A Compliance-Preserving Retrieval System for Aircraft MRO Task Search

Byungho Jo
AI Convergence Research Center, Inha University
Incheon, South Korea
bhjo12@inha.ac.kr

Abstract

Aircraft Maintenance Technicians (AMTs) spend up to 30% of work time searching manuals—a documented efficiency bottleneck in MRO operations where every procedure must be traceable to certified sources. We present a compliance-preserving retrieval system that adapts LLM reranking and semantic search to aviation MRO environments by operating alongside, rather than replacing, certified legacy viewers. The system constructs revision-robust embeddings from ATA chapter hierarchies and uses vision-language parsing to structure certified content, allowing technicians to preview ranked tasks and access verified procedures in existing viewers. Evaluation on 49k synthetic queries achieves >90% retrieval accuracy, while bilingual controlled studies with 10 licensed AMTs demonstrate 90.9% top-10 success rate and 95% reduction in lookup time—from 6-15 minutes to 18 seconds per task. These gains provide concrete evidence that semantic retrieval can operate within strict regulatory constraints and meaningfully reduce operational workload in real-world multilingual MRO workflows.

1 Introduction

Aircraft Maintenance Technicians (AMTs) regularly rely on certified maintenance manuals to locate the exact tasks required to inspect or repair aircraft systems (FAA Regulations). Despite their centrality to aviation safety, these manuals have grown into extremely large and intricate information sources—often exceeding tens of thousands of pages organized through multi-level Air Transport Association (ATA) structures (Avers et al., 2012; Commerce, 2018). As a result, field reports (Taylor, 2008) indicate that up to 30% of a AMTs' work time is spent searching for the correct procedure. This challenge is not merely an industry inefficiency but represents a fundamental information-retrieval bottleneck—technicians must translate informal, problem-driven queries into highly structured, deeply nested documentation that was never designed for natural-language access.

```
Chapter 21: Air Conditioning
Chapter 22: Auto Flight (120 tasks)
  ... (10 chapters omitted)
Chapter 32: Landing Gear (155 tasks)
  +- 32-09 Main Landing Gear (100 tasks)
  +- 32-41 Brake System (55 tasks)
      +- 32-41-20 Brake Disconnect
      +- ... (6 components)
      +- 32-41-31 Gear Brake
          +- 401 Removal
          |    +- 32-41-41-000-801 Removal
          |    +- 32-41-41-400-801 Installation
          +- 601 Inspection
  ... (15 chapters omitted)
Chapter 72: ENGINE (180 tasks)
```

Figure 1: ATA chapter-based manual structure illustrating the hierarchical complexity that AMTs must navigate to locate specific maintenance tasks. A representative example, Ch.32 → 32-41 → 32-41-31 → 401 → 32-41-41-000-801, demonstrates a five-level navigation path with over fifty branching options. Numbers in blue indicate the task counts at each level.

As illustrated in Figure 1, certified manuals impose an additional structural burden: their ATA chapter hierarchy a deep, tree-structured index that technicians must manually navigate. Reaching a single end-task often requires traversing four to six nested levels, each containing dozens of branching options, before encountering several candidates with near-identical titles. For example, tasks such as "Brake Valve Removal" and "Brake Shuttle Valve Removal" appear across different ATA substructures with minimal lexical distinction. This combination of hierarchical depth, dense branching, and high lexical ambiguity makes keyword-based search fundamentally unreliable, frequently forcing technicians to open and compare multiple candidates before determining the correct procedure.

Prior work in AI for aircraft maintenance largely

Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 5: Industry Track), pages 317–329
March 25-27, 2026

sidesteps this retrieval bottleneck. Research on predictive maintenance, and AR/VR-based training systems (Jo et al., 2014; Tuğçe, 2025) has focused on optimizing maintenance execution rather than helping AMTs locate the correct procedure in the first place. Meanwhile, recent NLP efforts in aviation—such as safety-report generation (Tikayat Ray et al., 2023) therefore do not engage with the rigid, hierarchical structure of certified manuals. Standard RAG frameworks typically present rewritten or synthesized text to the user, but MRO regulations require technicians to read the certified manual itself—even when the regenerated content is semantically identical. This regulatory constraint prevents direct adoption of standard RAG pipelines and motivates compliance-preserving retrieval approaches. To the best of our knowledge, no existing work addresses semantic task retrieval under this unique combination of constraints: immutable documentation, deep hierarchical indexing, and AMTs-generated natural-language queries.

To address this gap, we introduce a compliance-preserving assistive retrieval system that enables natural-language task lookup without modifying OEM manuals or viewers. The key idea is to exploit the stability of ATA metadata: task titles and hierarchy paths change far less frequently than full text. Our system builds revision-robust task embeddings exclusively from this metadata while using a vision–language model (VLM) only to structure page-level content for previews—not for retrieval—thus avoiding any generated or altered text. At query time, an LLM re-ranks a candidate set but never sees or produces procedural content, preserving certification boundaries.

We evaluate the system through two complementary studies. (1) A large-scale synthetic benchmark of 49k AMT-style queries tests robustness to paraphrasing, synonyms, and typos. (2) A bilingual human study with ten licensed AMTs examines real-world performance using English and Korean queries over English-only manuals, reflecting common multilingual MRO environments.

The contributions of this study are as follows:

- We formalize task lookup in certified maintenance manuals as a semantic retrieval problem while respecting immutability, traceability, and revision-control constraints.
- We demonstrate a scalable manual-to-knowledge conversion pipeline using ATA metadata and VLM-based structuring that requires only minimal post-editing.
- Through synthetic and human evaluations, we show that the method achieves >90% retrieval accuracy and reduces lookup time by over 95%—from minutes to seconds, suggesting a practical path for adoption in airline MRO workflows.

2 Related Work

2.1 Artificial Intelligence in Aircraft MRO

Prior AI research in aircraft MRO has focused on operational execution (AR/VR-guided maintenance) and predictive analytics (failure forecasting) but has not addressed the fundamental bottleneck of locating correct procedures within certified manuals. Augmented Reality systems (Jo et al., 2014; Tuğçe, 2025) assume technicians have already identified the correct task, while predictive maintenance (Yang et al., 2022) forecasts component failures without addressing manual navigation complexity. The challenge of semantic task retrieval under regulatory constraints—where manuals cannot be modified and every procedure must be traceable—remains unexplored.

2.2 Large Language Models in Aviation

Recent studies have explored domain-adapted LLMs in aviation for Q&A for pilot training (Wang et al., 2024), safety report summarization (Tikayat Ray et al., 2023), and traffic management (Abdulhak et al., 2024). However, MRO task retrieval requires mapping queries to exact certified procedures with full audit trails—a regulatory constraint that prohibits LLM-generated content. To our knowledge, no prior work addresses semantic retrieval under these constraints.

3 Proposed System Architecture

To address the dual challenge of regulatory compliance and operational efficiency, our system introduces an assistive retrieval system that functions independently of certified OEM viewers, which cannot be modified under aviation regulations. As illustrated in Figure 2, the system operates in two main stages: an offline knowledge structuring stage and an online retrieval stage.

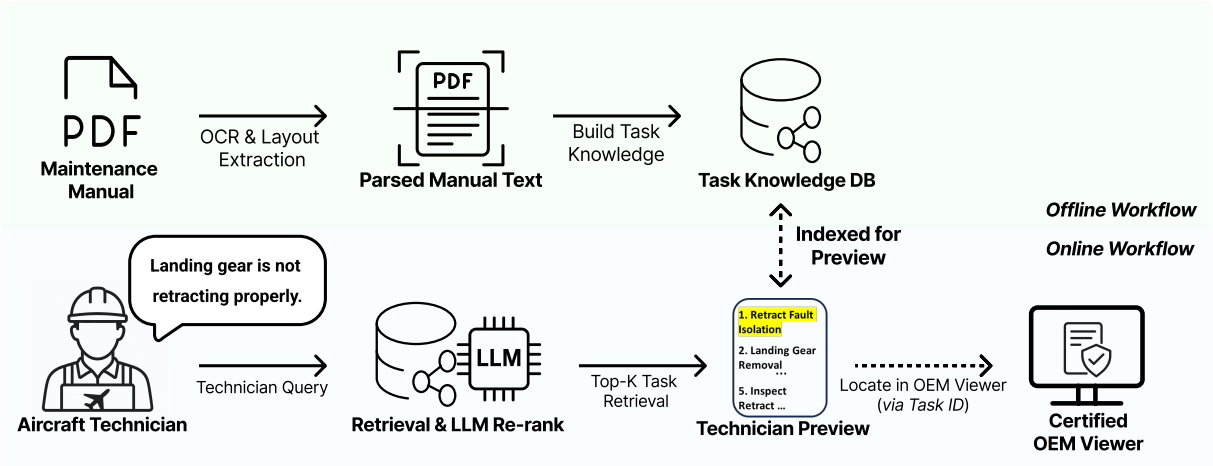

Figure 2: Offline workflow extracts and structures tasks from maintenance PDF manuals into a Task Knowledge DB. During the online workflow, a technician query triggers Top-K retrieval with LLM re-ranking, previews the ranked tasks, and opens the certified procedure in the official viewer, maintaining full compliance while reducing lookup time from minutes to seconds.

3.1 Offline Maintenance Task Knowledge Structuring

The knowledge representation pipeline runs once per manual revision cycle and consists of two complementary processes:

Revision-robust Task Embedding. To minimize re-indexing frequency across manual revisions, we construct task embeddings from stable semantic components: ATA chapter hierarchy titles concatenated with final task titles (e.g., "Landing Gear → Brake System → Gear Brake → Removal"). We exclude task procedural text (which changes frequently across revisions).

Manual-to-Knowledge Conversion. We extract structured task-level representations from PDF manuals using a vision-language model that captures verbatim text and layout information. Given the manuals' clear hierarchical structure (Section → Sub-task → Step), the extracted text is transformed into structured records using rule-based parsing that preserves original identifiers and metadata, requiring only minimal post-editing. Technicians can preview these structured tasks before accessing the procedure in the certified viewer.

3.2 Manual Retrieval Pipeline Architecture

When a technician submits a query (*e.g.*, "Landing gear is not retracting properly"), the system retrieves the top-N semantically relevant candidate tasks from the embedding database.

LLM-assisted Re-ranking. To refine accuracy, the LLM receives a structured prompt containing: (1) the technician's natural-language query, (2) the top-50 candidate tasks with their ATA IDs, hierarchy paths, and titles, and (3) an instruction to output only a JSON-formatted array of re-ranked indices based on semantic relevance (e.g., [3, 15, 7, 1, ...]). This strict output format prevents content generation and hallucination—the LLM performs ranking only, without access to or ability to modify procedural content.

Fail-safe Fallback. If the LLM fails to return valid JSON, the system defaults to the baseline dense retrieval rankings, ensuring robustness in safety-critical workflows.

Structured Task Presentation. The Top-N re-ranked tasks are presented with their ATA IDs, titles, and metadata. This reflects real-world MRO workflows where technicians routinely review 5-10 similar tasks due to functional overlap across subsystems (e.g., "Brake Valve Removal" may exist in multiple brake assemblies), enabling them to identify the correct task before accessing the certified OEM viewer.

4 Experiments

Our evaluation followed a two-phase design to progressively validate the proposed system under increasing realism. Phase 1 established controlled baseline performance using synthetic benchmark queries, while Phase 2 validated real-world simulated utility through a human study with practicing maintenance technicians.

4.1 Synthetic Benchmark

Synthetic Query Generation. We constructed a large-scale benchmark using publicly available Boeing 737 AMM and FIM indices (wtruib, 2024a,b), covering 8,229 tasks across major aircraft systems. Using GPT-4o (OpenAI et al., 2024), we generated six queries per task in both full-sentence and keyword styles. To simulate field conditions, we created typo-injected variants, resulting in 49,643 evaluation queries total. Query design was informed by experienced AMTs to reflect realistic workplace search patterns.

Evaluation Metric. We report Hit@k (k = 1, 5)—the percentage of queries where the ground-truth task appears within the top-k results. We focus on Hit@5 as the primary metric, reflecting operational requirements identified through AMT interviews: technicians routinely review 5-10 candidate procedures due to functional overlap between similar tasks (e.g., "Brake Assembly Removal" across multiple landing gear positions). While Hit@1 measures exact-match precision, Hit@5 captures the system's ability to deliver a manageable candidate set—the actual deployment criterion. The automated scale of synthetic evaluation allows us to measure both metrics, with Hit@5 performance >90% indicating reliable operational utility.

4.2 Human Study Design

Participants. We conducted a controlled study with 10 licensed aircraft maintenance technicians currently employed at a commercial airline in Korea. The participant group comprised technicians with diverse experience levels ranging from 1 to 10 years, including both junior technicians and senior experts, ensuring the generalizability of our findings across different skill levels commonly found in real-world MRO operations.

Experimental Protocol. We designed a controlled retrieval evaluation using 10 AMM maintenance tasks spanning diverse ATA chapters (landing gear, fuel systems, flight controls, etc.) and action types (removal, installation, inspection, lubrication). To simulate the airline's cloud-based PDF viewer environment, we deployed a web-based interface that mirrored their operational workflow: query submission, ranked task preview, and direct PDF access for verification.

Participants were provided with official AMM task titles (e.g., "Escape Slide Pack and Cover Removal") and instructed to reformulate them into natural workplace language without directly copying. For example, a participant might query "how to remove escape slide" or "slide pack cover disassembly procedure." This tested the system's ability to bridge the semantic gap between certified documentation and technicians' everyday phrasing.

Each participant completed ten retrieval tasks twice—once in English and once in Korean—resulting in 197 searches total. This bilingual design evaluated cross-lingual performance, as the knowledge base contained only English ATA structures and task titles. A multilingual embedding model (BGE-M3 (Chen et al., 2024)) enabled Korean queries to retrieve English task embeddings in the same semantic space, with Qwen3-8B-FP16 (Yang et al., 2025) used for re-ranking.

For each query, the system presented the top-10 ranked candidate tasks with metadata (ATA ID, title, chapter). Participants clicked on candidates to open the corresponding AMM PDF pages directly in the viewer, replicating the intended deployment workflow where technicians preview ranked results before accessing certified procedures in their existing system. They verified whether the correct target task appeared within the top-10 results and recorded the outcome as Success or Failure.

The system automatically logged task completion times from query submission to final verification. Participants also reported their estimated times for locating the same manuals using conventional workplace methods and during their early-career (junior) period, enabling comparative analysis across experience levels. Details of the system implementation are provided in the Appendix.

Evaluation Metrics. We collected: (1) Retrieval success rate—whether the target task appeared within the top-5 and top-10 results. (2) Task completion time (TCT, from query submission to final verification in PDF viewer). (3) Cross-lingual performance (English vs. Korean accuracy and TCT). (4) Comparative time efficiency (system TCT vs. self-reported manual lookup times for current and junior-level experience).

Unlike synthetic evaluation where Hit@1 and Hit@5 can be precisely measured, the human study prioritizes ecological validity: technicians interact with the system as they would in deployment, reviewing the top-5 and top-10 ranked list and clicking through to verify the correct procedure—mirroring real-world operational workflow.

Table 1: Task retrieval accuracy (Hit@k) across manual types and query conditions. Hit@5 (bold) represents the primary operational metric, as technicians routinely review 5-10 candidates in practice. Hit@1 is reported for reference but understates operational utility due to functional overlap between similar tasks.

Model	Overall		AMM		AMM-typo		FIM		FIM-typo	
	Hit@1	Hit@5	Hit@1	Hit@5	Hit@1	Hit@5	Hit@1	Hit@5	Hit@1	Hit@5
BM25	46.79	73.06	54.68	87.57	32.50	63.38	66.46	90.26	49.01	73.63
Dense Retrieval	60.65	85.34	66.94	90.59	49.29	78.84	66.91	89.49	59.88	82.69
Llama3.3-70B	**79.24**	91.64	**79.23**	91.75	**76.76**	88.96	**78.06**	93.05	**82.96**	**92.89**
Qwen3-32B	78.25	**91.81**	78.18	**92.70**	76.22	**89.20**	76.79	**93.10**	81.82	92.33
Qwen3-14B	77.65	91.58	78.42	92.68	74.86	88.96	76.44	92.78	80.87	92.00
Phi4-14B	77.91	90.38	77.96	91.04	76.10	87.93	76.61	91.44	80.97	91.21
Qwen3-8B	76.41	90.81	76.80	91.75	73.99	88.27	75.05	91.78	79.82	91.52
Qwen3-4B	72.58	91.02	73.25	92.13	66.49	88.08	74.17	92.55	76.61	91.43

4.3 Synthetic Benchmark Analysis

Quantitative Analysis. Table 1 presents retrieval accuracy across 49,643 evaluation samples. With LLM re-ranking (Grattafiori et al., 2024), the system achieves 91.64% Hit@5, representing a 6.3 percentage-point improvement over the nomic dense retrieval baseline. (Nussbaum et al., 2025).

Notably, compact models maintain near-identical performance to large-scale counterparts, confirming that 4B-8B models are sufficient for practical MRO deployment enabling reduced computational cost without measurable degradation in retrieval quality. Performance patterns across manual types reveal that FIM queries consistently achieve slightly higher Hit@5 compared to AMM queries, which can be attributed to FIM's more focused fault-isolation vocabulary, in contrast to the broader procedural scope of AMM content.

The system demonstrates robust performance under noisy input conditions—a critical requirement for field operations where technicians input queries under time pressure or adverse conditions. Under typo-injected query conditions, LLM-based re-ranking maintains over 88% Hit@5 across all manual types, whereas lexical baselines degrade sharply. For example, BM25 (Robertson and Zaragoza, 2009) drops from 87.57% on clean AMM queries to 63.38% under typo perturbations.

Failure Case Analysis. Despite achieving over 90% Hit@5, the evaluation exposes a consistent failure mode involving tasks with near-identical titles but divergent procedural content. Such cases arise, for instance, in cleaning tasks that differ in maintenance stage, tooling, or execution steps—differences not captured by title-based or metadata-only representations. This highlights a key limitation of title-centric embeddings in procedurally dense maintenance domains. These cases suggest that incorporating fine-grained procedural representations beyond task titles is necessary for further improving retrieval robustness.

Table 2: Overall and cross-lingual retrieval performance in the human study. The knowledge base contains only English-language manuals.

	English	Korean	Overall
Top-5 SR (%)	88.7	84.0	86.3
Top-10 SR (%)	95.9	86.0	90.9
95% CI	89.9–98.4	77.9–91.5	86.0–94.1
TCT (s)	14.2	22.2	18.0

4.4 Human Study Analysis

Overall, the system achieved 90.9% top-10 success rate (179/197 queries, 95% CI: 86.0–94.1%) and 86.3% top-5 success rate (170/197 queries, 95% CI: 80.8-90.4%), with mean Task Completion Time (TCT) of 18.0 seconds (95% CI: 12.5–23.6s). The 90.9% top-10 success rate aligns with synthetic Hit@5 performance (91.64%), validating that controlled benchmark evaluation translates to real-world operational utility. The quantitative results are shown in Table 2

Cross-lingual Performance. As shown in Table 2, cross-lingual performance revealed a 9.9-point gap between English (95.9% top-10 SR) and Korean (86.0% top-10 SR) queries, with mean TCTs of 14.2 and 22.2 seconds, respectively. This disparity reflects both embedding limitations and the linguistic characteristics of aviation maintenance: certified terminology is standardized in English, and technicians commonly use English task names in practice. Several participants reported that Korean phrasing felt less natural and often reverted to English terminology when formulating precise queries. Nevertheless, the 86.0% Korean success rate demonstrates

Table 3: Time efficiency gains compared to traditional manual lookup

Metric	Experienced	Junior
Traditional Method	6.35 min	15.41 min
Our System	$\sim$0.30 min	$\sim$0.30 min
Time Reduction	95.3%	98.1%
Absolute Savings	6.1 min	15.1 min

Table 4: Cumulative distribution of task completion times (successful queries)

Time Threshold	Success Rate (Cumulative)
$\leq$ 10 seconds	57.5%
$\leq$ 20 seconds	79.3%
$\leq$ 30 seconds	88.3%
$\leq$ 60 seconds	96.6%

that multilingual embeddings provide a viable path forward, even in domains where English remains the dominant operational language.

Time Efficiency Gains. Compared to conventional manual lookup methods, our system delivered substantial efficiency improvements. As shown in Table 3, traditional lookup required an average of 6.35 minutes for experienced AMTs and 15.41 minutes for juniors. In contrast, our system reduced lookup times to approximately 18 seconds on average, corresponding to a 95.3% reduction for experienced and 98.1% for junior AMTs—absolute time savings of 6.1 and 15.1 minutes, respectively.

Task Completion Time Distribution. As shown in Table 4, among successful queries, 57.5% were resolved within 10 seconds, 79.3% within 20 seconds, 88.3% within 30 seconds, and 96.6% within 60 seconds. These results indicate that the majority of lookups can be completed in real time, supporting the suitability of the system for operational deployment in maintenance environments.

Operational Impact. The combination of high retrieval accuracy (90.9%) and dramatic time reduction (>95%) directly addresses the critical inefficiency identified by domain experts—namely, that manual lookup can consume up to 30% of technician work time. While our evaluation measured the time required to locate a single manual entry, technicians typically perform multiple lookups during a maintenance session, so the cumulative savings scale proportionally. These findings provide strong empirical evidence that our compliance-preserving system delivers substantial productivity gains while safeguarding the precision and regulatory compliance required in aviation MRO operations.

Failure Case Analysis. While the system achieves high accuracy, the human study reveals two primary failure patterns. First, retrieval failures occur when technicians issue information-sparse shorthand queries, often combining position codes and abbreviated component names with only partial task intent. For example, a technician searched for "l2 ceiling pnl remove" (where "L2" denotes the Left Door 2 location) when seeking the certified task "Aft Entry Ceiling Panel Removal." Because the query provides limited procedural context beyond a location cue and an abbreviated noun phrase, the system faces substantial lexical and semantic underspecification during matching. Similarly, "apu fuel drain mast" failed to retrieve "APU Fuel Feed Line Shroud Drain Mast Installation" because key component qualifiers (e.g., "feed line shroud") were omitted, leaving multiple plausible procedures. These cases highlight an inherent limitation under information-sparse queries: without sufficient contextual signals, retrieval models cannot reliably disambiguate technician intent in time-pressured maintenance settings.

Second, cross-lingual failures in Korean queries arise from translation ambiguity and code-switching. Aviation-specific English terms lack standardized Korean equivalents, leading technicians to alternate between semantic translations and phonetic borrowings. For example, "Escape Slide Pack and Cover Removal" was retrieved using the semantic translation "비상탈출 슬라이드 커버," but failed when queried as "이스케잎 슬라이드 교환," which introduces both orthographic variation and potential intent drift ("교환" vs "removal"). More broadly, technicians intermix borrowed terms (e.g., "브레이크," "밸브") with native or descriptive translations (e.g., "제동장치," "차단밸브"), creating high lexical variability for the same underlying concept. This variability can weaken cross-lingual embedding alignment and reduce the effectiveness of downstream LLM-based reranking.

These failure patterns highlight the challenge of bridging the gap between informal workplace terminology and formal certified documentation, underscoring the need for context-aware retrieval mechanisms that can handle position codes, abbreviations, and cross-lingual code-switching.

Table 5: Text extraction accuracy of Qwen 2.5-VL on A320 and B737 family PDF-based aircraft maintenance manuals.

Dataset	Precision↑	Recall↑	F1↑	CER↓
A320-Family	99.39	99.82	99.57	1.14
B737-Family	99.64	99.27	99.45	2.57
Total	99.51	99.54	99.51	1.85

4.5 Knowledge Structuring Quality

To validate the offline knowledge structuring pipeline, we evaluated the vision-language parsing accuracy on production manuals. Using Qwen 2.5-VL-72B (Bai et al., 2025) with rule-based post-processing, we achieved >99% precision/recall with <3% character error rate across 20 A320 and 20 B737 manuals, as shown in Table 5. These results confirm the VLM as a dependable, low-overhead front end for automated text extraction across aircraft types with minimal post-editing. This fidelity is crucial for maintaining knowledge-base integrity, the foundation for downstream retrieval and re-ranking. Detailed extraction methodology is provided in Appendix E.

5 Conclusion

We present a compliance-preserving retrieval system that enables AMTs to locate maintenance tasks using natural-language queries without modifying certified systems. Our evaluation demonstrates over 90% retrieval accuracy across both synthetic benchmarks (>90% Hit@5 on 49k queries) and real-world validation (90.9% top-10 success rate with 10 licensed AMTs in bilingual English/Korean queries), reducing lookup time by over 95%—from 6-15 minutes to approximately 18 seconds. These results validate the practical utility of LLM-augmented retrieval in safety-critical MRO workflows while maintaining full regulatory compliance. Future work should incorporate procedural context to resolve ambiguities between similar task titles, and extend capabilities to multimodal queries and cross-document linking for a comprehensive MRO cognitive assistant.

Limitations

While our study demonstrates the feasibility of LLM-assisted task retrieval for certified aircraft manuals, several limitations should be noted. First, the human evaluation was conducted in a controlled lab setting using a mock viewer rather than an operational maintenance environment. As a result, factors such as line-maintenance time pressure, interface latency, and device constraints were not fully captured. Baseline lookup times were also self-reported because the airline's certified viewer could not be instrumented, and thus may differ from actual performance. Second, the current prototype supports only text-based queries. Although effective for desktop use, real-world deployment will require multimodal interfaces—particularly voice input—to support hands-free operation. Finally, the retrieval pipeline does not yet incorporate technician context (e.g., task stage, aircraft configuration), which limits disambiguation when multiple procedures share similar titles. These limitations outline a clear path for transitioning the system from a research prototype toward operational MRO deployment.

Acknowledgments

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No.RS-2022-00155915, Artificial Intelligence Convergence Innovation Human Resources Development (Inha University)). We extend our sincere gratitude to Sungmin Son, an Aircraft Maintenance Engineer, for his enthusiastic participation in the human study and for his valuable assistance in coordinating the study. We also thank all participating maintenance engineers who dedicated their time to this research despite their demanding operational commitments.

References

Sinan Abdulhak, Wayne Hubbard, Karthik Gopalakrishnan, and Max Z. Li. 2024. Chatatc: Large language model-driven conversational agents for supporting strategic air traffic flow management. In *International Conference on Research in Air Transportation*. ArXiv:2402.14850.

Katrina B. Avers, William B. Johnson, Joy O. Banks, and Brenda Wenzel. 2012. Technical documentation challenges in aviation maintenance: A proceedings report. Technical Report DOT/FAA/AM-12/16, Federal Aviation Administration, Office of Aerospace Medicine, Washington, DC.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei

Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. Qwen2.5-vl technical report. *Preprint*, arXiv:2502.13923.

Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2318–2335, Bangkok, Thailand. Association for Computational Linguistics.

Aircraft Commerce. 2018. Aircraft analysis & fleet planning — issue no. 121: The 737 max.

FAA Regulations. Title 14 Code of Federal Regulations § 43.13(a): Performance rules (general). https://www.ecfr.gov/current/title-14/section-43.13.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Geun-Sik Jo, Kyeong-Jin Oh, Inay Ha, Kee-Sung Lee, Myung-Duk Hong, Ulrich Neumann, and Suya You. 2014. A unified framework for augmented reality and knowledge-based systems in maintaining aircraft. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 28, pages 2990–2997.

Zach Nussbaum, John Xavier Morris, Andriy Mulyar, and Brandon Duderstadt. 2025. Nomic embed: Training a reproducible long context text embedder. *Transactions on Machine Learning Research*.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, and 401 others. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389.

Myles Taylor. 2008. TATEM—Technologies and Techniques for New Maintenance Concepts (Publishable Summary).

Archana Tikayat Ray, Anirudh P. Bhat, Ryan T. White, Van Minh Nguyen, Olivia J. Pinon Fischer, and Dimitri N. Mavris. 2023. Examining the potential of generative language models for aviation safety analysis: Case study and insights using the aviation safety reporting system (asrs). *Aerospace*, 10(9):770.

Nur Tuğçe. 2025. Enhancing aviation maintenance training through augmented reality: A case study on ar-based engine maintenance simulation. *International Journal for Multidisciplinary Research*, 7(1):1–12.

Liya Wang, Jason Chou, Alex Tien, Xin Zhou, and Diane Baumgartner. 2024. Aviationgpt: A large language model for the aviation domain. In *AIAA AVIATION Forum and ASCEND 2024*.

wtruib. 2024a. Aircraft Maintenance Manual Boeing 737 Documentation.

wtruib. 2024b. Fault Isolation Manual (FIM) B737-800 CHAPTER LIST.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Hong Yang, Aidan LaBella, and Travis Desell. 2022. Predictive maintenance for general aviation using convolutional transformers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 12636–12642.

A System Implementation

The evaluation system was deployed on an NVIDIA Titan RTX GPU (24GB VRAM). We employed the BGE-M3 embedding model for zero-shot cross-lingual retrieval over 7,834 pre-embedded AMM tasks. A Flask-based web interface presented the top-10 candidate tasks with clickable PDF links, and Qwen3-8B-FP16 (Yang et al., 2025) was used for re-ranking.

B Additional Retrieval Performance Analysis

We provide detailed retrieval performance for each LLM used in our re-ranking pipeline. This analysis demonstrates the robustness and consistency of re-ranking performance under different candidate retrieval sizes.

Specifically, we report re-ranking accuracy under retrieval candidate pool sizes of $k = 10, 20, 30, 40, 50$, across the following LLMs:

- LLaMA3.3-70B (Table 6)
- Qwen3-32B (Table 7)
- Qwen3-14B (Table 8)
- Qwen3-8B (Table 9)
- Qwen3-4B (Table 10)
- Phi-4-14B (Table 11)

These tables collectively illustrate that re-ranking accuracy consistently improves as the retrieval candidate pool size increases, demonstrating the stability and scalability of our retrieval + re-ranking pipeline across different LLMs in real-world MRO scenarios.

Table 6: llama3.3-70B Re-ranking Performance.

Retrieval Top-k	**Re-ranking k**				
	k=1	k=2	k=3	k=4	k=5
10	76.11	83.79	86.28	87.43	88.18
20	77.92	85.82	88.43	89.64	90.35
30	78.65	86.58	89.25	90.49	91.16
40	79.08	86.92	89.58	90.81	91.49
50	79.24	87.13	89.81	90.98	91.64

Table 7: Qwen3-32B Re-ranking Performance.

Retrieval Top-k	**Re-ranking k**				
	k=1	k=2	k=3	k=4	k=5
10	75.41	83.22	85.84	87.12	87.94
20	77.16	85.35	88.04	89.36	90.23
30	77.80	86.09	88.93	90.27	91.04
40	78.11	86.45	89.33	90.76	91.54
50	78.25	86.74	89.60	91.00	91.81

Table 8: Qwen3-14B Re-ranking Performance.

Retrieval Top-k	**Re-ranking k**				
	k=1	k=2	k=3	k=4	k=5
10	75.14	83.27	85.92	87.17	87.99
20	76.76	85.14	87.91	89.33	90.08
30	77.29	85.82	88.68	90.05	90.83
40	77.57	86.17	89.05	90.47	91.32
50	77.65	86.43	89.33	90.75	91.58

Table 9: Qwen3-8B Re-ranking Performance.

Retrieval Top-k	**Re-ranking k**				
	k=1	k=2	k=3	k=4	k=5
10	74.68	83.05	85.74	87.07	87.87
20	76.05	84.75	87.61	89.01	89.84
30	76.26	85.34	88.24	89.59	90.44
40	76.52	85.58	88.53	89.97	90.74
50	76.41	85.55	88.50	89.98	90.81

Table 10: Qwen3-4B Re-ranking Performance.

Retrieval Top-k	**Re-ranking k**				
	k=1	k=2	k=3	k=4	k=5
10	69.31	81.31	85.07	86.78	87.83
20	71.83	83.14	87.09	88.78	89.77
30	72.02	84.06	87.85	89.55	90.52
40	72.40	84.33	88.15	89.90	90.89
50	72.58	84.56	88.32	90.02	91.02

Table 11: Phi4-14B Re-ranking Performance.

Retrieval Top-k	**Re-ranking k**				
	k=1	k=2	k=3	k=4	k=5
10	71.20	81.98	85.21	86.74	87.74
20	77.02	84.60	87.15	88.54	89.46
30	77.51	85.06	87.70	89.11	90.02
40	77.70	85.41	88.06	89.43	90.30
50	77.91	85.72	88.22	89.56	90.38

C Synthetic Query Generation Details

We provide the system prompts used to generate realistic, diverse English queries that aircraft maintenance technicians (AMTs) might enter to locate a specific task in either the Aircraft Maintenance Manual (AMM) or the Fault Isolation Manual (FIM), including typo-focused variants ensures transparency and reproducibility of our synthetic query generation process.

C.1 Prompt Used for AMM Query Generation

```
You are an assistant helping to create a question-answer dataset for an Aircraft
Maintenance Manual (AMM) reference system. Aircraft Maintenance Manual (AMM) structure:
- Chapter: A major subject area, following ATA iSpec 2200 format (e.g., Landing Gear, Engine)
- Subchapter: A subdivision of the chapter that groups related content or components
- Subject: A specific topic or component within the subchapter
- Task Group: A collection of related maintenance tasks
- Task: A specific, action-oriented maintenance procedure

Given:
- Chapter Name: {chapter_name}
- Subchapter Name: {sub_chapter_name}
- Subject Name: {subject_name}
- Task Group Name: {task_group_name}
- Task Title: {task_name}

Generate diverse, realistic English search queries that a technician might enter to locate
this Task in the AMM. Use a variety of different question formats and structures. Create
realistic search patterns that technicians would actually use:

1. Full-sentence queries (3 examples):
   - Use varied question formats, such as:
     * "How to remove landing gear wheel"
     * "What's the procedure for engine oil change"
     * "Steps for brake pad replacement"
   - Vary the starting phrases
   - Keep these concise as technicians typically write brief queries

2. Keyword-based queries (3 examples):
   - For tasks with long titles, use only the most important parts (e.g., "wheel removal"
     instead of "Main Landing Gear Wheel - Removal/Installation")
   - Include variations with:
     * Just the main component (e.g., "oil filter")
     * Component + action (e.g., "replace brake pads")
     * Partial matches and abbreviations where appropriate (e.g., "MLG wheel install")
   - Avoid perfect, complete queries that use the entire task title

Make these queries authentic - as if a real technician is quickly typing at a keyboard
while working on an aircraft.
```

C.2 Prompt Used for AMM Typo Query Generation

```
You are an assistant helping to create a question-answer dataset for an Aircraft
Maintenance Manual (AMM) reference system.

Aircraft Maintenance Manual (AMM) structure:
- Chapter: A major subject area, following ATA iSpec 2200 format (e.g., Landing Gear, Engine)
- Subchapter: A subdivision of the chapter that groups related content or components
- Subject: A specific topic or component within the subchapter
- Task Group: A collection of related maintenance tasks
- Task: A specific, action-oriented maintenance procedure

Given:
- Chapter Name: {chapter_name}
- Subchapter Name: {sub_chapter_name}
- Subject Name: {subject_name}
- Task Group Name: {task_group_name}
- Task Title: {task_name}

Generate diverse, realistic English search queries WITH COMMON ERRORS that a technician
might enter to locate this Task in the AMM. We already have perfectly typed queries
in our dataset, so focus ONLY on creating queries that contain various types of
```

```
realistic typing mistakes.

Each entry should be a JSON object with:
- "question": the query text with realistic errors

Create realistic search patterns with errors that technicians would actually make:

1. Full-sentence queries with errors (3 examples):
   - Include common typing errors such as:
     * Transposed letters (e.g., "How to remvoe landing gear wheel")
     * Missing letters (e.g., "Whats the procedre for engine oil chage")
     * Wrong letters (e.g., "Stepps fpr brake pad replasement")
   - Spacing issues:
     * Missing spaces (e.g., "howto remove landinggear wheel")
     * Extra spaces (e.g., "what's  the procedure  for engine oil change")
     * Run-on words (e.g., "stepsfor brakepads replacement")
   - Capitalization inconsistencies or all lowercase

2. Keyword-based queries with errors (3 examples):
   - Include technical terms with common misspellings:
     * Component name errors (e.g., "landin gear weel" or "oil filtr")
     * Action word errors (e.g., "replce brake pads" or "instal wheel")
   - Spacing and abbreviation errors:
     * Run-together technical terms (e.g., "MLGwheel")
     * Incorrect abbreviations (e.g., "MGL" instead of "MLG")
     * Inconsistent spacing with hyphens or slashes

Remember that technicians might be:
- Typing quickly on tablets or mobile devices
- Working with dirty/greasy hands or wearing gloves
- In poorly lit areas or awkward positions
- Distracted by the maintenance environment
- Using speech-to-text that misinterprets technical terms

The errors should be realistic but should still allow the search to
function - queries should remain recognizable and related to the task.
```

C.3 Prompt Used for FIM Query Generation

```
You are an assistant helping to create a question-answer dataset for a Fault Isolation Manual
(FIM) reference system.

Fault Isolation Manual (FIM) structure:
- Chapter: A major subject area, following ATA iSpec 2200 format (e.g., TIME LIMITS/MAINTENANCE CHECKS)
- Subchapter: A subdivision of the chapter that groups related content or components
- Task: A specific fault isolation procedure

Given:
- Chapter Name: {chapter_name}
- Subchapter Name: {subchapter_name}
- Task Title: {task_title}

Generate diverse, realistic English search queries that a technician might enter to locate
this Task in the FIM. Use a variety of different question formats and structures.

Each entry should be a JSON object with:
- "question": the query text

Create realistic search patterns that technicians would actually use:

1. Full-sentence queries (3 examples):
   - Use varied question formats, such as:
     * "How to fix lightning damage"
     * "What causes extreme dust condition"
     * "Steps for lightning strike inspection"
     * "Procedure for dust troubleshooting"
   - Vary the starting phrases
   - Keep these concise as technicians typically write brief queries

2. Keyword-based queries (3 examples):
   - For tasks with long titles, use only the most important parts (e.g., "lightning fault" instead
```

```
   of "Lightning Strike - Fault Isolation")
   - Include variations with:
     * Just the main problem (e.g., "dust condition")
     * Problem + action (e.g., "lightning troubleshooting")
     * Partial matches and abbreviations where appropriate
   - Avoid perfect, complete queries that use the entire task title

Make these queries authentic - as if a real technician is quickly typing at a keyboard
while working on an aircraft.
```

C.4 Prompt Used for FIM Typo Query Generation

```
You are an assistant helping to create a question-answer dataset for a Fault Isolation Manual (FIM)
reference system.

Fault Isolation Manual (FIM) structure:
- Chapter: A major subject area, following ATA iSpec 2200 format (e.g., TIME LIMITS/MAINTENANCE CHECKS)
- Subchapter: A subdivision of the chapter that groups related content or components
- Task: A specific fault isolation procedure

Given:
- Chapter Name: {chapter_name}
- Subchapter Name: {subchapter_name}
- Task Title: "{task_title}"

Generate diverse English search queries WITH REALISTIC ERRORS that a technician might enter to
locate this Task in the FIM. We already have perfectly typed queries in our dataset, so focus ONLY
on creating queries with various types of errors.

Each entry should be a JSON object with:
- "question": the query text with realistic errors

Create search patterns with various types of errors that technicians might make:

1. Full-sentence queries with errors (3 examples):
   - Common typing errors:
     * Transposed letters (e.g., "ligthning" instead of "lightning")
     * Missing letters (e.g., "lightng damage")
     * Extra letters (e.g., "lighttning damagee")
     * Wrong letters (e.g., "loghting damafe")
   - Spacing issues:
     * Missing spaces (e.g., "howto fix lightningdamage")
     * Extra spaces (e.g., "how  to fix  lightning damage")
     * Inconsistent spacing (e.g., "how tofix lightning damage")
   - Capitalization errors (e.g., all lowercase or inconsistent caps)

2. Keyword-based queries with errors (3 examples):
   - Misspelled technical terms (e.g., "lightening strike" or "dust conditon")
   - Phonetic spelling errors (e.g., "lytning")
   - Abbreviations mixed with spelling errors
   - Run-on words or fragmented phrases

Errors should be realistic and reflect how technicians might actually type when:
- Working quickly on a maintenance task
- Using mobile devices with small keyboards
- Working with gloves or dirty hands
- Using speech-to-text that misinterprets technical terms
- Working in noisy or poorly lit environments

Important: The errors should be realistic but should still allow the search to function
(queries should remain recognizable and related to the task).

Make these queries authentic - as if a real technician is quickly typing at a keyboard
while working on an aircraft.
```

C.5 Query Output Examples

Below is one example set of queries generated for the task "Wing Dry Bay Tank Vapor Seal - Leak Check" (AMM):

- How do I perform a leak check on the wing dry bay tank vapor seal?
- What are the steps for inspecting the wing dry bay tank vapor seal?

- Where can I find the procedure for a vapor seal leak check on the wing dry bay tank?
- vapor seal inspection
- wing dry bay tank check leak
- check vapor seal on wing

D Query Prompt and Output Details

We detail the prompt used for LLM-based re-ranking of top-k retrieved AMM tasks. Given a user query and a list of retrieved documents with metadata (task title, chapter, subchapter, similarity scores), the LLM is instructed to select and return the indices of the top-5 most relevant tasks in order of relevance, ensuring alignment with maintenance context. The LLM returns:

```
You are an expert assistant for Aircraft Maintenance Manual (AMM) ranking.
You will be given a user question and a list of retrieved AMM documents with similarity scores.
Your task is to rank these documents and return the numbers of the most relevant documents in the specified format.

Instructions:
1. Analyze the user's question to understand their intent.
2. Consider each document's:
   - Task title relevance to the question
   - Chapter/subchapter context appropriateness
3. Based on the question above, select the 5 most relevant items from the list.
Return the numbers of the selected items in order of relevance (most relevant first) as a JSON array.

Focus on AMM maintenance procedures, inspections, and repairs.

User Question: airworthiness limitations task precautions

Retrieved Documents (ranked by similarity):
Document 1:
Chapter: STANDARD PRACTICES
Subchapter: STANDARD PRACTICES
Subject: 20-00-00 STANDARD PRACTICES
Task Group: PB.201  STANDARD PRACTICES - MAINTENANCE PRACTICES
Task: Airworthiness Limitation Precautions

Document 2:
Chapter: FLIGHT CONTROLS
Subchapter: FLIGHT CONTROLS
Subject: 27-09-91 FLIGHT CONTROLS SURFACES
Task Group: PB.601  FLIGHT CONTROLS SURFACES - INSPECTION
Task: Aileron - Inspection

... (48 additional retrieved documents)
```

LLM Output as Json Format (example)

```
{
  "selected_items": [1, 4, 2, 5, 3]
}
```

The JSON object above indicates that Document 1 is judged most relevant, followed by Documents 4, 2, 5, 3.
This output is directly used to display the ranked document list to technicians, allowing them to quickly access the most relevant manuals in the suggested order during their workflow.

E VLLM Utilization

We employ the Qwen 2.5-VL-72B vision–language model (Bai et al., 2025) to extract text from AMMs PDF, giving it the prompt:

```
“Please extract the text content from this PDF. Output regular text as is, but when you identify content in a table format, wrap it with \texttt{<table>} and \texttt{</table>} tags to distinguish it.
Within table tags, try to maintain the original structure while clearly indicating row and column relationships.”.
```

Then, the output is passed through a two-stage rule-based parser that first isolates section headers and segments individual subtasks.