromising the semantic grounding ensured by $\mathcal{N}$. The complete IU space is given by:

$$\mathcal{I} = \mathcal{N} \cup \mathcal{O}$$

This pool ($\mathcal{I}$) of necessary and optional IUs forms the content grounding against which the generated questions are evaluated.

Each IU is associated with a metadata dictionary that supports downstream operations such as answer tracing and coverage propagation. This metadata includes the *type*, which specifies whether the IU is necessary or optional; the *subtype*, indicating its structural form such as paragraph, list, window, or bullet; a unique *chunk_id* assigned to every IU; and a *parent_chunk_id*, which links optional IUs to their corresponding necessary IUs when applicable.

B.1 Semantic Similarity For Optional Units

To semantically align optional IUs with their originating content, we establish subtype-specific linkages to necessary IUs using sentence-level cosine similarity. For optional IUs of subtype *window*, each constituent sentence is compared individually against all sentence-type necessary IUs (i.e., paragraph-derived units). If a sentence in the optional window surpasses a similarity threshold when compared to a necessary IU sentence, the corresponding chunk ID is added to the optional IU's *parent_chunk_id* list.

For optional IUs of subtype *bullet*, which represent decomposed items from original list blocks, we first break down each list-type necessary IU into individual items. Each optional bullet is then matched to these decomposed items, and if similarity exceeds the threshold, the chunk ID of the original necessary IU list block is mapped as the parent. This linkage process is conditional on the structural subtype of both optional and necessary IUs and results in a set of *parent_chunk_id* annotations that support hierarchical coverage propagation.

B.2 Dynamic Refinement of List Answers

Certain IUs, especially those of subtype *list*, may include several loosely related points, making it difficult for a cross-encoder to assign high relevance to the entire span. To address this, we implement a dynamic refinement mechanism that performs windowed span extraction over list items, but only when the cross-encoder score falls below a threshold. This ensures finer-grained retrieval for otherwise diffuse content blocks.

Given a list of items $L = [l_1, l_2, ..., l_m]$,we generate overlapping candidate windows using a predefined size w. The windows are constructed as:

$$W = \{l_i \oplus \cdots \oplus l_{i+w-1} \mid i = 1, ..., m - w + 1\}$$

Each window is scored using the same cross-encoder as in Section 4.2. The highest scoring window is retained as the refined answer span if it outperforms the original full list in terms of relevance to the question. This strategy allows the system to zoom in on the most contextually appropriate subset of a long list.

C Human Evaluation and Agreement Computation

Figure 4 provides the precise instructions given to human annotators. Given the short nature of the task, the annotators voluntarily agreed to participate without payment.

To validate our automatic evaluation pipeline, we conducted a human study with 7 independent annotators. Each annotator was presented with a passage and two question sets (generated by different systems) and was asked to select the set they preferred based on clarity, informativeness, and answerability. All annotators were South Asian and postgraduate level researchers affiliated with the Indian Institute of Technology Bombay.

Majority Vote: For each example, we recorded the votes from all 7 annotators. The question set with the majority of votes was treated as the human-preferred set. If there was no clear majority (e.g., a tie), the case was resolved by a separate tie-breaker annotator to ensure a definitive preference for agreement analysis.

Pipeline Prediction: The same pairs of question sets were scored using our automated evaluation pipeline. For each pair, the set with the higher aggregate score across metrics was selected as the pipeline's preferred set.

Agreement Computation: We then compared the pipeline's selected set to the majority human vote.

- If the pipeline's choice matched the majority vote, the example was counted as an **agreement**.
- For each metric individually, we computed how many times it assigned a higher score to the human-preferred question set. We reported both the raw count and the corresponding fraction across all examples.
- Finally, we counted how often the *aggregate score* (across all metrics) was higher for the human-preferred set. This overall fraction reflects how frequently the pipeline ranked the

You will be shown a series of .md files — each containing

- A passage
- Two sets of questions generated from that passage

Your task is to **carefully read the passage** and both sets of questions, then select **which question set is better** based on the following four criteria:

Evaluation Criteria (Keep These in Mind)

1. **Answerability:** *Do the questions have clear, answerable responses based on the passage?* Choose the set with more questions that have **clear, direct answers** in the passage.
2. **Redundancy:** *Do the questions repeat the same information in different ways?* Prefer the set with **less repetition** and **more diverse focus** across questions.
3. **Coverage:** *How well do the questions collectively cover the passage's key points?* A better set asks about **most of the important facts, ideas, or sections** in the passage.
4. **Structural Entropy:** *How varied are the question types?* Look for a mix of **factual (what/where), reasoning (why/how), and list-based** questions — not just one kind repeated.

Please read each passage and question set carefully. **Make your judgment holistically**, but try to keep all four criteria in mind. There are no right or wrong answers — we are collecting your subjective preferences to evaluate the quality of generated questions.

Figure 4: Instruction provided to human evaluators

human choice higher, even without access to human votes.

For disagreement cases, we also computed the mean human agreement with the pipeline-selected set, capturing how controversial or borderline those decisions were.