A Multimodal and Dynamically Updatable Benchmark for Aviation Question Answering with Large Language Models

Liu He[1,2,*], Shuyan Liu[1], Xiaorui Qin[1], Ran An[1] and Jianghui Zeng[1]

[1]*Department of Standard and Data Technology Research, China Aero-Polytechnology Establishment, Beijing 100028, China*

[2]*School of Computer Science, Fudan University, Shanghai 200438, China*

Abstract: With the rapid advancement of artificial intelligence, large-scale language models (LLMs) have demonstrated strong capabilities in open-domain question answering, knowledge retrieval, and decision support. However, in safety-critical and knowledge-intensive industries such as aviation, existing evaluation benchmarks fall short in domain adaptation, comprehensiveness, and dynamic updating. As aviation increasingly integrates intelligent automation and robotic systems for maintenance, inspection, and manufacturing, reliable language-model evaluation becomes crucial for ensuring the safety and autonomy of such systems. This paper proposes a multimodal, multi-level benchmark dataset tailored to aviation QA tasks, alongside an automated updating mechanism and a multi-dimensional evaluation framework. The methodology integrates knowledge extraction from multimodal aviation documents, diverse QA pair generation, iterative complexity enhancement, and quality validation. Furthermore, dynamic updating is achieved via a hybrid strategy combining imitation and expansion, complemented by differentiated filtering and prompt optimization. To ensure rigorous assessment, a ten-dimension evaluation framework is introduced, covering accuracy, completeness, relevance, explainability, and safety, among others. By providing a reliable and dynamically evolvable benchmark, this work supports the integration of LLMs into robotic and automated decision-support systems in aviation, enabling more intelligent, autonomous, and safety-assured operations. Experimental results using aviation textbooks confirm the effectiveness of the proposed approach in generating high-quality, dynamically evolvable QA datasets. This work provides both methodological innovation and practical tools for the evaluation of LLMs in aviation, with potential extension to other knowledge-intensive domains.

Keywords: Large language models, Multimodal dataset generation, Dynamic dataset updating, Multi-dimensional evaluation framework.

1. INTRODUCTION

With the rapid advancement of artificial intelligence technologies, large-scale language models (LLMs) have demonstrated remarkable capabilities in open-domain question answering, knowledge retrieval, and decision support [1]. In the aviation industry—an inherently knowledge-intensive and safety-critical domain—the demand for intelligent question answering systems emphasizes higher precision and reliability. Modern aviation systems are becoming increasingly automated, with robotic technologies applied in manufacturing, inspection, and maintenance processes. These automation trends highlight the need for LLM-based systems that can understand and reason over technical documentation, assisting robotic control, workflow automation, and fault diagnosis. In scenarios such as process design optimization, material selection assistance, and efficient retrieval of tool and equipment information, models that can deliver accurate and timely responses to domain-specific queries play a crucial role in enhancing research and development efficiency while ensuring operational safety [2]. However, the scientific evaluation of these models requires high-quality, aviation-oriented benchmark datasets.

Existing evaluation approaches largely rely on two types of resources: domain-specific datasets manually constructed through expert annotation, which are costly and limited in coverage [3]; and general-purpose benchmarks such as MMLU or ARC [4], which, despite widespread adoption, fail to capture the unique knowledge structures and application requirements of the aviation domain. Consequently, evaluation results based on such benchmarks often lack credibility in specialized contexts, making it difficult to faithfully reflect model performance in aviation-specific question answering tasks. Moreover, most current evaluation datasets remain static and text-centric, overlooking the multimodal information embedded in aircraft design manuals, aviation textbooks, technical standards, and operational handbooks [5]. This limits the comprehensiveness and realism of performance assessment.

To address the practical demands of the aviation sector, there is an urgent need to construct a benchmark dataset that not only covers the specialized knowledge system but also evolves dynamically with model development and domain knowledge updates. From the perspective of robotics and automation, such a benchmark will also provide an essential foundation

*Address correspondence to this author at the Department of Standard and Data Technology Research, China Aero-Polytechnology Establishment, Beijing 100028, China;
E-mail: heliu1219@126.com

Received on 08-09-2025 Accepted on 24-10-2025 Published on 29-10-2025

E-ISSN: 2409-9694/25

for developing intelligent assistants that support autonomous decision-making and information retrieval within automated aviation workflows. Such a dataset should meet three core requirements: (1) the ability to automatically parse and extract multimodal materials (*e.g.*, diagrams, schematics, and regulatory texts) to ensure comprehensive knowledge coverage; (2) the generation of diverse question–answer pairs across multiple cognitive levels and question types to enable a holistic assessment of model capabilities; and (3) an automatic updating and quality-control mechanism to prevent obsolescence or error propagation, thereby ensuring long-term validity and reliability.

In response to these challenges, this study proposes a multimodal LLM-based method for the generation and automatic updating of a multi-level aviation QA dataset. Specifically, we introduce a multimodal knowledge extraction method that automatically parses structured knowledge points from aviation textbooks and professional examination papers; design a knowledge-constrained QA generation mechanism to produce diverse high-quality QA pairs spanning different types and cognitive levels; establish dynamic updating and quality-control strategies to ensure iterative reliability; and present the first aviation-domain multimodal QA benchmark supporting automated updating.

The main contributions of this work can be summarized as follows:

1. **Multimodal and domain-specific dataset construction:** We propose the first benchmark for aviation QA that integrates multimodal sources (*e.g.*, textbooks, schematics, procedural documents) and generates diverse QA pairs across multiple cognitive levels and question types.

2. **Dynamic updating mechanism:** A novel hybrid strategy combining imitation and Bloom ' s taxonomy-based expansion is introduced, together with differentiated filtering and prompt optimization, enabling continuous dataset evolution with high reliability.

3. **Multi-dimensional evaluation framework:** We design a comprehensive ten-dimension evaluation system — including accuracy, completeness, relevance, explainability, and safety — that overcomes the limitations of single-metric benchmarks and provides fine-grained insights.

4. **Automation-oriented application value:** The proposed benchmark facilitates the deployment of LLMs within aviation's intelligent automation and robotic systems—supporting tasks such as maintenance assistance, automated inspection reasoning, and control decision support—and offers a transferable methodology for other knowledge-intensive and automation-driven domains such as healthcare, law, and energy.

2. RELATED WORK

2.1. General Question Answering Benchmarks

Open-source benchmarks such as MMLU [6], HellaSwag [7], BIG-bench [8], ARC [9], and KoLA [10] have been widely used to evaluate LLMs across knowledge and reasoning tasks. While valuable for standardizing evaluation, they remain static, domain-agnostic, and unable to capture aviation-specific requirements such as numerical reasoning, procedural knowledge, and safety compliance.

These benchmarks have been instrumental in standardizing model evaluation. However, their limitations are equally evident: (1) most are static datasets lacking mechanisms for dynamic updating, thus failing to keep pace with rapid LLM development; (2) the evaluation dimensions primarily emphasize general knowledge and reasoning, offering little adaptation to specialized domains such as aviation; and (3) they do not cover essential domain-specific aspects such as numerical computation, procedural operations, and safety-critical requirements. As a result, these general-purpose benchmarks cannot adequately reflect model performance in aviation-oriented intelligent QA scenarios, highlighting the urgent need for a domain-specific evaluation framework.

2.2 Multimodal Dataset Construction

The rise of Visual Language Models has motivated multimodal datasets such as COCO Captions [11], VQA [12], TabFact [13], Table QA [14], and Science QA [15]. These resources advance cross-modal reasoning evaluation but mostly focus on general or educational domains. They are static, lack mechanisms for dynamic updating, and do not cover aviation-specific multimodal content such as schematics, standards, or operational manuals.

2.3 Domain-Specific Evaluation Methods

Specialized benchmarks such as MedQA [16], PubMedQA [17], LegalBench [18], and finance. However, aviation lacks a systematic, standardized benchmark. Existing efforts are fragmented, costly, and limited in scalability, while failing to capture aviation's multimodal and safety-critical characteristics. This

highlights the need for a dynamic, multimodal benchmark tailored to aviation QA.

By contrast, the aviation domain still lacks a systematic and standardized evaluation benchmark. Existing efforts are fragmented, often relying on costly expert annotations with limited scalability and without dynamic updating mechanisms. Moreover, aviation scenarios inherently involve multimodal resources (*e.g.*, textbook illustrations, equipment schematics, and procedural tables) and multi-dimensional competencies (*e.g.*, safety compliance, numerical computation, complex reasoning), which existing domain benchmarks fail to capture. Thus, the construction of a multimodal, dynamically updatable aviation-specific QA dataset is both a necessary step for deploying intelligent QA systems in aviation and an important gap in current research.

2.4. AI in Robotics and Automation

Recent advances in artificial intelligence have profoundly shaped robotics and automation. Large language and vision–language models are increasingly integrated into robotic systems for task planning, semantic understanding, and human–robot interaction. EmbodiedQA situates visual question answering in embodied settings, coupling perception with navigation and action [19]. RoboVQA targets multimodal, long-horizon reasoning for robotics, requiring temporal and causal inference over visual observations and textual instructions [20]. RoboMM introduces an all-in-one multimodal large model tailored to robotic manipulation and control, together with standardized evaluation protocols for manipulation-oriented reasoning [21]. Complementing these efforts, Balcı *et al*. benchmark LLM reasoning in indoor robot navigation, providing task-specific protocols and metrics for assessing planning and decision-making in realistic layouts [22].

These efforts collectively highlight the trend toward evaluating AI models not only for linguistic accuracy but also for their ability to support autonomous perception, reasoning, and decision-making in real-world systems. The proposed aviation QA benchmark aligns with this direction by introducing a structured, domain-specific evaluation paradigm that parallels emerging robotic evaluation frameworks. In particular, the integration of multimodal documents (*e.g.*, schematics, manuals, and operational standards) mirrors the needs of robotic systems that must interpret procedural and technical knowledge for automated maintenance, inspection, and control. Thus, this work complements and extends existing AI evaluation practices in robotics and automation by focusing on the aviation domain's knowledge-driven and safety-critical characteristics.

3. METHOD

3.1. Overall Framework

To systematically address the challenges of domain adaptation, lack of dynamic updating, and high costs associated with manual dataset construction in aviation-specific question answering (QA) tasks, we propose a multimodal LLM-based framework for multi-level dataset generation and automatic updating. As illustrated in Figure **1**, the framework consists of three major modules: (1) multi-category QA dataset generation, (2) dataset automatic updating and expansion, and (3) multi-dimensional model evaluation.

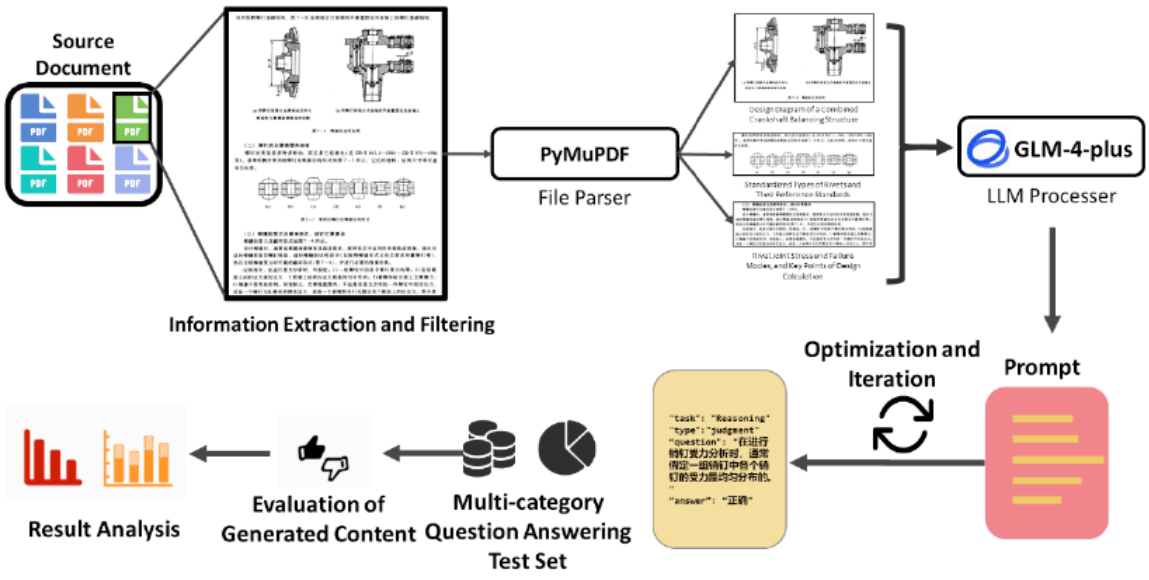


Figure 1: Domain evaluation data generation pipeline diagram.

First, in the dataset generation module, we leverage multimodal knowledge extraction and LLM-based QA generation mechanisms applied to aviation textbooks and related sources. The goal is to construct a high-quality dataset covering different cognitive levels (*e.g.*, comprehension, reasoning) and diverse question types (*e.g.*, multiple-choice, Completion, true/false, calculation, scenario analysis, and open-ended questions). An iterative complexity strategy is incorporated to enrich diversity and difficulty, ensuring that the dataset captures both domain knowledge mastery and advanced reasoning ability.

Second, in the dataset updating module, we design a hybrid strategy combining imitation and expansion. The imitation strategy generates new samples by preserving the style and knowledge points of existing QA pairs, while the expansion strategy, grounded in Bloom's taxonomy, produces QA pairs of varying difficulty. Furthermore, the dataset is annotated with sub-domain tags such as numerical computation, system principles, procedural rules, and scenario handling, ensuring adaptability to evolving domain knowledge and long-term usability.

Finally, the evaluation module introduces a comprehensive ten-dimension framework to assess model performance: accuracy, completeness, relevance, information richness, explainability, question-type specific correctness, reference conformance, usefulness, timeliness, and safety. Driven by an automated evaluation process, this system overcomes the limitations of single-metric accuracy and provides nuanced insights for capability diagnosis and iterative model improvement.

In summary, the proposed framework establishes a closed-loop process encompassing dataset generation, dynamic updating, and benchmark evaluation, enabling efficient construction of aviation QA datasets while ensuring reliability and adaptability in the face of knowledge evolution and model iteration.

3.2. Multi-Category QA Dataset Generation

The dataset generation phase is central to ensuring coverage, quality, and diversity. To this end, we design a pipeline consisting of knowledge extraction, QA pair generation, complex question construction, and quality validation, enabling efficient creation of multi-category, multi-level QA datasets from multimodal aviation materials.

3.2.1. Knowledge Extraction

Knowledge extraction forms the foundation of dataset construction, aiming to identify high-value information units from domain-specific documents such as aviation textbooks and standards. Unlike text-only extraction, we employ a multimodal large language model (LLM) to process complete document pages—integrating text, diagrams, tables, and formulas as unified inputs.

Specifically, each document page is treated as a single semantic unit and fed to the multimodal LLM in its entirety (page image with embedded text). Rather than fragmenting the page into isolated elements, the model jointly attends to layout structure, figures, tables, formulas, and captions, leveraging spatial arrangement and cross-modal cues to build a unified page-level representation. Lightweight signals (*e.g.*, OCR/layout tags) are used only as soft anchors to guide attention, not as hard segmentation. This holistic encoding enables the model to infer relational and procedural knowledge that emerges from the co-occurrence, ordering, and layout of elements across the page—information that is often lost in element-wise pipelines.

For example, a wiring diagram page may yield knowledge points such as "current-limiting resistor placement in EWIS harness design" or "safety margin for aluminum conductor connections," derived from both diagram annotations and accompanying text. Similarly, tables describing material properties are parsed into comparative knowledge statements, and equations within manuals are transformed into parametric rules (*e.g.*, tolerance constraints, load–temperature relationships).

Four extraction principles are followed:

1. Completeness – ensuring inclusion of essential details, examples, and data;
2. Accuracy – strictly adhering to original content without introducing external knowledge;
3. Conciseness – avoiding redundancy and low-value information;
4. Structured output – representing knowledge points in formats suitable for downstream use.

To further enhance reliability, each extracted knowledge point is assigned a confidence score ranging from 0 to 10, reflecting its perceived value. Only knowledge points above a threshold (*e.g.*, ≥7) are retained, ensuring both quality and trustworthiness of the resulting knowledge base.

3.2.2. QA Pair Generation

Following extraction, QA pairs are generated using domain-specific prompts with LLMs. To evaluate comprehension and reasoning capabilities, we distinguish between comprehension questions (*e.g.*, multiple-choice, Completion, short-answer) and reasoning questions (*e.g.*, true/false, calculation, scenario analysis). For transparency, each QA pair

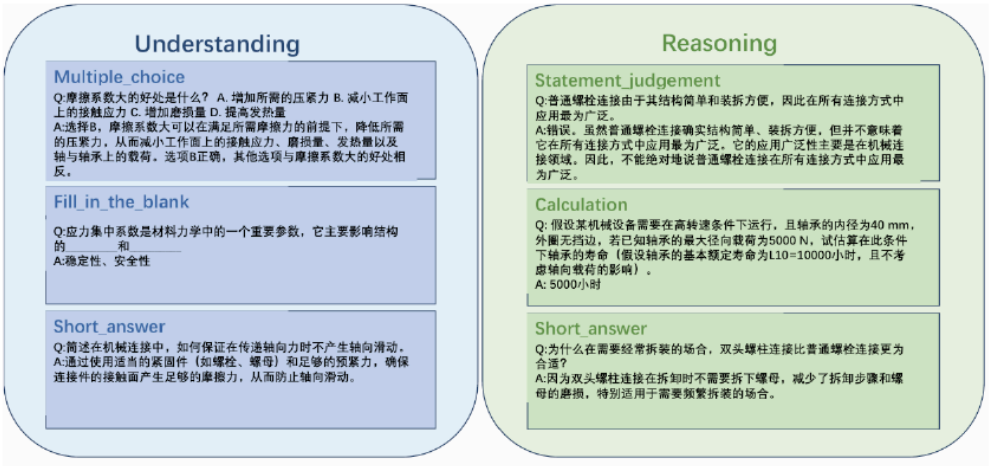


Figure 2: Examples of multi-category, multi-level questions: understanding-type questions are generally simple and straightforward, whereas reasoning-type questions require further inference grounded in domain knowledge.

includes not only the question and reference answer but also a detailed reasoning process (analysis), allowing traceability of logic and enhancing interpretability. The resulting dataset thus spans multiple question types and cognitive levels, offering a holistic representation of aviation knowledge.

3.2.3. Complex Question Construction

To better evaluate higher-order reasoning, we introduce an iterative complexity-enhancement strategy. Based on predefined dimensions—constraint addition, issue deepening, concept refinement, and multi-step reasoning—existing QA pairs are transformed into more challenging forms by adding constraints, refining contexts, or extending reasoning chains. This significantly improves both difficulty and diversity, enabling assessment of advanced reasoning and integrative capabilities within aviation contexts.

3.2.4. Quality Validation

To guarantee dataset quality, we employ an automated validation mechanism in which LLMs score generated QA pairs on a 0–10 scale, based on accuracy, answer–question alignment, and consistency with original knowledge points. Experiments indicate that setting the threshold at 8 effectively filters out low-quality pairs while retaining ~93% of high-quality data. The resulting dataset achieves both semantic coherence and wide question-type coverage, providing a reliable foundation for dynamic updating and evaluation.

3.3. Dataset Automatic Updating

To maintain timeliness and domain adaptation, we design an automatic updating module composed of generation strategies, filtering mechanisms, and prompt optimization.

3.3.1. Generation Strategies

As Figure **3** illustrated, two strategies are combined as follows.

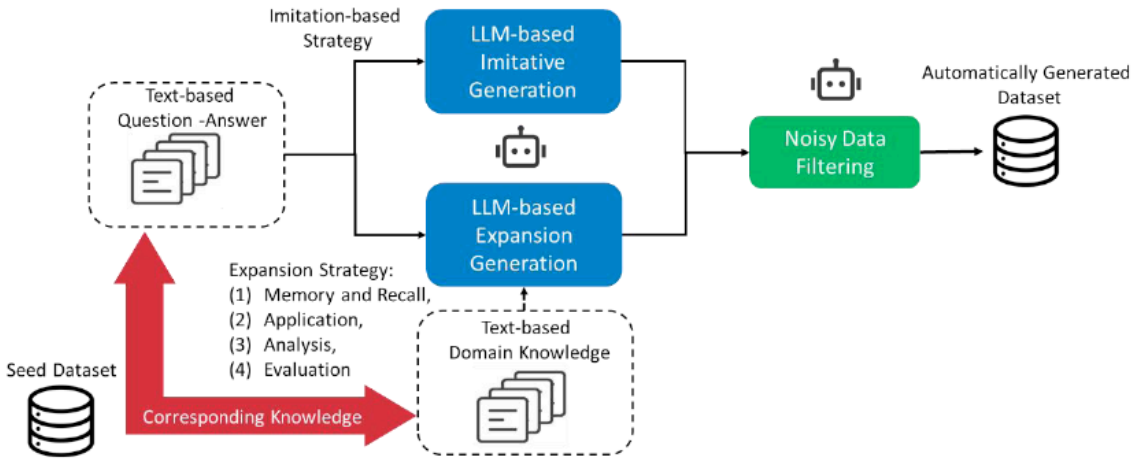


Figure 3: Automated update strategy for the evaluation dataset.

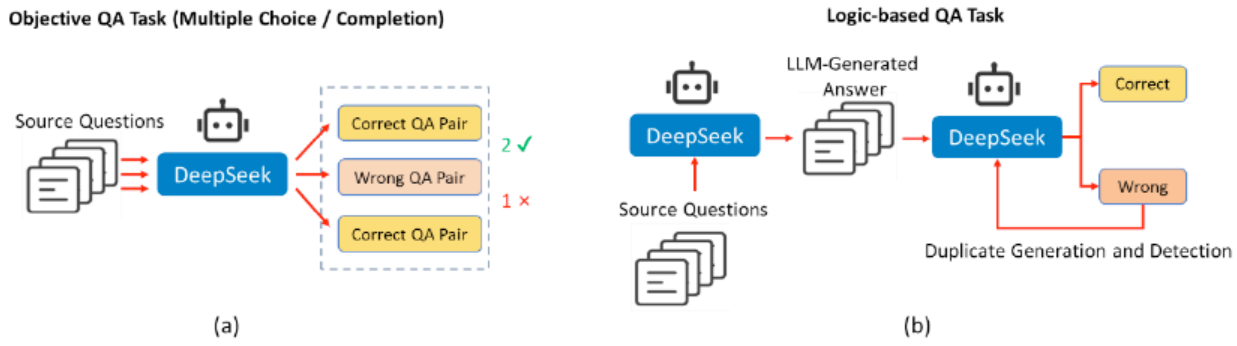


Figure 4: Data Filtering Strategy. (**a**) For objective QA (MC/fill-in), we use majority voting over multiple LLM generations to reduce randomness. (**b**) For reasoning QA (short-answer), we apply LLM self-evaluation to improve reliability given the lack of mature process-oriented metrics.

Imitation: Generates alternative QA pairs by preserving original style and knowledge points, improving redundancy and diversity.

Expansion: Based on Bloom's taxonomy, generates questions across progressive cognitive levels (memory → application → analysis → evaluation). Sub-domain labels (*e.g.*, numerical computation, system principles, procedural rules, scenario handling) are incorporated to ensure domain relevance.

3.3.2. Filtering Mechanisms

Ensuring correctness and consistency is critical. We adopt a differentiated filtering mechanism.

For objective questions (multiple-choice, true/false, Completion), multiple independent generations are compared through majority voting(in Figure **4** (**a**)).

For subjective questions (short-answer, calculation, scenario analysis), LLM self-assessment is combined with human sampling-based review for logical consistency and reasoning validity(in Figure **4** (**b**)). Specifically, approximately 15–20% of the generated subjective QA pairs are randomly selected for manual inspection by domain experts with aviation and engineering backgrounds. Each sampled pair is independently reviewed by at least two annotators to verify factual correctness, reasoning coherence, and adherence to reference materials. Disagreements are resolved through discussion or third-party arbitration. The resulting feedback is also used to refine the LLM's self-assessment criteria and filtering thresholds.

This hybrid filtering strategy substantially enhances reliability and prevents error propagation.

3.3.3. Prompt Optimization

Given the central role of prompts, we employ an iterative optimization cycle: candidate prompts are tested with repeated generations (*e.g.*, 100 trials) to measure accuracy and error rates. Prompts with superior performance are further refined through manual review. Results show that optimized prompts improve first-pass generation accuracy to over 95%, significantly boosting efficiency and robustness.

Overall, the automatic updating method forms a closed loop combining generation, validation, and optimization. This ensures timeliness, scalability, and long-term applicability of the aviation QA dataset, while supporting benchmark evolution in line with domain knowledge updates.

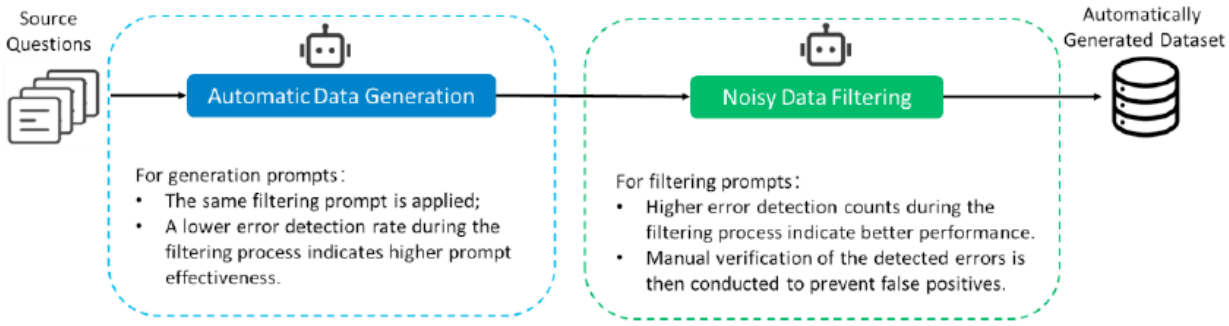


Figure 5: Prompt Optimization Strategy. We optimize prompts by combining accuracy-driven tuning with human proofreading; for each optimized prompt, we run 100 generations on a fixed sample to assess its effectiveness.

3.4. Multi-Dimensional Evaluation Framework

After dataset construction, comprehensive evaluation is crucial for characterizing model performance. Traditional single-metric accuracy is insufficient in capturing the nuanced abilities required in aviation contexts. To address this, we propose a systematic multi-dimensional evaluation framework that integrates aviation-specific requirements and LLM application needs.

The framework includes ten dimensions.

Accuracy – correctness of answers, especially critical for objective questions.

Completeness – coverage of all key points required by reference answers.

Relevance – focus on the core of the question, avoiding unnecessary divergence.

Information Richness – inclusion of examples, data, or step-by-step reasoning.

Explainability – provision of logical reasoning processes, vital for inference and scenario-based tasks.

Question-Type Specific Correctness – adherence to format rules (*e.g.*, multiple-choice must include options, calculations must show steps).

Reference Conformance – consistency with reference answers in facts, conclusions, and reasoning steps.

Usefulness – practical value of the answer in solving the problem.

Timeliness – incorporation of the most up-to-date regulatory or technical knowledge.

Safety – compliance with aviation safety standards and ethical norms, avoiding misleading or risky outputs.

These dimensions are complementary, forming a robust framework that captures knowledge correctness, logical soundness, domain adaptation, and safety compliance.

To derive a final assessment score, we adopt a weighted aggregation strategy. Each dimension is assigned a weight $w_i$reflecting its relative importance in aviation applications, and individual scores s_i(ranging from 0 to 1) are obtained through expert review or automatic metrics where applicable. The overall performance score S is computed as a weighted sum:

$$S = \sum_{i=1}^{10} w_i \times s_i, where \ \sum_{i=1}^{10} w_i = \quad (1)$$

In our implementation, higher weights are assigned to **Accuracy (0.20)**, **Safety (0.15)**, and **Explainability (0.15)**—dimensions most critical for aviation reliability and compliance—while others such as **Completeness**, **Relevance**, and **Reference Conformance** receive moderate weights (0.10 each). **Usefulness**, **Information Richness**, **Timeliness**, and **Question-Type Specific Correctness** share the remaining proportion.

4. EXPERIMENTS

4.1. Experimental Design

4.1.1. Data Sources

For the experimental stage, aviation textbooks were selected as the primary data source, covering both foundational and specialized courses such as Principles of Aviation and Aircraft Systems Engineering. Textbooks were chosen because of their well-structured knowledge systems and clear logical organization, making them reliable materials for knowledge extraction and QA generation. The raw data were imported in PDF format and processed through multimodal parsing, which transformed them into structured knowledge point files in JSONL format. These were further used to generate a multi-category, multi-level QA dataset. Although this experiment focused on textbooks, the proposed methodology is extensible and can be generalized to other data sources such as aviation standards, technical manuals, and procedural documents, thereby offering broader support for future research and applications.

4.1.2. Environment and Tools

Experiments were conducted in a high-performance computing environment. The hardware setup included an NVIDIA A800 GPU (80 GB memory) with CUDA 12.7. The software environment was Python 3.12, with major dependencies including transformers, openai, and zhipuai, supported by auxiliary libraries such as pandas and matplotlib for data processing and visualization. The main model employed was the official vision-language large model ChatGLM-4v-plus provided by Zhipu AI.

4.2. Dataset Analysis

The proposed approach produced three datasets: dataset_mixed (18,747 samples), public_dataset (3,930 samples), and update_dataset (7,387 samples). As shown in Table **1**, the generated set contains rich metadata (*e.g.*, task, sub_type, analysis), the public set

Table 1: Overview of the Constructed Datasets

Dataset	Samples	Key Fields	Domain Focus
dataset_mixed	18747	source, task, sub_type, question, answer, analysis	Aviation QA (generated)
public_dataset	3930	question, answer, subject, source	General-purpose (public)
update_dataset	7387	question, answer, field, method	Aviation QA (auto-updated)

is categorized by general subjects, while the updated set is annotated with aviation-specific fields and generation methods.

4.2.1. Scale and Structure

The generated dataset balances comprehension (~59%) and reasoning (~41%) tasks, with diverse subtypes such as multiple-choice, Completion, true/false, calculation, and short-answer questions (Figure **6**, **7**). In contrast, the public dataset is dominated by logic and coding tasks, highlighting its general-purpose nature. Our generated and updated datasets instead emphasize aviation-relevant knowledge, including system principles and numerical computation (Figure **8**).

The updated dataset demonstrates the effectiveness of the hybrid update mechanism. Figure **9** shows broad coverage across aviation sub-domains, while Figure **10** confirms that imitation ensures consistency (2,680 samples) and expansion enriches cognitive levels (≈1.1–1.2k per Bloom's category). Representative samples further illustrate these differences: dataset_mixed emphasizes calculation with reasoning analysis, public_dataset provides a general logic puzzle, and update_dataset contrasts

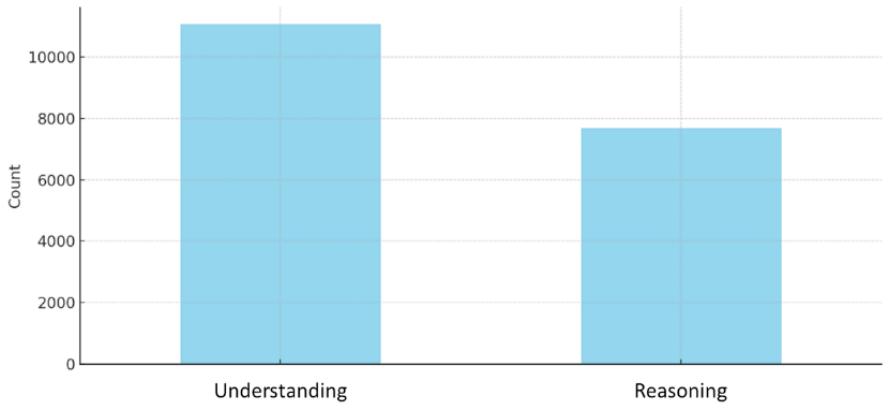


Figure 6: Distribution of Task Types in dataset_mixed.

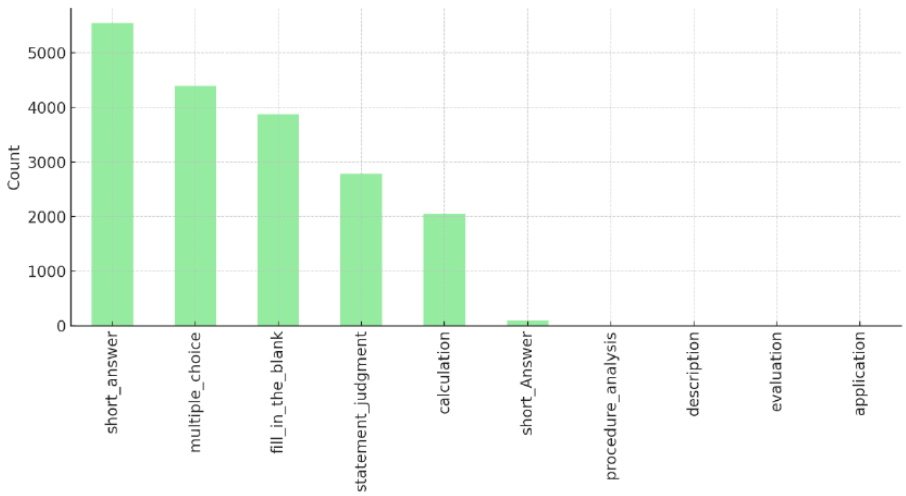


Figure 7: Distribution of Question Subtypes in dataset_mixed.

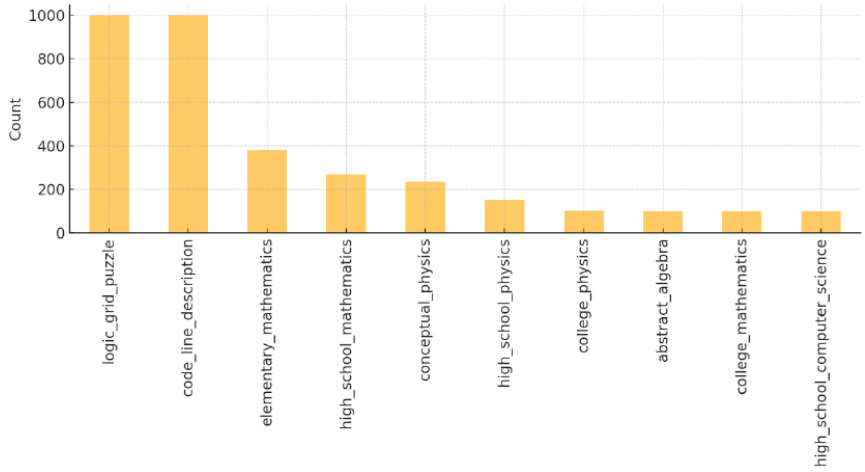


Figure 8: Top-10 Subject Distribution in public_dataset.

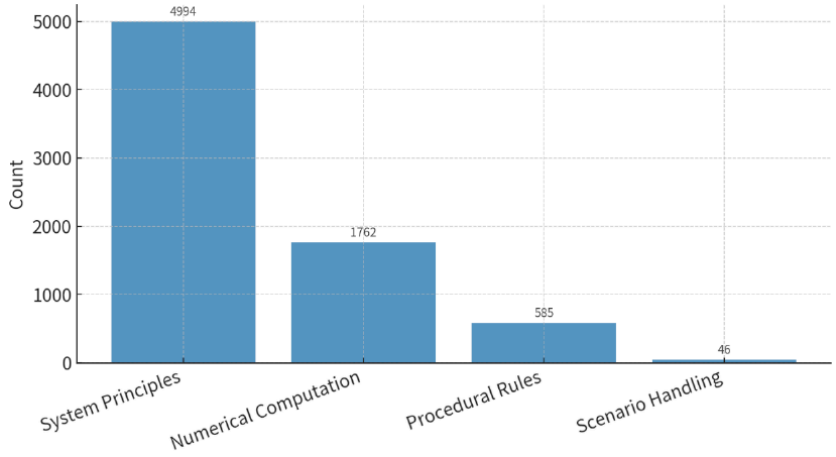


Figure 9: Aviation-Specific Fields in update_dataset.

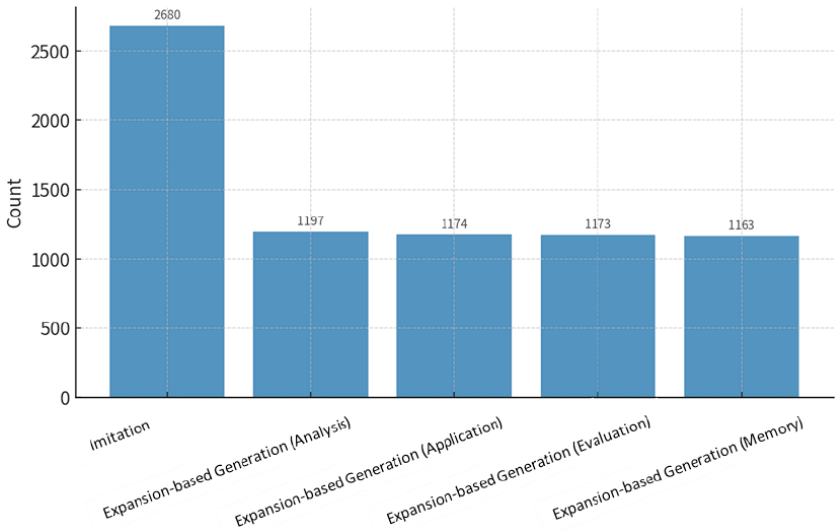


Figure 10: Generation Method Distribution in update_dataset.

imitation versus expansion for the same field, demonstrating diversity and difficulty enhancement.

Finally, quality validation indicates that with a scoring threshold of 8/10, approximately 93% of generated QA pairs are retained, ensuring reliability while filtering out noise. Together, these results show that our datasets not only complement public benchmarks but also deliver domain-specific, multimodal, and dynamically evolving resources tailored for aviation QA evaluation.

4.2.1. Knowledge Extraction Performance

In the knowledge extraction phase, PDF pages from aviation textbooks were parsed to produce structured knowledge point files. The model was instructed to output in JSON format with two fields: facts (extracted knowledge points) and confidence (scored from 0 to 10). Technically, Zhipu AI's BatchAPI was used to submit batch requests, process them on the platform, and download results locally for further filtering. A Python script was implemented to parse and retain only those results that satisfied format requirements and exceeded a confidence threshold of 7. The final output was a JSONL knowledge base file, which served as the foundation for downstream QA generation tasks.

4.2.2. QA Pair Generation Performance

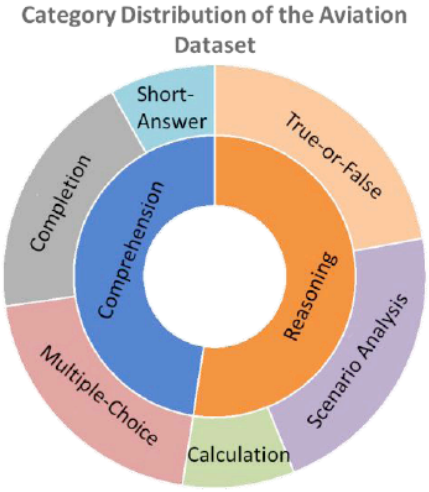
Based on the extracted knowledge points, the system generated QA pairs spanning multiple question types and cognitive levels. The results indicated that comprehension-type and reasoning-type questions were generated in proportions of approximately 48%:52%. Multiple-choice and true/false questions were the most frequent, while calculation and scenario-based questions provided stronger tests of reasoning depth. After quality validation, the average score reached 95, with an overall pass rate above 93%. This demonstrates that the generated QA pairs achieved high semantic coherence and alignment with domain knowledge. Figure **11** illustrates the distribution of question types, and reports the average score for each category.

4.2.3. Dataset Updating Performance

In the dataset updating experiments, both imitation and expansion strategies were applied to extend QA pairs. Results showed that the imitation strategy achieved strong semantic consistency, with a first-pass accuracy of 96%, making it suitable for diversifying existing questions. By contrast, the expansion strategy significantly increased question difficulty and differentiated cognitive levels, with a first-pass accuracy of approximately 95%-98%. Table **2** compares the accuracies of the two strategies across different question types, demonstrating their complementary strengths. Overall, the proposed updating mechanism effectively enabled dynamic dataset evolution while maintaining high reliability and quality.

4.3. Human Expert Validation

To validate the reliability of LLM self-assessment, we conduct a small-scale expert evaluation on a stratified random sample of subjective QA pairs. Concretely, we sample n = 100 items (covering short-answer, calculation, and scenario analysis in proportion to their dataset shares). Each item is independently reviewed by 3 aviation/engineering experts (≥ 5 years domain experience) using a



Category	Count	Question Type
Multiple-Choice	732(19.2%)	Comprehension 1817(47.6%)
Completion	775(20.2%)	
Short-Answer	310(8.2%)	
True-or-False	845(22.1%)	Reasoning 2004(52.4%)
Scenario Analysis	832(21.8%)	
Calculation	327(8.5%)	
Total	3821(100%)	—

Figure 11: Category Composition of Evaluation Data in the Aviation Standards and Quality Domain.

Table 2: Initial Generation Accuracy of Different Generation Methods

Generation Method	Imitation-based Generation	Expansion-based Generation (Memory)	Expansion-based Generation (Application)	Expansion-based Generation (Analysis)	Expansion-based Generation (Evaluation)
Initial Generation Accuracy (Approx.)	96%	95%	97%	98%	98%

4-dimension rubric: (i) factual correctness, (ii) reasoning coherence, (iii) reference alignment (with standards/manuals/textbooks when applicable), and (iv) safety compliance. Each dimension is rated on a 5-point scale (0–4) with anchors. Disagreements are resolved by discussion or a third reviewer. The expert feedback is then used to adjust self-assessment thresholds and refine prompts for subsequent iterations.

Rubric anchors: 0 = incorrect/unsafe; 1 = partially correct with major gaps; 2 = partly correct with moderate gaps; 3 = mostly correct, minor issues; 4 = fully correct, clear reasoning, compliant.

This expert validation complements LLM self-assessment, providing external reliability evidence and guiding threshold/prompt refinement for subsequent updates.

4.4. Domain RAG Enhancement Study

We evaluate the effectiveness of a domain-specific Retrieval-Augmented Generation (RAG) pipeline built from aviation standards and manuals. Using our benchmark, we compare base LLMs versus their RAG-enhanced counterparts across model scales (Qwen2.5-VL-3B / 7B / 32B) and question categories (Understanding: MC/Fill-in/Short-answer; Reasoning: True–False/Calculation/Scenario).

4.4.1. Knowledge Base and Retrieval Setup

- **Corpus.** Aviation standards, technical manuals, and regulatory/operational documents in PDF/Word.
- **Text extraction.** Extracted text is normalized and preserved with hierarchical markers (standard → chapter → clause) and page ranges.
- **Chunking.** We apply semantics-aware segmentation aligned to headings/clauses. Each chunk carries metadata {standard_id, clause_id, title_path, page_range}. Tables/figure captions are linearized into plain text and merged into the corresponding clause chunk. No image or geometric layout features are used.
- **Indexing.** Pure text dense retrieval, stored in a vector database FAISS. Optionally, a BM25 + Dense hybrid pipeline improves recall.
- **Context assembly.** Retrieved chunks are concatenated under a context budget 3000 tokens, with clause IDs and citations preserved. Near-duplicates are removed. We order snippets by topical relevance and coverage, prioritizing chunks with key terms and linearized formulas.
- **Prompting.** Type-specific templates are used for the six question types (Multiple-Choice MC, Completion Co, Short-Answer ShA, True or False TF, Calculation Ca, Scenario Analysis SA). Prompts require: (i) step-by-step reasoning where applicable, (ii) formulas/units for calculations, (iii) clause citations when standards are used, and (iv) conservative responses under uncertainty.

4.4.2. Models and Conditions

We test six conditions per scale:

- **Base:** Qwen2.5-VL-3B / 7B / 32B (no external context).
- **RAG:** Same models with retrieval context injected.

4.4.3 Evaluation Protocol

- **Dataset split.** Same test set as Section 4.4; balanced across six question types.
- **Metrics.** Ten dimensions from Section 3.4 + **weighted overall score** S. We also log

Table 3: Base vs RAG Results By Using Benchmark

Model	RAG	Comprehension				Reasoning				Total
		MC	Co	ShA	Total	TF	Ca	SA	Total	
3B	×	4.21	4.88	6.60	4.42	2.41	4.85	7.17	4.55	4.71
	√	5.68	6.22	7.20	6.25	3.01	4.89	7.10	4.81	5.53
7B	×	5.72	4.28	6.88	5.49	6.01	5.55	7.52	6.47	5.98
	√	6.18	6.64	7.72	6.72	7.38	5.78	7.74	7.23	6.98
32B	×	6.55	5.68	7.58	6.49	6.98	6.37	8.05	7.25	6.87
	√	7.44	7.11	8.34	7.54	7.86	6.18	8.27	7.71	7.63

hallucination rate, **unsafe advice rate**, and **reference usage**.

- **Quality controls.** Spot-check with the expert rubric from Section 4.3 on a stratified subsample to validate improvements in reasoning and safety.

4.4.4. Results and Discussion

From this experiment, the RAG framework delivers markedly greater gains on understanding-type questions than on reasoning-type questions. Across model scales, especially for understanding-type items, the magnitude of improvement decreases as the generator size increases: the 3B model benefits the most, while the 7B model shows the smallest gains. RAG also improves reasoning (calculation/scene) performance and safety/reference compliance, with the effects being most pronounced for 3B/7B models whose prior knowledge is limited. For the 32B model, improvements are smaller but still meaningful — primarily in explainability and usefulness — owing to stronger knowledge anchoring and citation. Typical failure cases stem from retrieval mismatch (irrelevant clauses) or context overload (attention dilution); both are mitigated by reranking and thresholding strategies.

Overall, this study provides a quantitative validation of domain RAG using our benchmark; the evaluation framework and dataset are readily applicable to model assessment in other domains, offering a robust basis for tracking progress and guiding system design.

5. DISCUSSION AND FUTURE WORK

5.1. Research Significance

This study introduces a professional aviation QA benchmark that spans multiple question types and difficulty levels while supporting dynamic updates. The proposed multi-dimensional evaluation framework goes beyond accuracy-based metrics, capturing broader model capabilities in professional contexts and improving the rigor and credibility of assessing open-ended and reasoning tasks.

5.2. Practical Value

The benchmark provides a unified tool for aviation stakeholders in model selection, optimization, and deployment. Its dataset generation and updating methods can integrate with aviation knowledge bases, standards, and manuals to support intelligent QA services in R&D, operations, and training. Moreover, the approach is generalizable, offering pathways for domain-specific evaluation in other knowledge-intensive sectors such as healthcare, energy, and transportation.

5.3. Limitations

The current work relies mainly on aviation textbooks, with limited validation on standards and manuals. While filtering and prompt optimization are effective for objective tasks, subjective questions still require human oversight. In addition, experiments were restricted to a few mainstream LLMs, and broader testing is needed to confirm generality and robustness.

5.4. Future Directions

Future work will focus on several key directions.

First, expanding data sources by incorporating multimodal aviation documents—such as standards, technical manuals, and incident reports—to enhance dataset coverage and domain relevance. In particular, we plan to establish a standardized integration pipeline for these materials, including (1) automatic parsing and structural segmentation of standards and manuals using OCR and natural language processing techniques, (2) mapping extracted clauses and procedural content to the existing knowledge taxonomy, and (3) generating question–answer pairs aligned with regulatory and operational contexts. This will ensure consistent, traceable, and automation-ready data integration for benchmark updates.

Second, optimizing the update and filtering mechanisms by introducing more efficient and reliable automated validation methods to reduce manual intervention.

Third, pursuing cross-domain adaptation to evaluate the generalizability of the proposed approach in other knowledge-intensive domains, including healthcare and law.

Finally, developing a practical aviation QA evaluation platform based on this framework to provide deployable tools and services for real-world industry applications.

5.5. Broader Impact for Robotics and Automation

The proposed benchmark has broader implications for the fields of robotics and automation. As intelligent robotic systems and automated workflows increasingly rely on large language models for knowledge retrieval, task planning, and decision support, reliable evaluation of such models becomes a foundation for system safety and autonomy.

In robotics, the benchmark can facilitate the development of AI copilots, inspection robots, and maintenance assistants capable of understanding technical documentation, interpreting procedural

manuals, and providing natural-language guidance in complex operational environments. In automation, it can support adaptive control and fault diagnosis systems by enabling LLMs to reason over structured knowledge extracted from aviation standards and manuals.

Furthermore, the benchmark's dynamic updating mechanism provides a template for maintaining up-to-date, regulation-aligned knowledge bases in rapidly evolving industrial settings. By bridging domain-specific evaluation with automation intelligence, this work contributes to building more transparent, explainable, and trustworthy AI systems for the next generation of autonomous and robotic technologies.

CONFLICTS OF INTEREST

No potential conflict of interest was reported by the author(s).

REFERENCES

[1] Brown T, Mann B, Ryder N, *et al*. Language models are few-shot learners. Adv Neural Inf Process Syst. 2020; 33: 1877-1901.

[2] Ahmed W. Artificial Intelligence in Aviation: A Review of Machine Learning and Deep Learning Applications for Enhanced Safety and Security. Premier J Artif Intell. 2025; 3: 100013.

[3] Kim CY, Kim SY, Cho SH, Kim YM. Bridging the Language Gap: Domain-Specific Dataset Construction for Medical LLMs. In: Guo J, *et al*., editors. Generalizing from Limited Resources in the Open World. Communications in Computer and Information Science. Vol. 2160. Singapore: Springer; 2024. (IJCAI 2024). https://doi.org/10.1007/978-981-97-6125-8_11

[4] Hendrycks D, Burns C, Basart S, *et al*. Measuring massive multitask language understanding. Int Conf Learn Represent. 2021.

[5] Clark P, Cowhey I, Etzioni O, *et al*. Think you have solved question answering? Try ARC, the AI2 reasoning challenge. arXiv preprint arXiv:1803.05457. 2018.

[6] Zellers R, Holtzman A, Bisk Y, Farhadi A, Choi Y. HellaSwag: can a machine really finish your sentence? Proc 57th Annu Meet Assoc Comput Linguist. 2019: 4791-4800. https://doi.org/10.18653/v1/P19-1472

[7] Srivastava A, Rastogi A, Rao A, *et al*. Beyond the imitation game: quantifying and extrapolating the capabilities of language models (BIG-bench). arXiv preprint arXiv:2206.04615. 2022.

[8] Liu H, Jiang H, Zhang S, Zhang H, Sun M. KoLA: benchmarking large language models for higher-order cognition with Bloom's taxonomy. arXiv preprint arXiv:2306.01874. 2023.

[9] Lin TY, Maire M, Belongie S, *et al*. Microsoft COCO: common objects in context. Eur Conf Comput Vis. 2014: 740-55. https://doi.org/10.1007/978-3-319-10602-1_48

[10] Antol S, Agrawal A, Lu J, *et al*. VQA: visual question answering. Proc IEEE Int Conf Comput Vis. 2015: 2425-33. https://doi.org/10.1109/ICCV.2015.279

[11] Chen W, Wu H, Zeng W, Li H. TabFact: a large-scale dataset for table-based fact verification. Int Conf Learn Represent. 2020.

[12] Zhong V, Xiong C, Socher R. TableQA: a large-scale dataset for question answering on tabular data. arXiv preprint arXiv:2006.06434. 2020.

[13] Lu P, Mishra S, Xia T, *et al*. Learn to explain: multimodal reasoning via thought chains for science question answering (ScienceQA). Adv Neural Inf Process Syst. 2022; 35: 2507-2521.

[14] Jin Q, Dhingra B, Liu Z, Cohen WW, Lu X. PubMedQA: a dataset for biomedical research question answering. Proc 28th ACM Int Conf Inf Knowl Manag. 2021:2577-85.

[15] Jin Q, Dhingra B, Liu Z, Cohen WW, Lu X. What disease does this patient have? A large-scale open-domain question answering dataset from medical exams (MedQA). arXiv preprint arXiv:1909. 00229. 2019.

[16] Guha N, Danks D, Hajian S, *et al*. LegalBench: a collaboratively built benchmark for measuring legal reasoning in large language models. arXiv preprint arXiv: 2308.11462. 2023.

[17] Chen Z, Chen W, Xu Z, Wang WY. FinQA: a dataset of numerical reasoning over financial data. Proc Conf Empir Methods Nat Lang Process. 2021: 3697-3711. https://doi.org/10.18653/v1/2021.emnlp-main.300

[18] Lozano Tafur C, Camero RG, Aldana Rodríguez D, Daza Rincón JC, Rativa Saenz E. Applications of artificial intelligence in air operations: a systematic review. Results Eng. 2025; 25: 103742. Available from: https://doi.org/10.1016/j.rineng.2024.103742

[19] Das A, Kottur S, Moura JMF, Lee S, Batra D, Parikh D. Embodied question answering. Proc IEEE Conf Comput Vis Pattern Recognit. 2018. https://doi.org/10.1109/CVPR.2018.00008

[20] Sermanet P. RoboVQA: Multimodal Long-Horizon Reasoning for Robotics. arXiv preprint arXiv:2311.00899. 2023. https://doi.org/10.1109/ICRA57147.2024.10610216

[21] Yan F. RoboMM: All-in-One Multimodal Large Model for Robotic Manipulation. arXiv preprint arXiv:2412.07215. 2024.

[22] Balcı E, Sarıgül M, Ata B. Benchmarking large language model reasoning in indoor robot navigation. 33rd Signal Processing and Communications Applications Conference (SIU). 2025: 1-4. https://doi.org/10.1109/SIU66497.2025.11111749

https://doi.org/10.31875/2409-9694.2025.12.05