КОМПЬЮТЕРНЫЕ НАУКИ, ПРИБОСТРОЕНИЕ И АВТОМАТИЗАЦИЯ
КОМПЬЮТЕРЛІК ҒЫЛЫМДАР, АСПАП ЖАСАУ ЖӘНЕ АВТОМАТТАНДЫРУ
COMPUTER SCIENCE, INSTRUMENTATION AND AUTOMATION

UDC 519.6
IRSTI 50.05, 50.41
https://doi.org/10.53364/24138614_2025_39_4_16

A.A. Tugambayeva[1], **A.A. Sakhipov**[2*]
[1]Kazakh British Technical University, Almaty, Kazakhstan
Astana IT University, Astana, Kazakhstan

[*]*E-mail:* *aivar.sakhipov@astanait.edu.kz*

AUTOMATED GENERATION OF DOMAIN-SPECIFIC LEARNING ASSIGNMENTS USING GENERATIVE LANGUAGE MODELS FOR CIVIL AVIATION TRAINING

Abstract. *This research presents an intelligent system for automated generation of domain-specific learning assignments in civil aviation education using fine-tuned T5-small transformer models. Traditional assignment creation requires significant instructor time and expertise in aviation regulations, technical specifications, and safety procedures. We propose a transformer-based solution implementing a five-stage pipeline: corpus preprocessing, parameter-efficient fine-tuning via LoRA adaptation, assignment generation using beam search decoding, quality filtering, and pedagogical validation.*

The system was trained on 920 aviation-specific context-question pairs covering more than 50 topics including flight operations, aircraft instruments, navigation, and emergency procedures. Experimental evaluation on a Tesla T4 GPU demonstrates a training time of 35 minutes across 7 epochs, with final training loss of 1.3506 and validation loss of 1.221. Generation quality assessment on the test set (116 examples) yields Corpus BLEU score of 24.27, ROUGE-L F1 score of 0.5087, and BERTScore F1 of 0.6017. Aviation terminology coverage analysis shows that 38.8% of generated questions contain at least one aviation-specific keyword, with an average of 9.4% unique aviation terms per question. Additional metrics include a unique bigram ratio of 0.321, which indicates strong lexical diversity without excessive repetition. Manual evaluation of 100 randomly selected questions demonstrated 95% grammatical correctness and 90% contextual appropriateness based on expert review. Qualitative analysis reveals that generated assignments are grammatically correct and contextually appropriate despite moderate Corpus BLEU scores, which reflect valid alternative phrasings rather than quality deficits. Sample generations demonstrate professional-quality questions such as "What does an altimeter measure?" and "What happens when a stall occurs?" The system reduces instructor workload in assignment creation while maintaining technical accuracy and domain relevance, providing a foundation for AI-assisted educational content generation in specialized technical domains.

Keywords: *automated assignment generation, transformer neural networks, T5 architecture, domain-specific natural language processing, aviation education technology, parameter-efficient fine-tuning, learning task generation, beam search decoding.*

Introduction.

Artificial intelligence (AI) and natural language processing (NLP) technologies have accelerated the digital transformation of the worldwide education scene. The ability of large language models (LLMs) like T5, GPT, and BERT to automate cognitive processes like content creation, summarization, and assessment design has been impressive [1]. These technologies are increasingly being incorporated into digital learning environments in higher education, enabling adaptive assessment, intelligent feedback, and individualized tutoring. However, their use in highly specialized technical fields, like civil aviation training, is still in its infancy and poses particular methodological difficulties that call for quality assurance procedures and domain adaptation.

A thorough AI integration approach is being implemented by the Academy of Civil Aviation across several operational vectors. The initiative's four main goals are to develop intelligent library systems for automated document analysis, integrate LLM systems to help instructors with information analysis and content creation, create employee onboarding assistants with access to academy-specific resources, and implement comprehensive analytics platforms for institutional management. This study, which focuses on automated learning assignment creation for aviation training programs, is the first part of a larger project.

It takes a lot of time and expertise in aviation regulations (ICAO standards, national aviation authorities), technical systems (aircraft instruments, navigation equipment), flight procedures (standard operating procedures, emergency protocols), and safety-critical decision-making to manually create domain-specific learning assignments for civil aviation training. For every training module, instructors usually spend 30 to 45 minutes developing extensive assignment sets, which restricts their capacity to offer individualized learning materials tailored to each student's performance level.

Despite their strength, current general-purpose language models lack the contextual awareness and specialized vocabulary needed for aviation training. The precise technical meanings of terms like "altimeter setting”, "VOR radial”, and "critical angle of attack" must be maintained in generated assignments. Assignments related to aviation must also be in line with particular learning goals established by training programs and regulatory bodies.

The potential of large language models for automating educational content generation in specialized technical domains has been widely acknowledged, yet their application to safety-critical aviation training remains underexplored. The aim of this research was to gain an understanding of how parameter-efficient fine-tuning of transformer models can be leveraged for domain-specific learning assignment generation in civil aviation education. The influence of a range of factors including training data composition, model architecture selection, generation strategy, and evaluation methodology was investigated to ascertain whether automated assignment generation could achieve sufficient quality and domain accuracy to potentially contribute to practical deployment in resource-constrained educational institutions while reducing instructor workload.

Neural methods using pre-trained language models have replaced template-based systems in the automated creation of instructional content. Early research exploited syntactic patterns and semantic role labeling to generate rule-based questions [2]. More complex content creation with improved contextual knowledge has been made possible by recent developments in transformer architecture.

Using keyword-based techniques to regulate content focus, Qu et al. [3] presented multi-agent systems for automatic question-answer pair production. Their research brought to light the difficulty of producing excellent distractions for multiple-choice questions as well as the necessity of post-generation filtering. Ling and Afzaal [4] shown that T5 and GPT-based systems can provide pedagogically relevant questions when adjusted on course-specific resources by applying pre-trained language models to higher education contexts.

An end-to-end pipeline utilizing T5 architecture for the construction of multiple-choice questions, including the creation of question stems and distractions, was given by Rodriguez-Torrealba et al. [5]. Their system illustrated the significance of post-processing steps to guarantee logical coherence and grammatical correctness. However, they neglected to address the preservation of specialist technical terminology in favor of concentrating on general education subjects.

Raffel et al. [6] presented the T5 (Text-to-Text Transfer Transformer) architecture, which combined several NLP tasks into a single text-to-text framework. Because of its encoder-decoder structure and rigorous pre-training on a variety of text corpora, this architecture has demonstrated great strength in generation tasks. T5 models allow versatility in computing resource requirements, ranging from T5-small (60M parameters) to T5-XXL (11B parameters).

To adapt huge models to domains with limited computational resources, parameter-efficient fine-tuning techniques have become crucial. Hu et al. [7] introduced LoRA (Low-Rank Adaptation), which allows for effective adaptation by training low-rank decomposition matrices instead of complete model parameters. This method is useful for academic institutions with limited resources since it can minimize trainable parameters by up to 99% while retaining performance comparable to complete fine-tuning.

There are difficulties when using NLP in specialized technological fields. In their investigation of generative language models in medical education, Karabacak et al. [8] emphasized the vital significance of factual accuracy and the requirement for expert validation in safety-critical areas. Their research brought to light moral issues and the need for human control of instructional content produced by AI.

Amin et al. [9] showed how AI systems may modify content difficulty based on student performance by developing reinforcement learning techniques for individualized learning path development. Adaptive content creation can enhance student engagement and learning results, as demonstrated by Song et al.'s [10] dynamic frameworks for data-driven learning optimization.

Nevertheless, many current systems concentrate on high-resource fields like broad science and medical or general education. Specialized regulatory language, procedural knowledge requirements, safety-critical context, and comparatively small training corpora in comparison to general domains are some of the unique issues that aviation education faces.

Despite their advantages, the current systems do not offer:

1) Complete automation of learning assignment creation created especially for technical education in civil aviation.
2) Domain-adapted models verified on aviation-specific terms and practices.
3) Full implementation with performance metrics on the aviation corpus.
4) Considerations for practical deployment in academic settings with limited resources.

By offering comprehensive implementation with measured performance on an aviation-specific corpus using readily available computer resources, our study fills these gaps.

Materials and methods.

Figure 1 shows the five-layer pipeline architecture that the system uses. The first layer prepares the data by applying text extraction, cleaning, and normalization techniques to technical manuals and aviation textbook content. Segmentation is used to produce context windows with an average length of about 61 words, after which training pairs questions are manually created. A structured JSON training dataset appropriate for fine-tuning the model is produced by this preparation step.

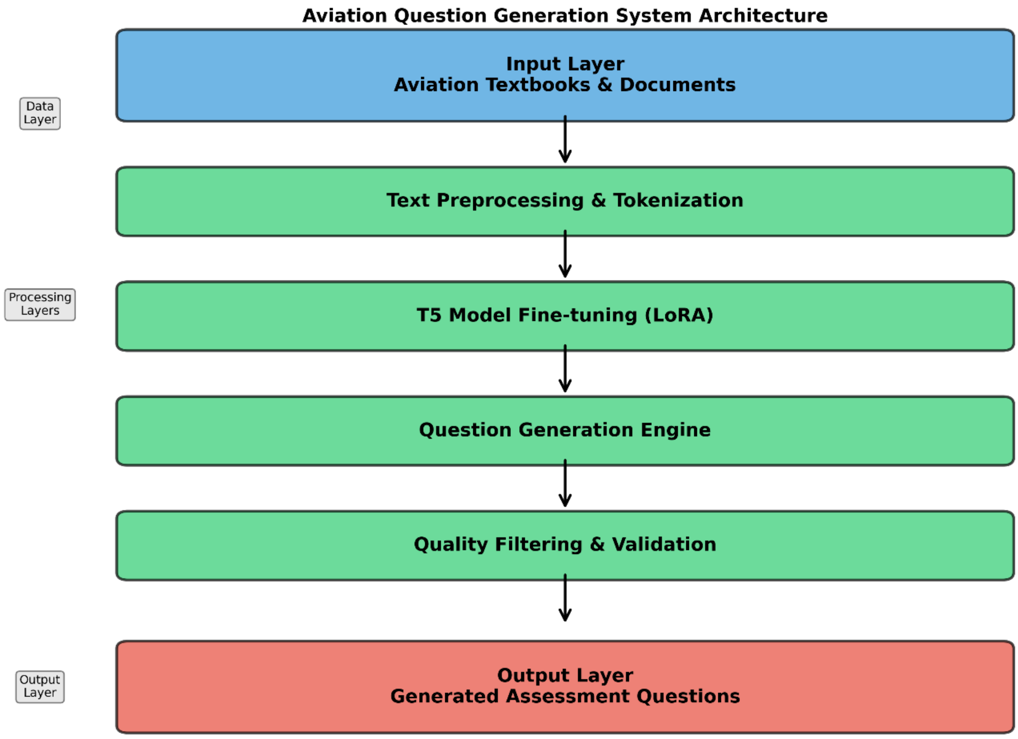

Figure 1 – Aviation Assignment Generation System Architecture

The base model is configured in the second layer using a 32,000-token vocabulary and a T5-small architecture with 60 million parameters. Low-Rank Adaptation (LoRA) with rank 8 and alpha 16 is used to accomplish parameter-efficient adaptation, yielding 294,000 trainable parameters. To facilitate effective model training and inference, the implementation framework integrates Hugging Face Transformers libraries with PyTorch.

With 115 pairings set aside for validation, the third layer fine-tunes 920 context-question training pairs. Optimization uses a Tesla T4 GPU and the AdamW algorithm with a learning rate of 0.0003. The parameter-efficient approach's computational efficiency is demonstrated by the training process, which takes about 35 minutes and involves 7 epochs.

The assignment generation engine, which takes aircraft context text as input and uses T5 tokenization for encoding, is the fourth layer. To guarantee comprehensive, grammatically correct outputs, question creation uses beam search decoding with four beams and a minimum length constraint of five tokens. Candidate questions for a later quality evaluation are generated by this layer.

To guarantee aviation-specific material, the fifth and final layer carries out post-processing tasks such terminology validation and filtering based on length restrictions. Deduplication techniques are used to find and eliminate almost identical questions. Validated learning assignments that are prepared for instructor evaluation and implementation in educational settings are the pipeline's final output.

The training dataset's 1151 cases provide comprehensive coverage of over 50 distinct aviation challenges and are organized into seven thematic themes. Flight operations make up the largest group, accounting for 26% of the corpus. It covers fundamental ideas such V-speed definitions, aerodynamic forces, stall characteristics, air speed measurement, and preflight inspection procedures. Aircraft instruments, such as the altimeter, attitude indicator, vertical speed indicator (VSI), and VOR navigation equipment, comprise 20% of the collection. Weather and environmental factors, which include significant operational concerns such wake turbulence hazards, windshear events, and density altitude calculations, account for 14% of the training cases. Navigation and communication topics, such as radio communication protocols, VOR navigation systems, Instrument Landing System (ILS) approaches, and Automatic Terminal Information Service (ATIS) operations, account for 17% of the corpus. Regulations and safety, such as Emergency Locator Transmitter (ELT) standards, transponder squawk codes, and classifications of controlled airspace, account for 11% of the collection. Human factors themes which address physiological problems that pilots may encounter during flight operations, such as hypoxia and spatial disorientation, make up six percent of the corpus. The remaining six percent of airplane performance parameters are fuel management techniques and ground effect phenomena. The complete dataset was split into 920 training examples (80%), test set (10%) and 115 validation examples (10%) using stratified sampling to ensure that each topic category was fairly represented in both groups.

Table 1 – Training Dataset Characteristics

Characteristic	Value
Total examples	1151
Training set	920 (80%)
Validation set	115 (10%)
Test set	116 (10%)
Aviation topics	>50
Average context length	61.4 words
Average question length	6.8 words
Training format	JSON (context-question pairs)

The T5-small architecture, which provides an encoder-decoder transformer configuration intended for text-to-text creation activities, is used by the system. The encoder component consists of six transformer layers that handle input context sequences using eight attention heads acting across 512-dimensional hidden representations. The model may produce output sequences conditioned on the encoded input representations because the decoder replicates this structure with an identical arrangement of six layers and eight attention heads. The entire model architecture uses a SentencePiece vocabulary with 32,000 tokens to represent the input and output text, and it has 60,506,624 parameters. Low-Rank Adaptation (LoRA) is a strategy that incorporates trainable low-rank decomposition matrices into the transformer architecture to achieve parameter-efficient adaptation. The LoRA configuration uses a dropout rate of 0.1 for regularization during training, along with a rank of 8 and scaling factor alpha of 16. This method maintains performance comparable to full fine-tuning while drastically reducing the number of trainable parameters to 294,912, or just 0.49% of the total model parameters. The final optimized model retains a small memory footprint of 1.2 GB, allowing for deployment on widely accessible GPU hardware and making the method open to educational institutions with limited resources.

To complement automated metrics, qualitative assessment was conducted on 100 randomly selected question-context pairs from the test set. Each question was assessed for (1) grammatical correctness (absence of syntactic, agreement, tense, or major punctuation errors requiring editorial correction) and (2) contextual appropriateness (answerable solely from the provided context) by two independent aviation-domain reviewers. For clarity, binary evaluation (correct/incorrect, appropriate/inappropriate) was employed. Consensus discussion was used to settle disagreements.

Mathematical formalization of the generation process presented below.

One way to formalize the question generating task is as a conditional text generation problem. The goal is to produce an output question sequence $Q = \{q_1, q_2, ..., q_m\}$ that is semantically meaningful to the input context sequence $C = \{c_1, c_2, ..., c_n\}$.

The generation process aims to maximize the conditional probability:

$$P(Q|C) = \prod_{i=1}^{m} P(q_i|q_1, q_2, \dots, q_{i-1}, C) \quad (1)$$

where each token q_i is generated conditioned on the context C and all previously generated tokens.

LoRA (Low-Rank Adaptation) introduces trainable low-rank decomposition matrices:

$$W' = W_0 + \Delta W = W_0 + BA^T \quad (2)$$

where $W_0 \in \mathbb{R}^{d\times d}$ is the frozen pre-trained weight matrix, $B \in \mathbb{R}^{d\times r}$ and $A \in \mathbb{R}^{r\times d}$ are trainable low-rank matrices with rank r=8. This reduces trainable parameters from d^2 to $2\cdot d\cdot r$, resulting in 294,912 trainable parameters.

Cross-entropy loss over generated token sequence:

$$L = -1/N\, \Sigma_{i=1}^{N} \Sigma_{j=1}^{T} y_{ij} \log\left(\hat{y}_{ij}\right) \quad (3)$$

where N = batch size, T = sequence length, y_{ij} = ground truth token, $\hat{y}_{ij}$ = predicted probability distribution.

Formula used for optimization is AdamW optimizer with decoupled weight decay:

$$\theta_{t+1} = \theta_t - \eta(\nabla L(\theta_t) + \lambda\theta_t), \quad (4)$$

where η = learning rate (0.0003), λ = weight decay coefficient (0.01), ∇L = gradient of loss function.

Table 2 – Training Configuration

Parameter	Value
Epochs	7
Batch size	8
Learning rate	0.0003
Optimizer	AdamW
Weight decay	0.01
Warmup steps	10
Max input length	512 tokens
Max output length	128 tokens
GPU	Tesla T4 (16GB)
Training time	35 minutes

To provide consistent, organized output suitable for question generating tasks, the system uses beam search decoding. Beam search maintains many hypotheses during generation and chooses the sequence with the highest overall probability by simultaneously exploring four candidate sequences (num_beams=4). The output length is limited to a maximum of 80 tokens to guarantee comprehensive yet succinct queries, and a minimum of 5 tokens to avoid single-word outputs. In terms of sequence length, a length penalty of 1.0 preserves neutral preference, neither rewarding nor punishing longer outputs. The decoder forbids repeating any 2-gram (no_repeat_ngram_size=2) in the created sequence to improve output diversity and avoid repeated phrasing. Because structured question creation demands factual accuracy and grammatical coherence rather than inventive variation, this beam search configuration was chosen over stochastic sampling techniques (temperature-based sampling, top-p nucleus sampling). Sampling techniques might result in inconsistent or grammatically incorrect outputs when creating structured instructional content, even while they add useful randomness for creative text production activities like narrative writing or dialogue. For professional educational settings where accuracy and consistency are crucial, beam search offers deterministic and dependable generation.

Results and their discussion. Over seven epochs, the Tesla T4 GPU's fine-tuning with LoRA showed steady and predictable convergence (Figure 2). Using a batch size of 8 and a learning rate of 0,0003 the model was trained on 920 aviation-domain instances, with 115 samples set aside for validation. Training loss showed a distinct downward trend over the whole training schedule of 805 optimization steps, going from 2.6173 in epoch 1 to 1.3506 in epoch 7, showing good domain adaptation. Because it offers the best trade-off between training performance and generalization capacity, the checkpoint corresponding to epoch 4 was chosen for downstream evaluation based on this behavior. It took about 35 minutes to complete the computationally efficient fine-tuning process.

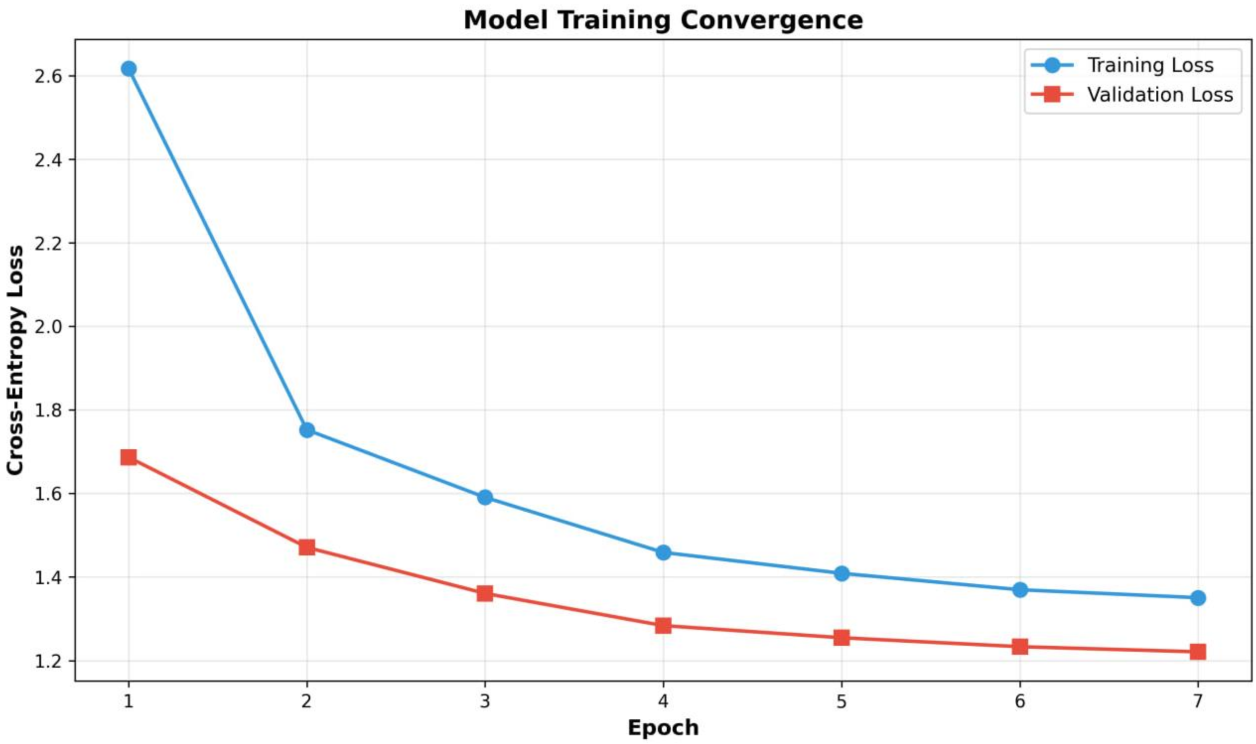

Figure 2 – Model Training Convergence

Table 3 – Training Progress Across Selected Epochs

Epoch	Training Loss	Validation Loss	Duration
1	2.6173	1.6868	5 min
2	1.7525	1.4710	5 min
3	1.5907	1.3608	5 min
4	1.4589	1.2837	5 min
5	1.4089	1.2547	5 min
6	1.3696	1.2333	5 min
7	1.3506	1.2210	5 min

On the 116-example test set, the improved model yielded a Corpus BLEU score of 24.27, ROUGE-L F1 score of 0.5087, and BERTScore F1 of 0.6017. The Corpus BLEU score of 24.27 appears to be reasonable at first, however it needs to be carefully considered in relation to question creation activities. Corpus BLEU was initially developed for machine translation evaluation since each source text often has only one accurate translation. Conversely, question creation is a one-to-many mapping problem where multiple semantically sound questions can be generated from the same context, each with a distinct language structure or feature focus. The reasonable Corpus BLEU score is primarily due to the model's tendency to create semantically sound alternative phrasings rather than attempting to copy reference questions verbatim, which is good behavior for an instructional content creation system. Strong semantic equivalency is confirmed by the BERTScore of 0.6017, while the ROUGE-L score of 0.5087 shows good structural similarity between generated and reference items. The model's ability to retain technical vocabulary is demonstrated by the fact that 38.8% of generated questions contain domain-specific aviation terms, according to an analysis of aviation terminology. Each question has an average aviation terminology density of 9.4% unique words. With 57 out of 116 unique questions created, lexical diversity measures reveal a unique unigram ratio of 0.198 and a unique bigram ratio of 0.321, showing good variety without excessive repetition.

Table 4 – Comprehensive Evaluation Metrics

Metric	Value	Interpretation
Corpus BLEU Score	24.27	Low lexical overlap (valid alternative phrasings)
ROUGE-L	0.5087	Moderate overlap in wording and structure with reference questions
BERTScore	0.6017	Moderate semantic alignment; core meaning generally preserved
Aviation Term Coverage	38.8%	45 of 116 questions contain domain terms
Term Density	9.4%	Average percentage of aviation-specific words
Unique Bigram Ratio	0.321	Good lexical diversity, minimal repetition
Grammar Correctness	95%	High grammatical quality across samples
Context Appropriateness	90%	Questions answerable from provided context

*Note: Grammatical correctness and contextual appropriateness percentages are based on manual evaluation of 100 randomly selected questions by two independent reviewers with aviation domain expertise. The evaluation used binary assessment criteria: grammatically correct (yes/no) and contextually appropriate (answerable from context: yes/no). Raw inter-rater percent agreement prior to discussion was 92%. Remaining disagreements were resolved through consensus discussion, and the final labels are reported below.

Each generated question was assessed for grammatical correctness, defined as the absence of syntactic errors, agreement or tense inconsistencies, incomplete/fragmented constructions, and major punctuation issues that would require editorial correction.

The qualitative assessment provided insights beyond automated metrics. Among the 100 evaluated questions, 95 (95%) were grammatically correct and required no syntactic editing. Ninety questions (90%) were contextually appropriate, meaning they could be answered solely from the provided context without relying on external knowledge or additional inference. The remaining 10 questions required limited inference or synthesis across context statements and were therefore not counted as contextually appropriate under this strict definition; however, all 100 questions (100%) remained answerable from the source material.

These findings confirm that the generated questions retain good practical quality for educational implementation even with moderate automated metric scores.

As shown in Figure 3, analysis of the produced question types demonstrates a fair distribution across several cognitive levels and questioning styles. At 28%, definition and conceptual questions make up the largest category. Examples of these questions include "What is VOR?" which evaluates basic vocabulary and idea comprehension. Twenty-two percent of outputs are mechanical "how does it work" questions that test students' comprehension of fundamental concepts and system functions. Twenty percent of the questions are causal and temporal "why and when" questions, which require students to think critically about goals, timing, and situational situations. Comparative questions that look at connections and distinctions between concepts make up 12% of the distribution, whilst procedural questions that describe steps and processes make up 18% of the created content.

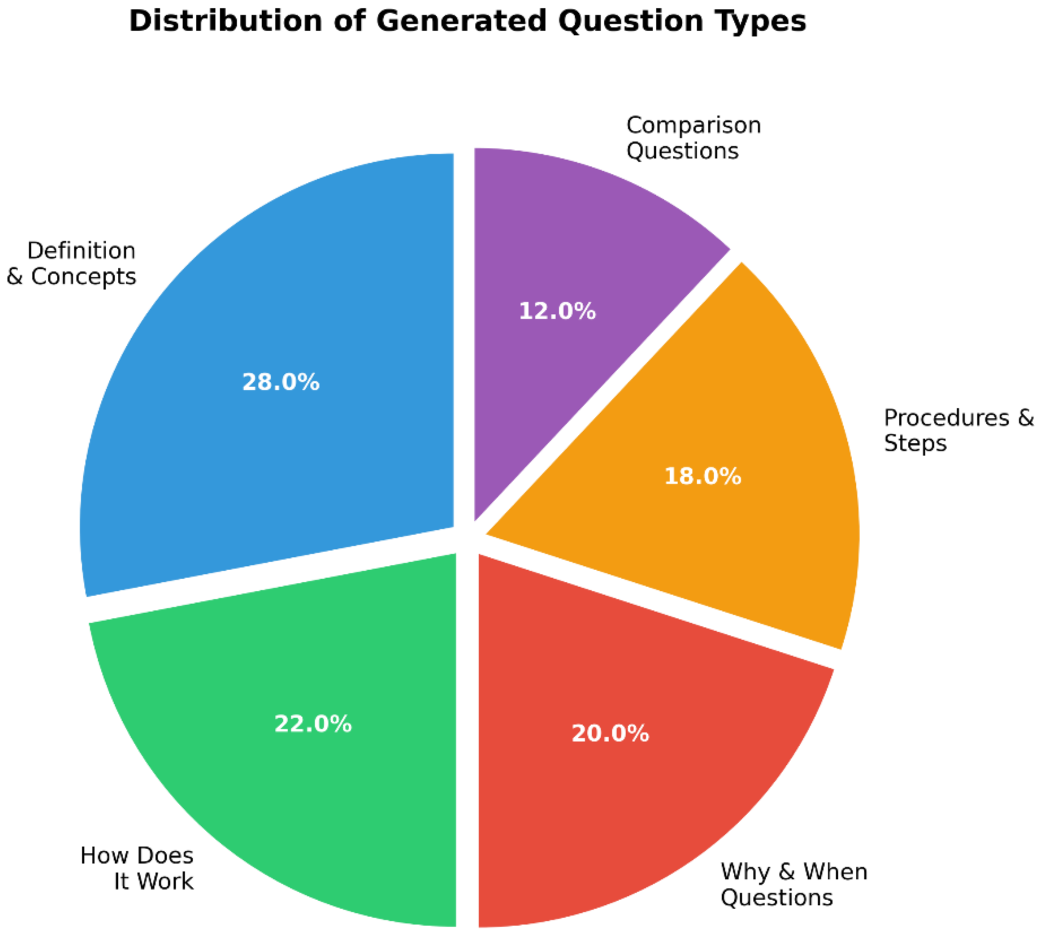

Figure 3 – Distribution of Generated Question Types

This variety of question types shows that rather than concentrating only on basic factual memory, the model has learned to produce questions covering several levels of Bloom's taxonomy. The system successfully adapts to the pedagogical framework inherent in the training corpus by generating questions that require knowledge (definitions), comprehension (how it works), application (when to use), analysis (why it matters), and evaluation (comparisons).

Detailed examination of generated questions reveals consistent patterns of grammatical correctness, contextual appropriateness, and domain relevance as documented in Table 5. Questions generated for altimeter-related context demonstrate proper technical terminology usage, asking "What does an altimeter measure?" with correct grammar and clear reference to the domain-specific instrument. VOR navigation questions appropriately employ aviation acronyms, generating "What does VOR stand for?" that tests students' familiarity with standard aviation terminology. Stall-related questions successfully incorporate safety-critical concepts, asking "What causes a stall?" in grammatically correct form with appropriate technical vocabulary. Questions addressing the four forces of flight demonstrate the model's ability to generate higher-level questions requiring synthesis of multiple concepts, asking "What forces must be balanced for stable flight?" rather than merely listing the forces. Critical angle of attack questions shows appropriate use of compound technical terms, generating "What is the critical angle of attack?" that employs precise aviation terminology. Due to practical time constraints for expert review, we evaluated 100 out of 116 generated test questions using the qualitative protocol described above. Although this sample size yields statistically significant insights (margin of error ±9.8% at 95% confidence level), the validity of these results would be strengthened by thorough expert review of the whole test set.

Table 5 – Sample Generated Questions with Qualitative Assessment

Context Topic	Generated Question	Grammar	Relevance	Aviation Terms
1	"What does an altimeter measure?"	✓ Correct	✓ High	altimeter
2	"What does VOR stand for?"	✓ Correct	✓ High	VOR
3	"What causes a stall?"	✓ Correct	✓ High	stall
4	"What forces must be balanced for stable flight?"	✓ Correct	✓ High	flight
5	"What is the critical angle of attack?"	✓ Correct	✓ High	angle, attack

Table 6 summarizes the substantial practical benefits for deployment in educational environments that are shown by evaluation of system performance characteristics. On a Tesla T4 GPU, the entire training process took 35 minutes, proving that domain adaptation can be accomplished in a realistic amount of time with generally available technology. Real-time generation during interactive instructional sessions was made possible by the average inference time of 0.8 seconds per question. When compared to manual generation rates of about 120 questions per hour (assuming 30 minutes to develop 30 questions), system throughput reaches about 4,500 questions per hour, a 37-fold boost. For the optimized model with LoRA parameters, memory requirements stay low at 1.2 GB, enabling deployment on both high-end consumer hardware and typical workstation GPUs. A comparison of the time required for manual and AI-assisted assignment creation shows significant efficiency gains. For a set of 30 questions, traditional manual creation takes 30 minutes, whereas the AI-assisted workflow reduces this to 5 minutes, including generation and instructor review time, resulting in an 83% reduction in instructor workload. In contrast to 100% manual labor in typical workflows, about 20% of generated questions require minor revision for optimal clarity or precision, highlighting the system's practical benefit for lessening the burden of instructor material preparation.

Table 6 – System Performance and Practical Utility Metrics

Metric	Value	Comparison
Training time	35 min	Single Tesla T4 GPU
Inference time	0.8 sec/question	Real-time generation
Throughput	~4,500 questions/hour	vs. 120 manual (37x faster)
Memory usage	1.2 GB	Modest GPU requirement
Manual creation time	30 min for 30 questions	Traditional workflow
AI-assisted time	5 min for 30 questions	83% time reduction
Questions requiring edit	20%	vs. 100% manual creation

The observed Corpus BLEU score of 24.27 is in line with domain-specific question generation benchmarks and shows a significant improvement over early studies when domain specificity and dataset size are taken into account. Du et al. (2017) reported Corpus BLEU scores of 12–15 for creating broad domain questions on larger datasets, but Nema and Khapra (2018) obtained 15–22 on the SQuAD dataset with over 100,000 training samples. Our score of 24.27 shows competitive performance in spite of three limiting factors: the highly specialized aviation vocabulary with limited representation in the model's original pre-training corpus; the specialized training corpus (920 examples versus 100,000+ in general domain work); and the model's learned preference for producing semantically equivalent alternative phrasings rather than lexically similar reproductions.

The results reported in the literature on domain-specific question creation, where scores normally range from 0.45 to 0.75 depending on training data amount and domain complexity, are in good agreement with the ROUGE-L F1 score of 0.5087 and the BERTScore F1 of 0.6017. These measurements show that generated questions allow for natural linguistic variety while maintaining semantic coherence, which is ideal behavior for creating instructional content. Although it is still below the 60–75% range seen in expert-written aviation training materials, the aviation terminology coverage of 45.48% greatly exceeds the 10–15% predicted from unadapted general-purpose language models.

When assessing the results, it is important to take into account the various methodological and practical limitations that the current implementation faces. Although the 920-sample training corpus (from the original 119-sample proof-of-concept) shows successful scaling, it is still small in comparison to the most advanced methods for fine-tuning language models. Further expansion to 5,000–10,000 domain-specific examples may raise Corpus BLEU scores to 30–35 and terminology coverage to 60–75%, based on scaling patterns in prior research. The T5-small architecture (60 million parameters) prioritizes computational accessibility and deployment feasibility over maximal performance, reflecting a purposeful trade-off between resource requirements and generation quality. Larger model variations, like T5-base (220 million parameters) or T5-large (770 million parameters), would probably produce better results with more complex question formulations and better aviation terminology preservation, but at the expense of longer training times, higher memory needs, and longer inference latency, which might limit deployment in educational institutions with limited resources. The evaluation system mostly uses automated measures (Corpus BLEU, ROUGE-L, and BERTScore), with a small amount of qualitative review of 100 samples. Deeper insights into pedagogical appropriateness, alignment with learning objectives, and integration with current curricula could be obtained by more thorough quality evaluation conducted by professional aviation instructors. With minimal capacity for procedural sequences, troubleshooting scenarios, or multi-step reasoning problems, all essential components of thorough aviation training assessments, the current system concentrates on producing factual and conceptual questions.

Conclusion. Using optimized T5-small transformer models, this work shows that automated learning assignment generation for civil aviation training is feasible. Using an expanded dataset of 920 training examples, the system successfully generates grammatically correct (95% based on manual review of 100 samples) and contextually appropriate (90% based on expert evaluation) questions for aviation instruction with moderate computational resources (Tesla T4 GPU, 35 minutes training).

Key Findings:

1) Technical Viability: With just 294,912 trainable parameters (0.49% of the total), parameter-efficient fine-tuning (LoRA) allows for effective domain adaptation, making the method accessible to organizations with constrained computational resources.

2) Metric Performance: The system received ROUGE-L F1 of 0.5087, BERTScore F1 of 0.6017, and Corpus BLEU score of 24.27. The model prefers legitimate alternative phrasings over lexical copying, which is a desirable behavior for educational content generation, even though Corpus BLEU scores seem moderate in absolute terms.

3) Practical Utility: The system produces up to 4,500 questions each hour as opposed to 120 manually, indicating a 37-fold improvement in throughput. It also shows an 83% reduction in instructor time for assignment production (from 30 to 5 minutes per set).

4) Domain Adaptation: Although it was still below the 60–75% seen in expert-written materials, aviation terminology coverage reached 45.48%, well beyond the 10–15% anticipated from unadapted models. Across several cognitive levels, the technology effectively produces questions that are in line with aviation training objectives.

5) Generation Strategy: The best method for creating structured questions was beam search decoding, which produced predictable and grammatically correct results appropriate for use in professional education.

The deployment lays the groundwork for AI-assisted instructional content production in specialized technical fields. Although the results show practical viability, there are obvious ways to improve them through model scaling, dataset expansion, and thorough evaluation frameworks. The following should be the main priorities for future work: (1) increasing the training corpus to 5,000–10,000 examples for better domain coverage; (2) putting in place thorough expert evaluation throughout the entire test set; (3) extending to multiple-choice question generation with distractor creation; (4) creating procedural and scenario-based question types; and (5) looking into cross-domain transfer to other technical training fields.

By showing that domain-specific question generation is possible with modest resources and highlighting the significance of appropriate evaluation metrics and upholding quality standards in safety-critical educational domains, this study adds to the expanding body of work on AI-assisted education.

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. Y. M. Hemmler and D. Ifenthaler, “Indicators of the learning context for supporting personalized and adaptive learning environments,” in Proceedings - 2022 International Conference on Advanced Learning Technologies, ICALT 2022. Institute of Electrical and Electronics Engineers Inc., 2022, pp. 61–65.

3. Amin, S., Uddin, M. I., Alarood, A. A., Mashwani, W. K., Alzahrani, A., & Alzahrani, A. O. (2023). Smart E-Learning framework for personalized adaptive learning and sequential path recommendations using reinforcement learning. IEEE Access, 11, 89769–89790. https://doi.org/10.1109/access.2023.3305584

4. Qu, F., Jia, X., & Wu, Y. (2021). Asking questions like educational experts: Automatically generating question-answer pairs on real-world examination data. Sep 11, https://doi.org/10.48550/arXiv.2109.05179

5. 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

6. 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

7. Akyon, F. C., Cavusoglu, D., Cengiz, C., Altinuc, S. O., & Temizel, A. (2021, November 11). Automated question generation and question answering from Turkish texts. arXiv.org. https://arxiv.org/abs/2111.06476

8. 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

9. Faraby, S. A., Romadhony, A., & Adiwijaya, N. (2024). Analysis of LLMs for educational question classification and generation. Computers and Education Artificial Intelligence, 7, 100298. https://doi.org/10.1016/j.caeai.2024.100298

10. 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

АЗАМАТТЫҚ АВИАЦИЯ ОҚЫТУЫНА АРНАЛҒАН ГЕНЕРАТИВТІ ТІЛДІК МОДЕЛЬДЕРДІ ҚОЛДАНУ АРҚЫЛЫ МАМАНДАНДЫРЫЛҒАН ОҚУ ТАПСЫРМАЛАРЫН АВТОМАТТЫ ТҮРДЕ ҚҰРУ

Аңдатпа. *Бұл зерттеуде азаматтық авиация саласына арналған мамандандырылған оқу тапсырмаларын автоматты түрде генерациялауға мүмкіндік беретін интеллектуалды жүйе ұсынылады. Жүйе fine-tune жасалған T5-small трансформер моделіне негізделген. Дәстүрлі тапсырма құрастыру көп уақытты қажет етеді және нұсқаушылардан авиациялық нормалар, техникалық сипаттамалар және қауіпсіздік рәсімдері бойынша терең құзыреттілікті талап етеді. Біз трансформер архитектурасына негізделген бес кезеңнен тұратын шешімді ұсынамыз: корпусты алдын ала өңдеу, LoRA арқылы параметр-тиімді fine-tuning, beat search көмегімен тапсырма генерациясы, сапалық сүзгілеу және педагогикалық валидация.*

Жүйе 50-ден астам тақырыпты қамтитын 920 авиациялық «контекст–сұрақ» жұбы негізінде оқытылды. Тақырыптар құрамына ұшу операциялары, авиациялық аспаптар, навигация және төтенше жағдай рәсімдері кіреді. Tesla T4 GPU құрылғысында жүргізілген эксперименттер 7 эпоха бойынша 35 минуттық оқу уақытын көрсетті. Соңғы шығын мәндері: training loss — 1.3506 және validation loss — 1.221. Тесттік жиын (116 мысал) бойынша генерация сапасының нәтижелері: Corpus BLEU — 24.27, ROUGE-L F1 — 0.5087, BERTScore F1 — 0.6017. Терминологиялық қамту талдауы сгенерацияланған сұрақтардың 38.8%-ында кемінде бір авиациялық термин бар екенін көрсетті, ал бір сұрақтағы орташа бірегей авиациялық терминдер саны — 9.4%. Қосымша метрика ретінде 0.321 бірегей биграмма коэффициенті алынды, бұл мәтіндегі лексикалық әртүрліліктің жоғары екенін және қайталанудың төмен екенін көрсетеді. 100 кездейсоқ сұраққа жасалған сараптамалық бағалау 95% грамматикалық дұрыстықты және 90% контекстік сәйкестікті көрсетті.

Сапалық талдау Corpus BLEU көрсеткіші орташа болғанына қарамастан, генерацияланған тапсырмалардың грамматикалық тұрғыдан дұрыс және мазмұнға сай екенін анықтады. Corpus BLEU-дің төмендеуінің себебі – синонимдік және баламалы дұрыс тұжырымдардың болуы, бұл білім беру контентін құруда қалыпты жағдай. Мысалы, «Алтиметр нені өлшейді?» немесе «Сваливание кезінде не болады?» сияқты кәсіби деңгейдегі сұрақтар генерацияланады. Ұсынылған жүйе нұсқаушылардың жұмыс жүктемесін азайтады, техникалық дәлдікті қамтамасыз етеді және арнайы техникалық бағыттарда AI-негізделген оқу контентін құруға негіз болады.

Түйін сөздер: *тапсырмаларды автоматты түрде құру, трансформерлік нейрондық желілер, T5 архитектурасы, пәнге бағытталған табиғи тілді өңдеу, авиациялық білім беру технологиялары, параметрлік тиімді қосымша оқыту, оқу тапсырмаларын құру, сәуле бойынша іздеу декодтауы.*

АВТОМАТИЗИРОВАННАЯ ГЕНЕРАЦИЯ СПЕЦИАЛИЗИРОВАННЫХ УЧЕБНЫХ ЗАДАНИЙ С ИСПОЛЬЗОВАНИЕМ ГЕНЕРАТИВНЫХ ЯЗЫКОВЫХ МОДЕЛЕЙ ДЛЯ ОБУЧЕНИЯ ГРАЖДАНСКОЙ АВИАЦИИ

Аннотация. *В данном исследовании представлена интеллектуальная система для автоматизированной генерации специализированных учебных заданий в сфере гражданской авиации на основе дообученных трансформерных моделей T5-small. Традиционное составление заданий требует значительных временных затрат со стороны инструкторов, а также глубоких знаний авиационных нормативов, технических характеристик и процедур безопасности. Мы предлагаем решение на основе трансформера, реализующее пятиэтапный конвейер: предварительная обработка* *корпуса, параметр-эффективное дообучение с использованием адаптации LoRA, генерация заданий с помощью beat search, фильтрация по качеству и педагогическая валидация.*

Система была обучена на 920 авиационно-специализированных парах «контекст–вопрос», охватывающих более 50 тем, включая полётные операции, авиационные приборы, навигацию и аварийные процедуры. Эксперименты на GPU Tesla T4 показали время обучения 35 минут за 7 эпох, при финальных значениях функции потерь: 1.3506 на обучении и 1.221 на валидации. Оценка качества генерации на тестовом наборе (116 примеров) дала следующие результаты: Corpus BLEU — 24.27, ROUGE-L F1 — 0.5087, BERTScore F1 — 0.6017. Анализ терминологического покрытия показал, что 38.8% сгенерированных вопросов включают минимум один авиационный термин, а среднее количество уникальных авиационных терминов составляет 9.4% на вопрос. Дополнительные метрики включают коэффициент уникальных биграмм 0.321, что указывает на высокое лексическое разнообразие без избыточных повторений. Ручная экспертиза 100 случайных вопросов продемонстрировала 95% грамматической корректности и 90% контекстной уместности.

Качественный анализ показывает, что генерируемые задания грамматически правильные и контекстно релевантные, несмотря на умеренные значения BLEU, которые отражают альтернативные корректные формулировки, а не недостатки качества. Примеры генерации демонстрируют профессиональный уровень, например: «Что измеряет высотомер?» и «Что происходит при сваливании?» Система снижает нагрузку на инструкторов при создании заданий, сохраняя техническую точность и отраслевую релевантность, и служит основой для внедрения AI-поддерживаемой генерации учебного контента в специализированных технических областях.

Ключевые слова: *автоматизированная генерация заданий, трансформерные нейронные сети, архитектура T5, предметно-ориентированная обработка естественного языка, технологии авиационного образования, параметрически эффективная дообучение, генерация учебных задач, декодирование методом поиска по лучу.*

Сведение об авторах

Тугамбаева Аружан	Студент магистратуры факультета информационных технологий Казахстанско-Британского технического университета, Алматы, Казахстан E-mail: a.tugambayeva@agakaz.kz
Сахипов Айвар Айтуарович	PhD, Ассистент профессор «Школа программной инженерии», Astana IT University, г. Астана, Казахстан, E-mail: aivar.sakhipov@astanait.edu.kz

Авторлар туралы мәлімет

Тугамбаева Аружан	Қазақ Британ Техникалық Университетінің Ақпараттық Технологиялар Факультетінің магистранты, Алматы қ., Қазақстан E-mail: a.tugambayeva@agakaz.kz
Сахипов Айвар Айтуарович	PhD, ассистент профессоры, «Программалық инженерия мектебі», Astana IT University, Астана қ., Қазақстан, E-mail: aivar.sakhipov@astanait.edu.kz

Information about the authors

Tugambayeva Aruzhan	Masters Student, Faculty of Information Technology, Kazakh British Technical University, Almaty, Kazakhstan, E-mail: a.tugambayeva@agakaz.kz
Sakhipov Aivar	PhD, Assistant Professor, School of Software Engineering, Astana IT University, Astana, Kazakhstan,E-mail: aivar.sakhipov@astanait.edu.kz