

Which Questions Improve Learning the Most? Utility Estimation of Questions with LM-based Simulations

Dong-Ho Lee*, Hyundong Cho*, Jonathan May†, Jay Pujara†

Information Sciences Institute, University of Southern California

dongho.lee@usc.edu, {jcho, jonmay, jpujara}@isi.edu

Abstract

Asking questions is fundamental for deeper understanding, yet evaluating and generating questions that improve downstream outcomes remains a challenging problem. To address this gap, we introduce QUEST (Question Utility Estimation with Simulated Tests), a framework that offers a new outcome-driven paradigm for question evaluation and generation, instead of indirect metrics that are based on the underlying content of the question and its answer. QUEST uses language models to simulate learners that ask questions while studying a textbook chapter and quantifies the utility of a question as its direct contribution to understanding as measured by the learner’s exam performance. To analyze the effectiveness of QUEST, we curate TEXTBOOK-EXAM, a benchmark that aligns textbook sections with exam questions across five academic disciplines. Using QUEST, we filter for high-utility questions and fine-tune question generators via rejection sampling and show that questions generated by QUEST-trained models improve simulated exam scores by 18.5% compared to the next strongest baseline. Furthermore, we demonstrate the practical real-world value of QUEST by showing that simulation-computed utility of a question is moderately positively correlated with its usefulness as perceived by human evaluators.

1 Introduction

Asking questions is a crucial skill for improving understanding. Effective teachers pose questions to stimulate critical thinking (Bloom et al., 1956; Tofade et al., 2013), while high-performing students actively ask clarifying questions (Abbott, 1980; Brassell and Rasinski, 2008; Davoudi and Sadeghi, 2015). But among the many questions one could ask, how can we identify those that are beneficial for improving downstream learning out-

*Authors contributed equally

†Equal advising contribution

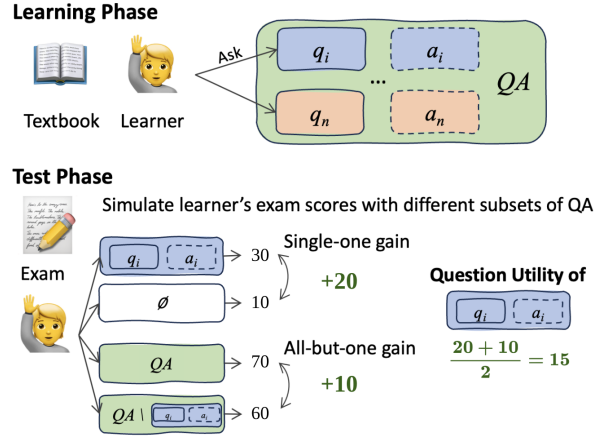


Figure 1: **Question utility** is defined as the direct impact of a question-answer (QA) pair on a learner’s understanding, measured by simulated exam performance improvement. QUEST estimates this marginal benefit by comparing simulated learners with and without access to the QA pair, enabling question evaluation and generation based on direct outcome rather than indirect metrics.

comes, and how can we systematically generate such questions?

The psychology literature offers general heuristics (Graesser et al., 2010; Vale, 2013) while recent NLP works have focused on quantifying the quality of *inquisitive questions*, which are learner-generated and curiosity-driven queries (Scialom and Staiano, 2020), using indirect metrics such as salience (Wu et al., 2024) or expected information gain (Rao and Daumé III, 2018; Keh et al., 2024). Although such proxies reflect properties like relevance or informativeness, they do not measure a question’s impact on downstream outcomes.

We propose a new direction: measuring a question’s utility by its *direct* impact on downstream outcomes. Inspired by recent work leveraging language models (LMs) as simulators of human behavior (Zelikman et al., 2023; Park et al., 2024; Zhang et al., 2024; He-Yueya et al., 2024), we propose QUEST (Question Utility Estimation with

Simulated Tests), a framework that estimates question utility – how much a question helps a learner by simulating its contribution to a downstream task – and uses high-utility questions to fine-tune question generators.

In QUEST, as shown in Figure 1, an LM role-plays as a novice learner who poses questions while studying learning materials and receives answers from another LM that acts as a teacher (§2.3). We estimate the *utility* of each question by simulating exam performance taken with different QA subsets, without access to the original text, to isolate their marginal benefit (§2.4). QUEST can then fine-tune question generators via rejection sampling (Bai et al., 2022) with high-utility questions that surpass a certain threshold (§2.5).

To evaluate QUEST, we curate TEXTBOOK-EXAM, a dataset in which each sample consists of a chapter with multiple sections and corresponding exam questions (§3). Using TEXTBOOK-EXAM, we experiment with various question generation strategies and quality metrics. We find that QUEST-trained question generators produce questions that lead to an average improvement of 18.5% in simulated exam scores across five academic subjects (§5.1). In addition, correlation analysis reveals that indirect metrics such as salience and expected information gain are not correlated with utility and that optimizing for them yields suboptimal results (§5.2). While exam question similarity is weakly correlated with utility, models fine-tuned to predict exam questions from the chapter content produce low-utility questions. Lastly, through a human evaluation study with a setup similar to QUEST, we find that simulation-computed utility is moderately correlated with the usefulness of the question perceived by human evaluators, thus validating the practical value of QUEST (§5.5).

In summary, our main contributions are: (i) QUEST, a framework that quantifies the utility of a question by estimating its direct impact on downstream outcomes through LM-based simulations of a learner; (ii) TEXTBOOK-EXAM, a benchmark that aligns textbook sections with exam questions across five academic disciplines, to evaluate QUEST; (iii) demonstrating that QUEST-trained question generators produce questions that result in significant improvements in simulated exam scores on TEXTBOOK-EXAM; (iv) establishing the practical and unique value of QUEST by showing that simulation-computed utility of a question is moderately correlated with the usefulness of the question

perceived by human evaluators but not correlated with indirect metrics.

2 QUEST

In this section, we present QUEST,¹ a framework for measuring a question’s utility based on its impact on downstream learning outcomes and fine-tuning question generators with high-utility questions. We first formulate the problem and define our core evaluation metrics. Then, we provide an overview of the QUEST framework and describe each component in detail.

2.1 Problem Formulation

Our objective is to generate questions that effectively support learners studying a given document D . We assume access to a document D consisting of sequential sections $S_1, \dots, S_k$, and a set of exam questions $E = \{e_1, \dots, e_m\}$ designed to evaluate understanding of the content in D .

Given this setup, we develop a question generator M_q that produces a set of questions $Q = \{q_1, \dots, q_n\}$. Each question q_i is paired with an answer a_i , generated independently using a fixed answer generator M_a based on general parametric knowledge. We assume that all answers are approximately correct and consistent across questions, so that any observed difference in learner performance can be attributed to the quality of the question itself. The resulting set of question–answer pairs $QA = \{(q_1, a_1), \dots, (q_n, a_n)\}$ is used to support a simulated learner, who attempts the exam E having access only to these QA pairs. We exclude the document D to isolate the instructional value of the questions, avoiding confounding effects from the document content. The learner’s performance on the exam serves as a proxy for the instructional quality of the generated questions.

To quantify this, we introduce two evaluation metrics: (1) *exam score* is defined as the performance of the simulated learner M_l on E , using only the QA pairs as study material. This score, measured as average accuracy across exam questions, ranges from 0 to 100 and reflects how well the questions prepare the learner; (2) *utility* u_i of each question q_i is defined as its marginal contribution to the *exam score*. Unlike traditional educational frameworks such as Item Response Theory (IRT), which model latent traits such as learner

¹All our code and data are available at <https://github.com/usc-isi-i2/quest>.

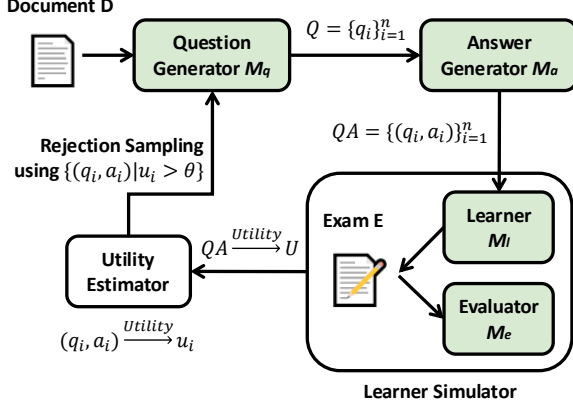


Figure 2: **QUEST Framework.** M_q generates questions and M_a provides answers. A simulated learner M_l studies the QA pairs and takes an exam E , which is scored by evaluator M_e . The utility estimator measures each question’s impact on exam performance, and only high-utility questions are retained to refine M_q .

ability or test item difficulty (Harvey and Hammer, 1999; Sumita et al., 2005; Lalor et al., 2016; Zhou et al., 2019), our formulation of utility focuses on the *interventional* effect of a question asked while learning on the outcome (i.e. exam performance) for a fixed simulated learner with no prior knowledge of D who learns only from the provided QA pairs.

2.2 Overview

QUEST consists of the following four components, corresponding directly to the outcome-based formulation above: (1) **question generator** (M_q) takes a document D and generates a set of questions $Q = \{q_1, \dots, q_n\}$, grounded in the document’s content; (2) **answer generator** (M_a) independently produces an answer a_i for each question q_i using general parametric knowledge, forming the set of QA pairs $QA = \{(q_1, a_1), \dots, (q_n, a_n)\}$; (3) **learner simulator**, composed of a learner model M_l and an evaluator M_e , assesses the effectiveness of QA by having M_l attempt an exam E using only the QA pairs, and M_e compute the resulting exam score; (4) **utility estimator** runs the learner simulator under different QA subsets to estimate each pair’s marginal contribution to the learner’s exam performance. This enables selection of high-utility questions for refining M_q via rejection sampling. See Figure 2 for a visual summary. Additional details, such as the prompts we use and validation of the reliability of answers by M_a and scores produced by M_e , are provided in Appendix A.

2.3 Generating Questions

The **Question Generator** (M_q) produces a set of questions Q from the input document D . For each section S_k , which represents the part of the document the learner is currently reading (the *anchor*), it considers both S_k and its preceding context $C_k = \{S_1, \dots, S_{k-1}\}$ to generate a set of questions $Q_k = M_q(S_k, C_k)$.

This approach ensures that the questions are informed by both local and global context, following prior work on inquisitive question generation (Wu et al., 2024). Each question q_i is then paired with an answer $a_i = M_a(q_i)$, generated independently by the **Answer Generator** (M_a) using general parametric knowledge. The resulting QA pairs $QA = \{(q_1, a_1), \dots, (q_n, a_n)\}$ serve as input for learner simulation and utility estimation.

2.4 Evaluating Question Generators via Exam Score

To isolate the effect of using different question generators M_q , the learner’s exam performance is simulated with access to only the generated QA pairs; i.e. M_l answers the exam E based solely on $QA = \{(q_1, a_1), \dots, (q_n, a_n)\}$, producing responses $P_{QA} = M_l(E, QA)$.

These responses are then scored by an evaluator model M_e , which either compares them to ground-truth answers or uses its own parametric knowledge when ground truth is unavailable or insufficient, such as for open-ended or free-form questions. This yields an exam score $s_{QA} = M_e(E, P_{QA})$, which serves as a direct measure of how effectively the generated questions support learning.

2.5 Improving Question Generator via Utility

Our objective is to improve the question generator M_q by selecting and learning from high-utility questions, those that most effectively enhance exam performance when studied by the learner simulator.

To do this, we estimate the utility u_i of each QA pair (q_i, a_i) based on its marginal contribution to exam performance. The utility estimator runs the learner simulator under two perturbations: (1) the *single-one gain*, defined as the exam score when only (q_i, a_i) is provided: $s_{\{(q_i, a_i)\}} - s_\emptyset$, which captures how much a learner can benefit from the pair in isolation; and (2) the *all-but-one gain*, defined as the score drop when (q_i, a_i) is removed from the full QA set: $s_{QA} - s_{QA \setminus \{(q_i, a_i)\}}$, which captures how essential the pair is in context. While $s_\emptyset$ ide-

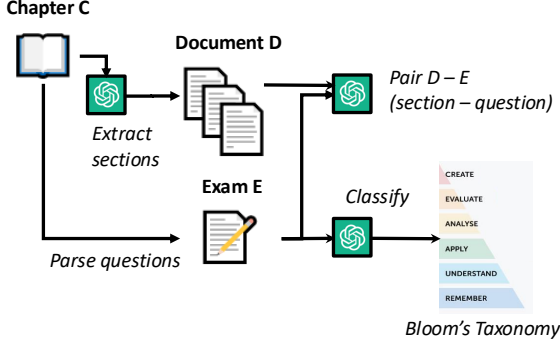


Figure 3: **TEXTBOOK-EXAM Curation Pipeline.** Given a chapter C , we segment its content into document sections $D = \{S_1, \dots, S_k\}$ and extract exam questions to form the exam E . Each exam question is classified by Bloom’s taxonomy and aligned to relevant sections.

ally reflects a zero-knowledge baseline, it may be non-zero in practice due to the learner’s parametric knowledge (e.g., general world knowledge or memorized patterns from pretraining). We treat this as a background prior and compute utility relative to this baseline. The final utility is computed as the weighted average, with α as a hyperparameter, of the two gains:

$$u_i = \alpha(s_{\{(q_i, a_i)\}} - s_0) + (1 - \alpha)(s_{QA} - s_{QA \setminus \{(q_i, a_i)\}})$$

We retain only the QA pairs with $u_i \geq \theta$, and update M_q using these high-utility examples via a rejection sampling strategy (Bai et al., 2022). We set $\alpha = 0.5$ unless mentioned otherwise.

3 TEXTBOOK-EXAM: A Dataset for Outcome-Based Question Evaluation

To evaluate the effectiveness of QUEST in diverse learning settings, we construct TEXTBOOK-EXAM, a dataset composed of documents D and corresponding exam questions E , derived from real-world textbooks. Each entry is designed to simulate a realistic learning task where a learner studies D and is later evaluated on E .

3.1 Data Processing

We start with college-level textbooks from the OpenStax repository.² Each textbook is divided into chapters, and for each chapter C , we extract the main content to form the document D and the end-of-chapter exam questions to form the exam E .

²<https://github.com/philschatz/textbooks>

Subject	# C	Split	# E/C	% E w/ A	# S/C
Microbiology	20	Train	12.4	64%	16.4
	5	Test	13.4	58%	17.0
Chemistry	20	Train	14.2	51%	11.0
	5	Test	16.2	49%	6.4
Economics	20	Train	12.2	23%	14.1
	5	Test	12.2	23%	14.4
Sociology	20	Train	10.4	62%	16.6
	5	Test	11.2	67%	19.0
US History	20	Train	7.2	51%	14.9
	5	Test	8.4	38%	13.2

Table 1: **Data Statistics** of TEXTBOOK-EXAM. # C : number of chapters, # E/C : avg. number of questions per chapter, % E w/ A: proportion of questions that have a reference answer, # S/C : avg. number of sections per chapter.

Sectioning the Document. To simulate a learner progressing incrementally through a document, we divide each D into sections $\{S_1, \dots, S_k\}$ using an LM-based segmentation method. We provide few-shot examples of sections for each subject for consistent structuring across subjects. This setup allows for evaluating how well questions generated from partial context can prepare learners for the full exam.

Forming the Exam. We extract exam questions to form E from each chapter’s end-of-chapter materials. To ensure sufficient exam coverage while maintaining computational feasibility, we retain only chapters with at least 10 exam questions and cap each set to a maximum of 25. We also use an LM to map each exam question to the sections relevant for answering it in order to use this information for baselines such as few-shot and supervised fine-tuning (SFT). To analyze the cognitive demands of the exam questions, we annotate each $e_j \in E$ with a Bloom’s taxonomy category (Krathwohl, 2002) using an LM. Refer to Appendix B.4 for details on question type diversity in TEXTBOOK-EXAM according to Bloom’s taxonomy.

Train-Test Split. For each subject, we select 25 chapters C arranged in natural curriculum order: the first 20 for training and the last 5 for evaluation. Preserving sequential structure helps avoid leakage of prerequisite knowledge and supports realistic evaluation of learning progression.

3.2 Data Statistics

The statistics for TEXTBOOK-EXAM resulting from data processing are shown in Table 1. The general distributions of the numbers of questions and sections per chapter are similar between the

training and test sets. Refer to Appendix B for further details on TEXTBOOK-EXAM.

4 Experimental Setup

To evaluate QUEST, we compare a variety of question generation (QG) methods and analyze how well *utility* correlates with other question quality metrics (*i.e.*, *salience*, *expected information gain*).

4.1 Baselines

No-Study baseline. $s_\emptyset$ denotes the exam score of a simulated learner without any study material, reflecting prior knowledge and serving as a lower bound.

Prompting baselines.

- **Zero-shot:** Prompts the base model to generate a question that helps a student understand a given document section S_k , without providing any examples.
- **Few-shot:** Augments the zero-shot prompt with five (section, exam question) pairs randomly sampled from the training set.
- **Chain-of-Thought (CoT):** Similar to zero-shot, but the model is asked to first generate a reasoning trace before writing the question (Wei et al., 2022; Zhou et al., 2024).
- **Bloom-based:** Samples a Bloom’s Taxonomy level (Krathwohl, 2002) from the distribution observed in the training set (Figure 7), then prompts the model to generate a question at that cognitive level for the current section.

Fine-tuning baselines.

- **Supervised Fine-tuning (SFT):** fine-tunes the base model to predict exam questions given (section, exam question) pairs.
- **QUEST (Ours):** applies rejection sampling to fine-tune the model on only high-utility QA pairs from simulations with the training data. We fix the utility threshold θ to 0.1 based on empirical performance observed on a development subject (*i.e.*, Chemistry), and use this value consistently across all subjects.

For SFT and QUEST, we experiment with two variants: (a) *Subject-specific*: model is trained and evaluated on the same subject; (b) *Cross-subject*: model is trained on all subjects and evaluated on each individually.

4.2 Question Quality Metric

To evaluate whether our measure of utility is a good indicator of question quality, we compare it to the following alternative quality metrics:

Salience. Measures a question’s relevance to a document D by a LM (Wu et al., 2024). It is rated on a Likert scale from 1 to 5: 1 indicates that the question for section k is unrelated to $S_{[1:k]}$ and contributes minimally to understanding, while 5 indicates that it is essential to comprehending S_k .

Expected Information Gain (EIG). Quantifies the reduction in uncertainty about a student’s knowledge state after answering a question (Lindley, 1956; Schaeffer and Presser, 2003; Rao and Daumé III, 2018; Yu et al., 2020; White et al., 2021; Keh et al., 2024). We estimate EIG using an LM by computing the entropy of the model’s token probability distribution before and after conditioning on the first token of the correct answer, capturing the shift in the model’s belief distribution.

4.3 Experiments and Implementation Details

All LMs used in data preprocessing and throughout the framework, including the question generator M_q , answer generator M_a , learner M_l , and evaluator M_e in the reader simulator, are based on gpt-4o-mini. For SFT in the baseline and rejection sampling in QUEST, we use OpenAI fine-tuning API to further train gpt-4o-mini. We limit question generation to a single question per section and run three trials on the test set due to computational cost constraints of running inference and fine-tuning. Refer to Appendix A.6 for a cost analysis of QUEST. For all other implementation details not covered here, refer to Appendix C for details on baselines and D for details on question quality metrics.

5 Experimental Results

We empirically evaluate QUEST along four dimensions: (1) overall exam performance; (2) relationship of utility with existing quality metrics and qualitative characteristics of high-utility questions; (3) the effects of design choices such as rejection threshold and model variants; and (4) human evaluation to validate QUEST results.

5.1 Overall Performance

Table 2 reports the simulated *exam scores* for different question generators. Our main findings are:

	Microbiology	Chemistry	Economics	Sociology	US History	Average
No-study s_0	45.8 _{1.0}	9.2 _{0.9}	0.3 _{1.1}	61.2 _{0.8}	3.2 _{1.2}	23.9 _{0.7}
Zero-shot	62.3 _{0.4} (+16.5)	39.7 _{0.3} (+30.5)	40.3 _{0.5} (+40.0)	60.7 _{0.2} (-0.5)	19.4 _{0.1} (+16.2)	44.5 _{0.4} (+20.6)
Few-shot	61.6 _{0.5} (+15.8)	45.6 _{0.2} (+36.4)	46.8 _{0.3} (+46.5)	62.4 _{0.1} (+1.2)	15.8 _{0.5} (+12.6)	46.4 _{0.2} (+22.5)
Chain-of-thought	61.4 _{0.3} (+15.6)	45.4 _{0.5} (+36.2)	45.9 _{0.2} (+45.6)	61.1 _{0.1} (-0.1)	18.6 _{0.4} (+15.4)	46.5 _{0.3} (+22.6)
Bloom-based	57.3 _{0.1} (+11.5)	36.6 _{0.5} (+27.4)	29.4 _{0.2} (+29.1)	61.5 _{0.4} (+0.3)	21.7 _{0.3} (+18.5)	41.3 _{0.5} (+17.4)
SFT (Subject-Specific)	65.4 _{0.2} (+19.6)	23.9 _{0.3} (+14.7)	46.1 _{0.1} (+45.8)	64.2 _{0.4} (+3.0)	20.2 _{0.5} (+17.0)	44.0 _{0.2} (+20.1)
SFT (Cross-Subject)	58.7 _{0.4} (+12.9)	21.1 _{0.1} (+11.9)	46.8 _{0.5} (+46.5)	63.4 _{0.3} (+2.2)	26.3 _{0.2} (+23.1)	43.3 _{0.1} (+19.4)
QUEST (Subject-Specific)	75.6_{0.5}* (+29.8)	45.7_{0.4} (+36.5)	58.2_{0.2}* (+57.9)	64.9_{0.3} (+3.7)	31.1_{0.1}* (+27.9)	55.1_{0.5}* (+31.2)
QUEST (Cross-Subject)	<u>73.3_{0.2}* (+27.5)</u>	41.3 _{0.3} (+32.1)	46.8 _{0.4} (+46.5)	64.8 _{0.5} (+3.6)	25.2 _{0.2} (+22.0)	50.3 _{0.1} * (+26.4)

Table 2: **End-of-chapter exam scores** for different question generation methods, averaged over three trials. QUEST achieves the highest scores across all subjects. Gains (in parentheses) are relative to the no-study base score s_0 . Standard errors are reported as subscripts. Top and second-best scores are **bolded** and underlined. *Improvements over non-QUEST baselines that are statistically significant at $p < 0.01$ with paired t-test.

Metric 1	Metric 2	Correlation ρ	p -value
Utility	Salience	0.097	0.003
	EIG	-0.022	0.512
	Exam Embedding Sim.	0.250	0.001
	Exam ROUGE	0.040	0.010
	Bloom’s Taxonomy	0.120	0.001

Table 3: **Correlation** between utility and other metrics.

(1) **Prompting methods** (Few-shot, CoT, Bloom-based) improve over the no-study baseline, leveraging better task understanding and reasoning (Brown et al., 2020; Wei et al., 2022; Zhou et al., 2024). However, they do not explicitly optimize for learning outcomes, limiting their effectiveness; (2) **SFT** shows mixed results—improving in some subjects but underperforming in others—and does not consistently outperform prompting on average; and (3) **QUEST** achieves the highest exam scores across all subjects, with an average gain of 18.5% over the next strongest baseline (CoT). Its subject-specific variant consistently outperforms the cross-subject version, highlighting the importance of domain-specific optimization. These results show that outcome-based filtering and training yields the most effective instructional questions.

5.2 Question Utility Analysis

Correlation with indirect metrics. To analyze whether indirect metrics align with our proposed *utility*, we compute all three metrics (*utility*, *salience*, and *expected information gain (EIG)*) on generated questions from the training set. As shown in Table 3, utility has only a weak correlation with salience ($\rho = 0.097$) and no significant correlation with EIG ($\rho = -0.022$).

This suggests that utility captures a distinct signal from existing indirect metrics, which may not

reliably reflect a question’s actual impact on learning outcomes. Table 4 makes that gap concrete for Chemistry: a high-salience, high-utility question asks how combining elements yields new properties; a low-salience, high-utility question asks for a specific list of green-chemistry principles that happens to help on the exam; and a high-salience, low-utility question asks a generic definition of ionic charge. The same High/High, Low/High, and High/Low pattern appears across subjects (Appendix E.1), with no consistent surface-level cue that separates high- from low-utility questions.

Overlap with exam questions. To evaluate the relationship between generated high-utility questions and exam questions, we measure their semantic and lexical similarity. For each generated question, we compute cosine similarity using embeddings with text-embedding-3-small from OpenAI for semantic overlap and the ROUGE score (Lin, 2004) for lexical overlap with all exam questions in the same chapter. We then assess the correlation between utility and the most similar exam question based on these measures. The correlation between utility and semantic similarity is 0.25 ($p < 0.001$), indicating a weak positive relationship, while the correlation with ROUGE is nearly zero at 0.04 ($p < 0.01$). These findings suggest that high-utility questions are not simple paraphrases of exam questions but may introduce novel concepts that enhance learning beyond surface-level similarity.

Bloom’s taxonomy analysis. An interesting observation is that Bloom’s taxonomy, which categorizes cognitive depth based on question type—where “what” questions typically involve simple recall, while “why” and “how” questions

Sal.	Util.	Question
High	High	How does the combination of elements result in properties that are different from those of the individual elements in their uncombined states, and can you provide more examples of such transformations?
Low	High	What are the ‘Twelve Principles of Green Chemistry’?
High	Low	What happens to the charge of an atom when it gains or loses electrons?

Table 4: **Chemistry examples** of generated questions by salience and utility. Additional subjects are in Appendix E.1.

	Microbiology			Chemistry			Economics			Sociology			US History		
Filter	E	Sal.	EIG	E	Sal.	EIG	E	Sal.	EIG	E	Sal.	EIG	E	Sal.	EIG
<i>utility</i> > 0.1	76	4.27	-.18	46	4.65	-.20	58	4.70	-.04	65	4.49	-.02	31	4.65	-.01
<i>salience</i> = 5	73	4.42	-.24	39	4.46	-.22	46	4.66	-.08	64	4.49	-.03	23	4.68	-.02
<i>EIG</i> > 0	61	4.21	-.17	32	4.40	-.09	47	4.65	.01	62	4.46	.01	21	4.65	-.01

Table 5: **Performance of QUEST models trained on different filtering criteria.** Each group reports: exam score (**E**), average salience (**Sal.**), and expected information gain (**EIG**).

require deeper processing—also does not strongly correlate with utility. Using Bloom’s taxonomy as a cognitive depth scale (Likert 1-6), the correlation between utility and cognitive depth is 0.12 ($p < 0.001$), indicating a weak positive relationship.

Optimization on indirect metrics. To assess whether indirect metrics can serve as effective training objectives, we compare QUEST models trained on datasets filtered by three different criteria: (1) questions with *utility* > 0.1 (Ours), (2) questions with *salience* = 5, and (3) questions with *EIG* > 0. We evaluate each model’s output using three metrics: *exam score* (utility), average salience, and average EIG. Results are shown in Table 5.

We observe that training on utility-optimized questions consistently yields the highest exam performance across all subjects. In contrast, optimizing for salience or EIG improves the respective metric, but fails to enhance utility.

These results underscore a key distinction: indirect metrics may capture intrinsic quality (e.g., relevance or informativeness), but do not translate into improved downstream outcomes. In contrast, direct optimization on utility, measured via simulated exam performance by modeling a learner’s mental model, leads to more effective question generation for downstream outcomes.

5.3 Rejection Sampling Analysis

We examine the effect of the utility threshold on rejection sampling results by varying it for chemistry as a representative example. Figure 4 illustrates that increasing the utility threshold leads to gen-

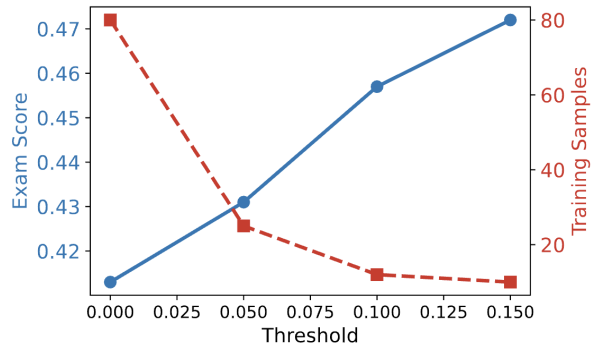


Figure 4: **Impact of threshold** in QUEST on exam scores for Chemistry.

erated questions that enable higher exam scores, confirming that stricter filtering selects more educationally valuable questions. However, higher thresholds also reduce the number of training examples, potentially limiting model learning. This trade-off highlights the importance of balancing quality and quantity. The results suggest that expanding the training pool by sampling more questions per section while maintaining a high threshold may yield further improvements.

5.4 Model Variants

To test whether QUEST’s gains depend on a particular model size, we vary the model used for the question generator (M_q), answer generator (M_a), and reader simulator (M_l), changing one module at a time in the zero-shot setting. For the M_a and M_l variants, we reuse the same questions (and answers) from the zero-shot baseline and re-run only the affected stage. Results are shown in Table 6.

Upgrading M_q or M_a to gpt-4o yields moderate gains (5.7% and 7.1% on average), but both still

	QG (M_q)	AG (M_a)	RS (M_l)	Microbiology	Chemistry	Economics	Sociology	US History
Zero-shot	4o-mini	4o-mini	4o-mini	62.0	41.4	39.8	60.9	23.3
	4o	4o-mini	4o-mini	68.1	45.7	46.6	63.4	18.0
	4o-mini	4o	4o-mini	68.2	42.2	48.0	63.4	23.2
	4o-mini	4o-mini	4o	71.0	17.3	47.6	56.4	26.3
QUEST	4o-mini	4o-mini	4o-mini	75.6	45.7	58.2	64.9	31.1

Table 6: **Model-size ablations** of QUEST modules on TEXTBOOK-EXAM. Each zero-shot row changes one of M_q , M_a , or M_l while holding the others fixed. QUEST remains strongest across subjects.

trail the QUEST-trained gpt-4o-mini generator by more than 11%. Using gpt-4o as the reader simulator is mixed: some subjects improve, while Chemistry drops sharply, suggesting that a larger learner can change scoring behavior rather than uniformly raise exam scores. Overall, substituting a stronger model into a single module does not replace training on high-utility questions.

5.5 Human Evaluation

The validity of QUEST results, and thus QUEST’s practical value of estimating question utility and using it to improve question generation, hinges on whether utility reflects actual usefulness for downstream outcomes perceived by humans, and that it is distinct from other metrics such as salience and EIG. To validate that utility is indeed a reliable approximation of such usefulness, we conduct a human evaluation study³ where human participants take on the role of a learner in QUEST. For a single evaluation sample, we ask human participants to answer a set of related exam problems after reading a section of a chapter and a set of questions, one of each with exclusively high utility, high salience, and high EIG, and their corresponding answers. Next, they are instructed to rank questions without ties based on their usefulness for solving the exam problems and also rate their usefulness and interestingness (i.e. how intriguing the question is regardless of its usefulness) on a Likert scale from 1 to 5, and provide a brief explanation for their rankings and ratings.

This study is construct validation of the utility metric, not a head-to-head comparison of generators. An ideal check would compare QUEST-trained models against every baseline with human learners across many sections and subjects, but that design is impractical: solving exam items after multiple QA sets is cognitively demanding, so session length, fatigue, and variance grow quickly,

³The human evaluation study is available at <https://quest-human-eval.streamlit.app/>.

	EIG	Salience	Utility
ρ Useful. $\uparrow$	−0.219	−0.007	0.321
μ Useful. $\uparrow$	3.00 _{0.10}	2.97 _{0.09}	3.96 _{0.09} **
μ Interesting. $\uparrow$	3.13 _{0.08}	3.23 _{0.09}	3.26 _{0.09}
μ Rank $\downarrow$	2.25 _{0.06}	2.04 _{0.05}	1.71 _{0.05} *

Table 7: **Human evaluation results.** μ denotes the mean and ρ denotes correlation. Standard errors are reported as subscripts. * and ** denote statistically significant at $p < 0.05$ and $p < 0.01$ with paired t-test, respectively.

and fair comparisons require counterbalancing that multiplies conditions. Human sessions also cannot score a full exam the way QUEST does; they must be limited to a few items. We therefore test whether simulation-computed utility predicts human-perceived usefulness across models and subjects. The results are shown in Table 7.

The results show that utility is moderately positively correlated with usefulness perceived by humans ($\rho = 0.321$), while other metrics are weakly negatively correlated. In addition, high utility questions are consistently ranked and rated higher than high salience and high EIG questions for usefulness. These results establish that utility as measured by LM simulations can capture usefulness perceived by humans to some extent. It is also worth noting that high utility questions are rated as interesting as high salience and high EIG questions, indicating that optimizing for utility does not necessarily lead to a tradeoff with interestingness. Participant instructions, exclusivity sampling, and the evaluation interface are in Appendix A.5.

6 Related Work

Question Generation. Recent studies have explored generating *information-seeking* questions to enhance comprehension by reflecting the inquisitive nature of human question-asking. These include curiosity-driven inquisitive questions (Ko et al., 2020; Wu et al., 2024), confusion-driven

inquisitive questions (Chang et al., 2024), follow-up questions (Meng et al., 2023), and clarifying questions (Chen et al., 2018; Rao and Daumé III, 2018, 2019; Kumar and Black, 2020; Majumder et al., 2021). A widely adopted framework in this area is the concept of Questions Under Discussion (QUD), which views each sentence as the answer to an implicit or explicit question from prior context (Van Kuppevelt, 1995; Roberts, 2012; Onea, 2016; Benz and Jasinskaja, 2017). Building on this framework, recent works have applied question generation to tasks such as decontextualization (Newman et al., 2023), which recovers missing context to make snippets standalone, and elaboration (Wu et al., 2023), which generates additional details to enhance clarity. Other applications include discourse comprehension by linking sentences to their broader context (Ko et al., 2022), planning for summarization (Narayan et al., 2023), and modeling information loss during text simplification (Trientes et al., 2024).

Question Evaluation. Existing work evaluates information-seeking questions based on expected information gain, measuring the additional information provided by the answers (Lindley, 1956; Schaeffer and Presser, 2003; Rao and Daumé III, 2018; Yu et al., 2020; White et al., 2021; Keh et al., 2024), or salience, assessing a question’s relevance and importance in introducing new concepts or clarifying key ideas (Wu et al., 2024). Empirical findings show that QUDs—questions guaranteed to be addressed in the article—consistently receive high salience ratings (Ko et al., 2022), reflecting reader expectations and linking to the utility of the question (Van Rooy, 2003; Wu et al., 2024). In contrast, our work directly evaluates the utility of questions using an LM as a human simulator.

Another prominent line of work that evaluates question quality is test information functions in item response theory (Harvey and Hammer, 1999; Sumita et al., 2005; Lalor et al., 2016; Zhou et al., 2019), but its primary concern is measuring how well existing exam questions can differentiate learners of varying proficiency or estimating latent traits across learners, not the *interventional* effect of a question asked while learning on improving downstream performance.

LM as Human Simulation. LMs have been used to simulate real-world social interactions, starting with everyday human interactions (Park et al., 2024; Maharana et al., 2024; Lee et al., 2025) and ex-

panding to controlled environments such as market competition (Zhao et al., 2024), hospital settings where medical staff and patients engage in treatment simulations (Li et al., 2024; Schmidgall et al., 2024), news interviews (Lu et al., 2024), content moderation (Cho et al., 2024; Shaikh et al., 2024), and academic peer review (Jin et al., 2024). In education, LMs simulate classroom interactions between teachers and students (Zhang et al., 2024) and predict learning outcomes to improve educational materials (He-Yueya et al., 2024; Zelikman et al., 2023). Our work builds on He-Yueya et al. (2024), using simulated learning outcomes as an evaluation metric for question quality and as a reward signal to improve question generation.

7 Conclusion

We introduce QUEST, a framework for estimating question utility by measuring the question’s impact on a simulated learner’s performance on real exams and fine-tuning models to generate high-utility questions. We establish that question utility reflects usefulness perceived by humans to some extent, establishing QUEST’s practical value of estimating question utility and using it to improve question generation. We show that QUEST-trained models consistently outperform existing prompting and fine-tuning baselines. On the other hand, indirect metrics from previous work such as salience and expected information gain fail to capture actual learning outcomes, emphasizing the need for evaluation grounded in task-specific goals.

Limitations

We acknowledge that our scope of question utility and QUEST studied in this work is confined to an educational setting with textbooks that contain learning material with corresponding exam questions. However, the definition of question utility as the contribution a question and its resulting answer makes to improved outcomes for a downstream task is generalizable and can be applied to other scenarios where downstream performance can be concretely quantified.

One of the main technical limitations of QUEST is that we generate only one question per section due to budget constraints. Due to the stochastic nature of language models, the type of questions that get generated for each section may vary significantly. However, we believe this effect to be minimal and the variance to be sufficiently cap-

tured in our results, as the improvements of QUEST over other baselines were statistically significant at $p < 0.01$ on average and for three of the five subjects in TEXTBOOK-EXAM, as reported in Table 2. Due to the same constraints, we have not tested QUEST with rejection sampling that generates more than one question per section. However, the threshold analysis results shown in Figure 4 indicate that setting a higher threshold leads to larger improvements, and therefore sampling a larger set of questions from each section will most likely lead to further improvements for QUEST. Since QUEST is effective even with the minimal setup of generating a single question per section, the main value of our work remains valid in that we have conceptually demonstrated that rewards from LM-based simulations can lead to meaningful improvements over baselines.

A major aspect overlooked by our work is the effect of a learner’s persona and prior knowledge on the utility of a question. There is a wide variety in learning styles and the extent of prior knowledge that a learner may have, which all influence the utility of a question. As a preliminary exploration, we have kept the learner persona and prior knowledge fixed for all experiments. We leave it to future work to explore the effect of varying learner personas and prior knowledge on measuring question utility.

Lastly, QUEST simulations incur a significant number of API calls, which can be costly and time-consuming, making it difficult to scale when hyperparameters such as the number of questions and exam problems become large. However, as shown in our cost analysis in Appendix A.6, the cost is still relatively affordable given the high number of API calls and total tokens when using gpt-4o-mini. In addition, with the rapid improvements in language model inference efficiency, costs and time for QUEST simulations will continue to decrease. Most importantly, we argue that QUEST enables a scalable and controlled experimental setup to measure outcome-driven metrics that is otherwise impractical in the real world and have shown potential for practical value via human evaluation that the utility of a question as measured by QUEST reflects actual usefulness as perceived by humans to some extent. With further improvements in language model capabilities for simulating human behavior, frameworks like QUEST will become more reliable and applicable to more complicated and nuanced use cases. We hope this work inspires more work

that grounds utility in real human outcomes for complex tasks that were previously infeasible to model.

Ethical Considerations

Although our work has shown that LM-based simulators can provide an effective reward signal for training better question generators for a specific downstream task and that this reward signal reflects actual usefulness as perceived by humans, we make no claims to how QUEST correctly simulates a real human being’s learning process and behavior. Human learning is a complex process involving long-term memory, context, and numerous other factors that are not captured by QUEST. Therefore, we do not advocate using our measure of question utility beyond a proof of concept for developing metrics grounded in downstream outcomes. In other words, QUEST should not be used to directly assess the quality of questions that students ask in a classroom setting and provide feedback based on its assessment.

The main reason is that the core assumption of our work – that the exam problems contained in TEXTBOOK-EXAM suitably evaluate a learner’s understanding of the learning material – is not guaranteed to hold in other settings. The utility of a question as defined by QUEST is heavily dependent on what the exam problems are, and if the exam questions are misaligned with desirable educational outcomes, e.g., exam questions that only require rote memorization rather than critical thinking, a student’s question may be considered low utility despite being useful in other scenarios, such as brainstorming. While we take extensive measures to ensure that this is not the case in our experiments given the wide variety of questions across subjects and types of questions included in TEXTBOOK-EXAM as shown in Figure 7, we cannot guarantee similar diversity and comprehensiveness in other textbook datasets.

Acknowledgments

This work was funded by the Defense Advanced Research Projects Agency with award HR00112220046. We thank the volunteer participants who contributed to our human evaluation.

References

Gerry Abbott. 1980. Teaching the learner to ask for information. *TESOL Quarterly*, pages 5–16.

- Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. 2022. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.
- Anton Benz and Katja Jasinskaja. 2017. Questions under discussion: From sentence to discourse.
- Benjamin S Bloom, Max D Engelhart, EJ Furst, Walker H Hill, and David R Krathwohl. 1956. Handbook i: cognitive domain. *New York: David McKay*, pages 483–498.
- Danny Brassell and Timothy Rasinski. 2008. *Comprehension that works: Taking students beyond ordinary understanding to deep comprehension*. Teacher Created Materials.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.
- Yapei Chang, Kyle Lo, Tanya Goyal, and Mohit Iyyer. 2024. [Booookscore: A systematic exploration of book-length summarization in the era of LLMs](#). In *The Twelfth International Conference on Learning Representations*.
- Guanliang Chen, Jie Yang, Claudia Hauff, and Geert-Jan Houben. 2018. Learningq: a large-scale dataset for educational question generation. In *Proceedings of the international AAAI conference on web and social media*, volume 12.
- Hyundong Cho, Shuai Liu, Taiwei Shi, Darpan Jain, Basem Rizk, Yuyang Huang, Zixun Lu, Nuan Wen, Jonathan Gratch, Emilio Ferrara, and Jonathan May. 2024. [Can language model moderators improve the health of online discourse?](#) In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7478–7496, Mexico City, Mexico. Association for Computational Linguistics.
- Mohammad Davoudi and Narges Amel Sadeghi. 2015. A systematic review of research on questioning as a high-level cognitive strategy. *English Language Teaching*, 8(10):76–90.
- Art Graesser, Yasuhiro Ozuru, and Jeremiah Sullins. 2010. What is a good question?
- Robert J Harvey and Allen L Hammer. 1999. Item response theory. *The Counseling Psychologist*, 27(3):353–383.
- Joy He-Yueya, Noah D Goodman, and Emma Brunskill. 2024. Evaluating and optimizing educational content with large language model judgments. *arXiv preprint arXiv:2403.02795*.
- Yiqiao Jin, Qinlin Zhao, Yiyang Wang, Hao Chen, Kaijie Zhu, Yijia Xiao, and Jindong Wang. 2024. Agentreview: Exploring peer review dynamics with llm agents. *arXiv preprint arXiv:2406.12708*.
- Sedrick Keh, Justin Chiu, and Daniel Fried. 2024. [Asking more informative questions for grounded retrieval](#). In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 4429–4442, Mexico City, Mexico. Association for Computational Linguistics.
- Wei-Jen Ko, Te-yuan Chen, Yiyan Huang, Greg Durrett, and Junyi Jessy Li. 2020. [Inquisitive question generation for high level text comprehension](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6544–6555, Online. Association for Computational Linguistics.
- Wei-Jen Ko, Cutter Dalton, Mark Simmons, Eliza Fisher, Greg Durrett, and Junyi Jessy Li. 2022. [Discourse comprehension: A question answering framework to represent sentence connections](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11752–11764, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- DR Krathwohl. 2002. A revision bloom’s taxonomy: An overview. *Theory into Practice*.
- Vaibhav Kumar and Alan W Black. 2020. [ClarQ: A large-scale and diverse dataset for clarification question generation](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7296–7301, Online. Association for Computational Linguistics.
- John P. Lalor, Hao Wu, and Hong Yu. 2016. [Building an evaluation scale using item response theory](#). In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 648–657, Austin, Texas. Association for Computational Linguistics.
- Dong-Ho Lee, Adyasha Maharana, Jay Pujara, Xiang Ren, and Francesco Barbieri. 2025. Realtalk: A 21-day real-world dataset for long-term conversation. *arXiv preprint arXiv:2502.13270*.
- Junkai Li, Siyu Wang, Meng Zhang, Weitao Li, Yunghwei Lai, Xinhui Kang, Weizhi Ma, and Yang Liu. 2024. Agent hospital: A simulacrum of hospital with evolvable medical agents. *arXiv preprint arXiv:2405.02957*.
- Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.
- Dennis V Lindley. 1956. On a measure of the information provided by an experiment. *The Annals of Mathematical Statistics*, 27(4):986–1005.

- Michael Lu, Hyundong Justin Cho, Weiyan Shi, Jonathan May, and Alexander Spangher. 2024. [News-interview: a dataset and a playground to evaluate llms’ ground gap via informational interviews](#). Preprint, arXiv:2411.13779.
- Adyasha Maharana, Dong-Ho Lee, Sergey Tulyakov, Mohit Bansal, Francesco Barbieri, and Yuwei Fang. 2024. [Evaluating very long-term conversational memory of LLM agents](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13851–13870, Bangkok, Thailand. Association for Computational Linguistics.
- Bodhisattwa Prasad Majumder, Sudha Rao, Michel Galley, and Julian McAuley. 2021. [Ask what’s missing and what’s useful: Improving clarification question generation using global knowledge](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4300–4312, Online. Association for Computational Linguistics.
- Yan Meng, Liangming Pan, Yixin Cao, and Min-Yen Kan. 2023. [FollowupQG: Towards information-seeking follow-up question generation](#). In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 252–271, Nusa Dua, Bali. Association for Computational Linguistics.
- Shashi Narayan, Joshua Maynez, Reinald Kim Amplayo, Kuzman Ganchev, Annie Louis, Fantine Huot, Anders Sandholm, Dipanjan Das, and Mirella Lapata. 2023. Conditional generation with a question-answering blueprint. *Transactions of the Association for Computational Linguistics*, 11:974–996.
- Benjamin Newman, Luca Soldaini, Raymond Fok, Arman Cohan, and Kyle Lo. 2023. [A question answering framework for decontextualizing user-facing snippets from scientific documents](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3194–3212, Singapore. Association for Computational Linguistics.
- Edgar Onea. 2016. *Potential questions at the semantics-pragmatics interface*, volume 33. Brill.
- Joon Sung Park, Carolyn Q Zou, Aaron Shaw, Benjamin Mako Hill, Carrie Cai, Meredith Ringel Morris, Robb Willer, Percy Liang, and Michael S Bernstein. 2024. Generative agent simulations of 1,000 people. *arXiv preprint arXiv:2411.10109*.
- Sudha Rao and Hal Daumé III. 2018. [Learning to ask good questions: Ranking clarification questions using neural expected value of perfect information](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2737–2746, Melbourne, Australia. Association for Computational Linguistics.
- Sudha Rao and Hal Daumé III. 2019. [Answer-based Adversarial Training for Generating Clarification Questions](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 143–155, Minneapolis, Minnesota. Association for Computational Linguistics.
- Craige Roberts. 2012. Information structure: Towards an integrated formal theory of pragmatics. *Semantics and pragmatics*, 5:6–1.
- Nora Cate Schaeffer and Stanley Presser. 2003. The science of asking questions. *Annual review of sociology*, 29(1):65–88.
- Samuel Schmidgall, Rojin Ziaei, Carl Harris, Eduardo Reis, Jeffrey Jopling, and Michael Moor. 2024. Agentclinic: a multimodal agent benchmark to evaluate ai in simulated clinical environments. *arXiv preprint arXiv:2405.07960*.
- Thomas Scialom and Jacopo Staiano. 2020. [Ask to learn: A study on curiosity-driven question generation](#). In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2224–2235, Barcelona, Spain (Online). International Committee on Computational Linguistics.
- Omar Shaikh, Valentino Emil Chai, Michele Gelfand, Diyi Yang, and Michael S Bernstein. 2024. Rehearsal: Simulating conflict to teach conflict resolution. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pages 1–20.
- Eiichiro Sumita, Fumiaki Sugaya, and Seiichi Yamamoto. 2005. [Measuring non-native speakers’ proficiency of English by using a test with automatically-generated fill-in-the-blank questions](#). In *Proceedings of the Second Workshop on Building Educational Applications Using NLP*, pages 61–68, Ann Arbor, Michigan. Association for Computational Linguistics.
- Toyin Tofade, Jamie Elsner, and Stuart T Haines. 2013. Best practice strategies for effective use of questions as a teaching tool. *American journal of pharmaceutical education*, 77(7):155.
- Jan Trienes, Sebastian Joseph, Jörg Schlötterer, Christin Seifert, Kyle Lo, Wei Xu, Byron Wallace, and Junyi Jessy Li. 2024. [InfoLossQA: Characterizing and recovering information loss in text simplification](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4263–4294, Bangkok, Thailand. Association for Computational Linguistics.
- Ronald D Vale. 2013. The value of asking questions. *Molecular biology of the cell*, 24(6):680–682.
- Jan Van Kuppevelt. 1995. Discourse structure, topicality and questioning. *Journal of linguistics*, 31(1):109–147.

- Robert Van Rooy. 2003. Questioning to resolve decision problems. *Linguistics and Philosophy*, 26:727–763.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.
- Julia White, Gabriel Poesia, Robert Hawkins, Dorsa Sadigh, and Noah Goodman. 2021. [Open-domain clarification question generation without question examples](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 563–570, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Yating Wu, Ritika Mangla, Alexandros G Dimakis, Greg Durrett, and Junyi Jessy Li. 2024. Which questions should i answer? salience prediction of inquisitive questions. *arXiv preprint arXiv:2404.10917*.
- Yating Wu, William Sheffield, Kyle Mahowald, and Junyi Jessy Li. 2023. [Elaborative simplification as implicit questions under discussion](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5525–5537, Singapore. Association for Computational Linguistics.
- Lili Yu, Howard Chen, Sida I. Wang, Tao Lei, and Yoav Artzi. 2020. [Interactive classification by asking informative questions](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2664–2680, Online. Association for Computational Linguistics.
- Eric Zelikman, Wanjing Ma, Jasmine Tran, Diyi Yang, Jason Yeatman, and Nick Haber. 2023. [Generating and evaluating tests for k-12 students with language model simulations: A case study on sentence reading efficiency](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2190–2205, Singapore. Association for Computational Linguistics.
- Zheyuan Zhang, Daniel Zhang-Li, Jifan Yu, Linlu Gong, Jinchang Zhou, Zhiyuan Liu, Lei Hou, and Juanzi Li. 2024. Simulating classroom education with llm-empowered agents. *arXiv preprint arXiv:2406.19226*.
- Qinlin Zhao, Jindong Wang, Yixuan Zhang, Yiqiao Jin, Kaijie Zhu, Hao Chen, and Xing Xie. 2024. Competeai: Understanding the competition dynamics of large language model-based agents. In *Forty-first International Conference on Machine Learning*.
- Pei Zhou, Jay Pujara, Xiang Ren, Xinyun Chen, Heng-Tze Cheng, Quoc V Le, Ed H Chi, Denny Zhou, Swaroop Mishra, and Huaixiu Steven Zheng. 2024. Self-discover: Large language models self-compose reasoning structures. *arXiv preprint arXiv:2402.03620*.
- Wei Zhou, Renfen Hu, Feipeng Sun, and Ronghuai Huang. 2019. [An intelligent testing strategy for vocabulary assessment of Chinese second language](#)

[learners](#). In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 21–29, Florence, Italy. Association for Computational Linguistics.

A QUEST Details

QUEST consists of four key modules. (1) The **Question Generator** (M_q) generates a set of questions $Q = \{q_1, q_2, \dots, q_n\}$ from a document D (Section 2.3); (2) The **Answer Generator** (M_a) then produces corresponding answers, forming question-answer pairs $QA = \{(q_1, a_1), (q_2, a_2), \dots, (q_n, a_n)\}$ using parametric knowledge (Section 2.3). (3) The **Reader Simulator** consists of a **Learner** (M_l) and an **Evaluator** (M_e). The learner model M_l simulates a learner’s understanding by attempting the final exam using only the generated QA pairs, formulated as $P \sim M_l(E; QA)$. The evaluator model M_e then assesses the learner’s responses P by comparing them against ground-truth answers when available or using parametric knowledge to assign a score. In this section, we present the prompts that we use for each of the components in QUEST and validate their reliability for the role that they simulate.

A.1 Question Generator

Question Generator

Article: $S_{[1:k-1]}$
 Student is currently reading the section: S_k .
 Generate a question that helps the student understand the section better.

Output in the following JSON format:
 ```json  
 {"question": question}  
 ```

A.2 Answer Generator

Answer Generator

questions: $\{\{questions\}\}$

Answer each question shortly and output in following JSON format:
 ```json  
 {"qa\_pairs": [  
 {"question": question\_1, "answer":  
 answer\_1},  
 {"question": question\_2, "answer":  
 answer\_2},  
 ...  
 {"question": question\_n, "answer":  
 answer\_n}  
 ]}  
 ```

We manually evaluate whether the answers generated by our answer generator are mostly correct so that they can be used as reliable information for taking the exam E . We inspect 20 random samples of generated questions for each subject for 100 total samples, and find that the answer generator provides factually correct answers to all of them, albeit with some cases not being completely exhaustive. For example, for the question “*What are some examples of chemical properties, and how do they differ from physical properties in terms of changes in matter?*”, the answer is not an exhaustive list of all chemical properties, but the examples provided are all correct. The general correctness is not surprising given that the textbooks are for entry-level college courses and most subjects do not involve subjectivity.

A.3 Learner

Learner

You are now a learner participating in a structured learning simulation. Your task is to:

1. ****Study the Provided Learning Materials:**** Carefully read and understand the content enclosed in the [LEARNING MATERIALS] tags.
2. ****Answer the Exam Questions Using Only the Learning Materials:**** When you respond to the questions in the [EXAM], you must:
 - Base all answers solely on the information contained in the [LEARNING MATERIALS].
 - Clearly show how your reasoning follows from the [LEARNING MATERIALS].
 - If the question asks about something not covered in the [LEARNING MATERIALS], do not provide an answer or guess. Instead, respond exactly with:
3. ****No External Knowledge or Guessing:**** Provide no additional reasoning or information if the content is not in the [LEARNING MATERIALS].

Let's begin.

```
[LEARNING MATERIALS]
{{learning_material}}
[/LEARNING MATERIALS]
```

Now, proceed to the exam below and answer as instructed:

```
[EXAM]
{{exam}}
[/EXAM]
```

Respond with answers in the following JSON format (key: exam question number; value: your answer):

```
{
  "1": "< your answer to exam question 1 >",
  "2": "< your answer to exam question 2 >",
  ...
}
```

```
# TARGET TEXTBOOK CONTENT:
{{target_content }}
```

This is the main component of QUEST that is responsible for measuring the utility of a question and that we validate with human evaluators, results shown in Section 5.5 and details of the study in Appendix A.5.

A.4 Evaluator

Evaluator

You are a teacher who is evaluating a student's understanding of a document.

Here is the document:

```
{{document}}
```

Now, determine the correctness of the student's answers to the following question.

question:

```
{{question}}
```

ground truth:

```
{{answer}}
```

student's answer:

```
{{prediction}}
```

Please provide a score between 0 and 1, where:

- 0 indicates the student's answer is completely incorrect.
- 1 indicates the student's answer is completely correct.

If ground truth is not provided (e.g., None), determine the correctness of the student's answer based on your own understanding of the document.

Answer in the following JSON format:

```
{"score": <score>, "feedback": "<feedback>"}
```

To validate whether the evaluator provides a reasonable assessment for the correctness of the an-

swer, we manually examine 50 scores from each subject assigned to an answer when compared to the ground truth or based on the textbook content when one is not provided. Multiple choice or short answer questions are easy to assess, but answers to long free-form questions are more subjective and some questions involve multiple sub-questions. Despite the variety of question types, we find that our prompt works well with gpt-4o-mini in assigning appropriate credit to each answer. Out of 250 scores we examine, there are only two instances for the sociology subject that incorrectly assigned full credit for a wrong answer: subculture was given full credit when the correct answer is counterculture for the question *“The Ku Klux Klan is an example of what part of culture? 1. Counterculture 2. Subculture 3. Multiculturalism 4. Afrocentricity”*. Answers that contain multiple items are also appropriately weighted in most cases. For example, for the question *“What evidence would you use to support this statement: Ancient people thought that disease was transmitted by things they could not see.”*, the answer *“Ancient texts from civilizations such as the Egyptians, Greeks, and Romans indicate an awareness of contagious diseases. For example, Hippocrates noted the spread of illnesses in crowded areas, and the Romans implemented public health measures during plagues.”* is given a score of 0.8 because the statement about Romans implementing quarantine is an overstatement; they took public health measures but not to an extent comparable to modern quarantine practices. These results motivate using LMs as automated answer generators and scorers for QUEST.

A.5 Human Evaluation Details

Setup. The instructions for our human evaluation study are shown in Figure 6 and the interface is shown in Figure 5. Our human evaluation study is available at <https://quest-human-eval.streamlit.app/>. We recruit eight volunteer participants from the University of Southern California, all located in North America, to each complete 19 samples that span the five subjects. All participants provided informed consent before taking part. Participants span a wide variety of demographics, but are all fluent in English and have a computer science background. Participants are instructed to mark questions that they have prior knowledge of the answer to so that they are excluded from the evaluation. Each sample consists of a section of a chapter, prior sections that can be opened to

read if desired, and a set of questions and their corresponding answers, one of each with exclusively high utility, high salience, and high EIG. Exclusivity is enforced by making sure that a high utility question is not also a high salience or high EIG question such that its salience and EIG scores are not within the top 20% of the distribution. In practice, we find that questions with exclusively high utility, high salience, and high EIG are rare, mostly because salience is often high for most questions. Therefore, it requires nontrivial amount of sampling to get a set of questions that fulfill the exclusivity requirement, hence the relatively small number of evaluation samples. However, meaningfully replicating the learner simulation with human participants is a relatively cognitively loaded task that is not practical for a large number of samples.

With this setup, we collect a total of 107 samples, which we find to be sufficient for our analysis in getting statistically significant results for usefulness and ranking.

The rationale for treating this study as construct validation of utility, rather than a head-to-head comparison of generators, is discussed in §5.5.

A.6 Cost Analysis

QUEST requires generating questions that a simulated learner asks while reading through a textbook chapter and then approximating the learner’s performance on related exam problems for various subsets of generated questions, which requires generating answers that models the learner’s understanding and verifying their correctness. Therefore, the total number of input and output tokens for QUEST can get costly, but we argue that it is not impractical when using a smaller model with lower cost and worthwhile given that QUEST enables measuring a metric that is otherwise infeasible to measure. We provide the exact costs for using gpt-4o-mini for representative subjects in Table 8 (Microbiology as the most expensive and U.S. History as the least expensive). Although the number of API calls is high and the number of total tokens is substantial, the cost is still relatively affordable. In addition, many of the calls can be processed as a pipeline of parallel batch jobs (e.g., generating questions and answers for all sections, running the learner simulator for all subsets of questions, etc.), which can reduce the cost significantly. The bulk of the total costs with experimentation is actually incurred by fine-tuning the model with a custom dataset, which

Learning Material

> Prior sections (click to expand)

Current section

Properties of matter fall into one of two categories. If the property depends on the amount of matter present, it is an extensive property. The mass and volume of a substance are examples of extensive properties; for instance, a gallon of milk has a larger mass and volume than a cup of milk. The value of an extensive property is directly proportional to the amount of matter in question. If the property of a sample of matter does not depend on the amount of matter present, it is an intensive property. Temperature is an example of an intensive property. If the gallon and cup of milk are each at 20 °C (room temperature), when they are combined, the temperature remains at 20 °C. As another example, consider the distinct but related properties of heat and temperature. A drop of hot cooking oil splattered on your arm causes brief, minor discomfort, whereas a pot of hot oil yields severe burns. Both the drop and the pot of oil are at the same temperature (an intensive property), but the pot clearly contains much more heat (extensive property).

Questions

Q1.

Why is the amount of heat in a substance considered an extensive property, while temperature is an intensive property?

Answer

The amount of heat is extensive because it increases with the quantity of substance, whereas temperature is intensive as it does not change with the amount of substance present.

Usefulness (1-5); 5 = most useful

3

Interestingness (1-5; 5 = most interesting)

3

Preference rank (1 = most preferred)

1

Brief explanation for your rank [only need to fill in for one of the boxes]

helps with p2

Exam Problems

Exam P1. Explain the difference between extensive properties and intensive properties.

Your answer to Exam P1

Extensive changed based on amount of matter, intensive does not.

Exam P2. Identify the following properties as either extensive or intensive.

- (a) volume
- (b) temperature
- (c) humidity
- (d) heat
- (e) boiling point

Your answer to Exam P2

a - extensive
b - intensive
c - intensive
d - extensive

Figure 5: Human evaluation interface.

is a one-time expense but costs on average \$5 per fine-tuning job.

B TEXTBOOK-EXAM Data Processing Details

B.1 Data License

The openstax textbooks are licensed under the Creative Commons Attribution License. The textbooks used in this work are all in English.

B.2 Parsing Sections

Our prompt for parsing sections as part of creating TEXTBOOK-EXAM is shown below. In order to ensure consistency across subjects in how sections are divided, we provide manually annotated few-

shot examples. In addition, we use GPT-4o for this part as it is a one-time expense and data quality has important implications for all the experiments that we conduct in this work.

Instructions (click to expand)

Please follow these steps for each task:

1. Read the learning material on the left (expand Prior sections if needed). The current section is always visible.
2. Review the three Questions in the middle (they relate to the current section).
3. Answer the Exam Problems on the right with help from the questions in the middle. You will not be graded on correctness, but please put in a reasonable effort to answer them. We are most interested in how you value the Questions.
4. After all exam answers are filled, rate each Question on its usefulness (1-5) and interestingness (1-5) and choose a unique preference rank (1-3).
 1. Usefulness: how useful was the question for getting a deeper understanding of the content and answering the exam problems?
 - [5 = directly useful, 4 = somewhat useful, indirectly useful, 1 = not useful at all]
 2. Interestingness: how interesting was the question, regardless of the exam problems?
 - [5 = most interesting, 3 = somewhat interesting, 1 = not interesting]
 3. Ranking: provide a unique ranking (can't be tied) for each question based on its usefulness. Provide a brief explanation for your ranking.

Only need to fill in for one of the boxes

 - [1 = most preferred, 2 = somewhat preferred, 3 = least preferred]
5. Submit to move to the next task. A progress bar appears at the top.

You will complete one sample at a time. Progress bar will be shown at the top.

Figure 6: Human evaluation instructions.

Subject	Input Tokens	Output Tokens	Total Tokens	API Calls	Cost
US History	1.91M	0.40M	2.31M	5,479	\$0.53
Microbiology	3.59M	0.74M	4.33M	10,456	\$0.98

Table 8: Cost analysis for QUEST using gpt-4o-mini for the Microbiology (most expensive) and US History subjects (least expensive).

Extracting Sections

Instructions for extracting sections from the given textbook content:

1. Transform markdown for equations into LaTeX and remove all other markdown formatting to only keep the raw content.
2. Split the content into sections of uniform length and number each section.
3. Skip the learning objectives, key concepts, and summary content.
4. Ensure that all content, except skipped parts, is covered verbatim in at least one of the resulting sections.

```
# EXAMPLE
## INPUT:
{{ example_input }}

## OUTPUT:
{{ example_output }}
```

Produce only valid JSON with the following format:

```
{"section": {"1": {"content": "Verbatim section 1 content from chapter"}, ...}}
```

```
# TARGET TEXTBOOK CONTENT:
{{ target_content }}
```

B.3 Parsing Exam Questions

We parse the end-of-chapter review questions in OpenStax textbooks based on the markdown headers and formatting with BeautifulSoup.⁴ There are occasionally ill-formatted questions that we throw out, since the main goal here is to find a subset of chapters that are suitable for testing QUEST. We avoid chapters with too few (< 10) and too many questions (> 25) so that our simulations are conducted with enough questions for measuring question utility while also completing within a reasonable timeframe.

B.4 Bloom's Taxonomy Level Distribution

We share our prompt for categorizing questions based on the revised Bloom's taxonomy below:

⁴<https://www.crummy.com/software/BeautifulSoup/>

Bloom Classification

Classify the questions into one of the six main categories of Bloom’s Taxonomy based on the cognitive processes required for answering it correctly.

Bloom’s Taxonomy Categories:

1. **Remembering:** Producing or retrieving definitions, facts, or lists, or reciting previously learned information.
2. **Understanding:** Grasping the meaning of information by interpreting and translating what has been learned.
3. **Applying:** Using learned information in new and concrete situations.
4. **Analyzing:** Breaking down or distinguishing the parts of learned information.
5. **Evaluating:** Making judgments about information, validity of ideas, or quality of work based on a set of criteria.
6. **Creating:** Using information to generate new ideas or products.

Question 1: {{question 1}}

Question 2: {{question 2}}

...

Provide only the Bloom category and format your response in JSON with the following structure:

```
{
  "bloom_categories": [
    {
      "question": question,
      "bloom_category": bloom_category
    }
  ]
}
```

With our prompt, each question is mapped to one of six categories: *Remembering*, *Understanding*, *Applying*, *Analyzing*, *Evaluating*, or *Creating*, and aligned to relevant document sections. The distribution, shown in Figure 7, indicates that TEXTBOOK-EXAM includes a broad range of question types, supporting outcome-based evaluation across different cognitive levels. For instance, questions in Microbiology and Sociology primarily focus on *Remembering* and *Understanding*, whereas Chemistry and Economics exhibit a more varied distribution.

C Baseline Details

In this section, we provide additional details for the baselines that we compare against QUEST-trained models.

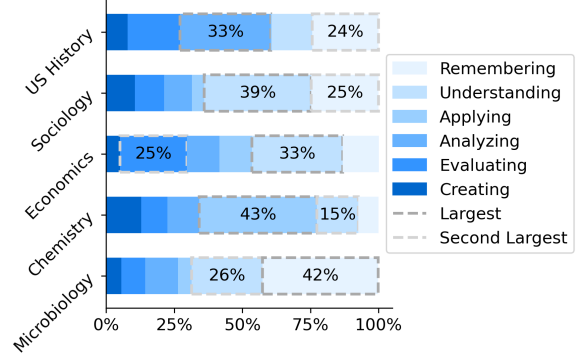


Figure 7: **Distribution of Bloom’s taxonomy labels in TEXTBOOK-EXAM.** TEXTBOOK-EXAM consists of questions that require a wide variety of cognitive levels and subjects vary in the cognitive skills emphasized in their exams.

C.1 Bloom-based Prompting

Generating Next Paragraph

Use the cognitive process of
 {{selected_bloom_level}}:
 {{bloom_definition}}
 to generate the next paragraph for the
 following
 text that will help the student understand
 the content better:

{{full_context}}

Output in the following JSON format:
 {"next_paragraph": next_paragraph}

Generating Question

Given the input context and the next
 paragraph,
 what is the key question that connects the
 two?

Input context: {{context}}
 Next paragraph: {{next_paragraph}}

Output in the following JSON format:
 {"question": question}

D Alternative Question Quality Metrics

D.1 Salience

Salience measures a question’s relevance and importance within the document D (Wu et al., 2024). It is rated on a Likert scale from 1 to 5, where a score of 1 indicates that the question in section k is unrelated to $S_{[1:k]}$ and contributes minimally to understanding, while a score of 5 indicates strong relevance to $S_{[1:k]}$ and essential comprehension support for S_k by clarifying key concepts or introducing new information. We use the QSalience

model released by Wu et al. (2024)⁵ to estimate salience. The model we use is a Mistral model fine-tuned with Peft⁶, which has a context window limit of 4096 tokens. Therefore, for inputs longer than the limit, we use gpt-4o-mini with the following prompt as a fallback:

Salience Prediction

Article: {article}
Question: {question}

System Instructions Imagine you are a curious reader going through the article. You come across a question and need to determine whether it should be answered within the article or not. Your task is to assign a score based on the relevance and necessity of answering the question.

Scoring Criteria

- Score = 1:** The question is **completely unrelated** to the article.
- Score = 2:** The question is **related but already answered** in the article.
- Score = 3:** The question is **related but answering it is not essential**, as it expands on a **minor or non-central idea**.
- Score = 4:** The question is **related and answering it enhances the reader's understanding** of the article.
- Score = 5:** The question is **related and must be answered**, as it expands on **central ideas of the article**.

Scoring Guidelines

- The score is based on the **information utility** of the answer.
- If a question is related but **not central or necessary**, **do NOT** assign it a high score.
- Assign **Score 3** if the question is unanswered but **not critical**, and **Score 2** if it has already been answered.
- Distinguishing Scores 4 and 5:**
 - If the article would feel **incomplete** without the answer, assign **Score 5**.
 - Otherwise, assign **Score 4**.
- A **Score of 4** is useful, but **other questions may be more important**.
- A **Score of 5** is reserved for **must-answer, central questions**.
- Avoid bias toward high scores** and carefully follow the instructions.

The score should strictly be an **integer between 1 and 5**.
Score:

D.2 Expected Information Gain (EIG)

Expected Information Gain (EIG) quantifies the reduction in uncertainty about a student's knowledge state after answering a question. We estimate EIG using an LLM by computing the entropy of the model's token probability distribution before and

after conditioning on the first token of the correct answer. EIG is computed as:

$$EIG(Q) = H(X) - H(X|A_1)$$

where $H(X)$ is the prior entropy, representing the model's uncertainty before seeing the answer, and $H(X|A_1)$ is the posterior entropy after conditioning on the first token A_1 . The entropy is given by:

$$H(X) = - \sum_{x \in V} P(x) \log P(x),$$

$$H(X|A_1) = - \sum_{x \in V} P(x|A_1) \log P(x|A_1)$$

where V is the vocabulary, and $P(x)$, $P(x|A_1)$ are token probabilities before and after observing A_1 , respectively. We estimate probability distributions by querying an LLM with the following prompts. Instead of conditioning on the full answer, we use only the first token, as most uncertainty reduction occurs at the start of an answer. This aligns with the autoregressive nature of LLMs and ensures EIG captures incremental belief updates rather than full-answer memorization.

Prior Entropy Estimation

Imagine you are a reader encountering a question in the article.
Article: {article}
Question: {question}
Answer:

Posterior Entropy Estimation

Imagine you are a reader encountering a question in the article.
Article: {article}
Question: {question}
Answer: { A_1 }

E Additional Results

E.1 Qualitative Examples

Table 9 extends the Chemistry excerpt in Table 4 with examples of questions that have high salience and high utility, high utility and low salience, and high salience and low utility across all five subjects. As we have shown with our results that compute a lower correlation between the two, there is no noticeable pattern based on a manual inspection. We encourage future work to refine question utility or develop new question quality metrics that provide more interpretable insight into the value of a question as it pertains to a specific downstream task.

⁵Available at github.com/ritikamangla/QSalience.

⁶huggingface.co/lingchensanwen/mistral-ins-generation-best-balanced.

Subject	Salience	Utility	Question
Chemistry	High	High	How does the combination of elements result in properties that are different from those of the individual elements in their uncombined states, and can you provide more examples of such transformations?
	Low	High	What are the ‘Twelve Principles of Green Chemistry’?
	High	Low	What happens to the charge of an atom when it gains or loses electrons?
Economics	High	High	Why does inelastic demand allow the government to pass the tax burden onto consumers without significantly affecting the quantity demanded?
	Low	High	What are the three different combinations of labor and physical capital mentioned for cleaning up a park?
	High	Low	How does José evaluate the trade-off between giving up the third T-shirt and purchasing two additional movies in terms of marginal utility, and what does this reveal about his preferences for consumption?
Sociology	High	High	How does the concept of carrying capacity relate to the tragedy of the commons as described by Garrett Hardin and William Forster Lloyd?
	Low	High	What are some reasons people often forget where their old electronics are disposed of or stored?
	High	Low	How do the challenges faced by students from low socioeconomic backgrounds, as described in the example, impact their ability to succeed in the educational system compared to their peers from higher social classes?
U.S. History	High	High	What were the main differences in the motivations and goals of the English migrants who settled in the Chesapeake Bay colonies compared to those who settled in New England?
	Low	High	What challenges did Venetian sailors face when transporting goods along the old Silk Road?
	High	Low	What were some of the challenges and dangers faced by serfs and their families in their daily lives during the Middle Ages?
Microbiology	High	High	What roles do NAD ⁺ , NADP ⁺ , and FAD play in cellular metabolism, and how do their oxidized and reduced forms differ in function?
	Low	High	What evidence did Robert Remak provide to support the idea that cells originate from other cells?
	High	Low	What was the significance of Dmitri Ivanovski’s use of the Chamberland filter in discovering the cause of tobacco mosaic disease?

Table 9: **Qualitative examples** of generated questions by salience and utility. Each subject includes one high-salience/high-utility, one low-salience/high-utility, and one high-salience/low-utility question.