Surface Mastery, Deep Failures: ATC-QA Benchmark Uncovers Critical Limitations of LLMs in Aviation Safety

Anonymous ACL submission

Abstract

We present ATC-QA, a novel benchmark for evaluating large language models (LLMs) in aviation safety applications. Derived from 43,264 qualified Aviation Safety Reporting System (ASRS) reports, our benchmark comprises 47,151 question-answer pairs spanning seven question types and four difficulty levels. Experimental evaluation of nine representative LLMs reveals a striking dichotomy: apparent mastery of classification tasks (up to 95% accuracy) coupled with profound failures in critical capabilities. We identify a pronounced "terminology generation bottleneck" where even top-performing models achieve only 20% accuracy on fill-in-the-blank questions—a 75 percentage point drop from their classification performance. Our analysis further uncovers systematic process-result discrepancies in calculation tasks, where models produce correct numerical answers (53-82% accuracy) through fundamentally flawed reasoning processes (8-55% correctness). Counter-intuitive performance patterns across difficulty levels, where models often perform better on more complex questions, suggest fundamental differences between human and machine-perceived complexity. Architecture choice significantly impacts performance beyond parameter count, with similar-sized models showing up to 4× performance gaps on domain-specific tasks. ATC-QA provides a framework for assessing LLM capabilities in safety-critical environments where domain expertise is essential, highlighting the need for specialized evaluation in high-stakes domains.

1 Introduction

Aviation safety directly impacts transportation security and passenger well-being. While traditional approaches have advanced safety, they struggle to manage the increasing complexity of modern air traffic operations and regulations.

Large language models (LLMs) show promise for transforming aviation safety management through improved natural language understanding. However, applying general-purpose LLMs to specialized domains presents challenges, as mainstream benchmarks like SQuAD (Rajpurkar et al., 2016), CEval (Huang et al., 2023), and MMLU (Hendrycks et al., 2021) lack the industry-specific context needed for meaningful aviation safety assessment (Fischer et al., 2024; Li et al., 2024).

Recent work on specialized benchmarks, such as GPQA (Rein et al., 2024), highlights the challenges of evaluating LLMs in technical domains. The inconsistent reasoning of LLMs about domain concepts (Sosa et al., 2024) raises concerns for aviation applications where precision is vital. Similar challenges have been observed in medical (Ouyang et al., 2024) and tool learning domains (Ye et al., 2024), where domain-specific knowledge and precision are critical.

ATC-QA provides a comprehensive multi-dimensional benchmark for evaluating LLM capabilities in aviation safety compliance. This work contributes a diverse dataset of 47,151 QA pairs across seven question types with systematic difficulty classification, accompanied by extensive evaluation across nine representative LLMs.

Our analysis reveals several striking phenomena: a significant gap between classification performance and terminology generation capability; discrepancies between numerical answers and reasoning processes in calculation tasks (Huang et al., 2024); and counter-intuitive performance patterns across difficulty levels. We also find that architecture choice impacts performance beyond parameter count (Zhou et al., 2024).

This benchmark establishes a framework for assessing LLM capabilities in environments where precise understanding of technical terminology and procedures is essential (Mou et al., 2024). The observed surface-level mastery coupled with deeper functional failures highlights the importance of

multi-dimensional evaluation approaches for specialized domains.

2 ATC-QA Benchmark

Our benchmark comprises 47,151 aviation safety QA pairs derived from 43,264 qualified ASRS incident reports. Each entry includes both question-answer content and structured metadata as outlined in Table 1.

Using Gemini 2.0 Flash (Google DeepMind, 2025), we transformed aviation incident narratives into seven distinct question types through a multi-stage process: report-to-QA conversion, question type diversification, quality optimization, and standardization. Questions were categorized into four difficulty levels ranging from basic recall (L1) to expert analysis (L4), with manual validation confirming classification quality.

Table 2 and Figure 1 show the distribution across question types and difficulty levels. True/False (30.6%) and Single Choice (22.2%) questions are most common, with L3 difficulty questions (67.6%) representing the majority of the benchmark.

3 Methodology

We developed a four-stage pipeline to create the ATC-QA benchmark from aviation safety reports. Figure 2 illustrates our data processing workflow, starting with ASRS report filtering and culminating in structured QA pairs.

3.1 Data Sources and Preprocessing

The Aviation Safety Reporting System (ASRS) dataset contains 47,723 de-identified incident reports submitted by aviation professionals. Each report includes a narrative description, expert synopsis, and structured metadata.

Our filtration process removed 4,459 reports (9.34%) with formatting issues or incomplete metadata, yielding 43,264 qualified narratives for benchmark development.

3.2 Generation Framework

Our generation framework consists of four main phases:

3.2.1 Phase 1: Report-to-QA Conversion

We used a prompt engineering approach to extract aviation safety knowledge from incident narratives. Gemini-2.0-Flash identified critical safety elements including event causality, safety procedures, preventative measures, and learning points.

This process generated 50,718 initial QA pairs with metadata classifying knowledge type and relevance.

3.2.2 Phase 2: Question Diversification

We transformed basic QA pairs into seven distinct question formats to evaluate different aspects of model capabilities. This involved analyzing semantic structure based on information density, technical terminology, numerical content, and scenario complexity.

For classification tasks, we implemented contextual negation strategies. Choice-based questions used semantic similarity algorithms to generate plausible distractors. Fill-in-the-blank questions targeted terminology recall by strategically removing domain terms. Calculation questions extracted numerical values and organized the calculation process. Comprehensive and analytical questions incorporated multi-step reasoning requirements.

3.2.3 Phase 3: Quality Optimization

Our validation process addressed content validity and structural consistency through automated checks and expert review. We enhanced question clarity while preserving domain terminology, applied format-specific validation rules, verified answer uniqueness, and ensured explanatory content aligned with aviation standards. This approach is similar to recent work in dataset cleansing using LLM-based annotation methods (Choi et al., 2024), but tailored specifically to aviation safety domain requirements.

3.2.4 Phase 4: Standardization

The final phase established a standardized framework for consistent model assessment. Each question received a unique identifier and type specification according to a defined taxonomy.

We implemented type-specific formatting solutions to preserve semantic integrity while enabling automated evaluation, with specialized handlers for complex question types.

After removing 4,095 questions that failed to meet final quality thresholds, the standardization process yielded 47,151 fully specified QA pairs.

4 Experiments and Evaluation

We evaluated nine representative LLMs on ATC-QA to assess their capabilities in aviation safety domain. Our analysis uncovered several signifi-

Field	Description
question ID	Unique identifier for each question
type	Question type (Single, Multi, Fill-in, T/F, Calc., Analysis, Comp.)
difficulty	Difficulty level (L1, L2, L3, L4) from basic to expert
question	Question content formatted according to type
answer	Expected correct response to the question
analysis	Explanation of the solution with detailed reasoning
knowledge points	Technical concepts relevant to the question

Table 1: Dataset structure showing the fields included in each QA pair

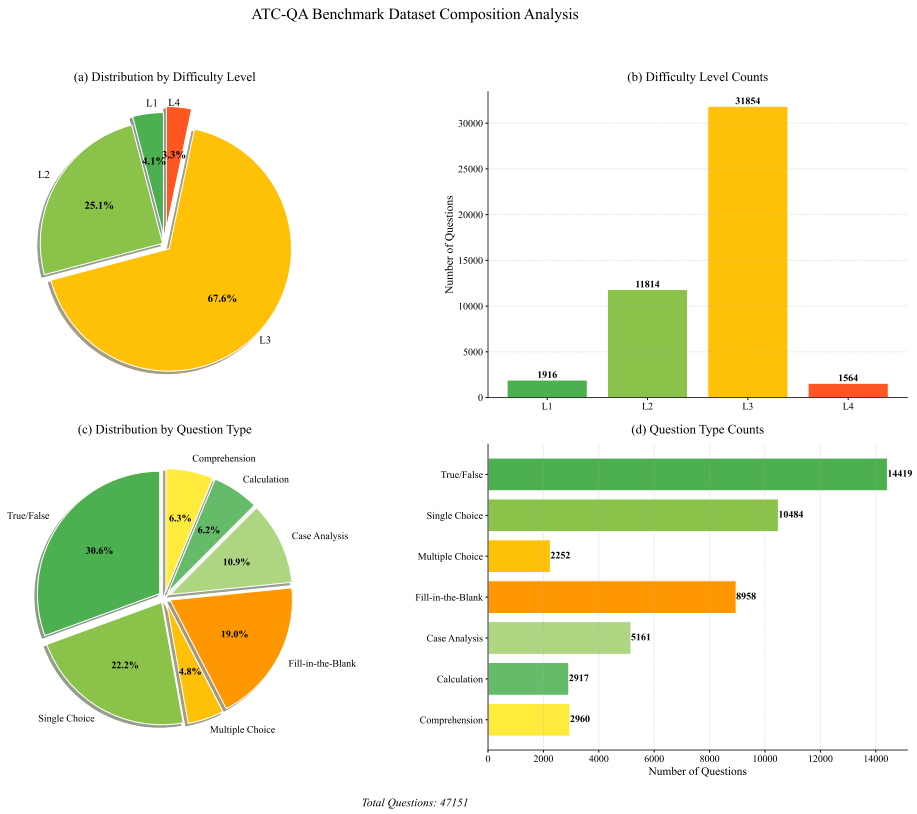

Figure 1: ATC-QA benchmark statistics showing: (a) percentage by difficulty level, (b) count per difficulty level, (c) percentage by question type, and (d) count per question type across all 47,151 questions

Question Type	Difficulty Level Distribution				Summary	
	L1 (4.1%)	L2 (25.1%)	L3 (67.6%)	L4 (3.3%)	Total	Proportion
True/False	1,324	4,803	7,052	1,240	14,419	30.6%
Single Choice	9	1,773	8,684	16	10,484	22.2%
Multiple Choice	1	383	1,866	1	2,252	4.8%
Fill-in-the-Blank	581	2,561	5,510	306	8,958	19.0%
Case Analysis	1	2,112	3,048	0	5,161	10.9%
Calculation	0	29	2,888	0	2,917	6.2%
Comprehension	0	153	2,806	1	2,960	6.3%
Total	1,916	11,814	31,854	1,564	47,151	100%

Table 2: Distribution of questions by type and difficulty level

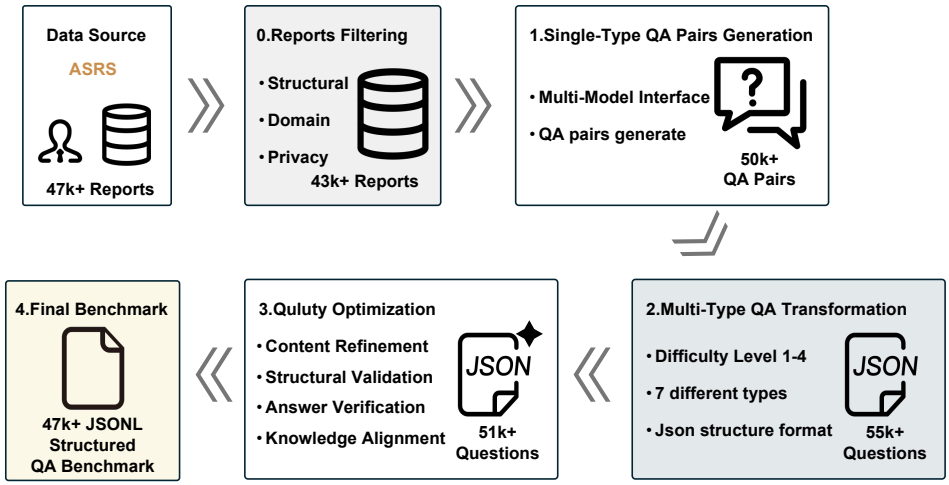

Figure 2: Data processing pipeline for creating the ATC-QA benchmark

cant patterns with important implications for safety-critical applications.

4.1 Experimental Setup

We selected nine models spanning different architectures and parameter scales:

Closed-source models: Grok-2, GPT-4o-mini, Gemini-1.5-Flash, and Gemini-2.0-Flash

Open-source models: Llama-3.1-8B, Llama-3.2-1B, Llama-3.2-3B, DeepSeek-R1-Qwen-7B, and DeepSeek-R1-Llama-8B

We tailored evaluation methods to each question type while maintaining comparability. Classification tasks used standard accuracy metrics, while multiple-choice questions were evaluated through precision, recall, and F1 scores to account for partial correctness.

For fill-in-the-blank questions, we implemented exact and partial matching protocols to account for terminology variations. Calculation questions received multi-dimensional assessment including answer accuracy, numerical proximity (±5% tolerance), and process correctness.

Complex reasoning tasks underwent human-in-the-loop evaluation supplemented by LLM-based judgment tools, using a 1-5 point scale measuring reasoning quality and content coverage.

All evaluations used standardized zero-shot prompting templates to assess fundamental model capabilities. Performance metrics represent averages across 1,000 randomly sampled questions.

4.2 Performance Analysis

Figure 3 presents performance across question types, revealing significant capability variations.

Classification Performance. Table 3 shows strong performance on classification tasks, with Gemini models and Grok-2 consistently leading.

Precision-recall trade-offs in multiple-choice questions reveal varying model strategies: high-precision models like Llama-3.1-8B (88% precision, 68% recall) adopt conservative approaches, minimizing incorrect selections at the cost of missing some correct answers; high-recall models like GPT-4o-mini (74% precision, 94% recall) more aggressively identify potential correct options but include more errors; while Gemini models achieve the most balanced performance (precision: 87%, recall: 89%, F1: 85%).

Terminology Generation Bottleneck. Table 4 reveals a significant limitation in fill-in-the-blank questions, where even top models achieve only 18-20% exact match accuracy—a 75 percentage point drop compared to their classification performance.

This "terminology generation bottleneck" persists even with partial matching (only 2-4 percentage points improvement), suggesting a fundamental limitation in producing precise aviation terminology. Notably, base architecture choice significantly impacts performance, with DeepSeek models having similar parameter counts but different foundations exhibiting a 3-4× performance gap on this task. This finding has critical implications for safety applications requiring precise technical vocabulary generation and aligns with broader research showing that LLMs struggle with consistent conceptual reasoning (Sosa et al., 2024), particularly when required to produce precise domain-specific terminology rather than recognize it.

Process-Result Gap in Calculations. Table 5 reveals a consistent gap between numerical accuracy and methodological correctness. Models achieve higher result accuracy (53-82%) than process correctness (8-55%), with gaps ranging from 20-45 percentage points.

This indicates that correct answers often emerge from incorrect reasoning paths— a critical concern for safety applications where procedural correctness is essential (Huang et al., 2024). Proximity scores further suggest that models may be leveraging pattern recognition rather than properly executing aviation calculation procedures, a finding that parallels observations in biomedical reasoning tasks where LLMs can generate plausible outputs without demonstrating sound methodological reasoning (Qi et al., 2024).

Complex Reasoning Capabilities. Table 6 shows substantial differences in complex reasoning capabilities. Gemini models achieved the highest overall scores (4.28-4.42) and content coverage (83-87%), while smaller models showed significant limitations, particularly in comprehensive reasoning where content coverage dropped to 47-50% for smaller Llama models.

4.3 Counter-Intuitive Difficulty Patterns

The relationship between question difficulty and model performance reveals counter-intuitive patterns that challenge common assumptions. As demonstrated in Figure 4, while all models show declining performance as difficulty increases (average drop of 15 percentage points from L1 to L4), the magnitude of this effect varies significantly by model type:

Monotonically increasing: Grok-2 shows steady improvement (67%→71%→76%→78%) as difficulty increases.

Peak-at-L3: Gemini-2.0-Flash peaks at L3 (65%→72%→83%→75%).

Non-monotonic: Llama-3.2-1B shows erratic patterns (52%→43%→47%→69%).

This counter-intuitive relationship reveals fundamental differences between human-perceived and machine-perceived question complexity. Our analysis suggests three contributing factors to this phenomenon. First, higher difficulty questions typically contain more detailed technical contexts and specific operational scenarios, providing LLMs with richer semantic networks for inference. Second, difficult aviation questions often incorporate specialized terminology with less lexical ambiguity than general language, reducing the model's uncertainty during reasoning. Third, complex questions frequently require explicit multi-step reasoning, which appears to activate more structured reasoning pathways in transformer architectures compared to seemingly simpler questions that may require implicit knowledge not directly stated in the context.

This explanation aligns with recent cognitive modeling research showing that transformer architectures process information differently from humans, with particular strength in structured pat-

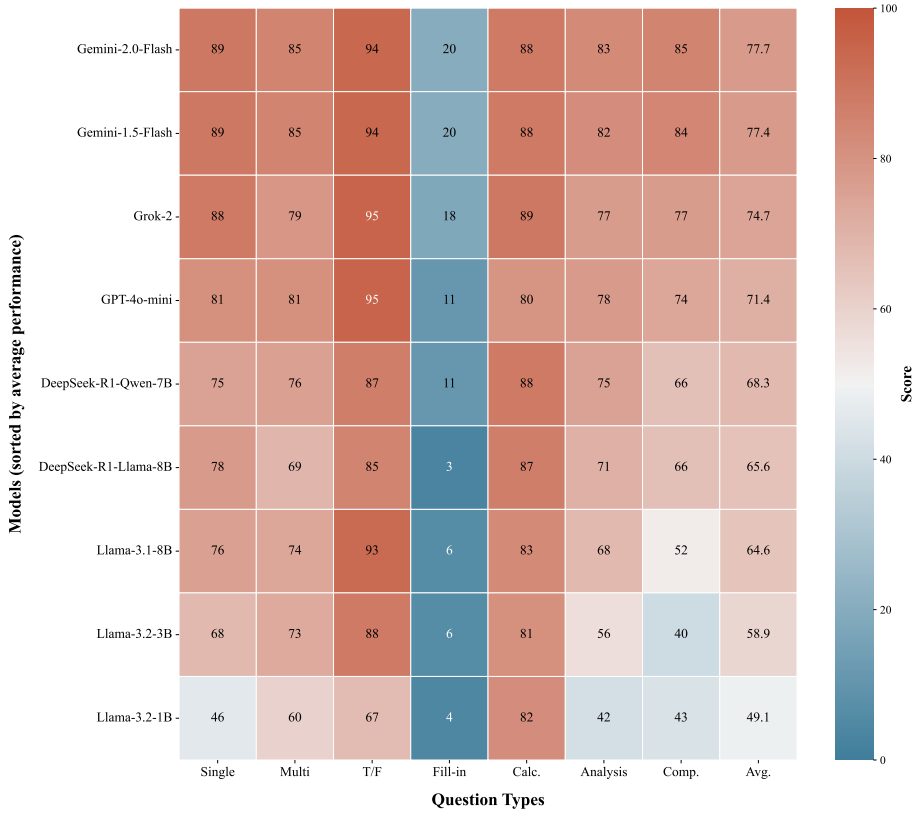

Figure 3: Performance comparison by question type showing strong classification performance (T/F: 67-95%) but weak terminology generation (Fill-in: 3-20%) across nine LLMs

	T/F	Single Choice	Multiple Choice		
Model	Acc.(%)	Acc.(%)	Precision	Recall	F1 Score
Grok-2	95	88	86	77	79
GPT-4o-mini	95	81	74	94	81
Llama-3.1-8B	93	76	88	68	74
Llama-3.2-1B	67	46	66	61	60
Llama-3.2-3B	88	68	84	70	73
Gemini-1.5-Flash	94	89	87	89	85
Gemini-2.0-Flash	94	89	87	89	85
DeepSeek-R1-Qwen-7B	87	75	74	89	76
DeepSeek-R1-Llama-8B	85	78	63	86	69

Table 3: Classification performance showing three distinct model strategies: high-precision (Llama-3.1-8B), high-recall (GPT-4o-mini), and balanced (Gemini models)

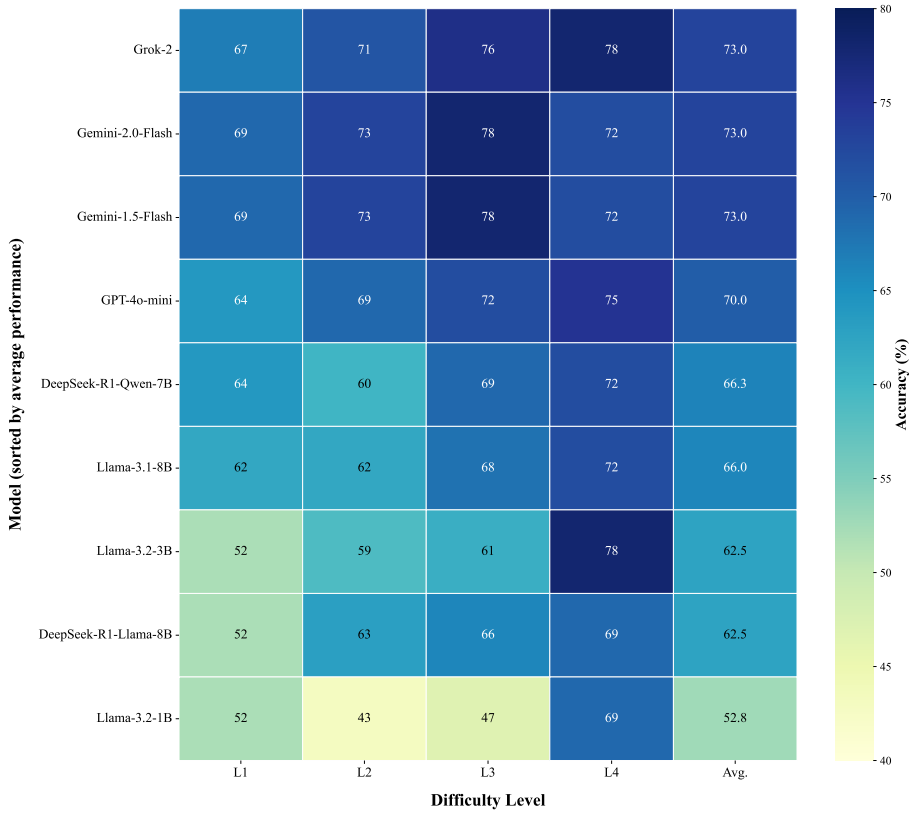

Figure 4: Performance across difficulty levels showing stronger results on L1 (92%+) and weaker performance on L4 (74% max), with 15% average performance drop from L1 to L4

	Match Rate (%)	
Model	Exact	Partial
Grok-2	18	22
GPT-4o-mini	11	14
Llama-3.1-8B	6	7
Llama-3.2-1B	4	4
Llama-3.2-3B	6	6
Gemini-1.5-Flash	20	24
Gemini-2.0-Flash	19	23
DeepSeek-R1-Qwen-7B	11	12
DeepSeek-R1-Llama-8B	3	4

Table 4: Fill-in-the-blank performance showing universally low accuracy (max 20% exact match) and strong architecture influence (DeepSeek-Qwen-7B outperforming DeepSeek-Llama-8B by 3-4×)

tern recognition rather than abstract concept formation (Sun et al., 2024). The stability difference between large and small models (9 versus 26 percentage point variation across difficulty levels) further suggests that increased parameter count enhances the model's ability to maintain consistent reasoning patterns across varying problem complexities—a finding with significant implications for safety-critical applications.

4.4 Model Typology Analysis

Building on our multi-dimensional evaluation across both question types and difficulty levels, we identify distinct behavioral patterns that characterize different model architectures. This cross-examination of performance metrics reveals three clearly differentiated LLM profiles:

Recognition-dominant models (e.g., Llama-3.1-8B) excel in classification (91-93% on T/F) but exhibit fundamental limitations in generative tasks (6% on fill-in) and procedural reasoning (15-22% on calculations).

Reasoning-oriented models (e.g., DeepSeek-R1-Qwen-7B) demonstrate superior process correctness (55%) while maintaining competitive classification performance.

Balanced performers (Gemini and Grok-2) maintain consistent performance across dimensions, with robust knowledge representation and multi-faceted reasoning capabilities.

This typology synthesizes the observed patterns

Model	Metrics (%)		
	Accuracy	Proximity	Process Correctness
Grok-2	82	89	45
GPT-4o-mini	70	80	42
Llama-3.1-8B	54	83	22
Llama-3.2-1B	53	82	8
Llama-3.2-3B	62	81	15
Gemini-1.5-Flash	80	88	50
Gemini-2.0-Flash	79	88	51
DeepSeek-R1-Qwen-7B	75	88	55
DeepSeek-R1-Llama-8B	74	87	45

Table 5: Calculation performance showing models achieving higher result accuracy (53-82%) than process correctness (8-55%), revealing correct answers often derived through incorrect reasoning

Model	Analysis			Comprehensive		
	Score (1-5)	Coverage (%)	Avg.	Score (1-5)	Coverage (%)	Avg.
Grok-2	4.09	78	77	4.07	76	77
GPT-4o-mini	4.11	79	78	3.96	78	74
Llama-3.1-8B	3.73	69	68	3.09	57	52
Llama-3.2-1B	2.68	46	42	2.73	50	43
Llama-3.2-3B	3.22	56	56	2.58	47	40
Gemini-1.5-Flash	4.28	83	82	4.35	85	84
Gemini-2.0-Flash	4.34	84	83	4.42	87	85
DeepSeek-R1-Qwen-7B	4.01	74	75	3.65	71	66
DeepSeek-R1-Llama-8B	3.84	69	71	3.64	68	66

Table 6: Complex reasoning performance showing larger models achieving superior performance while smaller models struggle with complex reasoning

from previous sections and carries significant implications for safety-critical deployments. The stark performance differences between recognition and generation tasks suggest that even advanced LLMs develop disparate capabilities in information processing. For aviation safety applications, this indicates that model selection should be task-specific rather than assuming universal capability across all operational contexts—a recognition-dominant model might be appropriate for classifying incident reports, while safety-critical calculation tasks would require reasoning-oriented architectures with stronger procedural correctness.

5 Limitations

While ATC-QA advances aviation safety assessment, several limitations merit acknowledgment. The ASRS dataset introduces North American-centric biases from its voluntary reporting mechanism, limiting global representativeness. Our QA generation's dependence on Gemini-2.0-Flash may also propagate inherent biases into the benchmark, similar to challenges observed in other AI-augmented benchmark generation approaches (Xia et al., 2024).

Methodologically, the uneven distribution of question types—with True/False questions (30.6%) significantly overrepresented compared to Multiple Choice (4.8%)—may skew performance metrics. The counter-intuitive relationship between human-assigned difficulty levels and model performance suggests fundamental differences in complexity perception. Our heterogeneous evaluation approaches across question types further introduces methodological inconsistencies affecting comparability, a challenge also noted in broader studies of domain-specific LLM evaluation (Sun et al., 2024).

The domain-specificity creates inevitable trade-

offs between aviation safety performance and broader generalizability, similar to findings in domain adaptation strategies for retrieval-augmented generation (Zhang et al., 2024) and specialized benchmarks in medical (Ouyang et al., 2024) and tool learning scenarios (Ye et al., 2024). Laboratory results may not translate to operational environments where additional contextual variables influence model behavior. Future work should expand data sources beyond ASRS, implement multi-model consensus approaches for question generation, develop unified evaluation frameworks, and engage aviation practitioners for validation.

References

Juhwan Choi, Jungmin Yun, Kyohoon Jin, and Young-Bin Kim. 2024. Multi-news+: Cost-efficient dataset cleansing via llm-based data annotation. *arXiv preprint*, arXiv:2404.09682.

Kevin Fischer, Darren Fürst, Sebastian Steindl, Jakob Lindner, and Ulrich Schäfer. 2024. Question: How do large language models perform on the question answering tasks? answer:. *arXiv preprint*, arXiv:2412.12893.

Google DeepMind. 2025. Gemini 2.0 flash. Technical report, Google DeepMind.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *ICLR*.

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. 2024. Large language models cannot self-correct reasoning yet. *arXiv preprint*, arXiv:2310.01798.

Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Jiayi Lei, Yao Fu, Maosong Sun, and Junxian He. 2023. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc.

Tianhao Li, Jingyu Lu, Chuangxin Chu, Tianyu Zeng, Yujia Zheng, Mei Li, Haotian Huang, Bin Wu, Zuoxian Liu, Kai Ma, Xuejing Yuan, Xingkai Wang, Keyan Ding, Huajun Chen, and Qiang Zhang. 2024. Scisafeeval: A comprehensive benchmark for safety alignment of large language models in scientific tasks. *arXiv preprint*, arXiv:2410.03769.

Yutao Mou, Shikun Zhang, and Wei Ye. 2024. Sg-bench: Evaluating llm safety generalization across diverse tasks and prompt types. *arXiv preprint*, arXiv:2410.21965.

Zetian Ouyang, Yishuai Qiu, Linlin Wang, Gerard De Melo, Ya Zhang, Yanfeng Wang, and Liang He. 2024. CliMedBench: A large-scale chinese benchmark for evaluating medical large language models in clinical scenarios. *arXiv preprint*, arXiv:2410.03502.

Biqing Qi, Kaiyan Zhang, Kai Tian, Haoxiang Li, Zhang-Ren Chen, Sihang Zeng, Ermo Hua, Hu Jinfang, and Bowen Zhou. 2024. Large language models as biomedical hypothesis generators: A comprehensive evaluation. In *First Conference on Language Modeling*. OpenReview.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392. Association for Computational Linguistics.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. 2024. GPQA: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling*. OpenReview.

Rosario Uceda Sosa, Karthikeyan Natesan Ramamurthy, Maria Chang, and Moninder Singh. 2024. Reasoning about concepts with LLMs: Inconsistencies abound. In *First Conference on Language Modeling*. OpenReview.

Chongyan Sun, Ken Lin, Shiwei Wang, Hulong Wu, Chengfei Fu, and Zhen Wang. 2024. Lalaeval: A holistic human evaluation framework for domain-specific large language models. In *First Conference on Language Modeling*. OpenReview.

Chunqiu Steven Xia, Yinlin Deng, and Lingming Zhang. 2024. Top leaderboard ranking = top coding proficiency, always? evoeval: Evolving coding benchmarks via LLM. In *First Conference on Language Modeling*. OpenReview.

Junjie Ye, Yilong Wu, Songyang Gao, Caishuang Huang, Sixian Li, Guanyu Li, Xiaoran Fan, Qi Zhang, Tao Gui, and Xuanjing Huang. 2024. RoTBench: a multi-level benchmark for evaluating the robustness of large language models in tool learning. *arXiv preprint*, arXiv:2401.08326.

Tianjun Zhang, Shishir G Patil, Naman Jain, Sheng Shen, Matei Zaharia, Ion Stoica, and Joseph E. Gonzalez. 2024. RAFT: Adapting language model to domain specific RAG. In *First Conference on Language Modeling*. OpenReview.

Xiyuan Zhou, Huan Zhao, Yuheng Cheng, Yuji Cao, Gaoqi Liang, Guolong Liu, Wenxuan Liu, Yan Xu, and Junhua Zhao. 2024. Elecbench: a power dispatch evaluation benchmark for large language models. *arXiv preprint*, arXiv:2407.05365:1–15.