Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

AeroSAR: Structure-Aware Retrieval-Augmented Generation for Instructional Aviation Manuals

Jingbo Zhao
Southeast University
Yang Zhang
State Key Laboratory of Air Traffic Management System
Chongchong Li
State Key Laboratory of Air Traffic Management System
Youyong Kong
kongyouyong@seu.edu.cn
Southeast University

Research Article

Keywords: Retrieval-augmented generation, Large language models, Structure-aware retrieval, Concept graph

Posted Date: August 5th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10441419/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

AeroSAR: Structure-Aware Retrieval-Augmented Generation for Instructional Aviation Manuals

Jingbo Zhao[1,3], Yang Zhang[1,2*], Chongchong Li[1,2], Youyong Kong[1,3*]

[1*]State Key Laboratory of Air Traffic Management System, Nanjing, 210007, Jiangsu, China.
[2]Nanjing Research Institute of Electronic Engineering, Nanjing, 210007, Jiangsu, China.
[3]Southeast University, Nanjing, 210000, Jiangsu, China.

*Corresponding author(s). E-mail(s): zhangyang16@cetc.com.cn; kongyouyong@seu.edu.cn;
Contributing authors: 220242377@seu.edu.cn; lichongchong@cetc.com.cn;

Abstract

Retrieval-augmented generation (RAG) typically treats a domain corpus as a flat collection of chunks ranked by topical similarity. This treatment is inadequate for instructional aviation manuals, in which topically similar passages can serve different pedagogical functions. We present **AeroSAR** (Aviation Structure-Aware RAG), which combines the Aviation Pedagogical Structure Schema (**APS²**), labeling chunks as CONCEPT, CONSTRAINT, PROCEDURE, or REFERENCE; Role-Aware Contrastive Embedding (**RACE**), which fine-tunes a dense retriever with role-conflict hard negatives; and the Cross-Manual Concept Graph (**CMCG**), which reranks candidates using concept sharing, narrative order, role complementarity, and explicit cross-references. A Query Intent Router sends constraint-dominant queries to lexical retrieval and routes the remainder through RACE and CMCG. In experiments, AeroSAR achieves strongest retrieval performance, while yielding consistent end-to-end gains across different generators. Ablations show cumulative component benefits and identify two-hop concept co-occurrence as the strongest graph signal.

Keywords: Retrieval-augmented generation, Large language models, Structure-aware retrieval, Concept graph

1 Introduction

Large language models have become capable open-domain question-answering engines, but in specialized professional domains their parametric memory remains an unreliable foundation: it is opaque, hard to update, and prone to fluent but unsupported answers. Retrieval-augmented generation (RAG) addresses this limitation by grounding generation in an external corpus of authoritative documents, retrieving evidence at inference time, and conditioning the generator on that evidence [1, 2]. Professional question answering is a natural application because correct answers must be traceable to trusted sources rather than synthesized from memorized statistical patterns. Aviation question answering is a representative high-stakes instance of this setting: its language is terminology-dense, its rules are explicit, and its operational constraints are strict. The instructional foundation of U.S. pilot training rests on several manuals published by the Federal Aviation Administration (FAA), including the Pilot's Handbook of Aeronautical Knowledge (PHAK), the Airplane Flying Handbook (AFH), the Aeronautical Information Manual (AIM), and the Instrument Procedures Handbook (IPH) [3–6]. Unlike ordinary reports or encyclopedic articles, these manuals are organized around learning, operation, and judgment tasks, so how a fact is presented is as informative as the fact itself.

This observation exposes a mismatch between how instructional manuals are written and how RAG systems process them. A conventional pipeline splits the corpus into a flat collection of chunks and ranks them by topical similarity, whether through lexical overlap or a dense semantic encoder [1, 7]. Yet within an instructional aviation manual, different passages serve different pedagogical functions. Some passages explain a concept or the mechanism behind a phenomenon; some state a numerical threshold, regulatory limit, or prohibition; some prescribe an ordered sequence of actions or a checklist; and some exist mainly to direct the reader to another chapter, figure, or manual. Crucially, the same aviation topic recurs across the four manuals in different roles. The topic of *stall recovery*, for example, may be explained as an aerodynamic phenomenon in the PHAK, presented as a recovery maneuver in the AFH, expressed as a regulatory or compliance condition in the AIM, and mentioned in a cross-reference in the IPH. A retriever that ranks passages only by topical similarity cannot reliably distinguish these functions and may rank an explanatory passage above one that specifies an operational limit. The resulting context concerns the right subject but serves the wrong function, weakening the usability and traceability of the generated answer.

The corrective principle we adopt is simple: effective retrieval from instructional manuals must match not only *topic* but also *role*. A query rarely asks for everything about a topic; it implicitly demands a particular kind of evidence. "What is the angle of attack?" requires an explanation, whereas "May I exceed the maximum flap-extended speed?" requires a constraint. Because the desired role is latent in the query and intrinsic to each chunk, we contend that chunk role should be treated as an explicit structural signal throughout retrieval rather than left to emerge as a byproduct of topical embeddings. This reframes instructional-manual RAG from a one-dimensional similarity-ranking problem into a two-dimensional matching problem over topic and pedagogical function.

We operationalize this principle in **AeroSAR** (Aviation Structure-Aware RAG), a structure-aware retrieval-and-generation framework with three methodological layers and a query-side router. The first layer is the Aviation Pedagogical Structure Schema (APS^2), a chunk-level ontology of four roles—CONCEPT, CONSTRAINT, PROCEDURE, and REFERENCE—that gives each passage an explicit structural coordinate. The second layer is Role-Aware Contrastive Embedding (RACE), which fine-tunes a dense retriever using role-conflict hard negatives so that topically similar but role-mismatched passages are separated in representation space. This approach extends the hard-negative tradition of dense retrieval [8, 9] from purely topical negatives to role-aware ones. The third layer is the Cross-Manual Concept Graph (CMCG), a heterogeneous graph over chunk and concept nodes that encodes concept sharing, narrative order, role complementarity, and explicit cross-references across the four manuals and reranks dense candidates strictly within their original pool. Finally, a lightweight Query Intent Router estimates the role-want distribution of each query and dispatches constraint-dominant queries to a robust lexical channel based on BM25 [10], while routing the remaining queries through the RACE encoder and CMCG reranker. Figure 1 summarizes how these components combine at inference time.

Our contributions are threefold. **First**, we formulate structure-aware retrieval for instructional aviation manuals as a problem of jointly matching topic and pedagogical role, and we make the case that role is a first-class retrieval signal rather than a byproduct of semantic similarity. **Second**, we propose APS^2, a four-way chunk-role ontology that captures the pedagogical function of a passage and can be assigned at scale through a multi-model voting protocol. **Third**, we design AeroSAR, a structure-aware framework that couples RACE (role-aware contrastive fine-tuning) with CMCG (a cross-manual concept-graph reranker) under a role-want query router.

The remainder of this paper is organized as follows. Section 2 reviews related work on retrieval-augmented generation, lexical and dense retrieval, contrastive training, graph-based RAG, and query routing. Section 3 formalizes the structure-aware retrieval problem and details the APS^2 ontology, the query intent router, RACE, and CMCG. Section 4 describes the AeroQA benchmark [11], knowledge-base and graph construction, baselines, evaluation metrics, and implementation. Section 5 reports the retrieval, role-alignment, end-to-end, ablation, sensitivity, and qualitative results. Section 6 discusses when role awareness helps, threats to validity, and generalization, and Section 7 concludes.

2 Related Work

2.1 Retrieval-Augmented Generation

RAG couples a parametric generator with a non-parametric retriever so that knowledge-intensive tasks can be grounded in an external corpus and the generator can condition its answer on retrieved evidence rather than rely only on memorized parameters [1, 2]. Subsequent work has refined nearly every stage of the pipeline. Self-RAG learns to decide when to retrieve and to critique its own generations with reflection tokens [12]; RAPTOR builds a recursive tree of summaries so that retrieval can operate at multiple levels of abstraction over long documents [13]. These methods

enrich retrieval control or granularity but do not explicitly model the pedagogical functions of passages. AeroSAR is complementary: it types each passage by its pedagogical function and makes that type an explicit retrieval signal.

2.2 Lexical and Dense Retrieval

Lexical retrieval, epitomized by BM25, scores documents through weighted term overlap and remains a strong, calibration-free baseline whenever answers hinge on exact terminology, numerals, or named entities [10]. Dense retrieval instead embeds queries and passages into a shared vector space and ranks by similarity: Dense Passage Retrieval established the dual-encoder recipe for open-domain QA [7], ColBERT introduced late interaction to retain token-level granularity at scale [14], Contriever showed that contrastive pre-training yields competitive unsupervised retrievers [15], and general-purpose embedding suites such as BGE/C-Pack now provide strong off-the-shelf backbones [16]. Such retrievers are routinely assessed for out-of-domain generalization on heterogeneous zero-shot benchmarks such as BEIR [17]. We adopt the `BGE-large-en-v1.5` encoder as the common backbone for all dense methods in our study. Because standard dense objectives emphasize semantic proximity, they can pull together passages that discuss the same subject even when those passages serve different pedagogical roles. AeroSAR retains a dense backbone for its semantic strengths but reshapes its objective and reranks its output so that role informs the final order alongside topic.

2.3 Contrastive Learning and Hard Negatives

Contrastive objectives underpin most modern text and retrieval encoders; SimCSE demonstrated how simple positive construction yields high-quality sentence embeddings [18], while a sustained line of retrieval work shows that the choice of *negatives* largely determines retriever quality. ANCE samples hard negatives globally from a periodically refreshed approximate-nearest-neighbor index to escape the weak gradients of in-batch negatives [8], and RocketQA combines cross-batch negatives, denoised hard-negative mining, and data augmentation to further sharpen the boundary between relevant and irrelevant passages [9]. These methods mine negatives that are topically confusable. RACE inherits this machinery but redefines the notion of a hard negative: its most informative negatives are passages that are topically close to the positive yet carry a *different* APS^2 role, so the encoder is explicitly trained to separate same-topic, different-role evidence rather than relying only on generic topically confusable negatives.

2.4 Graph-Based and Structure-Aware RAG

A growing body of work injects graph structure into RAG to capture relations that flat retrieval misses. Microsoft's GraphRAG builds an entity–relation graph and community summaries to answer global, corpus-level questions [19], while HippoRAG draws on a neurobiological analogy, running personalized PageRank over a knowledge graph to emulate associative long-term memory [20]. These approaches demonstrate that graph topology can surface evidence that pure vector similarity overlooks, but

they do not explicitly target the pedagogical function of instructional passages. Our CMCG departs from this tradition in two ways: its chunk nodes carry APS^2 roles, and it includes a `role_complement` relation linking same-topic, different-role passages across manuals. A graph-RAG baseline that applies personalized PageRank to the same CMCG underperforms the full AeroSAR system, suggesting that graph topology alone does not account for the gains from role-aware training and query routing.

2.5 Query Understanding, Routing, and Domain QA

The idea that queries carry latent intent has a long history in information retrieval, from taxonomies of search goals [21] to query-expansion techniques that rewrite a query to improve recall. HyDE, for instance, prompts an LLM to draft a hypothetical answer document and retrieves against its embedding, improving zero-shot dense retrieval [22]. Such methods reshape the query's *semantic* footprint but remain agnostic to the functional role of the evidence being sought, and as our experiments show they do not by themselves resolve the role heterogeneity of instructional manuals. A parallel line of work makes retrieval *adaptive* by routing each query to a different strategy: FLARE decides *when* to retrieve by anticipating upcoming low-confidence content [23], and Adaptive-RAG trains a lightweight classifier to route queries to no-retrieval, single-step, or multi-step pipelines according to their estimated complexity [24]. These routers select *how much* retrieval to perform, but they still treat the corpus as role-homogeneous. AeroSAR's Query Intent Router instead predicts a *role-want* distribution—the pedagogical role of the evidence sought by the query—and uses it to choose between a lexical channel and a structure-aware semantic channel, so that the routing decision is tied to the kind of evidence required rather than its quantity. To our knowledge, modeling chunk-level pedagogical role as an explicit end-to-end retrieval signal for instructional technical manuals, and routing queries by their role-want, has not been studied in prior RAG or domain-QA work.

3 Methodology

3.1 Problem Formulation

We formalize structure-aware retrieval over instructional manuals as follows. The knowledge base is a set of N chunks $\mathcal{C} = \{c_1, \ldots, c_N\}$ obtained by segmenting the four FAA manuals. Each chunk c has three associated attributes: its text, its provenance (source manual and chapter position), and a pedagogical role $r(c) \in \mathcal{R}$ drawn from the APS^2 ontology

$$\mathcal{R} = \{\text{CONCEPT, CONSTRAINT, PROCEDURE, REFERENCE}\}.$$

From the chunk texts we extract and normalize a vocabulary of aviation concepts $\mathcal{T}$, where a concept $t \in \mathcal{T}$ may be mentioned by many chunks and a chunk may mention many concepts.

A query q is associated with a set of ground-truth reference chunks $\mathcal{C}^+(q) \subseteq \mathcal{C}$ that suffice to answer it. Conventional retrieval seeks a ranking that places $\mathcal{C}^+(q)$ as high as

possible under a relevance score, typically a topical-similarity function $\mathrm{sim}(q, c)$. This objective is *role-blind*: two chunks with equal $\mathrm{sim}(q, c)$ are treated as interchangeable even if one explains a concept and the other states a constraint. Our central claim is that, for instructional manuals, role must enter the objective explicitly.

To that end, every query induces a *role-want* set

$$\mathcal{R}^+(q) = \{\, r(c) : c \in \mathcal{C}^+(q) \,\},$$

the pedagogical roles represented in the query's reference evidence, together with its soft counterpart, the role-want distribution $p(r \mid q)$. A structure-aware retriever returns a ranked list whose top-k prefix $\mathcal{L}_k(q)$ should be simultaneously (i) topically relevant, so that $\mathcal{C}^+(q)$ is recalled, and (ii) role-aligned, so that the roles present in $\mathcal{L}_k(q)$ cover $\mathcal{R}^+(q)$. We quantify the first goal with standard retrieval metrics and the second with Role-Recall (Section 4.4). The components introduced below operationalize this twofold objective: APS^2 supplies $r(c)$, the router estimates $p(r \mid q)$, RACE reshapes $\mathrm{sim}(q, c)$ to distinguish role-conflicting passages, and CMCG reranks candidates using concept and structural associations. The notation is summarized in Table 1.

Table 1 Summary of notation.

Symbol	Meaning
$\mathcal{C} = \{c_1, \ldots, c_N\}$	chunk set (knowledge base) and its size N
$r(c) \in \mathcal{R}$	APS^2 pedagogical role of chunk c
$\mathcal{R}$	role set {CONCEPT, CONSTRAINT, PROCEDURE, REFERENCE}
$\mathcal{T}$, t	concept vocabulary and a single concept
q, $\mathcal{C}^+(q)$	query and its ground-truth reference chunks
$\mathcal{L}_k(q)$	retrieved top-k list for query q
$\mathcal{R}^+(q)$, $p(r \mid q)$	role-want set and role-want distribution of q
$\mathrm{sim}(q, c)$	dense (cosine) similarity between query and chunk
$\mathcal{S}_q$, K	RACE top-K dense seed set and its size
τ, K'	contrastive temperature and hard negatives per positive
$\mathcal{G} = (\mathcal{V}_c \cup \mathcal{V}_t, \mathcal{E})$	CMCG: chunk nodes, concept nodes, edges
$\omega(e)$	base weight of edge e
α	dense/graph mixing coefficient in the final score

3.2 Framework Overview

At a high level, AeroSAR rests on the observation that the FAA's PHAK, AFH, AIM, and IPH are not a homogeneous collection of interchangeable passages but a corpus with a clear division of labor organized around instructional goals. One passage may explain a concept; another may prescribe an operating procedure, state a limiting condition, or merely point to another section. If a retriever orders passages by surface topical similarity alone, it tends to conflate passages that explain a phenomenon with those that specify operating limits, which reduces both the usability and the traceability of downstream generation.

AeroSAR comprises three methodological layers. First, APS^2 defines the four-way role ontology for instructional aviation manuals, so that a chunk is no longer an untyped point in topical-similarity space. Second, RACE adjusts the dense representation through role-aware contrastive learning, so that passages that are semantically close but role-mismatched become easier to distinguish. Third, CMCG models cross-manual concept complementarity over chunk and concept nodes, so that candidate reranking can exploit the fact that the same topic serves different pedagogical functions in different manuals. Figure 1 shows how these components operate online: constraint-dominant queries enter the BM25 channel, while the remaining queries enter RACE retrieval and are reranked by CMCG within the candidate set; the generation module then uses the retrieved top-k chunks to produce an answer.

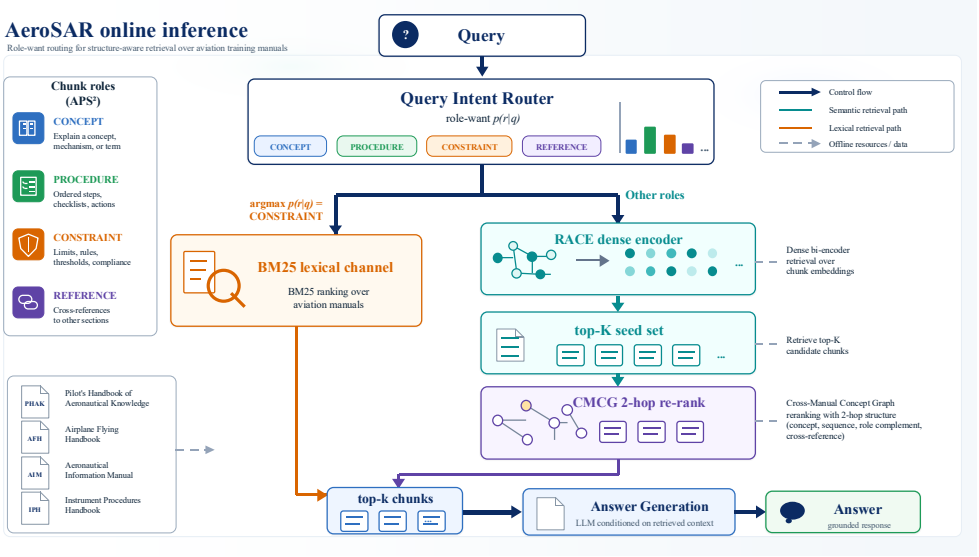


Fig. 1 The AeroSAR online retrieval-and-generation flow. Given an input query, the system first estimates the required role distribution and then selects either a lexical-matching channel or a structure-aware semantic channel; the returned top-k chunks provide context for answer generation.

3.3 APS^2: Role Ontology for Instructional Manuals

The core structural assumption of AeroSAR is that a passage in an instructional aviation manual has not only a topical attribute but also a pedagogical role. The PHAK primarily explains concepts, the AFH concentrates on flight procedures, the AIM presents both regulatory constraints and informational references, and the IPH deals extensively with instrument procedures and their applicability conditions. This division of labor implies that the same topic may appear in different roles across manuals.

Content related to *stall recovery*, for instance, may explain the aerodynamic principle in the PHAK, present recovery actions in the AFH, state a limit or compliance condition in the AIM, and appear as a cross-reference in the IPH.

Building on this observation, we propose the Aviation Pedagogical Structure Schema (APS^2), a chunk-level four-way role ontology for instructional aviation manuals, summarized in Table 2.

Table 2 The APS^2 chunk-role ontology for instructional aviation manuals.

Role	Meaning
CONCEPT	Passages that explain an aviation concept, a system mechanism, the meaning of a term, or the cause of a phenomenon.
CONSTRAINT	Passages that give limiting conditions, thresholds, applicability boundaries, compliance requirements, warnings, or prohibitive rules.
PROCEDURE	Passages that describe time-ordered operating steps, checklists, action sequences, or task flows.
REFERENCE	Passages whose primary function is to point to another section, figure, or manual.

The purpose of APS^2 is not to replace semantic similarity but to give it an interpretable structural coordinate. Two chunks may discuss the same aviation topic yet address different information needs; conversely, two chunks whose topics are not identical may still be equally useful for a class of queries because they share the same procedural or constraint role. AeroSAR therefore treats $r(c)$ as an intrinsic chunk attribute and uses it in graph construction, negative sampling, and query routing.

To obtain $r(c)$, we adopt a labeling protocol based on independent judgments from three models followed by voting-based fusion. Each model assigns one APS^2 label to a chunk under a shared role definition; unanimous labels are accepted directly, while disagreements are resolved by majority vote or, when no majority exists, a predefined conservative priority rule. The purpose of this protocol is to obtain a sufficiently stable role layer for subsequent structural modeling rather than to treat annotation itself as a primary contribution. Both RACE and CMCG use the resulting chunk-role labels.

3.4 Query Intent Router

Different queries require evidence with different roles. A question asking what a concept means usually requires CONCEPT evidence; a question asking whether an action is permitted or what maximum applies usually requires CONSTRAINT evidence; and a question asking how to perform an action depends more on PROCEDURE evidence. AeroSAR therefore introduces a Query Intent Router before retrieval to estimate the role-want scores $p(r \mid q)$. Here, "role-want" denotes the chunk role sought by the query; it is neither a dataset question-type label nor an answer-format label for the generation stage.

The router is modeled as a lightweight classifier trained with multi-label targets. Its supervision comes from the positive chunks of a query: if the reference evidence

set of a query is $\mathcal{C}^+(q)$, then the target role set of that query is defined as

$$\mathcal{R}^+(q) = \{r(c) : c \in \mathcal{C}^+(q)\}.$$

This weakly supervised definition keeps the routing signal close to the retrieval task itself and facilitates transfer to external data without additional query-level role annotation.

Online, the router output is used to select the retrieval channel. If $\arg\max_r p(r \mid q) = \text{CONSTRAINT}$, AeroSAR uses BM25 for term matching, because constraint-type answers usually rely on explicit terminology, numerical thresholds, and normative expressions, for which exact term matching tends to be more robust [10]. For queries dominated by other roles, the system obtains a candidate set with the RACE dense encoder and further reranks it through CMCG; such queries more often require cross-sentence semantic matching or cross-manual concept complementarity.

3.5 RACE: Role-Aware Contrastive Embedding

RACE (Role-Aware Contrastive Embedding) makes the dense retriever more sensitive to role differences at the representation-learning level. Ordinary dense retrievers tend to pull topically similar passages together, but in instructional manuals topical similarity does not imply the correct evidence role. A chunk that explains the principle of a stall and a chunk that gives the steps for stall recovery may share substantial vocabulary yet correspond to different query intents. RACE aims to preserve topical semantics while teaching the model to distinguish passages with the same topic but different roles.

Let e_q and e_c denote the vector representations of a query and a chunk, and let $\mathrm{sim}(q, c)$ denote their similarity. For each positive pair (q, c^+), RACE constructs a set of hard negatives $\{c_j^-\}$. These negatives are not sampled at random; they preferentially include chunks that are topically close to c^+ but carry a different APS[2] role. The negatives come from two complementary sources: neighboring passages in the CMCG of Section 3.6 that share a concept with c^+, and passages that rank highly under term matching but are role-inconsistent. The former provides hard negatives derived from cross-manual concept structure, while the latter provides hard negatives based on lexical similarity.

The training objective uses contrastive learning. For a batch $\mathcal{B}$, let

$$\mathcal{P}_i = \{c_i^+\} \cup \{c_{i,j}^- : 1 \le j \le K'\} \cup \{c_k^+ : k \ne i\}$$

denote the candidate pool of the i-th query, which contains explicit hard negatives together with in-batch negatives formed by the other positives in the batch. The query-to-chunk loss is

$$\mathcal{L}_{q\to c}(q_i) = -\log \frac{\exp(\mathrm{sim}(q_i, c_i^+)/\tau)}{\sum_{c\in\mathcal{P}_i} \exp(\mathrm{sim}(q_i, c)/\tau)},$$

where τ is a temperature coefficient. The symmetric direction $\mathcal{L}_{c\to q}$ is obtained by swapping query and chunk, and the overall objective is

$$\mathcal{L} = \tfrac{1}{2}\big(\mathcal{L}_{q\to c} + \mathcal{L}_{c\to q}\big).$$

This design places the methodological emphasis on role-aware negative construction: the model is required not only to find topically relevant passages but also to avoid ranking role-conflicting passages near the top. In practice, the encoder is fine-tuned using parameter-efficient low-rank adaptation [25] on top of the shared dense backbone.

At inference time, the RACE encoder first returns a top-K dense candidate set $\mathcal{S}_q$ for the query. This set is both the candidate pool for the subsequent CMCG reranking and the scope over which the graph-structural signal acts; thus RACE determines the initial semantic neighborhood that structural reranking can correct.

3.6 CMCG: Cross-Manual Concept Graph

APS^2 provides the role attribute of a chunk, but the role itself is still a local signal. To exploit concept complementarity across manuals, we build the Cross-Manual Concept Graph (CMCG), a heterogeneous graph composed of both chunk nodes and concept nodes:

$$\mathcal{G} = (\mathcal{V}_c \cup \mathcal{V}_t, \mathcal{E}),$$

where $\mathcal{V}_c$ corresponds to the chunk set $\mathcal{C}$ and $\mathcal{V}_t$ to the concept vocabulary $\mathcal{T}$. Chunk nodes carry manual provenance, chapter position, and APS^2 role; concept nodes represent aviation concepts or terms extracted and cleaned from the manuals.

CMCG contains four edge types. First, `has_concept` edges connect a chunk to the concepts it mentions, expressing membership between a passage and a concept. Second, `sequence` edges connect adjacent chunks within the same manual, preserving narrative order. Third, `role_complement` edges connect chunks that share a concept but carry different roles, capturing the complementary treatment of one topic across pedagogical functions. Fourth, `explicit_xref` edges represent explicit cross-references in the manual text, such as a passage directing the reader to another manual or section. The four edge types correspond to concept membership, local order, role complementarity, and explicit reference, respectively, allowing the graph to express the organization of instructional manuals at a finer grain than a pure vector space.

Each edge e carries a base weight $\omega(e)$. For a `has_concept` edge, if concept t occurs $\mathrm{tf}(c,t)$ times in chunk c, then

$$\omega(c,t) = \log\big(1 + \mathrm{tf}(c,t)\big),$$

using a logarithmic scaling that preserves term-frequency strength while suppressing the dominance of high-frequency terms. The `sequence` and `role_complement` edges indicate whether a structural relation exists and take a base weight of 1; `explicit_xref` edges are fewer and act more like an auxiliary cue, so they take a bonus weight below 1.

During retrieval, CMCG does not expand freely over the whole graph; it reranks within the candidate pool returned by dense retrieval. Given a query q, the RACE

encoder first returns the top-K candidate set $\mathcal{S}_q$. We also call $\mathcal{S}_q$ the seed set, because each seed $s \in \mathcal{S}_q$ is at once a candidate answer passage and a structural evidence source for graph reranking. For any candidate chunk $c \in \mathcal{S}_q$, AeroSAR combines the dense similarity with a graph score:

$$\text{score}_{\text{final}}(q, c) = \alpha \widetilde{\text{sim}}_{\text{dense}}(q, c) + (1 - \alpha) \widetilde{\text{graph}}(q, c),$$

where α controls the trade-off between semantic similarity and the graph-structural signal, and $\widetilde{\cdot}$ denotes normalization within the candidate pool. The graph score is defined as

$$\text{graph}(q, c) = \sum_{s \in \mathcal{S}_q} w(s, c) \, \text{sim}_{\text{dense}}(q, s).$$

Here $w(s, c)$ denotes the structural association strength between seed s and candidate c in CMCG, combining direct chunk-to-chunk relations with two-hop relations through concept nodes:

$$w(s, c) = \sum_{e \in \mathcal{E}^{\text{direct}}_{s,c}} \omega(e) + \sum_{t \in T(s) \cap T(c)} \omega(s, t) \, \omega(c, t),$$

where $\mathcal{E}^{\text{direct}}_{s,c}$ is the set of direct chunk-to-chunk edges between s and c, and $T(c)$ is the set of concepts connected to chunk c. The first term preserves direct structural relations such as `sequence`, `role_complement`, and `explicit_xref`; the second captures the two-hop association formed when two chunks share concepts. Thus, a candidate passage related by concepts or complementary structure to several high-similarity seeds receives an additional ranking bonus. Restricting the candidate pool to $\mathcal{S}_q$ prevents graph expansion from introducing passages far from the query into the final context. Concrete edge counts, thresholds, and filtering rules are reported with the graph statistics in Section 4.2.

3.7 Online Inference and Computational Cost

AeroSAR cleanly separates a one-time offline indexing stage from a low-cost online stage. Offline, we (i) segment the manuals into chunks and assign APS^2 roles, (ii) encode every chunk with the RACE encoder into a dense index, (iii) build the BM25 inverted index, and (iv) materialize the CMCG together with its precomputed edge weights. None of these steps depend on the query, and all are amortized across the entire query workload.

At query time, the router encodes q and predicts $p(r \mid q)$ with a lightweight classifier. A constraint-dominant query requires a single BM25 lookup. Every other query triggers one dense nearest-neighbor search that returns the seed set $\mathcal{S}_q$ of size K, after which CMCG reranks only those K seeds. Because the graph score sums over seed–candidate pairs strictly inside $\mathcal{S}_q$, reranking requires roughly $O(K^2)$ weighted lookups over precomputed edges. With K fixed at a few dozen and the pool never growing beyond $\mathcal{S}_q$, this design bounds the graph overhead and avoids the multi-hop traversal cost of full graph-walk methods such as personalized PageRank. The retrieved top-k chunks are then passed to the generator.

4 Experimental Setup

4.1 The AeroQA Benchmark

We evaluate AeroSAR on **AeroQA**, which contains 2,588 queries spanning three question types: multiple-choice and true/false items, which have a single objective answer, and open-ended (extended) items, which require a free-form explanatory answer. Every query is paired with a set of chunk-level ground-truth references $\mathcal{C}^+(q)$—the passages that are sufficient to answer it—rather than only a final answer string. This reference-level supervision is what makes structure-aware evaluation possible: because each reference chunk carries an APS^2 role, the role-want set $\mathcal{R}^+(q)$ of every query is determined directly from its references, and Role-Recall can be measured without any additional annotation.

The knowledge base is segmented from the four authoritative FAA instructional manuals listed in Table 3 [3–6], yielding 7,612 chunks, each assigned one APS^2 role label through three-model voting (Section 3.3). The four manuals differ systematically in pedagogical emphasis—the PHAK leans toward concept explanation, the AFH toward procedures, the AIM toward constraints and references, and the IPH toward instrument procedures and their applicability conditions—which is precisely the cross-manual role heterogeneity AeroSAR is designed to exploit. We emphasize that question type and role-want are distinct: a multiple-choice item may still require a CONSTRAINT chunk (for example, a question asking which option exceeds a maximum), so the router's target is not a restatement of the question format. To prevent train–test leakage at the chunk level, the development and test partitions are formed using a grouped split on `(handbook, chunk_idx)`, so that no chunk used as positive supervision during RACE training is reused as a ground-truth reference during evaluation.

Table 3 The four FAA instructional manuals that constitute the AeroQA knowledge base, and their dominant pedagogical emphasis.

Abbr.	Full title	Dominant APS^2 emphasis
PHAK	Pilot's Handbook of Aeronautical Knowledge	CONCEPT
AFH	Airplane Flying Handbook	PROCEDURE
AIM	Aeronautical Information Manual	CONSTRAINT, REFERENCE
IPH	Instrument Procedures Handbook	PROCEDURE, CONSTRAINT

4.2 Knowledge Base and Concept Graph Construction

Each manual is parsed from its source document and segmented into passage-level chunks that respect section and paragraph boundaries, so that a chunk is a coherent unit of explanation, a step list, or a rule rather than an arbitrary fixed-width window. Every chunk retains its provenance—source manual and section position—which the

CMCG later uses for `sequence` edges. The APS[2] role layer is produced by the three-model voting protocol of Section 3.3, and the fused labels are consumed by both RACE (for role-conflict negatives) and CMCG (for `role_complement` edges).

To build the concept layer, we extract candidate aviation terms and noun phrases from the chunk texts and normalize them—lowercasing, lemmatization, and alias merging—into the concept vocabulary $\mathcal{T}$. A `has_concept` edge then links a chunk to each concept it mentions, weighted by $\log(1 + \mathrm{tf})$ as in Section 3.6. `sequence` edges connect consecutive chunks within a manual; `role_complement` edges connect chunks from different manuals that share at least three concepts but differ in role; and `explicit_xref` edges are created when a chunk explicitly directs the reader to another section or manual. The resulting graph statistics are summarized in Table 4. Both the CMCG and the BM25 index are precomputed once, offline, and reused by every online query.

Table 4 Statistics of the Cross-Manual Concept Graph (CMCG).

Element	Count
Chunk nodes $\lvert\mathcal{V}_c\rvert$	7,612
Concept nodes $\lvert\mathcal{V}_t\rvert$	994
`has_concept` edges	67,996
`sequence` edges	7,608
`role_complement` edges	18,228
`explicit_xref` edges	67

4.3 Baselines and Comparison Protocol

We compare against four retrieval baselines spanning the main paradigms while holding shared ingredients constant where possible. **BM25** [10] is the lexical baseline, indexed over the same chunks with default parameters. **Vanilla Dense** is the zero-shot `BGE-large-en-v1.5` encoder [16] with cosine similarity and no fine-tuning. **HyDE** [22] is the query-expansion baseline: an LLM drafts a hypothetical answer document whose embedding is used for dense search with the same backbone. **HippoRAG** [20] is the graph-RAG baseline, running personalized PageRank over the same CMCG topology but without RACE or role-want routing. This comparison tests whether graph propagation alone can account for AeroSAR's gains. All dense methods share the same backbone and chunk index; retrieval metrics use the same top-5 cutoff.

Our own system is reported as an incremental sequence—**RACE** (encoder fine-tuning only), **RACE+CMCG** (adding graph reranking), and **AeroSAR** (adding the query intent router)—so that each component's marginal effect is visible. For the end-to-end evaluation we use a unified comparison set—Vanilla LLM (no retrieval), Naive RAG (BM25 retriever), HyDE, HippoRAG, and AeroSAR—and run it on three open-source instruction-tuned generators of different families and sizes: Qwen2.5-7B-Instruct [26], Llama-3-8B-Instruct [27], and Gemma3-4B-Instruct [28]. Every method

feeds the generator the same prompt template with its top-10 retrieved chunks as context, so end-to-end differences are attributable to retrieval quality rather than to prompt or decoding differences.

4.4 Evaluation Metrics

All retrieval metrics use a top-k evaluation cutoff with $k = 5$. For a query q with gold reference set $\mathcal{C}^+(q)$ and retrieved top-k list $\mathcal{L}_k(q)$, we report Hit Rate, Recall, Precision, F1, MRR, and nDCG [29], averaged over the query set $\mathcal{Q}$:

$$\text{Hit@}k = \frac{1}{|\mathcal{Q}|}\sum_q \mathbf{1}\big[\mathcal{L}_k(q) \cap \mathcal{C}^+(q) \neq \emptyset\big], \quad \text{Recall@}k = \frac{1}{|\mathcal{Q}|}\sum_q \frac{|\mathcal{L}_k(q) \cap \mathcal{C}^+(q)|}{|\mathcal{C}^+(q)|},$$

$$\text{Precision@}k = \frac{1}{|\mathcal{Q}|}\sum_q \frac{|\mathcal{L}_k(q) \cap \mathcal{C}^+(q)|}{k}, \quad \text{MRR} = \frac{1}{|\mathcal{Q}|}\sum_q \frac{1}{\text{rank}_q},$$

where rank_q is the rank of the first gold chunk. We compute F1 for each query as the harmonic mean of its precision and recall and then average over $\mathcal{Q}$; this macro-averaged F1 is therefore not necessarily the harmonic mean of the reported macro-averaged Precision and Recall. We define $\text{nDCG@}k = \frac{1}{|\mathcal{Q}|}\sum_q \text{DCG}_q/\text{IDCG}_q$, where $\text{DCG}_q = \sum_{i=1}^{k} \text{rel}_i/\log_2(i+1)$ and rel_i is binary relevance.

To measure role alignment, we report **Role-Recall** per APS[2] role ρ: among queries whose role-want includes ρ, the fraction whose top-k list contains at least one chunk of that role,

$$\text{Role-Recall}_\rho = \frac{\sum_{q:\,\rho \in \mathcal{R}^+(q)} \mathbf{1}[\exists\, c \in \mathcal{L}_k(q) :\, r(c) = \rho]}{\big|\{q : \rho \in \mathcal{R}^+(q)\}\big|}.$$

Whereas Recall measures chunk-level coverage, Role-Recall measures whether the retrieved list contains the *kind* of evidence represented in the references. Because it checks role presence rather than topical relevance within that role, it should be interpreted together with the standard retrieval metrics. For generation, objective items (choice and boolean) are scored by Accuracy (exact match), and open-ended items by token-level F1 and BERTScore [30]; the component-wise ablation additionally breaks out choice-only Accuracy.

4.5 Implementation Details

All dense retrieval and the RACE fine-tuning are performed on a single NVIDIA RTX 4090D GPU, and all generators are served through the vLLM inference framework [31]. The retriever backbone is `BGE-large-en-v1.5` (1024-dimensional embeddings, maximum sequence length 512, `[CLS]` pooling, cosine similarity); RACE adapts it with parameter-efficient low-rank adaptation (LoRA) [25] rather than full fine-tuning, which keeps training on a single GPU tractable. Generation uses deterministic decoding (temperature 0.0) with the same prompt template for all methods, and CMCG

Table 5 Configuration of AeroSAR. Left block: backbone and RACE training; right block: retrieval, CMCG, and generation.

Hyperparameter	Value	Hyperparameter	Value
Encoder	`BGE-large-en-v1.5`	Temperature τ	0.02
Embedding dim.	1024	Hard neg. K'/positive	7
Max seq. length	512	Seed-set size K	50
Similarity	cosine	Evaluation cutoff k	5
LoRA rank r	8	Mixing coefficient α	0.85
LoRA α	16	`explicit_xref` bonus	0.3
LoRA dropout	0.05	Generation engine	vLLM
Optimizer	AdamW	Decoding	greedy / temp. 0.0
Learning rate	2×10^{-5}	Max tokens	400
Epochs	3	Batch size	32
Trainable params	2.56M	Context chunks	top-10

reranking uses the selected mixing coefficient $\alpha = 0.85$. Table 5 lists the full configuration.

5 Results and Analysis

5.1 Main Retrieval Results

Table 6 reports the main retrieval comparison. AeroSAR achieves the best score on every retrieval metric and consistently improves on Vanilla Dense. BM25 reveals both the strengths and limitations of lexical matching: it remains competitive on Hit@5, showing that many questions contain useful exact lexical cues, but it falls behind AeroSAR on ranking and evidence-coverage metrics. This result indicates that lexical matching alone does not reliably distinguish passages that share a topic but serve different pedagogical roles. AeroSAR therefore does not replace BM25; the query intent router sends constraint-dominant questions to BM25 and the remaining queries to the RACE+CMCG channel, combining exact matching for short, threshold-oriented questions with dense semantic matching for conceptual and procedural content.

The graph-RAG baseline HippoRAG provides a useful contrast. Although it propagates over the same CMCG topology, its overall performance is lower than AeroSAR and, on several metrics, lower than Vanilla Dense. The result suggests that graph propagation alone is insufficient: without RACE and query-side role-want routing, propagation may promote topically related passages that do not match the evidence function sought by the query. HyDE improves modestly on Vanilla Dense, but generic query expansion does not directly address the role heterogeneity of instructional manuals, so it still trails AeroSAR.

5.2 Role Alignment Analysis

Role-Recall tests whether the retrieved results include the chunk role represented in the query's references, although it does not by itself establish that the same-role chunk

Table 6 Main retrieval results on AeroQA (top-5). Best result per metric in **bold**.

Method	Hit	Recall	Precision	F1	MRR	nDCG
BM25	0.9394	0.6470	0.2970	0.3871	0.7718	0.6160
Vanilla Dense	0.8909	0.5884	0.2642	0.3468	0.7191	0.5612
HyDE	0.8970	0.6076	0.2739	0.3590	0.7374	0.5768
HippoRAG	0.8606	0.5768	0.2691	0.3497	0.6234	0.5190
RACE	0.9030	0.6197	0.2861	0.3722	0.7600	0.6000
RACE + CMCG	0.9152	0.6333	0.2945	0.3823	0.7735	0.6110
AeroSAR	**0.9455**	**0.6687**	**0.3042**	**0.3977**	**0.8135**	**0.6467**

is topically relevant. As Table 7 shows, AeroSAR outperforms both BM25 and Vanilla Dense on the CONCEPT and REFERENCE roles; RACE is strongest on CONSTRAINT, while the full model is on par with BM25 there. The distribution indicates that AeroSAR's role-coverage gains are concentrated in concept explanations and reference fragments rather than being uniform across roles.

Table 7 Role-Recall on AeroQA (top-5), decomposed by APS^2 role. Best per column in **bold**.

Method	CONCEPT	PROCEDURE	CONSTRAINT	REFERENCE
BM25	0.9344	0.8846	0.9615	0.3750
Vanilla Dense	0.9344	0.8654	0.9423	0.3750
RACE	0.9426	0.8846	**0.9808**	**0.5000**
AeroSAR	**0.9590**	0.8846	0.9615	**0.5000**

The largest absolute Role-Recall increase occurs for REFERENCE. Pure reference chunks often contain little substantive semantic content, so Vanilla Dense and BM25 may overlook them; role-conflict hard negatives make RACE more sensitive to the distinction between content-bearing chunks and chunks that point elsewhere. The improvement for CONCEPT also supports the motivation of Section 3.3: role-aware training reduces errors caused by ranking only on topical similarity. The methods differ little on PROCEDURE, suggesting that procedural passages are comparatively easy to identify from lexical and semantic cues in the current data. For CONSTRAINT, BM25 is already strong, which motivates routing short, explicit, keyword-dense constraint questions to the lexical channel rather than sending every query through the dense path.

5.3 End-to-End QA Generation

The end-to-end results in Table 8 show that the retrieval-side gains carry through to generation. Across Qwen2.5-7B, Llama-3-8B, and Gemma3-4B, AeroSAR attains the highest objective Accuracy, extended F1, and extended BERTScore. Relative to

Naive RAG and HyDE, AeroSAR improves both objective and open-ended metrics for each generator. Although the gains are modest, their consistent direction across model families is compatible with an improvement in retrieved-context quality.

Table 8 End-to-end QA across three generators. Acc is objective accuracy (choice + boolean); extended F1/BERTScore are for free-response items. Best per block in **bold**.

Method	Acc	Ext. F1	Ext. BERTScore
Qwen2.5-7B-Instruct			
Vanilla LLM (no retrieval)	0.8070	0.3880	0.3207
Naive RAG (BM25 + top-10)	0.9532	0.4683	0.3826
HyDE	0.9532	0.4585	0.3748
HippoRAG (PPR on CMCG)	0.9455	0.4508	0.3687
AeroSAR	**0.9551**	**0.4737**	**0.3869**
Llama-3-8B-Instruct			
Vanilla LLM (no retrieval)	0.7661	0.3851	0.3217
Naive RAG (BM25 + top-10)	0.9415	0.4576	0.3884
HyDE	0.9181	0.4454	0.3766
HippoRAG (PPR on CMCG)	0.8978	0.4359	0.3674
AeroSAR	**0.9549**	**0.4743**	**0.3949**
Gemma3-4B-Instruct			
Vanilla LLM (no retrieval)	0.7602	0.3264	0.2478
Naive RAG (BM25 + top-10)	0.9298	0.4301	0.3336
HyDE	0.9181	0.4246	0.3257
HippoRAG (PPR on CMCG)	0.9078	0.4203	0.3195
AeroSAR	**0.9420**	**0.4431**	**0.3479**

The closed-book Vanilla LLM trails every RAG system, showing that all three generators benefit substantially from retrieved FAA evidence. The size of AeroSAR's gain differs across generators: Qwen2.5-7B already has a strong Naive RAG baseline, whereas Llama-3-8B and Gemma3-4B, which also have lower closed-book scores in this experiment, improve more clearly with AeroSAR. This pattern suggests that retrieved-context quality matters more when the generator provides a weaker task-specific prior. HyDE and HippoRAG do not automatically resolve role mismatch: HyDE expands semantic query coverage, while HippoRAG propagates graph associations without conditioning on the query's role-want. Their gap from AeroSAR is consistent with the value of aligning the evidence functions in the retrieved context with query intent.

5.4 Ablation Studies

We assess the contribution of each component through two analyses: a component-wise build-up and a CMCG edge-type ablation. Table 9 shows incremental gains as RACE, CMCG, and the router are added to Vanilla Dense. Recall and MRR increase at every step; under a fixed Qwen2.5-7B generation setting, choice accuracy and extended F1 increase in parallel. The aligned retrieval-side and generation-side trends associate improved upstream retrieval with better downstream answers.

Table 9 Component-wise build-up ablation (Qwen2.5-7B generator fixed). "Choice Acc." is choice-only accuracy. Full system in **bold**.

Configuration	Recall	MRR	Choice Acc	Ext. F1
Vanilla Dense	0.5884	0.7191	0.8966	0.4321
+ RACE	0.6197	0.7600	0.9138	0.4498
+ RACE + CMCG	0.6333	0.7735	0.9310	0.4612
+ RACE + CMCG + Router	**0.6687**	**0.8135**	**0.9483**	**0.4737**

The stepwise results match the method design. In this build-up order, adding RACE produces the largest single-step retrieval gain, indicating that role-aware hard negatives contribute most strongly to the observed improvement. Adding CMCG yields a further gain by adjusting candidate priorities within the RACE pool. Finally, adding the router improves the metrics again, indicating that switching between BM25 and RACE+CMCG by role-want is better suited to this mixture of constraint and non-constraint questions than a single retrieval path.

The edge-type ablation in Table 10 further explains the internal mechanism of CMCG. Removing `has_concept` edges causes the largest drop, showing that two-hop concept co-occurrence is the dominant signal of the graph score; removing `role_complement` edges causes the second-largest drop, supporting the claim that the "cross-manual, same-topic, different-role" relation contributes to retrieval reranking. The effects of `sequence` and `explicit_xref` are smaller but directionally consistent, matching their roles in the method as local-order and explicit-reference bonus signals. Together, the four edge types provide complementary signals, although concept co-occurrence contributes the most.

Table 10 CMCG edge-type ablation (RACE + CMCG, α=0.85). Δ is in percentage points (pp) relative to the baseline.

Removed edge	Recall	MRR	ΔRecall	ΔMRR
Baseline (RACE + CMCG)	0.6333	0.7735	—	—
− `has_concept`	0.6273	0.7660	−0.60	−0.75
− `role_complement`	0.6303	0.7695	−0.30	−0.40
− `sequence`	0.6321	0.7720	−0.12	−0.15
− `explicit_xref`	0.6329	0.7730	−0.04	−0.05

5.5 Hyperparameter Sensitivity

The sweep of the CMCG mixing coefficient α in Table 11 shows a broad high-performance region in the dense-leaning range. At $\alpha = 1.0$, the configuration reduces to pure RACE without graph reranking, while smaller values assign greater weight to the graph score. We select $\alpha = 0.85$ because it gives the highest Recall in the sweep;

$\alpha = 0.70$ gives a slightly higher MRR. Performance remains similar across the mid-to-high range, whereas small α values give the graph score greater influence and produce markedly lower results.

Table 11 Sensitivity to the CMCG mixing coefficient α (RACE encoder + CMCG dense-pool rerank). Selected value in **bold**.

α	**Recall**	**MRR**
0.5	0.5641	0.6647
0.6	0.6035	0.7632
0.7	0.6278	0.7759
0.8	0.6308	0.7711
0.85 (selected)	**0.6333**	**0.7735**
0.9	0.6303	0.7725
0.95	0.6242	0.7703
1.0 (= pure RACE)	0.6197	0.7600

The dense-pool design of AeroSAR uses CMCG to refine candidates that already satisfy a degree of topical relevance. A dense-leaning α range therefore matches the method's intended division of labor: dense similarity supplies semantic relevance, while the graph score adds cross-manual concept and role-complementarity signals. The relatively stable results from $\alpha = 0.70$ to 0.90 also show that the conclusion does not depend on a single sharply tuned value.

5.6 Qualitative Analysis

To illustrate the role-matching mechanism, consider three queries that share aviation vocabulary but differ in role-want. For a *constraint* query such as "What is the maximum flap-extended speed (V_{FE})?", the router selects BM25, allowing the exact term V_{FE} and its numerical context to be matched directly. For a *concept* query such as "Why can an airplane stall at any airspeed and attitude?", RACE favors the PHAK explanation of critical angle of attack over a topically similar recovery procedure, while CMCG's two-hop `has_concept` paths reinforce candidates sharing the *angle of attack* and *critical angle* concepts. Conversely, for a *procedure* query such as "How do I recover from a power-off stall?", RACE favors the AFH action sequence over the aerodynamic explanation. The `role_complement` edges encode cross-role associations that can support complementary evidence; discrimination between competing target roles is supplied primarily by RACE and the router.

These examples illustrate a failure mode of flat retrieval: explanatory, procedural, and constraint passages about one topic may rank closely, allowing a role-mismatched passage to occupy a top-5 slot. They also help explain the asymmetry in Table 7: the CONSTRAINT branch is already well served by the lexical channel, whereas the CONCEPT and REFERENCE branches, for which surface terms are less diagnostic of function, benefit more from role-aware training.

6 Discussion

6.1 When Does Role-Awareness Help?

The results indicate where structure-aware retrieval is most beneficial. The benefit is not uniform across roles: CONCEPT and REFERENCE passages, whose surface terms less clearly indicate their function, gain the most, while CONSTRAINT questions are already handled well by the lexical channel. A related pattern appears in generation: Llama-3-8B and Gemma3-4B, which have lower closed-book scores in our experiment, gain more from AeroSAR than Qwen2.5-7B. Taken together, these findings suggest that role-aligned context is most valuable when lexical cues do not disambiguate function and when the generator has weaker task-specific knowledge. Structure-aware retrieval is therefore better understood as a targeted context-quality mechanism than as a uniform accuracy multiplier.

6.2 Threats to Validity

Several factors bound the scope of our claims. First, AeroQA is built from four English FAA manuals; generalization to other languages, regulators, or technical domains remains to be demonstrated. Second, the APS^2 role layer is produced by multi-model voting rather than expert adjudication, so labeling noise and model bias may propagate into role-conflict negatives and `role_complement` edges. Reporting model-level agreement and conducting expert adjudication would strengthen this layer. Third, reference-based ground truth can miss valid but unlabeled supporting passages and thereby underestimate some retrievers. Role-Recall has the complementary limitation that it measures the presence of a desired role, not whether the retrieved same-role passage is topically relevant. Fourth, retrieval is evaluated at top-5, whereas generation uses a fixed top-10 context; behavior under other context budgets remains untested. Fifth, the experiments use one grouped split and report no confidence intervals or repeated-training variation, so small differences should be interpreted cautiously. Finally, no wall-clock latency or memory measurements are reported, and the end-to-end gains are modest for the strongest generator.

6.3 Generalization and Deployment

AeroSAR is designed so that most computation occurs offline: chunk encoding, BM25 indexing, and CMCG materialization are amortized, while the online path adds a lightweight router and within-pool reranking to standard dense retrieval (Section 3.7). This architecture should permit integration with an existing dense-retrieval stack, although deployment cost still requires direct measurement. More broadly, the formulation is not intrinsically aviation-specific: an explicit pedagogical-role layer, role-conflict contrastive training, a cross-document role-complement graph, and a role-want router could apply to corpora whose passages serve distinct didactic or regulatory functions. Medical guidelines, legal codes, and industrial maintenance manuals are therefore candidate domains for future evaluation, and the router could be extended with additional role-specialized channels.

7 Conclusion

We revisited retrieval-augmented generation for instructional aviation manuals and argued that a passage carries both a topic and a pedagogical role. AeroSAR instantiates this principle through three coupled components: the APS^2 ontology, which labels each chunk as CONCEPT, CONSTRAINT, PROCEDURE, or REFERENCE; RACE, which fine-tunes a dense retriever with role-conflict hard negatives; and CMCG, a cross-manual concept graph that reranks dense candidates within their original pool. A role-want query router dispatches constraint-dominant queries to a lexical channel and the remainder to the structure-aware semantic channel. On AeroQA's 7,612-chunk FAA knowledge base, AeroSAR achieves the best score on every reported retrieval metric, improves Role-Recall most clearly for CONCEPT and REFERENCE, and yields consistent end-to-end gains across Qwen2.5-7B, Llama-3-8B, and Gemma3-4B. The ablations show incremental benefits from the three components and identify two-hop concept co-occurrence as the strongest graph signal.

Looking ahead, the structure-aware formulation invites several extensions. A role tagger learned jointly with the retriever could replace the voting-based APS^2 layer and close the loop between role labeling and retrieval; a controlled, role-constrained expansion beyond the dense seed pool could recover complementary evidence that the initial dense stage misses; and the role-want router could grow into a multi-channel dispatcher with role-specialized retrievers. Because none of these directions are aviation-specific, we view AeroSAR primarily as evidence for a general principle—that in instructional and regulatory corpora, retrieval should match pedagogical role as well as topic—and we expect the same recipe to transfer to other high-stakes technical domains such as medical, legal, and industrial maintenance documentation.

Declarations

Funding

This work was supported by the Open Fund of the State Key Laboratory of Air Traffic Management System (Grant No. SKLATM202404).

Competing Interests

The authors have no relevant financial or non-financial interests to disclose.

Data Availability

The AeroQA benchmark data used in this study are publicly available from https://github.com/Frost1325/RADAR. No new dataset was created as part of this work. The source code is not publicly available because its release is subject to restrictions associated with the funding project.

Author Contributions

Jingbo Zhao: Conceptualization, investigation, methodology, software, validation, formal analysis, visualization, and writing—original draft. Yang Zhang and Chongchong Li: Data curation, validation, and writing— review and editing. Youyong Kong: Conceptualization, supervision, project administration, funding acquisition, and writing—review and editing. All authors read and approved the final manuscript.

References

[1] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., Riedel, S., Kiela, D.: Retrieval-augmented generation for knowledge-intensive NLP tasks. In: Advances in Neural Information Processing Systems (NeurIPS) (2020)

[2] Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, M., Wang, H.: Retrieval-augmented generation for large language models: A survey. arXiv preprint arXiv:2312.10997 (2023)

[3] Federal Aviation Administration: Pilot's Handbook of Aeronautical Knowledge (FAA-H-8083-25). Washington, DC, USA (2023). Federal Aviation Administration. U.S. Department of Transportation

[4] Federal Aviation Administration: Airplane Flying Handbook (FAA-H-8083-3). Washington, DC, USA (2021). Federal Aviation Administration. U.S. Department of Transportation

[5] Federal Aviation Administration: Aeronautical Information Manual (AIM): Official Guide to Basic Flight Information and ATC Procedures. Washington, DC, USA (2024). Federal Aviation Administration. U.S. Department of Transportation

[6] Federal Aviation Administration: Instrument Procedures Handbook (FAA-H-8083-16). Washington, DC, USA (2017). Federal Aviation Administration. U.S. Department of Transportation

[7] Karpukhin, V., Oğuz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., Yih, W.-t.: Dense passage retrieval for open-domain question answering. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 6769–6781 (2020)

[8] Xiong, L., Xiong, C., Li, Y., Tang, K.-F., Liu, J., Bennett, P.N., Ahmed, J., Overwijk, A.: Approximate nearest neighbor negative contrastive learning for dense text retrieval. In: International Conference on Learning Representations (ICLR) (2021)

[9] Qu, Y., Ding, Y., Liu, J., Liu, K., Ren, R., Zhao, W.X., Dong, D., Wu, H.,

Wang, H.: RocketQA: An optimized training approach to dense passage retrieval for open-domain question answering. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT), pp. 5835–5847 (2021)

[10] Robertson, S., Zaragoza, H.: The probabilistic relevance framework: BM25 and beyond. Foundations and Trends in Information Retrieval **3**(4), 333–389 (2009)

[11] AeroQA. https://github.com/Frost1325/RADAR. GitHub repository, accessed 13 July 2026 (2026)

[12] Asai, A., Wu, Z., Wang, Y., Sil, A., Hajishirzi, H.: Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In: International Conference on Learning Representations (ICLR) (2024)

[13] Sarthi, P., Abdullah, S., Tuli, A., Khanna, S., Goldie, A., Manning, C.D.: RAPTOR: Recursive abstractive processing for tree-organized retrieval. In: International Conference on Learning Representations (ICLR) (2024)

[14] Khattab, O., Zaharia, M.: ColBERT: Efficient and effective passage search via contextualized late interaction over BERT. In: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR), pp. 39–48 (2020)

[15] Izacard, G., Caron, M., Hosseini, L., Riedel, S., Bojanowski, P., Joulin, A., Grave, E.: Unsupervised dense information retrieval with contrastive learning. Transactions on Machine Learning Research (TMLR) (2022)

[16] Xiao, S., Liu, Z., Zhang, P., Muennighoff, N., Lian, D., Nie, J.-Y.: C-Pack: Packaged resources to advance general chinese embedding. In: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR) (2024)

[17] Thakur, N., Reimers, N., Rücklé, A., Srivastava, A., Gurevych, I.: BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In: Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks (NeurIPS Datasets and Benchmarks) (2021)

[18] Gao, T., Yao, X., Chen, D.: SimCSE: Simple contrastive learning of sentence embeddings. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 6894–6910 (2021)

[19] Edge, D., Trinh, H., Cheng, N., Bradley, J., Chao, A., Mody, A., Truitt, S., Larson, J.: From local to global: A graph RAG approach to query-focused summarization. arXiv preprint arXiv:2404.16130 (2024)

[20] Gutiérrez, B.J., Shu, Y., Gu, Y., Yasunaga, M., Su, Y.: HippoRAG: Neurobiologically inspired long-term memory for large language models. In: Advances in Neural Information Processing Systems (NeurIPS) (2024)

[21] Broder, A.: A taxonomy of web search. ACM SIGIR Forum **36**(2), 3–10 (2002)

[22] Gao, L., Ma, X., Lin, J., Callan, J.: Precise zero-shot dense retrieval without relevance labels. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL), pp. 1762–1777 (2023)

[23] Jiang, Z., Xu, F.F., Gao, L., Sun, Z., Liu, Q., Dwivedi-Yu, J., Yang, Y., Callan, J., Neubig, G.: Active retrieval augmented generation. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 7969–7992 (2023)

[24] Jeong, S., Baek, J., Cho, S., Hwang, S.J., Park, J.C.: Adaptive-RAG: Learning to adapt retrieval-augmented large language models through question complexity. In: Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT), pp. 7036–7050 (2024)

[25] Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: LoRA: Low-rank adaptation of large language models. In: International Conference on Learning Representations (ICLR) (2022)

[26] Qwen Team: Qwen2.5 technical report. arXiv preprint arXiv:2412.15115 (2025)

[27] Grattafiori, A., Dubey, A., et al.: The Llama 3 herd of models. arXiv preprint arXiv:2407.21783 (2024)

[28] Gemma Team: Gemma 3 technical report. arXiv preprint arXiv:2503.19786 (2025)

[29] Järvelin, K., Kekäläinen, J.: Cumulated gain-based evaluation of IR techniques. ACM Transactions on Information Systems (TOIS) **20**(4), 422–446 (2002)

[30] Zhang, T., Kishore, V., Wu, F., Weinberger, K.Q., Artzi, Y.: BERTScore: Evaluating text generation with BERT. In: International Conference on Learning Representations (ICLR) (2020)

[31] Kwon, W., Li, Z., Zhuang, S., Sheng, Y., Zheng, L., Yu, C.H., Gonzalez, J.E., Zhang, H., Stoica, I.: Efficient memory management for large language model serving with PagedAttention. In: Proceedings of the 29th Symposium on Operating Systems Principles (SOSP) (2023)