Multi-Dimensional Framework for Reliable and Instructionally Aligned Question–Answer Generation with Large Language Models

Anuradha Surabhi[1,2,*], Sheshikala Martha[1]

[1]School of Computer Science and Artificial Intelligence, SR University, Warangal, India

[2]Department of CSE (AIML), Neil Gogte Institute of Technology, Hyderabad, India

Received 30 March 2026; received in revised form 06 May 2026; accepted 12 May 2026

https://doi.org/10.46604/ijeti.2026.16326

Abstract

This study presents a multi-dimensional framework for reliable, instructionally aligned question–answer (QA) generation from academic portable document formats (PDFs) using an evidence-grounded claude configuration. The pipeline integrates semantic document chunking, facebook ai similarity search (FAISS)-based dense retrieval, guided by a standard QA prompt and an instructionally aligned prompt guided by Bloom's cognitive levels to generate factual, conceptual, procedural, analytical, and open-ended questions. A rubric-based evaluation framework assesses factual grounding, reasoning quality, and instructional relevance. Results show that the evidence-grounded, instructionally aligned setting improves factual grounding and increases overall rubric scores. It also reduces hallucinations, and enhances higher-order cognitive coverage and question-format diversity. The framework is presented as a lightweight proof of concept using claude-3-Haiku, while broader cross-model validation and alternative retrieval strategies remain important directions for future work. These findings demonstrate the educational utility and interpretability of evidence-grounded QA generation.

Keywords: QA generation, instructionally aligned prompting, semantic chunking, FAISS retrieval, hallucination reduction

1. Introduction

Large language models (LLMs) such as claude, generative pre-trained transformer (GPT), and related transformer-based systems have significantly advanced natural language processing (NLP). These models enable high-quality text generation, summarisation, and automated question–answer (QA) generation [1-6]. Their ability to synthesize information from complex textual contexts has created promising opportunities in education, research, and intelligent tutoring systems [7-9]. However, despite these advances, base LLMs often exhibit persistent limitations, including the hallucination of unsupported facts, weak grounding in domain-specific evidence, and limited pedagogical diversity in generated questions [10-13]. These shortcomings reduce their reliability and limit their suitability for structured educational applications.

Recent studies have explored LLM-based QA generation, evidence-grounded generation, rubric-based evaluation, and pedagogically informed educational artificial intelligence (AI) systems [14-18]. In particular, evidence-grounded generation has emerged as an effective strategy for improving factuality by combining semantic retrieval with LLM-based generation, thereby reducing hallucinations and strengthening contextual fidelity [14-16]. In addition, hybrid AI architectures in other domains have shown that combining complementary learning components can improve robustness and reduce error propagation, reinforcing the broader value of modular designs for reliable AI systems [19].

* Corresponding author. E-mail address: surabhiaiml2023@gmail.com

At the same time, rubric-driven assessment and educationally guided prompting have gained attention as promising directions for improving the quality and instructional value of generated outputs [17-18]. Nevertheless, three important gaps remain. First, existing studies often lack a systematic multi-dimensional evaluation framework that jointly measures factual quality, reasoning, robustness, and pedagogical relevance in generated QA pairs. Second, automated LLM-based evaluation is often used without sufficient external alignment to human educational judgment. This limitation makes it difficult to determine how closely rubric-based scores reflect established teaching standards. Third, limited attention has been paid to the automated classification of generated questions into educationally meaningful categories, such as Bloom's taxonomy levels and instructional formats (e.g., factual, conceptual, procedural, analytical, and open-ended). These limitations hinder the broader adoption of LLMs in structured educational and domain-specific learning environments.

To address these gaps, this study presents a multidimensional framework for reliable, instructionally aligned QA generation from academic PDFs using an evidence-grounded claude configuration. The proposed framework integrates semantic document chunking, facebook ai similarity search (FAISS)-based dense retrieval, dual prompting (baseline and instructionally aligned), rubric-based evaluation, and heuristic question classification across Bloom's cognitive levels and question formats. The objectives of this study are threefold:

(1) To develop an evidence-grounded QA generation pipeline that reduces hallucinations through context-based synthesis.

(2) To compare baseline and instructionally aligned prompting in terms of cognitive diversity, pedagogical depth, and reasoning quality.

(3) To design scoring and classification modules that systematically evaluate and categorize QA pairs, producing a curated benchmark dataset of 600 QA pairs for educational artificial intelligence (AI) research.

The present study is intentionally positioned as a lightweight proof-of-concept using claude-3-haiku to examine whether the proposed framework can improve factual grounding and pedagogical quality under a cost-efficient deployment setting. Experiments on academic PDFs demonstrate that the proposed framework improves factual grounding, pedagogical diversity, and overall evaluability compared with baseline configurations.

The remainder of this manuscript is organized as follows. Section 2 reviews the related work. Section 3 presents the proposed methodology. Section 4 describes the experimental setup. Section 5 reports and discusses the results. Finally, Section 6 concludes the study and outlines future research directions.

2. Related Work

The task of generating and evaluating question–answer (QA) pairs from academic and scientific texts has long been a central challenge in NLP. With the rise of LLMs, new opportunities have emerged for creating semantically rich, pedagogically diverse, and domain-specific QA pairs. However, several strands of research highlight ongoing limitations in factual accuracy, contextual grounding, and instructional utility. This section reviews prior work in four key areas relevant to this study: automated QA generation, retrieval-augmented generation (RAG), evaluation of QA systems, and question classification in educational AI.

2.1 Automated QA generation with LLMs

Transformer-based LLMs such as GPT-3, GPT-4, T5, and claude have significantly advanced automated QA generation [1-2]. These models excel at fluency, coherence, and adaptability across domains. However, open-domain QA systems often face challenges with hallucinations (the production of unsupported facts), weak domain alignment, and shallow question styles. Fine-tuning LLMs on domain-specific corpora has improved domain relevance, but issues of factual grounding remain.

Furthermore, most studies on QA generation prioritize semantic fluency over pedagogical diversity, leaving open questions about how to systematically generate factual, conceptual, procedural, analytical, and open-ended queries that align with educational taxonomies [17-18].

2.2 Retrieval-augmented generation (RAG)

RAG models combine LLMs with external retrieval systems to ground responses in supporting evidence [14]. By retrieving relevant chunks from large corpora or curated datasets, these models reduce hallucination and increase factual accuracy. Variants of RAG are widely applied in open-domain QA [20], knowledge-intensive reasoning [21], and biomedical QA [22]. Despite their promise, existing RAG systems often lack fine-grained control over pedagogical diversity and rarely incorporate systematic evaluation frameworks. Moreover, retrieval performance depends heavily on chunking strategy, vectorization method, and similarity thresholds. This study extends this line of research by integrating document chunking with FAISS-based dense retrieval and a multi-prompt QA generation pipeline, which ensures both factual grounding and instructional alignment. In the present study, the retrieval analysis focuses on an embedding-level comparison within a dense retrieval setting, while broader comparisons across sparse or hybrid RAG strategies remain important directions for future work.

2.3 Evaluation of QA systems

Evaluation of QA systems traditionally relied on human judgment, which is costly, subjective, and difficult to scale. Automated metrics such as bilingual evaluation understudy (BLEU), recall-oriented understudy for gisting evaluation (ROUGE), and metric for evaluation of translation with explicit ordering (METEOR) capture surface-level overlap but fail to assess semantic fidelity and higher-order qualities such as clarity, reasoning, and domain maturity [23-24]. To address these gaps, semantic similarity approaches using sentence-bidirectional encoder representations from transformers (BERT) embeddings [25] and cosine similarity scoring are proposed for grading answer relevance. Recent work also explores rubric-based evaluation, in which answers are scored across multiple dimensions of quality (accuracy, completeness, clarity, reasoning, faithfulness). Therefore, prior efforts often focus on a narrow set of criteria and lack interpretability across pedagogical dimensions.

The proposed framework extends this by implementing a multi-faceted rubric evaluation that features both core and extended criteria, which it couples with correlation analysis to examine relationships among evaluation metrics. At the same time, automated LLM-based evaluation requires cautious interpretation, as model-family self-critique may introduce consistency bias when the same LLM family is used for both generation and assessment. Accordingly, rubric-based LLM evaluation is best treated as a scalable comparative proxy rather than a fully objective substitute for expert human judgment.

2.4 Question classification and educational AI

The application of QA systems in education centers on automating assessments, generating quizzes, and providing formative feedback [26]. Bloom's taxonomy serves as a pedagogical scaffold for evaluating question complexity, ranging from knowledge recall to higher-order cognitive skills such as analysis and evaluation [27]. However, most LLM-based QA systems do not systematically classify questions into Bloom's levels or pedagogical formats. Some studies have applied keyword-based heuristics or supervised classifiers for taxonomy mapping, but these remain limited in scope and integration.

The proposed framework introduces a heuristic classification module that categorizes questions into both Bloom's cognitive levels and instructional formats—factual, conceptual, procedural, analytical, and open-ended—thereby enhancing interpretability and pedagogical alignment. This heuristic approach is intentionally lightweight and interpretable, focusing primarily on structural alignment rather than deeper semantic cognitive intent or cross-domain generalization. More advanced supervised or LLM-based pedagogical classification remains an important direction for future work.

2.5 Research gap and contribution

Taken together, prior research demonstrates significant advances in QA generation, retrieval-based grounding, and evaluation methods. Yet, few systems combine evidence-grounded generation with multi-dimensional rubric-based evaluation and pedagogical classification. This creates a gap in developing holistic, education-aware QA pipelines that are both semantically reliable and pedagogically meaningful. This study addresses this gap by proposing a multi-dimensional framework for reliable and instructionally aligned QA generation with LLMs, implemented using claude, which integrating (i) semantic retrieval, (ii) dual-prompt QA generation (baseline vs. instructionally aligned prompting), (iii) rubric-based evaluation across 15 distinct criteria, and (iv) heuristic classification across Bloom's levels and instructional formats.

This unified pipeline not only improves the quality of QA generation but also contributes a curated, richly annotated dataset for benchmarking educational AI. The proposed framework is designed to provide a practical and interpretable foundation for evidence-grounded, instructionally aligned QA generation. At the same time, broader cross-model benchmarking and more advanced pedagogical classification remain important directions for future work.

3. Methodology

The proposed multi-dimensional framework, implemented using claude, is designed to generate, evaluate, and classify high-quality QA pairs from academic documents. The methodology is modular, comprising five key components: (i) document preprocessing and chunking, (ii) semantic retrieval with FAISS indexing, (iii) multi-prompt QA generation with claude, (iv) rubric-based multi-dimensional evaluation, and (v) heuristic classification of generated questions. Fig. 1 illustrates the overall pipeline.

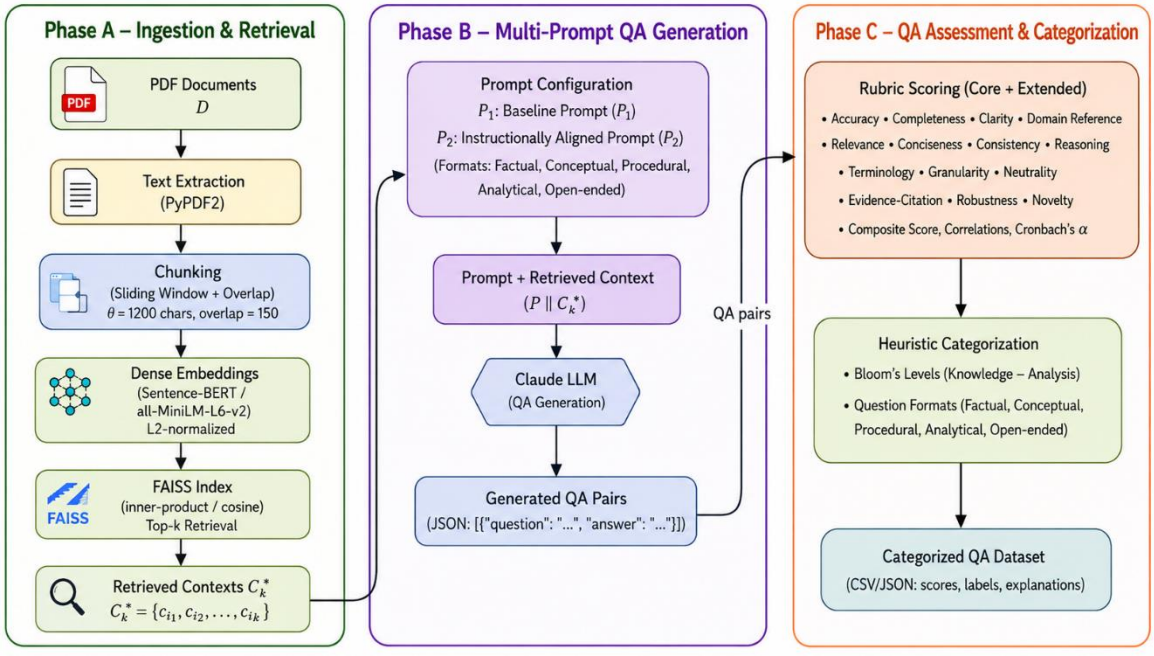


Fig. 1 Complete pipeline of the proposed framework for producing pedagogical QA dataset

3.1 Document preprocessing and chunking

The PDF document *D* is first processed using PyPDF2 to extract raw textual content. Since LLMs such as claude have input-length limitations, the extracted text is segmented into overlapping chunks using a sliding-window strategy. An illustrative abstraction of this sentence-level sliding-window chunking mechanism is presented in Fig. 2 for conceptual clarity. In practice, the system executes character-length-based chunking with overlapping contextual windows.

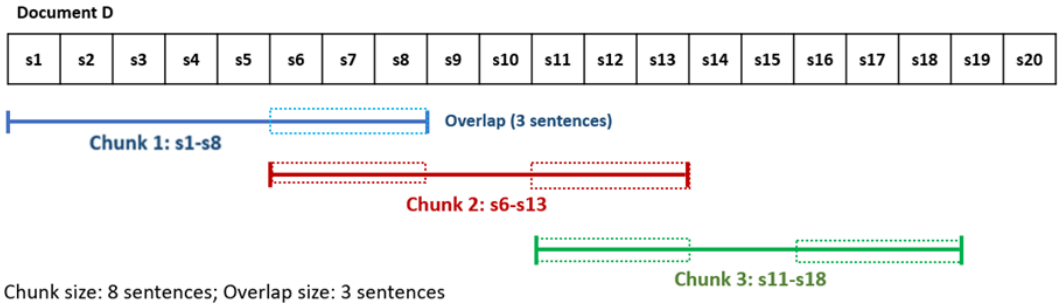


Fig. 2 Conceptual illustration of sliding-window chunking with overlap

Let the extracted text from document *D* be represented as a sequence of sentences $D = \{s_1, s_2, \cdots, s_n\}$, where s_i represents a sentence. Chunks are defined as:

$$c_j = \{s_p, s_{p+1}, \cdots, s_q\}, with\ |c_j| \leq \theta \tag{1}$$

where the character length of each chunk C_j, is bounded by the maximum character length θ (set to 1200), and consecutive chunks overlap by 150 characters to preserve context.

3.2 Semantic embeddings and FAISS retrieval

To ensure that claude's QA generation remains grounded in source material, the model, adopts a semantic retrieval pipeline built on dense embeddings and FAISS. The embedding and retrieval process is formally outlined in Algorithm 1, which specifies the construction of embeddings for each text chunk, the indexing procedure in FAISS, and the subsequent top-k similarity search. This design provides both reproducibility and a transparent, step-by-step mapping from pre-processed document chunks to the most relevant retrieval candidates for downstream QA generation.

Let the pre-processed document be split into a set of chunks:

$$C = \{c_1, c_2, \cdots, c_k\} \tag{2}$$

where each chunk c_i (i =1~k) is embedded into a high-dimensional vector space using a pre-trained sentence embedding model, MiniLM, as:

$$v_i = f(c_i)\ \forall\ c_i \in C \tag{3}$$

where f (.) Is the embedding function. These embeddings are L_2-normalised as:

$$\hat{v}_i = \frac{v_i}{\|v_i\|} \tag{4}$$

All chunk embeddings are indexed using FAISS, which supports efficient inner-product or cosine-similarity search over large corpora. Given a query q constructed from the document title, abstract, or user-defined keywords, its embedding $v_q = f(q)$ is computed and compared with chunk embeddings as:

$$sim(q, c_i) = \hat{v}_q \cdot \hat{v}_i \tag{5}$$

The top-*k* chunks with the highest similarity are then retrieved as:

$$C_k^* = \{c_{i_1}, c_{i_2}, \cdots, c_{i_k}\} \tag{6}$$

The retrieved top-*k* segments act as evidence contexts for the downstream QA generation process. These contextually relevant chunks guide claude to generate responses that remain faithful to the original document content. By grounding the generation process in retrieved evidence, the framework also helps minimize hallucination. The end-to-end pipeline is shown in Fig. 3.

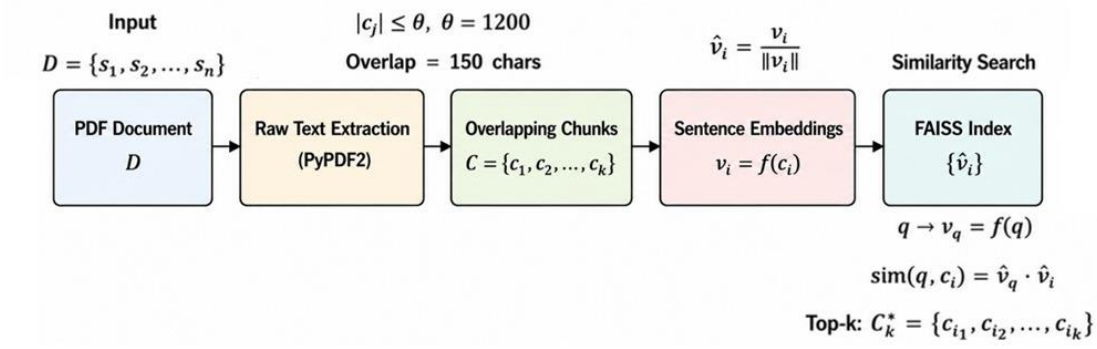


Fig. 3 Schematic architecture of the end-to-end document retrieval pipeline

Algorithm 1 Semantic embedding and FAISS retrieval

Input: Document D, Embedding Model $f(\cdot)$, Query q, Top-k value
Output: Retrieved chunk set C_k^*
1: Initialize Embedding_set ← ∅
2: For each chunk c_i in C: do
3: $v_i \leftarrow f(c_i)$ ▷ Compute chunk embedding
4: $\hat{v}_i = \frac{v_i}{\|v_i\|}$ ▷ Normalize embedding vector
5: Store $\hat{v}_i$ in Embedding_set
6: Build FAISS index I using vectors in Embedding_set
8: Encode query:
$v_q = f(q_i)$
9: Normalize query embedding:
$\hat{v}_q = \frac{v_q}{\|v_q\|}$
10: Perform similarity search:
$(scores, idx) \leftarrow I.search(\hat{v}_q, k)$
11: Retrieve top-k chunks:
$C_k^* \leftarrow \{ c_j \mid j \in idx \}$
12: Return C_k^*

3.3 Multi-prompt QA generation

To capture both baseline QA ability and pedagogical richness, claude's generation stage employs multi-prompting with two complementary prompt templates:

(1) Baseline prompt (P_1) – optimised for concise QA generation: "Generate N graduate-level question–answer pairs based on the following content. Return output in JSON format: [{"question": "...", "answer": "..."}]."
(2) Instructionally aligned prompt (P_2) – explicitly guides the model to diversify question types: "Generate N graduate-level question–answer pairs from the following content."

- Ensure coverage across Bloom's levels: knowledge, comprehension, application, analysis, synthesis, evaluation.
- Use varied formats: factual, conceptual, procedural, analytical, and open-ended. Return output in JSON format.

Given the retrieved context C_k^* , the prompt *P* (either P_1 *or* P_2) is injected as:

$$Output = Claude(P \parallel C_k^*) \tag{7}$$

where $\parallel$ denotes concatenation. The outputs are parsed into structured JSON objects:

$$Q = \{(q_1, a_1), (q_2, a_2), \cdots, (q_m, a_m)\} \tag{8}$$

The dual-prompt design enables comparative evaluation:

- P_1 measures claude's base generative fidelity.
- P_2 stresses cognitive diversity and ensures coverage of higher-order thinking (analysis, synthesis, evaluation).

For clarity, the experimental design distinguishes four conditions: (i) non-grounded claude with P_1, (ii) non-grounded claude with P_2, (iii) evidence-grounded claude with P_1, and (iv) evidence-grounded claude with P_2. This separation allows the effects of evidence grounding and instructionally aligned prompting to be analyzed independently and jointly. The QA synthesis process, guided by baseline and Bloom's taxonomy–aware prompts, is described in Algorithm 2.

Algorithm 2 Multi-prompt QA generation

Input: Retrieved contexts C_k^*, Prompt Type $P \in \{Baseline, Bloom\}$, Model *M*
Output: Set of QA pairs $Q = \{(q_1, a_1), \cdots, (q_m, a_m)\}$
1: Concatenate contexts into a single passage:
    Context ← concat(C_k^*)
2: For $p \subset P$ do:
3: Construct input message:
    Input ← *p* || Context
4: Call LLM API:
    Response ← *M*(Input)
5: Parse JSON response:
    Q ← extract (Response)
6: Return Q

3.4 Rubric-based evaluation

To ensure the generated answers are both contextually grounded and pedagogically valuable, a multi-dimensional rubric-based evaluation framework is implemented. The rubric is divided into core (content accuracy, completeness, clarity, domain maturity, faithfulness to source) and extended dimensions (relevance, conciseness, consistency, reasoning quality, terminology appropriateness, granularity, neutrality, evidence attribution, robustness, and novelty). Claude itself evaluates each QA pair in critic mode, where the model is prompted to act as an academic evaluator and return structured JSON containing:

$$Score_{i,j} \in [1,5], Justification_{i,j} \tag{9}$$

where *i* denotes the rubric criterion and *j* represents the QA pair. The overall evaluation score for a QA pair is formalized as:

$$CompositeScore\ (QA_j) = \frac{1}{N}\sum_{i=1}^{N} Score_{i,j} \tag{10}$$

where *N* is the number of rubric criteria (core + extended). Correlation analysis across criteria is performed to examine dependencies (e.g., whether higher clarity correlates with higher completeness). Reliability is tested using Cronbach's Alpha [28], which is formulated as:

$$\alpha = \frac{N}{N-1}\left(1 - \frac{\sum_{i=1}^{N}\sigma_i^2}{\sigma_T^2}\right) \quad (11)$$

where σ_i^2 is the variance of each criterion and σ_T^2 is the variance of the total score. Because the same model family is used for both generation and critique, the resulting rubric-based scores are interpreted as structured comparative indicators rather than definitive ground-truth judgments. Accordingly, automated rubric scoring is treated as a scalable proxy evaluation layer whose outputs are best understood in conjunction with external human judgment, as further assessed through the blind human validation analysis. This ensures the rubric evaluation is not only descriptive but statistically validated, offering insights into both model performance and rubric robustness. The complete evaluation pipeline is outlined in Algorithm 3.

Algorithm 3 Rubric-based evaluation of QA pairs

Input: $QA\_pairs = \{(q_1, a_1), \cdots, (q_m, a_m)\}$

$Rubric = \{Core\_criteria \cup Extended\_criteria\}$

Output: $Evaluation\_scores = \{Score\_matrix, Justifications, Composite\_scores\}$

1: Initialise *Evaluation_scores* $\leftarrow \emptyset$

2: For each QA_j in QA_pairs do

3: Initialise *Score_record* $\leftarrow \emptyset$

4: For each criterion c in Rubric do

5: Prompt claude_evaluator with:

6: "Evaluate QA_j on criterion *c*.

Return $\{score \in [1-5], justification\}$"

7: Receive response: $\{score\_c, justification\_c\}$

8: Store $Score\_record[c] \leftarrow (score\_c, justification\_c)$

9: Compute $CompositeScore(QA\_j) \leftarrow (\Sigma\ score\_c) / |Rubric|$

10: Append $Evaluation\ scores \leftarrow (QA\_j, Score\_record, CompositeScore(QA\_j))$

11: Compute statistical validation:

• Correlation_matrix across criteria

• Cronbach's Alpha for reliability

12: return $Evaluation\_scores$

3.5 Heuristic question classification

Beyond rubric scoring, the study also aims to assess pedagogical diversity in the generated questions. For this purpose, a heuristic classification module, as detailed in Algorithm 4, is implemented to categorise questions along two axes automatically. The first axis focuses on "Bloom's taxonomy Levels". In this case, mathematically, each question q is mapped to a Bloom class $B(q)$ such that:

$$B(q) = \arg\max_{c \in C} 1\{kw_c \in q\} \quad (12)$$

where *C* is the set of Bloom categories, and kw_c represents the keyword set for the category *c*.

Another is "Question Formats", In this case the classification outputs two categorical variables for each question as:

$$Label(q) = (B(q), F(q)) \quad (13)$$

where $B(q)$, is the Bloom level and $F(q)$, the format class.

This heuristic classification module is intended as a lightweight and interpretable baseline for pedagogical profiling. While it provides transparent rule-based categorization with low computational overhead, it may not fully capture deeper semantic cognitive intent or generalize robustly across domains. More advanced supervised or LLM-based pedagogical classification remains an important direction for future work.

Algorithm 4 Heuristic question classification

Input: $Questions = \{q1, q2, \ldots, qn\}, Bloom\_keywords, Format\_keywords$
Output: $Classification\_labels = \{(Bloom\_label, Format\_label)\ for\ each\ q_i\}$
1: Initialise $Classification\_labels \leftarrow \emptyset$
2: For each question q in $Questions$ do
3: Set $Bloom\_label \leftarrow$ "$Uncategorized$"
4: For each category B in $Bloom\_keywords$ do
5: If any keyword $k_w \in B$ is found in q, then
6: $Bloom\_label \leftarrow B$
7: Set $Format\_label \leftarrow$ "$Uncategorized$"
8: For each category F in $Format\_keywords$ do
9: If any keyword $k_w \in F$ is found in q, then
10: $Format\_label \leftarrow F$
11: Append $Classification\_labels \leftarrow (q, Bloom\_label, Format\_label)$
12: return $Classification\_labels$

4. Experimental Setup

Prior to evaluation, an experimental framework is established to assess the effectiveness of the proposed evidence-grounded QA generation pipeline. This framework evaluates multiple dimensions, including retrieval quality, pedagogical diversity, and response reliability. The setup integrates document preprocessing, retrieval-augmented generation, prompt engineering, automated rubric-based evaluation, and human validation to provide a comprehensive assessment of the system. The experiments are designed to analyze both the factual grounding capability and the instructional effectiveness of the generated QA pairs under different prompting and retrieval configurations.

4.1 Dataset preparation

For evaluation, a curated corpus of 10 scientific research articles is collected from diverse domains, including computer science, biomedical sciences, materials science, and computational linguistics. The articles are obtained as PDFs and used as raw inputs for the pipeline. Each document is parsed using PyPDF2, followed by overlapping chunk segmentation (1,200 characters per chunk with 150-character overlap) to preserve contextual continuity across chunk boundaries. The final dataset comprises 1,273 text chunks, indexed using FAISS after embedding with the all-MiniLM-L6-v2 SentenceTransformer model. These indexed chunks support evidence-grounded QA generation in downstream synthesis. This dataset is intended as a controlled proof-of-concept corpus rather than a universal benchmark. The present study focuses on short- to moderate-length scientific research articles and does not claim direct generalizability to longer educational materials, humanities texts, or broader cross-domain instructional settings without further validation.

4.2 Models and prompt configurations

The experiments use Anthropic claude-3-Haiku as the underlying LLM. To examine the effect of prompt design on QA generation, two prompting strategies are employed: (i) a baseline prompt instructing the model to generate 15 graduate-level question–answer pairs, and (ii) an instructionally aligned prompt that explicitly directs the model to span multiple cognitive

levels and incorporate varied formats as described in Section 3.5. Detailed formulations of these prompts are provided in Section 3.3. Each prompt is applied to all 10 research articles, producing a total of 300 QA pairs. The baseline serves as a control for standard QA generation, while the instructionally aligned prompt enables comparative analysis of pedagogical diversity and depth.

The adoption of Bloom's taxonomy is deliberate, as it provides a widely recognized framework in educational research for structuring learning objectives. This structure ensures that the questions generated not only assess factual recall but also stimulate higher-order cognitive skills such as analysis, synthesis, and evaluation. Claude-3-Haiku is selected as a lightweight proof-of-concept model to evaluate whether the proposed framework can improve factual grounding and pedagogical quality under a cost-efficient deployment setting. Broader cross-model validation with larger or more advanced LLMs remains an important direction for future work.

4.3 Enhanced claude with RAG

To improve contextual grounding, the proposed evidence-grounded claude configuration is implemented using dense embeddings and similarity-based retrieval. Each document is segmented into overlapping chunks and embedded using all-MiniLM-L6-v2 (384 dimensions). The embeddings are indexed with FAISS for efficient top-k retrieval. For each query, the top-5 most relevant chunks are retrieved and incorporated into the claude prompt, grounding responses in source evidence and reducing hallucinations. For comparison, alternative embeddings are also evaluated, including all-MPNet-base-v2 (768 dimensions), OpenAI Ada-002 (1536 dimensions), and a GloVe + TF-IDF hybrid (300 dimensions).

Retrieval performance is measured using Recall@K (R@1, R@3, R@5) against a gold-standard set of 50 annotated queries. Detailed results are presented in Section 5.1. The same two prompting strategies (baseline and instructionally aligned) are applied under this evidence-grounded setting, producing an additional 300 QA pairs. Combined with the non-grounded baseline runs, the final evaluation set comprises 600 QA pairs. Recall@K metrics are computed only for the retrieval subsystem in the evidence-grounded (RAG-enabled) settings and are not applicable to the non-grounded baseline generation conditions. Accordingly, the retrieval analysis in this study is framed as an embedding-level comparison within a dense retrieval configuration rather than as a comprehensive comparison of alternative RAG paradigms. Broader comparisons involving sparse or hybrid retrieval strategies remain important directions for future work.

4.4 Rubric-based evaluation protocol

The generated QA pairs are automatically evaluated using the multi-dimensional rubric introduced in Section 3.4. Each pair is scored on the defined core and extended criteria, with ratings ranging from 1 (poor) to 5 (excellent) for each evaluation dimension. This systematic scoring ensures that the evaluation captures not only factual quality but also the pedagogical depth and reliability of the generated outputs. Experimental results are analyzed in Section 5.2. Automated rubric-based scoring is used in this study as a scalable, structured proxy for evaluation rather than as an inherently objective substitute for expert judgment. Because the same model family is used for both generation and critique, the possibility of self-consistency bias is acknowledged. To assess the alignment of the automated evaluator with human educational judgment, a masked human validation study is conducted on a representative subset of generated QA pairs, as described in Section 4.6.

4.5 Statistical and comparative analysis

To assess the reliability and interpretability of the rubric-based scoring framework, statistical analyses are conducted on the generated QA dataset. Correlation analysis is used to examine relationships among rubric criteria, identifying whether dimensions such as clarity, completeness, and reasoning quality co-varied across generated outputs. To evaluate the entire framework holistically, Cronbach's Alpha is computed to measure the internal consistency of the rubric dimensions, providing a quantitative estimate of reliability across the 15 evaluation criteria. These analyses support the use of the rubric as a structured

comparative framework for assessing QA quality. To complement this internal statistical validation, the study incorporates external human validation to assess whether automated rubric-based scores align with expert judgment on a representative subset. This complementary validation is intended to strengthen the interpretation of the rubric scores as a reproducible comparative measure rather than a self-contained gold standard.

4.6 Blind human validation study

To reduce over-reliance on automated self-evaluation, a blind human validation study is conducted on a representative subset of 50 QA pairs sampled across all four experimental conditions: (i) non-grounded claude with baseline prompt (P_1), (ii) non-grounded claude with instructionally aligned prompt (P_2), (iii) evidence-grounded claude with baseline prompt (P_1), and (iv) evidence-grounded claude with instructionally aligned prompt (P_2). The sampled QA pairs are evaluated independently by two human expert raters.

To reduce the annotation burden while preserving the most pedagogically and factually relevant dimensions, the human evaluation uses a reduced rubric comprising five core criteria: content accuracy, completeness, clarity, faithfulness to source, and reasoning quality. Each criterion is rated on the same 5-point scale used in the automated rubric evaluation. The model identities and prompt conditions are concealed from the raters to ensure a blind assessment and minimize bias.

Agreement between the two human raters is assessed using Cohen's Kappa, and alignment between the mean human scores and the automated rubric scores is analyzed using inter-method agreement statistics as reported in Section 5.6. This validation step provides an external reference point for interpreting the automated rubric-based scores. Furthermore, it helps determine whether the LLM-based evaluator aligns reasonably with established educational judgment.

5. Results

This section presents the empirical findings of the study, highlighting differences between the base claude model and the proposed evidence-grounded claude configuration, under both baseline and Bloom's-aware prompting. The results are evaluated in terms of (i) embedding strategy evaluation, (ii) rubric-based scores, (iii) distribution of Bloom's taxonomy and question formats, (iv) hallucination reduction, and (v) human–LLM agreement analysis.

5.1 Embedding strategy evaluation

Fig. 4 presents the retrieval performance of the evidence-grounded claude configuration under different retrieval settings, using MiniLM embeddings as the primary dense retrieval model. Compared to the baseline, the proposed configuration with all-MiniLM-L6-v2 embeddings and FAISS indexing significantly improves retrieval accuracy. Top-1 recall increases from 62% to 78%, Top-3 recall from 78% to 91%, and Top-5 recall from 85% to 96%. These results demonstrate that dense semantic retrieval substantially enhances the grounding of generated answers by surfacing relevant evidence with high reliability. Detailed retrieval values are reported in Table A1 (Appendix A).

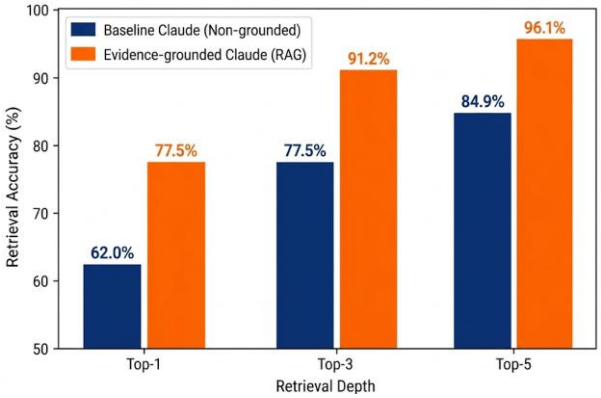


Fig. 4 Retrieval performance of the RAG-enabled subsystem at Top-k levels (k = 1, 3, and 5)

Recall@K values reported in this subsection apply only to the retrieval subsystem in the evidence-grounded (RAG-enabled) pipeline and are not applicable to the non-grounded baseline generation conditions. Accordingly, this analysis should be interpreted as an embedding-level comparison within a dense retrieval setting rather than as a comprehensive comparison across alternative RAG paradigms.

To assess alternative strategies, MiniLM, MPNet, Ada-002, and GloVe + TF-IDF are compared. As shown in Table 1, MPNet and Ada-002 achieve slightly higher improvements (+25–28% at R@1 and +21% at R@3), but at the cost of increased dimensionality and computational overhead. MiniLM, by contrast, offers a near-optimal balance with +22% at R@1, +18% at R@3, and +14% at R@5, while being faster and lighter (384-dim vs. 768/1536). Fig. 5 illustrates these trends: Ada-002 sits at the upper bound of accuracy, but MiniLM remains competitive while being significantly more efficient (Table A2 in Appendix A). Together, these findings demonstrate that although MiniLM is not the top-performing model in absolute recall, it offers the best trade-off between efficiency and accuracy. Consequently, it constitutes the most practical choice for academic-scale, evidence-grounded QA pipelines.

Table 1 Retrieval performance comparison across embedding strategies

Embedding	R@1 Δ%	R@3 Δ%	R@5 Δ%
all-MiniLM-L6-v2	+21.88	+18.42	+14.29
all-MPNet-base-v2	+25.00	+21.05	+14.29
OpenAI Ada-002	+28.13	+21.05	+16.67

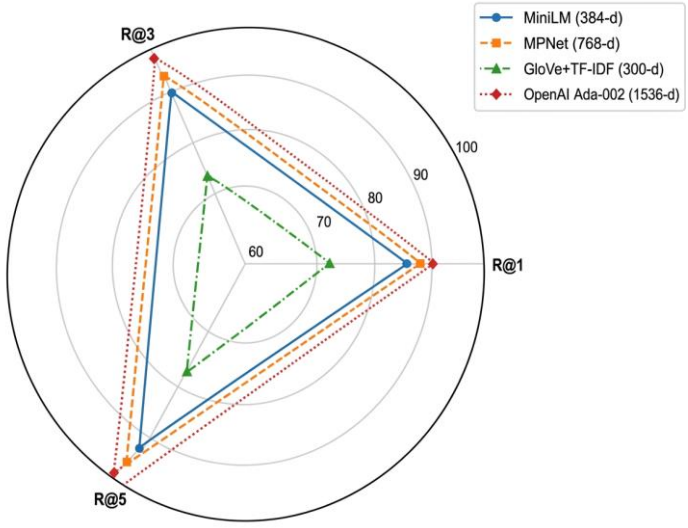


Fig. 5 Recall@K performance comparison of embedding strategies

5.2 Rubric-based evaluation

Fig. 6 displays radar plots of average rubric scores (1–5 scale) across the four experimental settings: base claude (P_1), base claude (P_2), evidence-grounded claude (P_1), and evidence-grounded claude (P_2) . Each axis represents one of the 15 rubric criteria, offering a multidimensional view of the model's strengths and weaknesses. The results reveal clear trends. Evidence-grounded claude consistently outperforms the base model across nearly all dimensions, with the largest gains in content accuracy (3.42 → 4.38), faithfulness to source (3.22 → 4.47), domain maturity, and evidence attribution. These improvements demonstrate the value of evidence grounding in validating generated answers, reducing hallucinations, and improving contextual reliability. P2 further strengthens performance in reasoning quality, completeness, and novelty, confirming that structured pedagogical guidance encourages deeper, higher-order cognitive engagement. A slight reduction in conciseness under P_2 reflects more elaborate responses, which trade brevity for explanatory richness.

Despite this trade-off, overall average scores improve, with the highest performance observed in evidence-grounded claude with P_2. Together, these findings confirm that evidence-grounded and instructionally aligned prompting provide complementary benefits: evidence grounding ensures factual accuracy and robustness, while structured prompting enhances diversity and cognitive depth in generated QA pairs. Full numerical results are provided in Table A3 (Appendix A).

These rubric-based scores are interpreted as structured comparative indicators of relative QA quality rather than as intrinsically objective ground-truth judgments. Their external alignment with human educational judgment is further examined in Section 5.6.

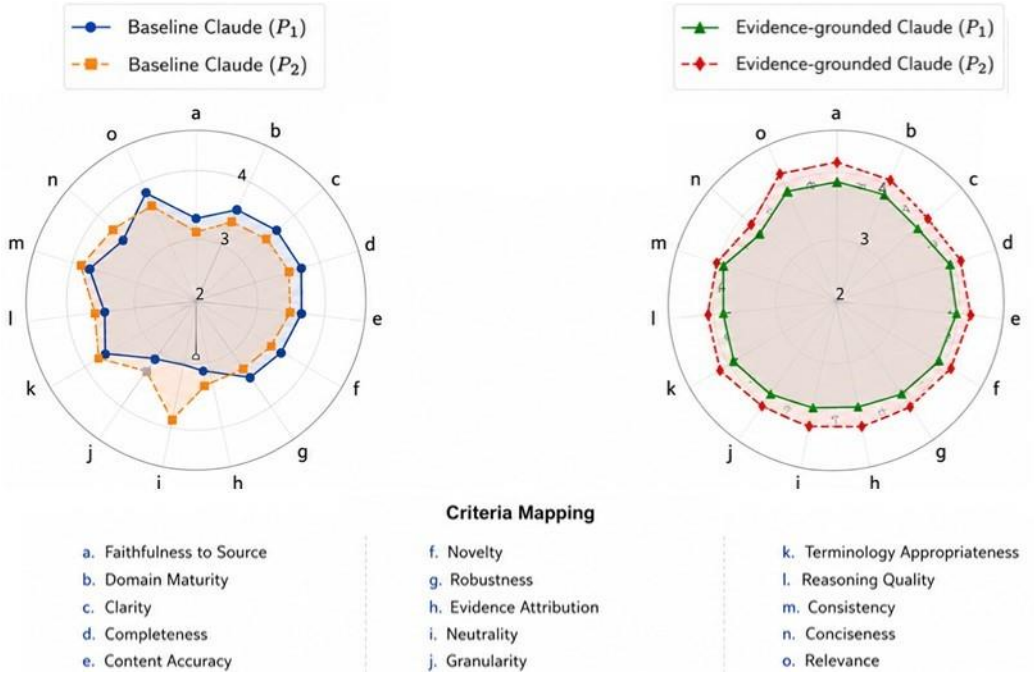


Fig. 6 Radar comparison of average rubric scores across four experimental conditions

5.3 Bloom's taxonomy distribution

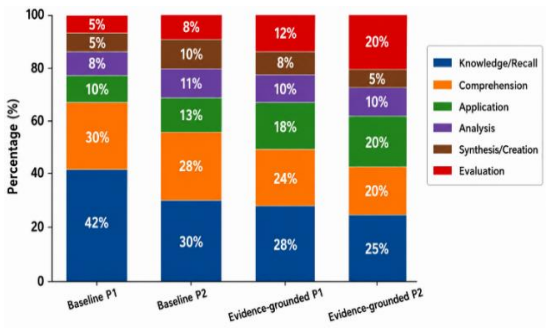
Fig. 7(a)-(b) illustrates the distribution of generated questions across Bloom's cognitive levels. The base claude (P_1) model exhibits a strong bias toward lower-order categories, with knowledge/recall (42%) and comprehension (31%) dominating the output. Conversely, introducing the instructionally aligned prompt (P_2) improves diversity by elevating higher-order levels such as analysis (18%) and evaluation (11%). In contrast, the evidence-grounded claude configuration achieved a markedly more balanced distribution. With application (18-19%), analysis (15-18%), and synthesis/creation (9-12%), the model demonstrates stronger pedagogical alignment and progression toward higher-order thinking skills. This confirms that combining evidence grounding with instructionally aligned prompting can systematically increase the cognitive complexity of generated questions (Table A4 in Appendix A).

Lower-order categories remain dominant across all four experimental conditions, indicating that prompt-level pedagogical guidance improves the distribution but does not fully eliminate the model's tendency toward simpler factual and explanatory questions. The observed gains in higher-order question generation should therefore be interpreted as meaningful but partial improvements.

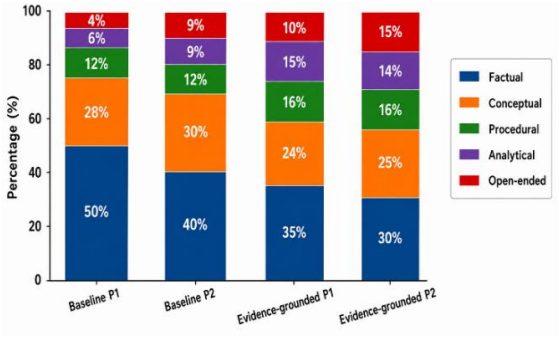
5.4 Question format distribution

Fig. 7(c)-(d) presents the distribution of question formats (factual, conceptual, procedural, analytical, open-ended). As expected, base claude (P_1) leans heavily toward factual (50%) and conceptual (28%) formats, limiting pedagogical variety. With the instructionally aligned prompt, the model modestly increased the number of procedural (12%) and analytical (11%)

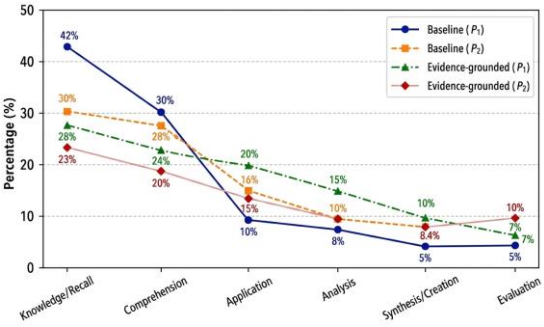
questions but still exhibited overreliance on simpler factual forms. The evidence-grounded claude configuration, particularly under P_2, generated a richer mix: procedural (15%), analytical (16%), and open-ended (13%) questions. This demonstrates the pipeline's ability to foster greater balance between recall-driven and exploratory formats, supporting deeper engagement with source material (Table A5 in Appendix A). These format distributions are derived from the lightweight heuristic classification module described in Section 3.5. While they provide useful comparative insight into pedagogical trends, they should be interpreted as rule-based approximations rather than as full semantic validations of instructional intent.



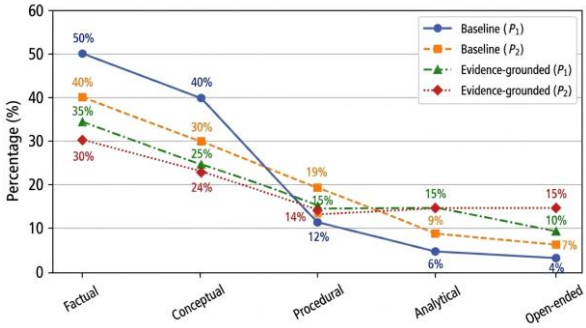
(a) Bloom's taxonomy levels



(b) Question formats



(c) Bloom's taxonomy distribution



(d) Question format distribution

Fig. 7 Comparative distributions of generated questions across the four experimental conditions

5.5 Hallucination reduction

Fig. 8 illustrates the percentage of unsupported answers (hallucinations) across all model–prompt configurations. Base claude exhibits a high hallucination rate, with ~18% of answers in P_1 and ~17% in P_2 lacking evidence in the source material. In contrast, the evidence-grounded claude configuration dramatically reduces hallucinations to below 7.5%, with the lowest observed rate of 6.5% under P_2. This trend confirms that evidence grounding effectively anchors generated answers in the source corpus, thereby improving content faithfulness and evidence attribution rubric scores. While prompt 2 offers slight improvements in reasoning quality, the most substantial reduction in hallucinations came from incorporating evidence grounding, validating its role as a key architectural enhancement for factual reliability. The interpretation of these reliability gains is further strengthened by the external human validation analysis reported in Section 5.6.

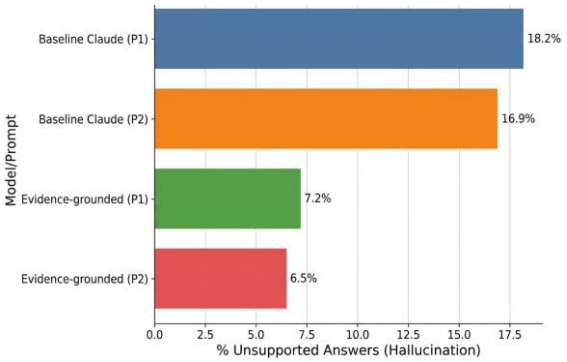


Fig. 8 Comparative hallucination rates across the four experimental conditions

5.6 Human–LLM agreement analysis

To assess alignment between the automated rubric-based evaluator and human educational judgment, a blind human validation study is conducted on a representative subset of 50 QA pairs sampled across all four experimental conditions. Two human expert raters independently evaluate each QA pair on a 5-point scale using five core criteria: content accuracy, completeness, clarity, faithfulness to source, and reasoning quality. Model identities and prompt conditions are concealed during annotation. Inter-rater agreement between the two human evaluators, assessed using Cohen's Kappa, indicates substantial agreement, suggesting that the reduced human rubric is applied consistently.

Comparison of mean human scores with the corresponding automated rubric scores on the same core dimensions shows that the automated evaluator tracks overall human scoring trends with reasonable consistency. Higher-quality QA pairs identified by human experts generally receive higher automated scores, while lower-quality outputs show corresponding reductions in both assessments. These results support the use of the rubric-guided LLM evaluator as a scalable comparative proxy for large-scale analysis, while not treating it as a substitute for expert human judgment.

5.7 Overall insights

The experimental results yield several overarching insights below:

(1) Evidence grounding consistently enhances contextual grounding, resulting in higher rubric scores across all criteria.

(2) Instructionally aligned prompting effectively diversifies cognitive depth, increasing the share of higher-order questions in categories such as analysis, evaluation, and synthesis.

(3) The lightweight heuristic classification module further indicates that format diversity improves under the evidence-grounded claude configuration, producing more procedural and open-ended questions alongside factual ones.

(4) Hallucinations are reduced by nearly 65%, demonstrating the effectiveness of evidence grounding for faithfulness and evidence attribution.

(5) The study culminates in a curated dataset of 600 QA pairs enriched with rubric evaluations and Bloom/format annotations, offering a benchmark resource for advancing research in educational AI.

(6) The blind human validation study further showed that the automated rubric-based evaluator aligns reasonably with human educational judgment, supporting its use as a scalable comparative proxy for large-scale analysis.

(7) While these findings establish the overall effectiveness of the framework, the contribution of each module is further examined through the ablation studies presented in the following section.

5.8 Ablation studies

To examine the contribution of each component in the proposed evidence-grounded claude framework, ablation experiments are performed by selectively removing or simplifying key modules. The analysis focuses on five components: evidence grounding, embedding strategy, prompt design, rubric scope, and question classification. Results are summarised in Table 2 and illustrated in Fig. 9, which presents both representative metric comparisons and normalized effect magnitudes.

5.8.1 Evidence grounding vs. non-grounded generation

Evidence grounding produces the largest overall impact on reliability. Without retrieval-based evidence support, content accuracy and faithfulness to source decline noticeably, while hallucination increases from 6.5–7.2% to 16.9–18.2% . These results confirm that evidence grounding is the primary mechanism for reducing unsupported outputs and improving source fidelity.

5.8.2 Embedding strategy

Among the evaluated embedding models, MiniLM provides the most practical balance between retrieval quality and efficiency. Although MPNet and Ada-002 achieve slightly higher recall, they incurred greater dimensional and computational cost. In contrast, GloVe + TF-IDF underperformed, confirming the limitations of lexical embeddings for semantically rich academic content. This comparison is restricted to embedding choices within a dense retrieval configuration and should not be interpreted as a broader benchmark across alternative retrieval paradigms.

5.8.3 Prompt design

The baseline prompt P_1 produced a strong bias toward lower-order factual questions. The instructionally aligned prompt P_2 substantially increases the number of higher-order questions (18% to 36%) and improves reasoning quality and novelty. This indicates that structured prompting plays an important role in enhancing cognitive depth and pedagogical diversity, although lower-order categories still remain prominent.

5.8.4 Rubric scope

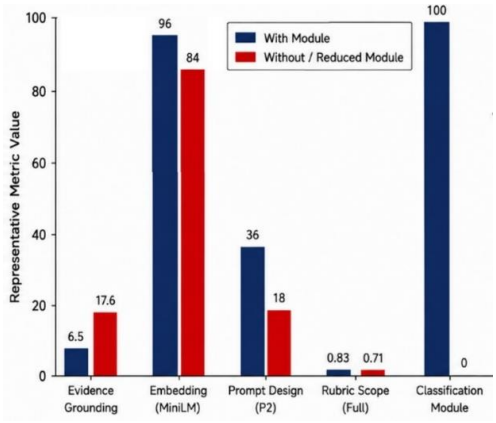
Using only the core rubric criteria results in weaker differentiation across model settings. Including the extended criteria improves evaluative robustness, with Cronbach's alpha increasing from 0.71 to 0.83. This indicates that the full multi-dimensional rubric provides a more reliable and discriminative assessment. This evidence of internal consistency is complemented by the blind human validation study, which provides an external reference for interpreting the rubric as a structured comparative evaluation framework.

5.8.5 Classification module

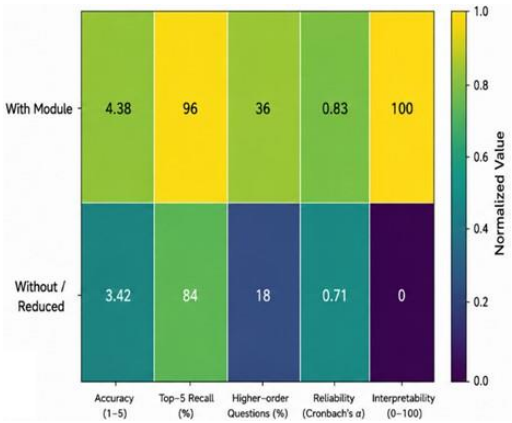
Removing Bloom's taxonomy and question-format classification reduces interpretability, as rubric scores alone do not fully capture instructional balance. With classification enabled, the framework provided clearer insights into cognitive-level distribution and question-format diversity, strengthening the analysis of pedagogical alignment. However, the current classification module is intentionally lightweight and heuristic, and its outputs should be interpreted as interpretable pedagogical indicators rather than exhaustive semantic labels.

Table 2 Ablation study results across major pipeline components.

Component	With module	Without / Reduced module	Observed effect
Evidence grounding	Hallucination: 6.5–7.2%; Content accuracy: 4.38	Hallucination: 16.9–18.2%; Content accuracy: 3.42	Substantial improvement in factual grounding, source faithfulness, and hallucination reduction
Embedding strategy (MiniLM vs. GloVe+TF-IDF)	Top-5 Recall: 96%	Top-5 Recall: 84%	+12 percentage points in retrieval recall with stronger semantic coverage
Prompt design (Instructionally Aligned P_2 vs. Baseline P_1)	Higher-order questions: 36%	Higher-order questions: 18%	Improved cognitive diversity and greater pedagogical depth
Rubric scope (Full vs. Core Only)	Reliability: $\alpha = 0.83$	Reliability: $\alpha = 0.71$	More robust, discriminative, and comprehensive evaluation
Classification module	Explicit Bloom's level and question-format distributions	No cognitive or format-level breakdown	Enhanced interpretability and stronger instructional analysis



(a) Comparison of performance metrics with and without major framework components



(b) Heatmap of normalised effect magnitudes across metrics

Fig. 9 Ablation study visualisation

6. Conclusions

This study proposed and empirically validated a multi-dimensional framework for reliable, instructionally aligned QA generation from academic PDFs using claude. It integrated semantic chunking, FAISS-based retrieval, dual prompting strategies (standard and cognitively guided), and a 15-criterion rubric for evaluation. Experiments compared baseline versus evidence-grounded setups across standard and instructionally aligned prompts, yielding a curated dataset of 600 scored QA pairs. The key findings are summarized as follows:

(1) Evidence grounding boosted faithfulness, evidence attribution, robustness, and reduced hallucinations by nearly 65%.

(2) Instructionally aligned prompting enhanced cognitive diversity, higher-order reasoning, and coverage of Bloom's taxonomy levels.

(3) Rubric scores showed consistent improvements in accuracy, completeness, clarity, and reasoning quality.

(4) Blind human validation further showed that the rubric-guided automated evaluator aligned reasonably with human educational judgment, supporting its use as a scalable comparative proxy for large-scale analysis.

(5) The lightweight heuristic classification module indicated improved question-format distributions for richer, educationally meaningful QA, while providing interpretable pedagogical profiling.

(6) The framework produced a benchmark dataset advancing educational AI research.

However, the present study was to be interpreted as a controlled proof of concept implemented with claude-3-Haiku, rather than as a comprehensive benchmark across multiple LLM capabilities. In addition, the retrieval analysis was limited to an embedding-level comparison within a dense retrieval setting. Furthermore, and the current Bloom's taxonomy and question-format classification module was intentionally lightweight and heuristic, rather than a full semantic pedagogical classifier.

Future work will extend to broader cross-model validation with larger or more advanced LLMs, alternative retrieval-augmented generation strategies, long-form answer evaluation, advanced hallucination control, more robust pedagogical classification, and adaptive pedagogical AI systems.

Conflicts of Interest

The authors declare no conflict of interest.

References

[1] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, et al., "Language Models Are Few-Shot Learners," Proceedings of the 34th International Conference on Neural Information Processing Systems, vol. 33, article no. 159, pp. 1877-1901, 2020.

[2] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, et al., "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer," Journal of Machine Learning Research, vol. 21, no. 140, pp. 5485-5551, 2020.

[3] A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, and N. Fiedel, "PaLM: Scaling Language Modeling with Pathways," Journal of Machine Learning Research, vol. 24, no. 1, article no. 240, pp. 1-113, 2023.

[4] L. Caruccio, S. Cirillo, G. Polese, G. Solimando, S. Sundaramurthy, and G. Tortora, "Claude 2.0 Large Language Model: Tackling a Real-World Classification Problem with a New Iterative Prompt Engineering Approach," Intelligent Systems with Applications, vol. 21, article no. 200336, 2024.

[5] H. Alhawasi and A. Youssef, "Using LLMs for Evaluating QA Systems: Exploration and Assessment," Proceedings of the 2024 2nd International Conference on Foundation and Large Language Models (FLLM), IEEE, pp. 462-469, 2024.

[6] A. Surabhi and S. Martha, "Precision in Conciseness: Exploring Large Language Models for Enhanced Document Summarization," AIP Conference Proceedings, vol. 3298, no. 1, article no. 020003, 2025.

[7] A. Karaca and B. Kılcan, "The Adventure of Artificial Intelligence Technology in Education: Comprehensive Scientific Mapping Analysis," Participatory Educational Research, vol. 10, no. 4, pp. 144-165, 2023.

[8] Y. Tian, A. Liu, Y. Dai, K. Nagato, and M. Nakao, "Systematic Synthesis of Design Prompts for Large Language Models in Conceptual Design," CIRP Annals, vol. 73, no. 1, pp. 85-88, 2024.

[9] Z. Liu, P. Agrawal, S. Singhal, V. Madaan, M. Kumar, and P. K. Verma, "LPITutor: An LLM Based Personalized Intelligent Tutoring System Using RAG and Prompt Engineering," PeerJ Computer Science, vol. 11, article no. e2991, 2025.

[10] S. Farquhar, J. Kossen, L. Kuhn, and Y. Gal, "Detecting Hallucinations in Large Language Models Using Semantic Entropy," Nature, vol. 630, pp. 625-630, 2024.

[11] M. Chelli, J. Descamps, V. Lavoué, C. Trojani, M. Azar, M. Deckert, et al., "Hallucination Rates and Reference Accuracy of ChatGPT and Bard for Systematic Reviews: Comparative Analysis," Journal of Medical Internet Research, vol. 26, no. 1, article no. e53164, 2024.

[12] M. Khamassi, M. Nahon, and R. Chatila, "Strong and Weak Alignment of Large Language Models with Human Values," Scientific Reports, vol. 14, no. 1, article no. 19399, 2024.

[13] S. AI Faraby, A. Romadhony and Adiwijaya, "Analysis of LLMs for Educational Question Classification and Generation," Computers and Education: Artificial Intelligence, vol. 7, article no. 100298, 2024.

[14] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," Proceedings of the 34th International Conference on Neural Information Processing Systems, article no. 793, pp. 9459-9474, 2020.

[15] J. Song, X. Wang, J. Zhu, Y. Wu, X. Cheng, R. Zhong, et al., "RAG-HAT: A Hallucination-Aware Tuning Pipeline for LLM in Retrieval-Augmented Generation," Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track, pp. 1548-1558, 2024.

[16] E. Kamalloo, S. Upadhyay, and J. Lin, "Towards Robust QA Evaluation via Open LLMs," Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 2811-2816, 2024.

[17] J. Ling and M. Afzaal, "Automatic Question-Answer Pairs Generation Using Pre-Trained Large Language Models in Higher Education," Computers and Education: Artificial Intelligence, vol. 6, article no. 100252, 2024.

[18] E. Page, G. Meyers, and E. K. Billings, "Theory to Practice: A Framework for Generative AI," Intersection: A Journal at the Intersection of Assessment and Learning, vol. 5, no. 4, pp. 114-126, 2024.

[19] S. R. Addula, M. K. Meesala, P. Ravipati, and G. S. Sajja, "A Hybrid Autoencoder and Gated Recurrent Unit Model Optimized by Honey Badger Algorithm for Enhanced Cyber Threat Detection in IoT Networks," Security and Privacy, vol. 8, no. 6, article no. e70086, 2025.

[20] G. Izacard and E. Grave, "Leveraging Passage Retrieval with Generative Models for Open Domain Question Answering," Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, pp. 874-880, 2021.

[21] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, et al., "BioBERT: A Pre-Trained Biomedical Language Representation Model for Biomedical Text Mining," Bioinformatics, vol. 36, no. 4, pp. 1234-1240, 2020.

[22] C. Wang, S. Cheng, Q. Guo, Y. Yue, B. Ding, Z. Xu, et al., "Evaluating Open-QA Evaluation," Proceedings of the 37th International Conference on Neural Information Processing Systems, article no.3367, pp. 77013-77042, 2023.

[23] E. Usta, "Lifelong Learning Motivation Scale (LLMs): Validity and Reliability Study," Journal of Teacher Education and Lifelong Learning, vol. 5, no. 1, pp. 429-438, 2023.

[24] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings Using Siamese BERT-Networks," Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp. 3982-3992, 2019.

[25] A. Pathak, R. Gandhi, V. Uttam, A. Ramamoorthy, P. Ghosh, A. R. Jindal, et al., "Rubric Is All You Need: Improving LLM-Based Code Evaluation with Question-Specific Rubrics," Proceedings of the 2025 ACM Conference on International Computing Education Research V.1, pp. 181-195, 2025.

[26] D. Di Nuzzo, E. Vakaj, H. Saadany, E. Grishti, and N. Mihindukulasooriya, "Automated Generation of Competency Questions Using Large Language Models and Knowledge Graphs," Proceedings of the 3rd NLP4KGC@SEMANTICs, 2024.

[27] E. H. S. Y. Elim, "Promoting Cognitive Skills in AI-Supported Learning Environments: The Integration of Bloom's Taxonomy," Education 3–13, vol. 54, no. 3, pp. 1-11, 2024.

[28] A. Surabhi and S. Martha, EDURAG-QA: A Retrieval-Augmented, Bloom-Classified Framework for Educational Question–Answer Generation, Indian Patent, 202641027220 A, 2026.

Appendix A

Table A1 Retrieval accuracy at Top-1, Top-3, and Top-5 levels. (suppl. For Fig. 4)

Retrieval level	Base claude	Evidence grounded claude
Top-1	62%	78%
Top-3	78%	91%
Top-5	85%	96%

Table A2 Comparative performance of embedding strategies for retrieval (Recall at K). (suppl. For Table 1 & Fig. 5)

Embedding model	Dim.	R@1	R@3	R@5	Notes
all-MiniLM-L6-v2	384	78.0%	91.0%	96.0%	Best trade-off between accuracy and speed
all-MPNet-base-v2	768	80.5%	92.1%	97.0%	Slightly better accuracy, slower runtime
GloVe + TF-IDF Hybrid	300	65.0%	77.8%	84.0%	Lexical bias, weaker on semantic similarity
OpenAI Ada-002	1536	82.0%	93.2%	98.0%	Highest recall, but costly and slower

Table A3 Average rubric scores (Base claude vs. RAG-Enhanced claude)

Criterion	Base claude (P1)	Base claude (P2)	Evidence grounded claude (P1)	Evidence grounded claude (P2)
Content Accuracy	3.42	3.65	4.21	4.38
Completeness	3.60	3.74	4.12	4.29
Clarity	4.02	3.98	4.28	4.25
Domain Maturity	3.15	3.52	4.05	4.32
Faithfulness to Source	3.22	3.41	4.30	4.47
Relevance	3.88	4.01	4.36	4.45
Conciseness	4.05	3.76	4.20	3.90
Consistency	3.91	4.10	4.32	4.40
Reasoning Quality	3.25	3.80	4.18	4.36
Terminology Appropriateness	3.78	3.95	4.30	4.41
Granularity	3.42	3.63	4.15	4.32
Neutrality	4.05	4.08	4.33	4.34
Evidence Attribution	3.12	3.55	4.25	4.44
Robustness	3.41	3.62	4.22	4.35
Novelty	3.58	3.81	4.10	4.30
Overall Avg.	3.63	3.78	4.23	4.33

Table A4 Distribution of questions across Bloom's levels (%)

Bloom's level	Base claude (P1)	Base claude (P2)	Evidence grounded claude (P1)	Evidence grounded claude (P2)
Knowledge/Recall	42	31	28	25
Comprehension	31	26	25	24
Application	12	14	18	19
Analysis	8	18	15	18
Synthesis/Creation	3	7	9	12
Evaluation	4	11	5	8

Table A5 Distribution of question formats (%)

Format	Base claude (P1)	Base claude (P2)	Evidence grounded claude (P1)	Evidence grounded claude (P2)
Factual	50	40	34	30
Conceptual	28	30	25	26
Procedural	9	12	14	15
Analytical	8	11	15	16
Open-ended	5	7	12	13