DOI: **10.37943/25VYNZ9998**

Aruzhan Tugambayeva
Master's student, Faculty of Information Technology
a.tugambayeva@agakaz.kz, orcid.org/0009-0006-4313-5652
Kazakh British Technical University, Kazakhstan
Aivar Sakhipov*
PhD, Assistant Professor, School of Software Engineering
aivar.sakhipov@astanait.edu.kz, orcid.org/0000-0003-1045-4199
Astana IT University, Kazakhstan

AI-BASED QUESTION GENERATION FOR AVIATION TRAINING: COMPARING RETRIEVAL-AUGMENTED GENERATION AND FINE-TUNED MODELS

Abstract: This study examines how retrieval-augmented and fine-tuned architecture influences the cognitive complexity, terminology usage, and pedagogical characteristics of automatically generated aviation-related questions. The objective is to determine how different modeling strategies affect not only linguistic quality but also the educational value of generated content. A retrieval-augmented generation pipeline was implemented by combining vector-based document retrieval using Facebook AI Similarity Search with the Mistral-7B language model, containing seven billion parameters, applied to a curated knowledge base of 238 aviation documents. In parallel, a T5-small language model, comprising 60 million parameters, was fine-tuned using the Low-Rank Adaptation method on a dataset of 920 aviation context–question pairs.

Both systems were evaluated on a test set of 116 examples. The evaluation framework included expert-based assessment aligned with Bloom's taxonomy of cognitive learning objectives, as well as domain-specific criteria such as aviation terminology coverage and lexical diversity. In addition, widely used text similarity metrics were employed, including Bilingual Evaluation Understudy, Recall-Oriented Understudy for Gisting Evaluation with the longest common subsequence variant, and Bidirectional Encoder Representations from Transformers Score.

The results reveal distinct differences in the cognitive profiles of the generated questions. All questions produced by the fine-tuned model corresponded to the Knowledge level of Bloom's taxonomy, indicating a strong emphasis on factual recall. In contrast, the retrieval-augmented system generated questions that more frequently addressed higher cognitive levels, particularly Comprehension (53.3%) and Application (40.0%). It also demonstrated broader coverage of aviation terminology (92.2% compared to 44.0%) and greater output diversity (112 unique questions versus 56). Conversely, the fine-tuned model achieved higher similarity scores and approximately five times faster inference speed.

Keywords: retrieval-augmented generation; question generation; fine-tuning; transformer models; aviation education; natural language processing; domain adaptation; low-rank adaptation; mistral.

Introduction

The integration of artificial intelligence in educational content generation has emerged as a significant research direction, with large language models (LLMs) demonstrating remarkable capabilities in text generation tasks [1]. However, applying these models to specialized technical domains such as civil aviation training presents unique challenges, including domain-specific terminology preservation, factual accuracy requirements, and the need for up-to-date knowledge integration [2].

Two primary approaches have emerged for adapting language models to specialized domains: fine-tuning and Retrieval-Augmented Generation (RAG). Fine-tuning involves training

Received: 14.01.2026 Accepted: 25.02.2026 Published: 30.03.2026

the model on domain-specific data, enabling it to internalize specialized knowledge and terminology [3]. RAG, alternatively, combines information retrieval mechanisms with generative models, allowing dynamic access to external knowledge bases during generation [4]. Each approach presents distinct trade-offs in terms of computational requirements, knowledge updatability, and output quality.

Civil aviation training represents a particularly demanding domain for automated content generation due to its safety-critical nature, extensive regulatory framework, and precise technical vocabulary. Training materials must accurately reflect current aviation standards and procedures while maintaining pedagogical effectiveness [5]. This creates a compelling use case for comparing fine-tuning and RAG approaches.

Automated question generation has evolved from rule-based systems to neural approaches leveraging pre-trained language models. Du et al. demonstrated that sequence-to-sequence models could generate contextually relevant questions, establishing benchmarks for the task [6]. The introduction of transformer architectures, particularly T5 [7], unified various NLP tasks under a text-to-text framework, enabling more flexible question generation approaches. Rodriguez-Torrealba et al. presented end-to-end pipelines for multiple-choice question generation using text-to-text transfer transformers, achieving strong results on educational datasets [8]. Recent work has explored parameter-efficient fine-tuning methods such as LoRA for domain adaptation [9], significantly reducing computational requirements while maintaining performance.

RAG was introduced by Lewis et al. as a method to augment language models with external knowledge retrieval [4]. The approach combines a retrieval component, typically based on dense vector representations, with a generative model that conditions its output on both the input and retrieved documents. This architecture addresses limitations of pure fine-tuning, including knowledge cutoff and inability to update information without retraining [10]. Practical RAG implementations rely on vector databases such as FAISS [11] for efficient similarity search and sentence transformers [12] for generating dense semantic representations.

Recent comparative studies have examined the trade-offs between RAG and fine-tuning approaches. Soudani et al. found that RAG surpasses fine-tuning for less popular factual knowledge in domain-specific applications [13]. In the aviation domain specifically, Wang et al. developed AviationGPT, demonstrating the feasibility of adapting large language models for aviation-specific NLP tasks [14]. Al Faraby et al. provided a comprehensive review of neural question generation for educational purposes, highlighting the importance of generating not only fluent but also pedagogically valuable questions [15].

Applications of language models in specialized domains have highlighted the importance of domain adaptation. Karabacak et al. explored generative models in medical education, emphasizing factual accuracy requirements [16]. Ling and Afzaal demonstrated the effectiveness of pre-trained language models for educational question generation [17]. Aviation-specific applications remain limited, though our previous scientific work demonstrated the feasibility of fine-tuned T5 models for aviation question generation, achieving Corpus BLEU of 24.27% on a custom aviation corpus [18].

This research addresses the gap in empirical comparisons between fine-tuning and RAG for domain-specific educational content generation. The aim of this study is to analyze how different generation architectures (retrieval-augmented and fine-tuned) produce questions with distinct cognitive and pedagogical characteristics in aviation education. The specific objectives are: (1) to implement a RAG-based question generation system utilizing FAISS vector retrieval and Mistral-7B; (2) to characterize the cognitive complexity of generated questions using Bloom's taxonomy; (3) to analyze domain-specific quality measures including aviation terminology coverage and question complexity; (4) to conduct preliminary expert assessment of pedagogical quality using Bloom's taxonomy; (5) to provide practical recommendations for method selection.

Methods and Materials

The experimental dataset comprises 1,151 aviation-specific context-question pairs covering more than 50 topics, including flight operations (26%), aircraft instruments (20%), navigation and communication (17%), weather phenomena (14%), regulations and safety (11%), human factors (6%), and aircraft performance (6%). The dataset was split into training (920 examples, 80%), validation (115 examples, 10%), and test (116 examples, 10%) sets using stratified sampling. Average context length is 61.4 words, and average reference question length is 5.3 words. The knowledge base for RAG contains 238 unique aviation contexts extracted from the training corpus.

The fine-tuning approach utilizes T5-small (60M parameters) with Low-Rank Adaptation (LoRA). LoRA configuration includes rank r=8, alpha=16, and dropout=0.1, applied to query and value attention matrices. This reduces trainable parameters to 294,912 (0.49% of total). Training was conducted for 7 epochs using AdamW optimizer with a learning rate 3×10^{-4}, batch size 8, and 10 warmup steps on a Tesla T4 GPU. Training was completed in approximately 35 minutes. Question generation uses beam search with 4 beams, minimum length 5, maximum length 80 tokens, and no-repeat n-gram size of 2.

LoRA reduces the number of trainable parameters by decomposing weight updates into low-rank matrices [9]. For a pre-trained weight matrix $W_0 \in \mathbb{R}^{d\times k}$, the adapted weight is defined as:

$$W' = W_0 + \Delta W = W_0 + BA, \tag{1}$$

where $B \in \mathbb{R}^{d\times r}$ and $A \in \mathbb{R}^{r\times k}$ are trainable low-rank matrices, and $r \ll \min(d, k)$ is the rank hyperparameter. During fine-tuning, W_0 remains frozen and only B and A are updated. The scaling factor α/r is applied to the update, yielding:

$$h = W_0 x + \frac{\alpha}{r} \cdot BAx, \tag{2}$$

where x is the input activation, and α is a scaling hyperparameter. In our configuration, r = 8 and α = 16, giving a scaling factor of α/r = 2. LoRA is applied to the query (Wq) and value (Wv) attention matrices of T5-small. The total number of trainable parameters is:

$$\left|\theta_{\text{LoRA}}\right| = 2r(d_q + d_v) \times L, \tag{3}$$

where d_q and d_v are the dimensions of query and value projections (512 for T5-small), and L is the number of transformer layers (6 encoder + 6 decoder). This yields 294,912 trainable parameters, representing 0.49% of the total 60M parameters.

The fine-tuned model generates questions by maximizing the conditional probability:

$$P_{\text{FT}}(y \mid x) = \prod_{t=1}^{T} P(y_t \mid y_{<t}, x; \theta_0 + \Delta\theta_{\text{LoRA}}), \tag{4}$$

where x is the input context, $y = (y_1, \ldots, y_T)$ is the generated question, θ_0 represents the frozen pre-trained parameters, and $\Delta\theta_{\text{LoRA}} = \{B_i, A_i\}$ are the learned low-rank updates. The model is optimized using cross-entropy loss:

$$\mathrm{L}_{\text{CE}} = -\sum_{t=1}^{T} \log P(y_t^* \mid y_{<t}^*, x; \theta_0 + \Delta\theta_{\text{LoRA}}), \tag{5}$$

where y* is the reference question from the training set.

The RAG system consists of three components: (1) a document store containing 238 unique aviation contexts indexed by dense embeddings, (2) a retrieval module using FAISS for efficient similarity search, and (3) a generation module utilizing Mistral-7B (7.3B parameters) through the Ollama framework. Document embeddings are generated using Sentence-BERT (all-MiniLM-L6-v2) [12], producing 384-dimensional vectors. FAISS IndexFlatIP is employed for cosine similarity search, retrieving the top 3 most relevant documents for each query context.

The generation prompt template instructs the model: "You are an aviation instructor creating test questions for pilot training. Based on the following aviation knowledge context, generate ONE clear, concise question that tests understanding of the key concept. Requirements: Generate only ONE question; Keep it under 15 words; Focus on the main concept; Use proper aviation terminology; Do not include the answer." The temperature parameter is set to default (0.7). Fig. 1 illustrates the architectural differences between the two approaches.

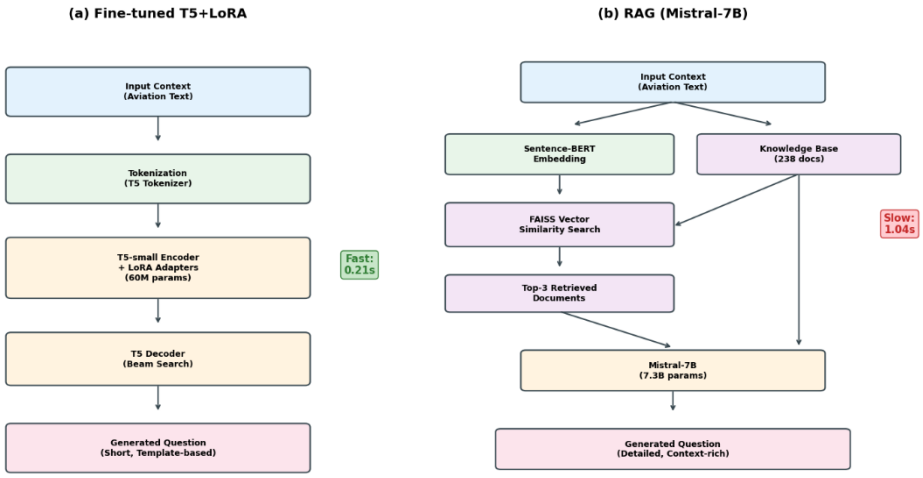

Fig. 1. Comparison of Fine-tuned T5+LoRA and RAG system architectures

The RAG pipeline formalizes question generation as a two-stage process: retrieval and conditional generation.

Retrieval stage. Given an input context q, the retrieval module computes cosine similarity between the query embedding and all document embedding in the knowledge base:

$$\text{sim}(q, d_i) = \frac{e(q) \cdot e(d_i)}{\| e(q) \| \cdot \| e(d_i) \|}, \tag{6}$$

where e(·) denotes the Sentence-BERT embedding function (all-MiniLM-L6-v2) producing 384-dimensional vectors, and d_i is the i-th document in the knowledge base D = {d_1, ..., d_{238}}. The top-k most similar documents are retrieved:

$$D_k(q) = \arg\max_{S \subset D, |S|=k} \sum_{d \in S} \text{sim}(q, d), \tag{7}$$

where k = 3 in our implementation. FAISS IndexFlatIP performs exact inner-product search after L2 normalization, which is equivalent to cosine similarity.

Generation stage. The retrieved documents are concatenated with the input context and passed to Mistral-7B. The generation probability is conditioned on both the input and retrieved documents:

$$P_{\text{RAG}}(y \mid x) = \prod_{t=1}^{T} P(y_t \mid y_{<t}, x, D_k(x); \theta_{\text{LLM}}), \tag{8}$$

where θ_{LLM} represents the frozen Mistral-7B parameters. Unlike the fine-tuned approach (Eq. 4), the model parameters are not modified; instead, the retrieved context $D_k(x)$ provides domain-specific knowledge at inference time. The key architectural distinction can be summarized as:

Fine-tuning: $P(y \mid x; \theta_o + \Delta\theta)$ is knowledge encoded in parameters.

RAG: $P(y \mid x, D_k(x); \theta_{LLM})$ is knowledge provided through context.

Evaluation employs both standard NLP metrics and domain-specific measures, formalized below.

Corpus BLEU measures n-gram precision between generated and reference questions [19]:

$$\text{BLEU} = BP \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n \right), \tag{9}$$

where p_n is the modified n-gram precision, w_n = 1/N are uniform weights (N = 4), and BP is the brevity penalty:

$$BP = \begin{cases} 1, & \text{if } c > r \\ \exp\left(1 - \frac{r}{c}\right), & \text{if } c \le r \end{cases}, \tag{10}$$

where c is the total length of generated output and r is the total length of reference questions. The brevity penalty penalizes overly short outputs, which is relevant given the significant length difference between fine-tuned (4.4 words) and RAG (15.2 words) outputs.

ROUGE-L F1 measures the longest common subsequence (LCS) between generated and reference questions:

$$R_{\text{lcs}} = \frac{\text{LCS}(y^*, \hat{y})}{|y^*|}, \tag{11}$$

$$P_{\text{lcs}} = \frac{\text{LCS}(y^*, \hat{y})}{|\hat{y}|}, \tag{12}$$

$$F_{\text{lcs}} = \frac{(1+\beta^2) \cdot P_{\text{lcs}} \cdot R_{\text{lcs}}}{R_{\text{lcs}} + \beta^2 \cdot P_{\text{lcs}}}, \tag{13}$$

where β is set to favor recall (β = 1.2).

BERTScore F1 computes semantic similarity using contextual embeddings from RoBERTa-large [20]:

$$R_{\text{BERT}} = \frac{1}{|y^*|} \sum_{y_i \in y^*} \max_{\hat{y}_j \in \hat{y}} e_i^{\cdot} e_j, \tag{14}$$

$$P_{\text{BERT}} = \frac{1}{|\hat{y}|} \sum_{\hat{y}_j \in \hat{y}} \max_{y_i \in y^*} e_i^{\cdot} e_j, \tag{15}$$

$$F_{\text{BERT}} = \frac{2 \cdot P_{\text{BERT}} \cdot R_{\text{BERT}}}{P_{\text{BERT}} + R_{\text{BERT}}}, \tag{16}$$

where e_i and e_j are the contextualized token embeddings from RoBERTa-large, and the cosine similarity between them captures semantic equivalence beyond surface-level n-gram matching.

Aviation terminology coverage measures the proportion of generated questions containing at least one domain-specific term from a curated vocabulary V of 65 aviation terms:

$$TC = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}\left(\hat{y}_i \cap V \neq \varnothing\right), \tag{17}$$

where N = 116 is the number of test samples and $\mathbb{1}(\cdot)$ is the indicator function.

Unique bigram ratio assesses lexical diversity across the generated corpus:

$$UBR = \frac{\left|\{\text{unique bigrams in } \hat{y}_1, \ldots, \hat{y}_N\}\right|}{\left|\{\text{total bigrams in } \hat{y}_1, \ldots, \hat{y}_N\}\right|}. \tag{18}$$

Higher values indicate greater lexical variety, with UBR = 1.0 meaning all bigrams are unique. Average question length and inference time are additionally reported to characterize output verbosity and practical deployment efficiency.

The writers, who have a joint background in artificial intelligence research and aviation education, carried out a preliminary expert review. Four criteria were used to evaluate a stratified sample of 15 questions per approach: (1) grammatical correctness (1–5 scale), (2) context relevance (1–5 scale), (3) terminology accuracy (1–5 scale), and (4) cognitive level classification using Bloom's taxonomy, classifying questions as Knowledge (recall of facts), Comprehension (understanding of concepts), or Application (use of knowledge in new situations) [21].

Results

Table 1 presents the comparative evaluation results on the 116-example test set.

Table 1. Comparative Evaluation Results

Metric	Fine-tuned T5+LoRA	RAG (Mistral-7B)
Corpus BLEU (%)	**24.77**	3.02
ROUGE-L F1	**0.499**	0.265
BERTScore F1	**0.932**	0.885
Aviation Term Coverage (%)	43.97	**92.24**
Unique Bigram Ratio	0.326	**0.546**
Unique Questions (of 116)	56	**112**
Avg. Question Length (words)	4.4	**15.2**
Avg. Inference Time (sec)	**0.208**	1.043

The fine-tuned T5+LoRA model achieved Corpus BLEU of 24.77% versus 3.02% for RAG, ROUGE-L F1 of 0.499 versus 0.265, and BERTScore F1 of 0.932 versus 0.885. The average question length was 4.4 words for fine-tuned outputs versus 15.2 words for RAG, compared to 5.3 words in reference questions. Fig. 2 presents the comparison of metrics.

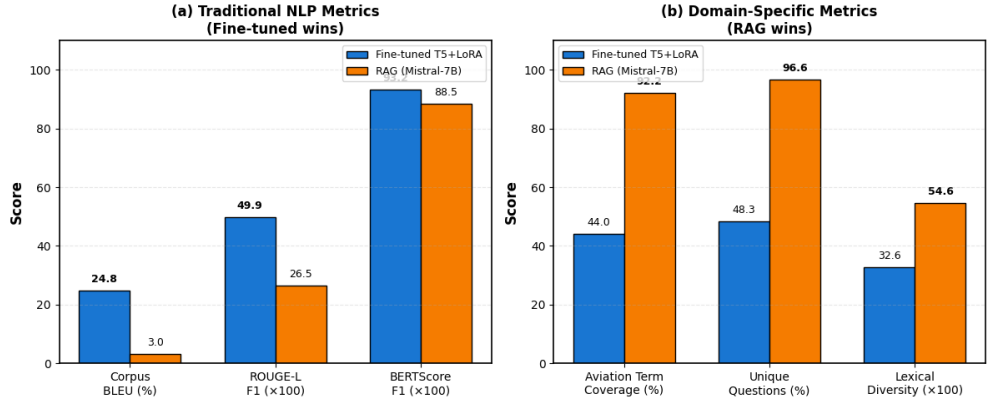

Fig. 2. Comparison of traditional NLP metrics (a) and domain-specific metrics (b)

The RAG approach achieved aviation terminology coverage of 92.24% versus 43.97% for fine-tuning. RAG achieved a unique bigram ratio of 0.546 compared to 0.326 for fine-tuning. RAG generated 112 unique questions out of 116 test samples (96.6%), while the fine-tuned model produced 56 unique questions (48.3%).

Analysis of question patterns showed that 73.3% of fine-tuned outputs followed the "What is X?" pattern, compared to 44.8% for RAG (Fig. 3).

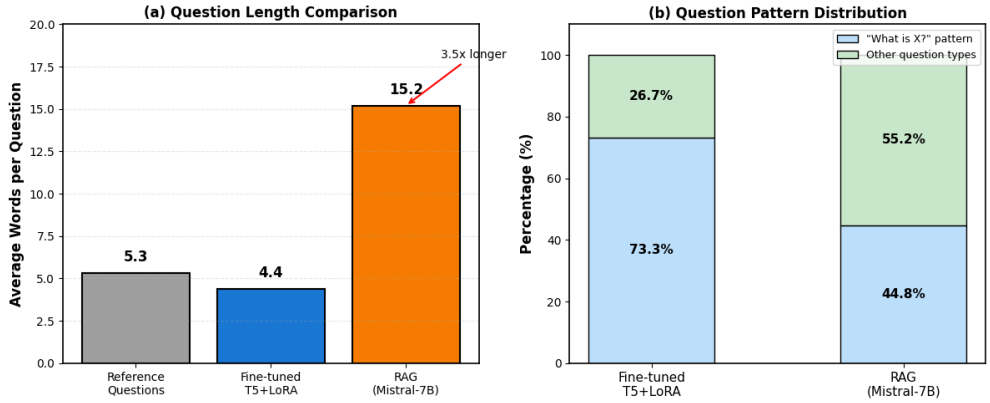

Fig. 3. Question length comparison (a) and question pattern distribution (b)

Table 2 presents sample-generated questions illustrating the differences between approaches.

Table 2. Sample Generated Questions

Context Topic	Fine-tuned T5	RAG (Mistral)
Maneuvering speed	What is VA?	What is the maximum speed at which full control deflection can cause structural damage?
Go-around	What is a go-around?	What action should be taken when a landing cannot be safely completed?
Four forces	What are the four forces?	What are the four forces acting on an aircraft, and what is each responsible for?
Transponder	What is a transponder?	What function requires a pilot to set the squawk code and select the appropriate mode?

Table 3 presents the results of a preliminary expert assessment conducted by the authors on a stratified sample of 15 questions per approach.

Table 3. Preliminary Expert Assessment Results (n=15 per approach)

Criterion	Fine-tuned T5+LoRA	RAG (Mistral-7B)
Grammatical Correctness	**5.00/5.00**	4.87/5.00
Context Relevance	3.47/5.00	**5.00/5.00**
Terminology Accuracy	2.33/5.00	**4.93/5.00**
Knowledge Level (%)	100.0%	6.7%
Comprehension Level (%)	0.0%	**53.3%**
Application Level (%)	0.0%	**40.0%**

According to the expert evaluation, all refined questions (100%) fell into the Knowledge level of Bloom's taxonomy, whereas RAG questions were split between the Comprehension (53.3%) and Application (40.0%) levels, with just 6.7% falling into the Knowledge level.

Discussion

Rather than determining which approach is superior, this study characterizes how different generation architectures produce questions with distinct cognitive and pedagogical properties. The substantial capacity difference between T5-small (60M) and Mistral-7B (7.3B) reflects practical deployment realities: educational institutions often cannot fine-tune large models, while RAG enables leveraging powerful models without domain-specific training. Our analysis, therefore, focuses on understanding the pedagogical trade-offs inherent to each architectural choice.

First, the model capacity asymmetry (60M vs 7.3B parameters) was a deliberate design choice reflecting real-world constraints: fine-tuning large models require computational resources unavailable to most educational institutions, while RAG democratizes access to large model capabilities. This study does not claim architectural superiority but characterizes how each approach shapes question properties.

The experimental results reveal a fundamental trade-off between traditional NLP metrics and domain-specific quality measures. The fine-tuned model's superior performance on BLEU, ROUGE, and BERTScore can be attributed to two factors: (1) its tendency to generate short questions that match the reference length (4.4 vs 5.3 words), and (2) the predominance of the "What is X?" template, which appears frequently in both generated and reference questions. However, this metric advantage masks important limitations in educational utility. Similar patterns were observed by Rodriguez-Torrealba et al., who noted that fine-tuned models tend to learn dominant question patterns from training data [8].

The dramatic difference in aviation terminology coverage (92.2% vs 44.0%) demonstrates RAG's ability to leverage retrieved context to incorporate domain-specific vocabulary. This finding aligns with Gao et al., who reported that RAG architectures excel at knowledge-intensive tasks requiring specialized terminology [10]. The advantage is particularly important for aviation education, where precise terminology is essential for safety communication. The fine-tuned model, constrained by its limited training corpus of 920 examples and small parameter count (60M), has not fully internalized the specialized vocabulary required for comprehensive question generation.

The diversity analysis reveals a critical limitation of the fine-tuned approach: it generates only 56 unique questions for 116 different contexts (48.3% uniqueness). This indicates significant pattern memorization, where the model defaults to generic question templates regardless of specific context content. RAG's 112 unique questions (96.6% uniqueness) demonstrate its ability to adapt question generation to specific input contexts, which is essential for creating varied assessment materials. This result is consistent with findings by Ling and Afzaal, who emphasized the importance of question diversity in educational applications [17].

Our findings align with Soudani et al., who demonstrated that RAG consistently outperforms fine-tuning for domain-specific factual knowledge, particularly when dealing with specialized terminology underrepresented in general training corpora [13].

The preliminary expert assessment provides critical insights into pedagogical quality that automated metrics cannot capture. While fine-tuned questions achieved perfect grammatical correctness (5.00/5.00), they scored significantly lower on context relevance (3.47/5.00) and terminology accuracy (2.33/5.00). Generic questions like "What is VA?" are grammatically correct but fail to test a deep understanding of maneuvering speed concepts. In contrast, RAG questions such as "What is the maximum speed at which full control deflection can be applied without risking structural damage?" directly assess comprehension of the underlying principle.

A thorough trade-off visualization across all evaluation dimensions is shown on Fig. 4.

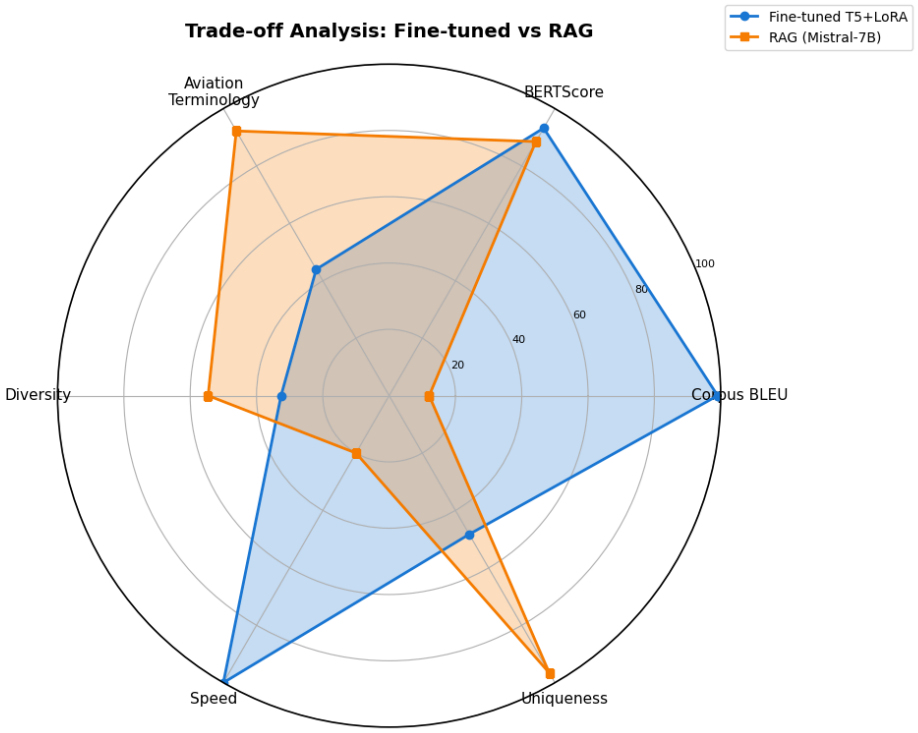

Fig. 4. Trade-off analysis radar chart comparing Fine-tuned T5+LoRA and RAG approaches

From a practical deployment perspective, the fine-tuned model offers a 5× inference speed advantage (0.21s vs 1.04s per question), making it suitable for real-time applications. However, RAG provides advantages in knowledge updatability, as new aviation regulations or procedures can be incorporated by updating the document store without model retraining, which is particularly valuable in the rapidly evolving aviation domain. Compared to our previous work [18], which achieved Corpus BLEU of 24.27%, the current fine-tuned model shows marginal improvement (24.77%), while the RAG approach offers a fundamentally different trade-off profile prioritizing pedagogical quality over metric similarity.

When interpreting these results, several limitations should be considered. First, the model capacity asymmetry (60M vs 7.3B parameters) was a deliberate design choice reflecting real-world constraints: fine-tuning large models require computational resources unavailable to most educational institutions, while RAG democratizes access to large model capabilities. This study does not claim architectural superiority but characterizes how each approach shapes question properties. Second, the fine-tuned model was trained on a relatively small dataset (920 samples), which may limit its ability to learn diverse question patterns beyond dominant templates. Third, the expert assessment is preliminary and was conducted by the authors; therefore, the pedagogical evaluation may be subject to assessor bias and should be interpreted cautiously. Fourth, automated metrics cannot fully capture educational quality, which ultimately requires broader human expert judgment.

Conclusion

This study presents a systematic comparison of RAG-based and fine-tuning approaches for automated question generation in civil aviation education, incorporating both automated metrics and preliminary expert assessment. Our findings reveal complementary strengths: fine-tuning achieves superior performance on traditional NLP metrics (Corpus BLEU 24.77%, BERTScore 0.932) and offers 5× faster inference, while RAG demonstrates dramatically better domain terminology coverage (92.2% vs 44.0%), generates substantially more diverse questions (112 vs 56 unique outputs), and produces pedagogically superior questions targeting higher cognitive levels (93.3% at Comprehension/Application level vs 0%).

For practical deployment, we recommend: (1) fine-tuning for applications requiring high throughput and consistent output format, such as automated quiz generation with fixed question templates; (2) RAG for applications prioritizing domain terminology accuracy and question diversity, such as comprehensive examination creation or adaptive learning systems; (3) hybrid approaches combining fine-tuned generation with retrieval-augmented post-processing for scenarios requiring both speed and domain coverage.

Acknowledgement

This research was conducted at the Academy of Civil Aviation, Kazakhstan, as part of the AI-SANA initiative for AI integration in educational processes.

References

[1] Maity, S., Deroy, A., & Sarkar, S. (2025). Can large language models meet the challenge of generating school-level questions? *Computers and Education: Artificial Intelligence*, 8, 100370. https://doi.org/10.1016/j.caeai.2025.100370

[2] Wang, H., Li, J., Wu, H., Hovy, E., & Sun, Y. (2022). Pre-trained language models and their applications. *Engineering*, 25, 51–65. https://doi.org/10.1016/j.eng.2022.04.024

[3] Howard, J., & Ruder, S. (2018). Universal language model fine-tuning for text classification. *Proceedings of the 56th Annual Meeting of the ACL*, 328–339. https://doi.org/10.18653/v1/P18-1031

[4] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems*, 33, 9459–9474. https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html

[5] ICAO. (2020). *Training development guide: Competency-based training methodology*. Doc 9941. International Civil Aviation Organization. https://www.icao.int/

[6] Du, X., Shao, J., & Cardie, C. (2017). Learning to ask: Neural question generation for reading comprehension. *Proceedings of the 55th Annual Meeting of the ACL*, 1342–1352. https://doi.org/10.18653/v1/P17-1123

[7] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140), 1–67. https://jmlr.org/papers/v21/20-074.html

[8] Rodriguez-Torrealba, R., Garcia-Lopez, E., & Garcia-Cabot, A. (2022). End-to-end generation of multiple-choice questions using text-to-text transfer transformer models. *Expert Systems with Applications*, 208, 118258. https://doi.org/10.1016/j.eswa.2022.118258

[9] Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. (2022). LoRA: Low-rank adaptation of large language models. *Proceedings of the ICLR*. https://doi.org/10.48550/arXiv.2106.09685

[10] Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., & Wang, H. (2024). Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*. https://doi.org/10.48550/arXiv.2312.10997

[11] Johnson, J., Douze, M., & Jégou, H. (2019). Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3), 535–547. https://doi.org/10.1109/TBDATA.2019.2921572

[12] Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. *Proceedings of EMNLP-IJCNLP*, 3982–3992. https://doi.org/10.18653/v1/D19-1410

[13] Soudani, H., Kanoulas, E., & Hasibi, F. (2024). Fine tuning vs. retrieval augmented generation for less popular knowledge. *SIGIR-AP 2024: Proceedings of the 2024 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region*, 12–22. https://doi.org/10.1145/3673791.3698415

[14] Wang, L., Chou, J., Tien, A., Zhou, X., & Baumgartner, D. (2024). AviationGPT: A large language model for the aviation domain. In *AIAA AVIATION FORUM AND ASCEND 2024* (p. 4250). https://doi.org/10.48550/arXiv.2311.17686

[15] Faraby, S. A., Adiwijaya, A., & Romadhony, A. (2023). Review on neural question generation for education purposes. *International Journal of Artificial Intelligence in Education*, 34(3), 1008–1045. https://doi.org/10.1007/s40593-023-00374-x

[16] Karabacak, M., Ozkara, B. B., Margetis, K., Wintermark, M., & Bisdas, S. (2023). The advent of generative language models in medical education. *JMIR Medical Education*, 9, e48163. https://doi.org/10.2196/48163

[17] Ling, J., & Afzaal, M. (2024). Automatic question-answer pairs generation using pre-trained large language models in higher education. *Computers and Education: Artificial Intelligence*, 6, 100252. https://doi.org/10.1016/j.caeai.2024.100252

[18] Tugambayeva, A., & Sakhipov, A. (2025). Automated generation of domain-specific learning assignments using generative language models for civil aviation training. *Vestnik AGAKAZ*, 4(39), 211–224. https://doi.org/10.53364/24138614_2025_39_4_16

[19] Sai, A. B., Tanber, A. K., & Khapra, M. M. (2022). A Survey of Evaluation Metrics Used for NLG Systems. *ACM Computing Surveys, 55*(2), 1–39. https://doi.org/10.1145/3485766

[20] Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2020). BERTScore: Evaluating text generation with BERT. *Proceedings of ICLR*. https://doi.org/10.48550/arXiv.1904.09675

[21] Larsen, T., Endo, B., Yee, A., Do, T., & Lo, S. (2022). Probing internal assumptions of the Revised Bloom's Taxonomy. *CBE Life Sciences Education, 21*(4), ar66. https://doi.org/10.1187/cbe.20-08-0170