KEO: Knowledge Extraction on OMIn via Knowledge Graphs and RAG for Safety-Critical Aviation Maintenance

Anonymous ACL submission

Abstract

We present Knowledge Extraction on OMIn (KEO), a domain-specific knowledge extraction and reasoning framework with large language models (LLMs) in safety-critical contexts. Using the Operations and Maintenance Intelligence (OMIn) dataset, we construct a QA benchmark spanning global sensemaking and actionable maintenance tasks. KEO builds a structured Knowledge Graph (KG) and integrates it into a retrieval-augmented generation (RAG) pipeline, enabling more coherent, dataset-wide reasoning than traditional text-chunk RAG. We evaluate locally deployable LLMs (Gemma-3, Phi-4, Mistral-Nemo) and employ stronger models (GPT-4o, Llama-3.3) as judges. Experiments show that KEO markedly improves global sensemaking by revealing patterns and system-level insights, while text-chunk RAG remains effective for fine-grained procedural tasks requiring localized retrieval. These findings underscore the promise of KG-augmented LLMs for secure, domain-specific QA and their potential in high-stakes reasoning.

Code: https://anonymous.4open.science/r/keo-9101.

1 Introduction

Large Language Models (LLMs) have demonstrated significant potential across various sectors, but they face challenges when applied to highly specialized and safety-critical domains such as aviation maintenance (Peykani et al., 2025; Zhang et al., 2025). These models are often limited by their reliance on parametric knowledge stored within their training data, which can lead to factual inaccuracies, outdated information, and an inability to handle domain-specific, real-world scenarios (Hu and Lu, 2024).

A prevailing strategy to address these knowledge deficiencies is Retrieval-Augmented Generation (RAG). However, conventional RAG, which relies on embedding-based retrieval of unstructured text chunks, often results in redundant and fragmented context, hindering performance on tasks requiring complex, multi-hop reasoning or global sensemaking (Ram et al., 2023; Shi et al., 2022). This fragmentation is particularly problematic in safety-critical environments, where the lack of verifiable and transparent context can compromise trust and reproducibility (Mealey et al., 2025).

To establish secure and trustworthy AI systems, research has pivoted toward Knowledge Graph (KG)-enhanced RAG. Prior work has shown that integrating KGs, which model knowledge as structured entities and relations (Hogan et al., 2021; Jiang et al., 2023), can effectively mitigate core LLM issues like hallucinations and factual inconsistencies by providing verifiable, domain-specific grounding (Wagner et al., 2025). Techniques leveraging relational structures and graph traversal have been explored to enhance reasoning in high-stakes areas like healthcare and risk analysis (Zhou et al., 2025; Bahr et al., 2025). Crucially, this structured approach enhances explainability and enables smaller language models to achieve performance comparable to that of larger models, supporting secure deployment in sensitive environments without reliance on external APIs (Perez-Cerrolaza et al., 2024).

To address these challenges, we introduce KEO (Knowledge Extraction on OMIn), a novel framework that integrates structured knowledge representation with RAG to enhance domain-specific knowledge extraction and reasoning. Our work is grounded in the aviation maintenance domain, leveraging the Operations and Maintenance Intelligence (OMIn) dataset (Mealey et al., 2024) to create a new Question Answering (QA) benchmark that evaluates both global sensemaking and fine-grained procedural tasks.

Our core contribution is a methodology that constructs a KG from the OMIn dataset and integrates

it into an RAG pipeline. This KG-augmented RAG approach enables more coherent, dataset-wide reasoning, allowing LLMs to uncover complex relationships and system-level insights that are difficult to extract using traditional text-chunk RAG methods. We systematically evaluate several locally deployable LLMs, and our experiments demonstrate that KEO significantly improves performance on complex reasoning tasks, while text-chunk RAG remains effective for localized, procedural QA. These findings highlight the potential of KG-augmented LLMs for secure, accurate, and domain-specific QA, paving the way for their responsible deployment in high-stakes reasoning environments.

2 Related Work

2.1 OMIn

The OMIn dataset provides a GS for maintenance data (Mealey et al., 2024) and emphasizes trust and reproducibility due to its focus on sensitive aviation and military information. Given the safety-critical nature of this domain, it is essential that all AI tools for knowledge extraction operate locally without reliance on external APIs (Perez-Cerrolaza et al., 2024). The dataset consists of 2,748 Aviation Incident records from the Federal Aviation Administration (FAA) (Federal Aviation Administration, 2024), each ranging from one to three sentences. Previous work has shown that NLP and LLM performance (e.g., zero-shot F1 scores) on OMIn is low because of its domain-specific content (Mealey et al., 2025). This suggests that the technology readiness levels (TRLs) (Mankins, 1995) of current tools for this data are at a low level (TRLs 1-2). Our goal is to enhance this TRL by applying a KG and RAG approach.

2.2 Enhancing LLMs with RAG

RAG enhances LLMs by incorporating external information beyond their context window, allowing models to remain lightweight while compensating for limited parametric memory. Approaches differ in how retrieval results are integrated: appending retrieved documents to prompts (Ram et al., 2023), injecting them into hidden layers via cross-attention (Borgeaud et al., 2022), or leveraging memory layers for entity-specific representations (De Jong et al., 2021; Févry et al., 2020). Others combine outputs with token-level distributions over nearest-neighbor tokens (Shi et al., 2022). Despite these variations, most rely on embedding-based retrieval of text chunks. Such *text-chunk RAG* often produces redundant, unstructured context, limiting support for complex reasoning. In contrast, our method KEO employs KGs to provide structured, dataset-wide context, enabling more coherent sensemaking.

2.3 Knowledge Graphs for RAG

KGs represent structured knowledge as collections of entities and relations, offering a flexible framework to model complex dependencies (Hogan et al., 2021; Jiang et al., 2023). In the context of LLMs, KGs serve as critical external knowledge sources to mitigate inherent limitations such as knowledge cutoff, factual inconsistencies, and hallucinations (Hu and Lu, 2024). By providing structured, verified knowledge, KGs can enhance factual accuracy, reduce hallucinations, and improve reasoning, while also offering explainability and domain-specific grounding without retraining (Wagner et al., 2025).

Recent work has explored integrating KGs into RAG to enhance factuality and reasoning. This includes leveraging relational structures (Procko and Ochoa, 2024; Gao et al., 2023; Fan et al., 2024), supporting KG construction and completion through triple extraction (Melnyk et al., 2022; Trajanoska et al., 2023), link prediction (Yao et al., 2025), and causal discovery (Zhang et al., 2024; Ban et al., 2023). KG-enhanced RAG methods have investigated subgraph or relational-path prompting (Baek et al., 2023; He et al., 2024; Sen et al., 2023), grounding outputs (Ranade and Joshi, 2023; Kang et al., 2023), and retrieval via graph traversal (Wang et al., 2024b). These approaches have been applied in domains such as healthcare (Zhao et al., 2025), customer service (Xu et al., 2024), and risk analysis (Bahr et al., 2025), where reasoning over structured knowledge is critical.

By incorporating KGs into RAG, LLMs can move beyond simple text-chunk retrieval, enabling more robust reasoning for complex, multi-hop queries that require navigating structured dependencies (Zhou et al., 2025). While GraphRAG (Edge et al., 2024) employs graph community summarization, our system KEO instead tailors KG navigation to fragmented aviation maintenance records by embedding entity mentions, expanding with multi-hop neighbors, and applying maximum spanning tree filtering (Zhu et al., 2025).

2.4 Using LLMs for Benchmark Construction and Evaluation

Existing benchmarks for RAG span open-domain (Yang et al., 2018; Chen et al., 2024; Tang and Yang, 2024), medicine (Xiong et al., 2024), finance (Wang et al., 2024a), and multilingual tasks (Lyu et al., 2025). However, these benchmarks largely assess factual retrieval, which suits text-chunk RAG. Recent work shows that LLMs can automatically generate benchmarks involving reasoning and summarization (Lin and Chen, 2023; Yuan et al., 2024; Wang et al., 2024a). Inspired by this, KEO creates benchmarks to evaluate global sensemaking by developing questions that require reasoning across records, rather than within single passages, similar to GraphRAG (Edge et al., 2024). Additionally, KEO constructs knowledge-to-action questions to assess knowledge transferability in aviation maintenance.

For evaluation, we adopt stronger LLMs as judges, following evidence of high alignment with human ratings (Zheng et al., 2023; Gu et al., 2024; Gebreegziabher et al., 2025). Prior work has validated ChatGPT as an effective evaluator (Wang et al., 2023), and RAGAS (Es et al., 2024) further formalizes LLM-based scoring for RAG. In KEO, LLM-based evaluation spans both fact-based problem–action questions and global sensemaking tasks, combining absolute scoring with pairwise comparisons guided by structured criteria.

3 Methodology

We present KEO, a domain-specific knowledge extraction and reasoning framework with LLMs in safety-critical settings. As illustrated in Figure 1, the framework integrates four core components: (1) KG creation from raw maintenance records, (2) a KG-based RAG pipeline for producing safe and reliable answers, (3) automatic creation of an aviation maintenance QA benchmark, and (4) LLM-based evaluation to assess both factual accuracy and higher-level reasoning. Together, these components enable structured sensemaking and actionable decision support in aviation safety applications.

3.1 LLMs for KG Creation

Building a KG traditionally depends on human-annotated gold-standard labels for core knowledge extraction tasks such as named entity recognition (NER), coreference resolution (CR), named entity linking (NEL), and relation extraction (RE). However, for the OMIn dataset (Mealey et al., 2025), no gold standard exists for RE (Appendix F.1). To explore the feasibility of automating this initial step, we experimented with both a stronger external model (GPT-4o) and a weaker locally deployable model (Phi-4-mini) in a zero-shot setting. Since knowledge extraction is performed as an offline preprocessing stage, it does not pose major security risks. In contrast, the operational RAG pipeline is carefully designed for secure, local execution, and final answer generation is restricted to lightweight models that can be safely deployed in sensitive environments.

The resulting KG can be represented as a set of weighted triplets:

$$G = \{(h, t, r, w) \mid h \in V,\ t \in V,\ r \in R\} \quad (1)$$

where h, t, r, and w denote the head entity, tail entity, relation, and weight (i.e., frequency of the triplet $\langle h, r, t\rangle$ in the corpus), respectively. V is the set of all entity mentions extracted from aviation maintenance records, which is also the set of nodes in KG, and R is the predefined set of allowed relation types.

Alternatively, this graph can be expressed in standard form as $G = (V, E)$, where V is the set of nodes and $E \subseteq V \times R \times V$ is the set of directed, labeled edges. Each edge $e = (h, r, t) \in E$ corresponds to a triplet in the original formulation, and can optionally carry a weight $w(e)$ representing its frequency.

As detailed in Appendix B, we construct the KG dynamically with an LLM, iteratively prompting it with the current set of nodes when generating new triplets. This strategy reduces redundancy by minimizing duplicate nodes for the same entity mentions. To evaluate scalability, we progressively build the KG using subsets of the corpus, beginning with 100 records and increasing in steps of 100 up to 500. We evaluate how the performance changes as the size of the underlying corpus grows while ensuring that all nodes can still be accommodated within the LLM's context window.

3.2 KG-based RAG Pipeline

Given the knowledge graph $G = \{(h, t, r, w)\}$ constructed from the aviation maintenance corpus, the proposed KEO pipeline aims to support global sensemaking and generalizable knowledge-to-action retrieval through a three-step pipeline: semantic-based node identification, importance-

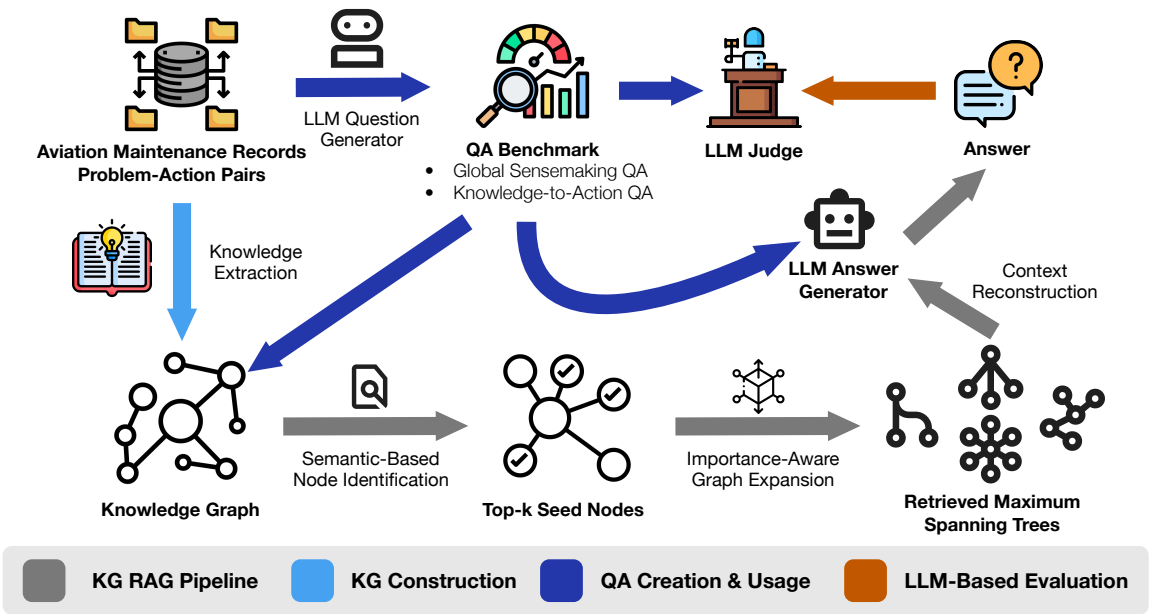

Figure 1: Overview of the KEO pipeline. Aviation maintenance records and problem–action pairs are first transformed into a QA benchmark covering both global sensemaking and knowledge-to-action questions. In parallel, KEO constructs a structured KG from raw maintenance data. A KG-based RAG workflow then leverages semantic node identification, importance-aware graph expansion, and structured context reconstruction to enhance LLM responses on these safety-critical questions. Finally, an LLM judge evaluates answers through both absolute and comparative scoring with carefully designed metrics.

aware graph expansion, and KG-based context reconstruction.

3.2.1 Semantic-Based Node Identification

In traditional *text-chunk RAG*, the corpus is divided into discrete chunks, denoted as $C = \{c_1, c_2, \ldots, c_n\}$. Each chunk is embedded into a high-dimensional vector using an embedding model, and the semantic similarity between the user query q and each chunk $c \in C$ is computed as:

$$S = \left\{ \frac{\text{emb}(q) \cdot \text{emb}(c)}{\|\text{emb}(q)\| \cdot \|\text{emb}(c)\|} \;\middle|\; c \in C \right\} \quad (2)$$

where $\text{emb}(\cdot)$ denotes the embedding function. The top-k most similar chunks are selected and concatenated into the prompt for LLM inference.

In contrast, KEO performs retrieval at the entity level rather than the chunk level. Specifically, we compute embeddings for each entity mention—that is, each node $v \in V$ in the knowledge graph—and measure semantic similarity between the user query and the graph nodes:

$$S = \left\{ \frac{\text{emb}(q) \cdot \text{emb}(v)}{\|\text{emb}(q)\| \cdot \|\text{emb}(v)\|} \;\middle|\; v \in V \right\} \quad (3)$$

A key limitation of *text-chunk RAG* is that retrieved chunks often lack intrinsic structure or semantic cohesion; they may originate from disparate parts of the corpus and present fragmented or redundant information. In contrast, our entity-centric approach retrieves the top-k most semantically relevant entities $V_k \subset V$ as seed nodes. These seeds serve as entry points for subsequent graph-based expansion and structured context reconstruction, allowing the model to reason over connected and contextually meaningful knowledge.

3.2.2 Importance-Aware Graph Expansion

The constructed KG encodes structured relationships among entities, allowing reasoning not only over direct links but also over shared neighbors. For instance, the entities *water in the fuel system* and *fuel tank sumps frozen* may not be directly connected but are both linked to *engine quit*, with edge weights of 8 and 21, respectively. This suggests that while both are plausible causes, *fuel tank sumps frozen* is more frequently observed and potentially more critical. Such patterns underscore the importance of incorporating multi-hop connectivity and edge weights when expanding from the initial seed entities.

Formally, given the initial set of top-k retrieved seed nodes $V_k \subseteq V$ and the full graph $G = (V, E)$, we define the m-hop expansion subgraph as:

$$G^{(m)} = (V^{(m)}, E^{(m)}) \quad (4)$$

where $V^{(m)}$ is the set of nodes reachable from V_k within m hops, and $E^{(m)} \subseteq E$ contains all edges connecting pairs of nodes in $V^{(m)}$. This expanded subgraph captures both direct and indirect connections, enabling the model to consider not just local matches but also structurally important contextual information.

We later apply importance-aware filtering over $G^{(m)}$ to prioritize salient and informative substructures before passing them into the context reconstruction module. To enable spanning tree computation, we first transform the directed multi-relational subgraph $G^{(m)}$ into an undirected weighted graph $U^{(m)} = (V^{(m)}, \tilde{E}^{(m)})$. Each undirected edge $\{u, v\} \in \tilde{E}^{(m)}$ is derived from its directed counterparts in $E^{(m)}$, where the edge weight is defined as the sum of the weights in both directions, and the relation label is the concatenation of the directed relations (if both exist):

$$\tilde{E}^{(m)} = \{ \left(\{u, v\},\ r',\ w'\right) \} \tag{5}$$

where $(u, v, r_{uv}, w_{uv}) \in E^{(m)}$ or $(v, u, r_{vu}, w_{vu}) \in E^{(m)}$, $w' = w_{uv} + w_{vu}$, and $r' = r_{uv} \| r_{vu}$.

Here, w_{uv} and w_{vu} are treated as zero if the corresponding directed edge is not present. The operator $\|$ denotes string concatenation. This unified representation allows for computing maximum spanning trees over undirected components while preserving relational semantics.

We then identify all l connected components in $U^{(m)}$, denoted as $\{H_1, H_2, \ldots, H_l\}$, using depth-first search (DFS). For each connected component H_i, we compute its corresponding maximum spanning tree (MST) T_i using Kruskal's algorithm:

$$T_i = \text{MST}(H_i), \quad \text{for } i = 1, 2, \ldots, l \tag{6}$$

These MSTs retain the most structurally significant links within each connected region of the expanded subgraph, ensuring that downstream reasoning is supported by a coherent, importance-aware knowledge structure.

3.2.3 KG-Based Context Reconstruction

Given the MSTs $T_1, T_2, \ldots, T_l$ obtained from the filtered subgraph $U^{(m)}$, we traverse each tree using depth-first search (DFS), starting from the node incident to the edge with the highest weight. For each visited edge, we record its corresponding entities and relation, converting the structured traversal path into textual descriptions. This process yields a KG-based context that retains the original graph structure, serving as input to the downstream LLM.

To further support global sensemaking, we augment the retrieved context with a hierarchical community summarization inspired by GraphRAG (Edge et al., 2024). Specifically, we apply community detection with the Leiden algorithm over the KG to identify densely connected subgraphs and recursively generate summaries for each community level. For leaf communities, we prioritize high-degree nodes and their relations to construct compact summaries. For higher-level communities, we compress summaries of their constituent sub-communities, ensuring a scalable and hierarchical abstraction of the KG. These summaries are concatenated with the graph traversal text to provide the LLM with both local detail and global structure.

3.3 Automatic QA Benchmark Creation

To construct a robust QA benchmark for aviation safety, we employed state-of-the-art closed-source LLMs, such as GPT-4o, to generate domain-specific questions grounded in aviation maintenance records and problem-action pairs from the OMIn dataset. The benchmark is composed of two major question types: global sensemaking questions and knowledge-to-action questions. The former are designed to require holistic reasoning beyond direct retrieval from the corpus, while the latter evaluate the transferability of knowledge derived from external sources to unseen maintenance problems. To further ensure reliability, we randomly reviewed a subset of the automatically generated questions and confirmed that they met our quality requirements, thereby validating their suitability before incorporating them into the final benchmark.

3.3.1 Global Sensemaking Questions

To prevent data leakage and ensure meaningful benchmark creation, we did not expose the question-generation LLMs to the raw maintenance records. Instead, we randomly sampled 500 of the 2,750 OMIn records to construct a knowledge graph, while the remaining 2,250 records were analyzed statistically to extract insights. These insights included failure patterns, component-level distributions, temporal trends, seasonal variations, and aircraft-specific maintenance behaviors.

Using this analytical summary as input, LLMs were prompted to generate high-level, domain-

specific questions that require synthesizing information across the dataset—thus assessing global sensemaking. We further classified these questions into three subtypes: **(1)** comprehensive questions, which require a holistic understanding of the dataset, **(2)** context-specific questions, which focus on patterns within similar maintenance scenarios, and **(3)** category-specific questions, which align with predefined analytical dimensions (e.g., failure mode, aircraft type). The detailed prompts used for each question type are provided in Appendix C. Examples of these global sensemaking questions are illustrated in Table 1.

3.3.2 Knowledge-to-Action Questions

The MaintNet dataset (Akhbardeh et al., 2020) offers targeted problem-action pairs derived from aviation maintenance procedures. To transform these into a QA benchmark, we rephrased each problem as a natural language question using the template: *"What action could be taken when: ..."*, where the corresponding action serves as the gold-standard answer.

While such questions can often be addressed using traditional RAG methods on text chunks, our objective is to evaluate whether knowledge derived from structured maintenance graphs can generalize to unseen cases. To this end, MaintNet records were excluded from the retrievable corpus. Instead, both baseline *text-chunk RAG* and our proposed KEO framework are restricted to retrieving from the OMIn dataset corpus only. This setup ensures that successful answering requires transferring structured maintenance knowledge to new, action-specific scenarios. Illustrative examples of such knowledge-to-action questions and their corresponding gold-standard answers are also shown in Table 1. Details regarding our constrained evaluation setup, including model choice and data limitations, are provided in Appendix G.

Table 1: Examples of benchmark questions and gold-standard answers of Knowledge-to-Action questions. GSM: Global SenseMaking.

Type	**Example Question (and Gold-Standard Answer if applicable)**
GSM (Comprehensive)	What recurring failure modes and seasonal trends emerge across aircraft types, and how might they inform proactive maintenance scheduling?
GSM (Context-Specific)	How do patterns of hydraulic leaks in landing gear systems vary across different aircraft classes and operational environments?
GSM (Category-Specific)	Which environmental factors are most commonly associated with electrical system failures in narrow-body aircraft?
Knowledge-to-Action	**Q:** What action could be taken when a recurring high-pressure fuel pump vibration is observed during pre-flight inspection? **A:** Replace the high-pressure fuel pump and run a fuel system vibration diagnostic.

4 Evaluation

4.1 Experimental Setup

We evaluate our method KEO on a benchmark of 133 questions, comprising 83 global sensemaking and 50 knowledge-to-action questions. To construct the KGs, we randomly sample 500 out of 2,748 OMIn records, reserving the remaining 2,248 for question generation. Global sensemaking questions are generated by GPT-4o using insights extracted from these records, while knowledge-to-action questions are created from fixed templates based on problem-action pairs from the MaintNet dataset.

We compare three different methods on this QA benchmark:

- **Vanilla LLM:** Direct few-shot prompting without any external context from the maintenance corpus.
- **Text-chunk RAG:** Retrieves top-matching text chunks from the maintenance corpus based on embedding similarity.
- **KEO (KG RAG):** Our proposed retrieval method using a KG constructed from OMIn maintenance records.

As detailed in Section 3.3, the benchmark questions span both global sensemaking and knowledge-to-action tasks, and are generated using GPT-4o, a strong LLM known for aligning well with human preferences (Achiam et al., 2023; Shankar et al., 2024; Wei et al., 2024). For KG construction, we compare GPT-4o and Phi-4-mini (3.8B, Microsoft) and vary the number of input data points from 100 to 500 to assess scalability.

To test downstream performance on aviation maintenance, we focus on lightweight, locally deployable LLMs for answer generation: Gemma-3-Instruct (27B, Google), Phi-4 (14B, Microsoft), and Mistral-Nemo-Instruct (12B, Mistral AI).

Generated answers are evaluated by two stronger LLM judges, GPT-4o (OpenAI) and Llama-3.3-Instruct (70B, Meta). The judges provide both absolute scores and pairwise comparison scores, following task-specific evaluation metrics in Appendix D and prompts in Appendix E.

Figure 2: Head-to-head win rate matrix of row method over column method (TC: text-chunk RAG, VN: vanilla LLM, KG: our method KEO) for 83 global sensemaking questions, evaluated by GPT-4o. The KG used is generated with GPT-4o from 100 records. Win rates are reported across five dimensions and overall. Green cells indicate a win, red cells indicate a loss. The proposed KEO method consistently outperforms text-chunk RAG when paired with stronger LLMs, but its performance may degrade with weaker backbone models.

Table 2: Overall evaluation by LLM-based evaluators on 83 global sensemaking questions. The knowledge graph used for KEO is the gold-standard version constructed from 100 records.

Model	Evaluator	Overall Score (1–5)		
		TC	VN	KG
gemma-3-it	GPT-4o	4.12 ± 0.42	3.70 ± 0.52	$\mathbf{4.31 \pm 0.29}$
phi-4	GPT-4o	3.90 ± 0.50	3.68 ± 0.51	$\mathbf{4.09 \pm 0.40}$
mistral-nemo-it	GPT-4o	3.84 ± 0.61	3.39 ± 0.60	$\mathbf{3.87 \pm 0.46}$
gemma-3-it	Llama-3.3-it	4.47 ± 0.21	4.29 ± 0.31	$\mathbf{4.87 \pm 0.25}$
phi-4	Llama-3.3-it	4.33 ± 0.24	4.32 ± 0.22	$\mathbf{4.44 \pm 0.17}$
mistral-nemo-it	Llama-3.3-it	4.32 ± 0.25	4.13 ± 0.42	$\mathbf{4.39 \pm 0.06}$

Table 3: Overall evaluation by LLM-based evaluators on 50 knowledge-to-action questions. The knowledge graph used for KEO is the gold-standard version constructed from 100 records.

Model	Evaluator	Overall Score (1–5)		
		TC	VN	KG
gemma-3-it	GPT-4o	3.84 ± 0.67	3.75 ± 0.80	$\mathbf{3.86 \pm 0.55}$
phi-4	GPT-4o	$\mathbf{4.17 \pm 0.46}$	4.01 ± 0.55	3.96 ± 0.49
mistral-nemo-it	GPT-4o	$\mathbf{3.72 \pm 0.66}$	3.70 ± 0.83	3.68 ± 0.81
gemma-3-it	Llama-3.3-it	$\mathbf{3.96 \pm 0.61}$	3.90 ± 0.71	3.93 ± 0.62
phi-4	Llama-3.3-it	$\mathbf{4.37 \pm 0.31}$	4.24 ± 0.43	4.18 ± 0.45
mistral-nemo-it	Llama-3.3-it	$\mathbf{4.00 \pm 0.57}$	3.78 ± 0.78	3.80 ± 0.72

Table 4: ROUGE-based quantitative evaluation on 50 knowledge-to-action questions. We report both ROUGE-1 F1 and ROUGE-L F1 scores. The knowledge graph used for KEO is the gold-standard version constructed from 100 records.

Model	Metric	TC	VN	KG
gemma-3-it	ROUGE-L	$\mathbf{0.276 \pm 0.231}$	0.273 ± 0.239	0.272 ± 0.240
	ROUGE-1	$\mathbf{0.296 \pm 0.240}$	0.293 ± 0.242	0.293 ± 0.245
phi-4	ROUGE-L	0.093 ± 0.057	0.100 ± 0.059	$\mathbf{0.106 \pm 0.069}$
	ROUGE-1	0.107 ± 0.068	0.110 ± 0.067	$\mathbf{0.124 \pm 0.077}$
mistral-nemo-it	ROUGE-L	0.192 ± 0.195	$\mathbf{0.271 \pm 0.246}$	0.268 ± 0.240
	ROUGE-1	0.213 ± 0.202	$\mathbf{0.299 \pm 0.254}$	0.292 ± 0.244

4.2 Results

Global Sensemaking Questions. As shown in Figure 2 and Table 2, KEO significantly outperforms both vanilla prompting and text-chunk RAG in global sensemaking tasks when evaluated by GPT-4o. The most notable improvements appear in the global perspective criterion and the overall evaluation score across all backbone LLMs. The advantage of KEO is especially pronounced when paired with stronger models such as Gemma-3-Instruct, suggesting that more capable LLMs benefit more

from the structured, concise context provided by the KG. In contrast, performance gains are attenuated when KEO is paired with smaller models like Mistral-Nemo-Instruct, likely due to limited reasoning ability to leverage the graph-based context.

Knowledge-to-Action Questions. In contrast, as shown in Table 3, KEO does not outperform text-chunk RAG on knowledge-to-action questions. Both RAG methods retrieve only from OMIn records to answer questions derived from MaintNet. Because these questions require specific procedural responses, directly retrieving semantically similar records often yields better results than providing abstracted insights from a KG. ROUGE-based evaluations in Table 4 show no statistically significant differences, likely due to the short and entity-sparse nature of gold-standard answers, which penalize more comprehensive responses.

Table 5: Evaluation results on 83 global sensemaking questions using GPT-4o as the evaluator. The knowledge graphs employed by the KEO method are constructed by GPT-4o using different numbers of aviation records.

Model	# of records	TC	VN	KG
gemma-3-it	100	4.08 ± 0.55	3.65 ± 0.68	4.34 ± 0.24
	200	4.14 ± 0.44	3.71 ± 0.47	4.35 ± 0.26
	300	4.11 ± 0.41	$\mathbf{3.73 \pm 0.49}$	4.30 ± 0.31
	400	$\mathbf{4.16 \pm 0.41}$	3.65 ± 0.50	$\boxed{\mathbf{4.38 \pm 0.19}}$
	500	4.11 ± 0.44	3.65 ± 0.61	4.37 ± 0.23
phi-4	100	$\mathbf{3.90 \pm 0.49}$	3.73 ± 0.51	4.08 ± 0.39
	200	3.88 ± 0.51	3.70 ± 0.54	4.09 ± 0.40
	300	3.87 ± 0.61	$\mathbf{3.71 \pm 0.55}$	$\boxed{\mathbf{4.11 \pm 0.41}}$
	400	3.80 ± 0.59	3.67 ± 0.55	4.11 ± 0.51
	500	3.89 ± 0.50	3.69 ± 0.53	4.09 ± 0.53
mistral-nemo-it	100	3.79 ± 0.62	3.44 ± 0.63	3.77 ± 0.65
	200	3.80 ± 0.64	$\mathbf{3.48 \pm 0.53}$	$\boxed{\mathbf{3.90 \pm 0.50}}$
	300	3.80 ± 0.65	3.46 ± 0.56	3.88 ± 0.45
	400	$\mathbf{3.85 \pm 0.56}$	3.36 ± 0.58	3.88 ± 0.53
	500	3.76 ± 0.60	3.45 ± 0.58	3.86 ± 0.55

Ablation Studies. We study three factors affecting KEO's performance: KG size, the LLM used for KG construction, and the choice of evaluator. Tables 5 and 6 show GPT-4o's evaluation across global sensemaking and knowledge-to-action tasks with varying KG sizes. Results indicate that KEO consistently surpasses baselines on global sensemaking, with performance peaking when the KG is built from roughly 200–300 records. In contrast, text-chunk RAG remains more competitive for knowledge-to-action tasks, where precise procedural retrieval is crucial.

Further analysis in Appendix A.1 shows that KGs produced by weaker models such as Phi-4-mini achieve slightly lower quality than those generated by GPT-4o, but KEO still retains a clear advantage over both vanilla prompting and text-chunk RAG. Evaluator choice also affects relative gains: both GPT-4o and Llama-3.3-70B-Instruct prefer KEO for global sensemaking, while for knowledge-to-action tasks, both evaluators exhibit a mild bias toward text-chunk RAG (Appendix A.2). Overall, these findings underscore the robustness of KEO for dataset-wide reasoning, while also revealing the limits of structured abstraction for fine-grained procedural tasks.

Table 6: Evaluation results on 50 knowledge-to-action questions using GPT-4o as the evaluator. The knowledge graphs employed by the KEO method are constructed by GPT-4o using different numbers of aviation records.

Model	# of records	TC	VN	KG
gemma-3-it	100	3.87 ± 0.66	$\mathbf{3.84 \pm 0.66}$	3.80 ± 0.75
	200	$\boxed{\mathbf{3.88 \pm 0.65}}$	3.82 ± 0.62	3.78 ± 0.70
	300	3.84 ± 0.69	3.83 ± 0.66	3.78 ± 0.73
	400	3.84 ± 0.67	3.74 ± 0.74	$\mathbf{3.80 \pm 0.70}$
	500	3.87 ± 0.67	3.78 ± 0.68	3.80 ± 0.77
phi-4	100	4.21 ± 0.41	4.00 ± 0.54	3.91 ± 0.55
	200	4.18 ± 0.44	4.02 ± 0.49	3.88 ± 0.53
	300	4.18 ± 0.46	4.03 ± 0.49	$\mathbf{3.97 \pm 0.50}$
	400	$\boxed{\mathbf{4.22 \pm 0.39}}$	4.04 ± 0.49	3.89 ± 0.59
	500	4.17 ± 0.45	$\mathbf{4.05 \pm 0.52}$	3.93 ± 0.59
mistral-nemo-it	100	3.72 ± 0.67	$\mathbf{3.72 \pm 0.79}$	3.64 ± 0.83
	200	3.71 ± 0.68	3.71 ± 0.81	3.70 ± 0.82
	300	3.68 ± 0.70	3.69 ± 0.80	3.70 ± 0.83
	400	$\boxed{\mathbf{3.74 \pm 0.71}}$	3.64 ± 0.90	3.63 ± 0.82
	500	$\boxed{\mathbf{3.74 \pm 0.71}}$	3.70 ± 0.76	$\mathbf{3.73 \pm 0.85}$

5 Conclusion

We introduced KEO, a framework integrating KGs into RAG for safety-critical QA. To evaluate its effectiveness, we constructed a benchmark covering both global sensemaking and knowledge-to-action tasks from the OMIn dataset. Experiments with locally deployable LLMs, judged by stronger models, show that KEO substantially improves dataset-wide reasoning and pattern discovery, while text-chunk RAG remains better suited for localized procedural actions. Our results indicate that structured knowledge is crucial for scaling LLMs to high-stakes domains, and that different retrieval paradigms may complement each other depending on task demands.

Limitations

While our evaluation of KEO focused on the aviation maintenance domain, the framework is transferable to other safety-critical areas such as healthcare, power systems, and defense, where structured reasoning with LLMs can provide actionable insights. Future work should investigate domain adaptation strategies, scaling to larger corpora, and integration with multimodal data sources such as schematics and sensor logs. Due to the scope of work, we limited our study to locally deployable LLMs and had limited public data to use, as noted in Appendix G.

Another limitation lies in evaluation. Although we report quantitative metrics such as ROUGE F1 score, these alone are insufficient to fully capture reasoning quality. To address this, we also adopted LLMs as judges with diverse instruct-tuned models, specialized prompting strategies, and pairwise comparison, which have been shown to improve the robustness of evaluation (Shankar et al., 2024; Wei et al., 2024). However, recent work highlights that LLM-as-a-Judge remains neither fully valid nor reliable: judgments can be highly sensitive to the choice of model, prompt template, and even evaluation order (Chehbouni et al., 2025). Moreover, adversarial studies reveal that LLM judges are vulnerable to prompt injection and manipulation (Li et al., 2025), raising concerns about their robustness in practical settings. These limitations suggest that human-in-the-loop evaluation is essential to complement automated judging and ensure reliable assessment of KEO in safety-critical domains.

Finally, while we relied on locally deployed LLMs to avoid dependency on external APIs, systematic security and robustness checks are needed to mitigate risks such as adversarial prompts, data leakage, and misuse in high-stakes environments.

Ethical Considerations

This work, operating in a safety-critical domain, prioritizes ethical deployment by designing the RAG pipeline exclusively for secure, local deployment using smaller, government-approved LLMs to mitigate risks associated with external APIs, data leakage, and potential misuse. We acknowledge that the LLM-constructed KG relies on the sensitive OMIn corpus, which has been processed under strict access controls and de-identification procedures to protect proprietary and personal information. The use of the KG is specifically intended to mitigate the LLM's risk of hallucination and improve factual grounding. However, we emphasize that the system serves only as a decision-support tool, necessitating human validation for all safety-critical applications.

As part of ethical disclosure, we note that this paper made limited use of ChatGPT as a writing assistant to improve clarity and presentation; all research design, implementation, and analysis were conducted by the authors.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.

Farhad Akhbardeh, Travis Desell, and Marcos Zampieri. 2020. Maintnet: A collaborative open-source library for predictive maintenance language resources. *arXiv preprint arXiv:2005.12443*.

Jinheon Baek, Alham Fikri Aji, and Amir Saffari. 2023. Knowledge-augmented language model prompting for zero-shot knowledge graph question answering. In *Proceedings of Workshop on Natural Language Reasoning and Structured Explanations*.

Lukas Bahr, Christoph Wehner, Judith Wewerka, José Bittencourt, Ute Schmid, and Rüdiger Daub. 2025. Knowledge graph enhanced retrieval-augmented generation for failure mode and effects analysis. *Journal of Industrial Information Integration*, 45:100807.

Taiyu Ban, Lyvzhou Chen, Xiangyu Wang, and Huanhuan Chen. 2023. From query tools to causal architects: Harnessing large language models for advanced causal discovery from data. *arXiv preprint arXiv:2306.16902*.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, and 1 others. 2022. Improving language models by retrieving from trillions of tokens. In *Proceedings of International Conference on Machine Learning*, pages 2206–2240.

Khaoula Chehbouni, Mohammed Haddou, Jackie Chi Kit Cheung, and Golnoosh Farnadi. 2025. Neither valid nor reliable? investigating the use of LLMs as judges. *arXiv preprint arXiv:2508.18076*.

Jiawei Chen, Hongyu Lin, Xianpei Han, and Le Sun. 2024. Benchmarking large language models in retrieval-augmented generation. In *Proceedings of AAAI Conference on Artificial Intelligence*, volume 38, pages 17754–17762.

Michiel De Jong, Yury Zemlyanskiy, Nicholas FitzGerald, Fei Sha, and William Cohen. 2021. Mention memory: Incorporating textual knowledge into transformers through entity mention attention. *arXiv preprint arXiv:2110.06176*.

Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, Dasha Metropolitansky, Robert Osazuwa Ness, and Jonathan Larson. 2024. From local to global: A graph RAG approach to query-focused summarization. *arXiv preprint arXiv:2404.16130*.

Shahul Es, Jithin James, Luis Espinosa Anke, and Steven Schockaert. 2024. RAGAs: Automated evaluation of retrieval augmented generation. In *Proceedings of Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 150–158.

Wenqi Fan, Yujuan Ding, Liangbo Ning, Shijie Wang, Hengyun Li, Dawei Yin, Tat-Seng Chua, and Qing Li. 2024. A survey on RAG meeting LLMs: Towards retrieval-augmented large language models. In *Proceedings of ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, page 6491–6501.

Federal Aviation Administration. 2024. FAA Accident and Incident Data System. `https://www.asias.faa.gov/apex/f?p=100:189:::NO`.

Thibault Févry, Livio Baldini Soares, Nicholas FitzGerald, Eunsol Choi, and Tom Kwiatkowski. 2020. Entities as experts: Sparse memory access with entity supervision. In *Proceedings of Conference on Empirical Methods in Natural Language Processing*, pages 4937–4951.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yixin Dai, Jiawei Sun, Haofen Wang, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2(1).

Simret Araya Gebreegziabher, Kuangshi Ai, Zheng Zhang, Elena L Glassman, and Toby Jia-Jun Li. 2025. Leveraging variation theory in counterfactual data augmentation for optimized active learning. In *Proceedings of Findings of the Association for Computational Linguistics: ACL*.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, and 1 others. 2024. A survey on LLM-as-a-judge. *arXiv preprint arXiv:2411.15594*.

Xiaoxin He, Yijun Tian, Yifei Sun, Nitesh V. Chawla, Thomas Laurent, Yann LeCun, Xavier Bresson, and Bryan Hooi. 2024. G-Retriever: Retrieval-augmented generation for textual graph understanding and question answering. In *Proceedings of Advances in Neural Information Processing Systems*, pages 132876–132907.

Aidan Hogan, Eva Blomqvist, Michael Cochez, Claudia D'amato, Gerard De Melo, Claudio Gutierrez, Sabrina Kirrane, José Emilio Labra Gayo, Roberto Navigli, Sebastian Neumaier, Axel-Cyrille Ngonga Ngomo, Axel Polleres, Sabbir M. Rashid, Anisa Rula, Lukas Schmelzeisen, Juan Sequeda, Steffen Staab, and Antoine Zimmermann. 2021. Knowledge graphs. *ACM Computing Surveys*, 54(4):1–37.

Yucheng Hu and Yuxing Lu. 2024. RAG and RAU: A survey on retrieval-augmented language model in natural language processing. *arXiv preprint arXiv:2404.19543*.

Pere-Lluís Huguet Cabot and Roberto Navigli. 2021. REBEL: Relation extraction by end-to-end language generation. In *Proceedings of Findings of the Association for Computational Linguistics: EMNLP*, pages 2370–2381.

Xuhui Jiang, Chengjin Xu, Yinghan Shen, Xun Sun, Lumingyuan Tang, Saizhuo Wang, Zhongwu Chen, Yuanzhuo Wang, and Jian Guo. 2023. On the evolution of knowledge graphs: A survey and perspective. *arXiv preprint arXiv:2310.04835*.

Minki Kang, Jin Myung Kwak, Jinheon Baek, and Sung Ju Hwang. 2023. Knowledge graph-augmented language models for knowledge-grounded dialogue generation. *arXiv preprint arXiv:2305.18846*.

Songze Li, Chuokun Xu, Jiaying Wang, Xueluan Gong, Chen Chen, Jirui Zhang, Jun Wang, Kwok-Yan Lam, and Shouling Ji. 2025. LLMs cannot reliably judge (yet?): A comprehensive assessment on the robustness of LLM-as-a-judge. *arXiv preprint arXiv:2506.09443*.

Yen-Ting Lin and Yun-Nung Chen. 2023. LLM-Eval: Unified multi-dimensional automatic evaluation for open-domain conversations with large language models. *arXiv preprint arXiv:2305.13711*.

Yuanjie Lyu, Zhiyu Li, Simin Niu, Feiyu Xiong, Bo Tang, Wenjin Wang, Hao Wu, Huanyong Liu, Tong Xu, and Enhong Chen. 2025. CRUD-RAG: A comprehensive Chinese benchmark for retrieval-augmented generation of large language models. *ACM Transactions on Information Systtems*, 43(2):41:1–41:32.

John C Mankins. 1995. Technology readiness levels. Technical report, NASA.

K. Mealey, J. Karr, P. Saboia Moreira, D. Finch, A. Riter, P. Brenner, and C. Vardeman II. 2024. nd-crane/trusted_ke: Omin dataset v1.0.0 - initial public release (v1.0.0-omin).

Kathleen Mealey, Jonathan A Karr Jr, Priscila Saboia Moreira, Paul R Brenner, and Charles F Vardeman II. 2025. Trusted knowledge extraction for operations and maintenance intelligence. *arXiv preprint arXiv:2507.22935*.

Igor Melnyk, Pierre Dognin, and Payel Das. 2022. Knowledge graph generation from text. In *Proceedings of Findings of the Association for Computational Linguistics: EMNLP*, pages 1610–1622.

NASA Aviation Safety Reporting System. NASA ASRS Dataset. https://asrs.arc.nasa.gov/search/database.html.

NASA Prognostics Center of Excellence. 2023. NASA Prognostics Center of Excellence. https://www.nasa.gov/content/nasa-prognostics-center-of-excellence.

Jon Perez-Cerrolaza, Jaume Abella, Markus Borg, Carlo Donzella, Jesus Cerquides, Francisco J Cazorla, Cristofer Englund, Markus Tauber, George Nikolakopoulos, and Jose Luis Flores. 2024. Artificial intelligence for safety-critical systems in industrial and transportation domains: A survey. *ACM Computing Surveys*, 56(7):1–40.

Pejman Peykani, Fatemeh Ramezanlou, Cristina Tanasescu, and Sanly Ghanidel. 2025. Large language models: A structured taxonomy and review of challenges, limitations, solutions, and future directions. *Applied Sciences*, 15(14).

Tyler Thomas Procko and Omar Ochoa. 2024. Graph retrieval-augmented generation for large language models: A survey. In *Proceedings of Conference on AI, Science, Engineering, and Technology*, pages 166–169.

Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. 2023. In-context retrieval-augmented language models. *Transactions of the Association for Computational Linguistics*, 11:1316–1331.

Priyanka Ranade and Anupam Joshi. 2023. Fabula: Intelligence report generation using retrieval-augmented narrative construction. In *Proceedings of International Conference on Advances in Social Networks Analysis and Mining*, pages 603–610.

Priyanka Sen, Sandeep Mavadia, and Amir Saffari. 2023. Knowledge graph-augmented language models for complex question answering. In *Proceedings of Workshop on Natural Language Reasoning and Structured Explanations*, pages 1–8.

Shreya Shankar, JD Zamfirescu-Pereira, Björn Hartmann, Aditya Parameswaran, and Ian Arawjo. 2024. Who validates the validators? aligning LLM-assisted evaluation of llm outputs with human preferences. In *Proceedings of ACM Symposium on User Interface Software and Technology*, pages 1–14.

Weijia Shi, Julian Michael, Suchin Gururangan, and Luke Zettlemoyer. 2022. Nearest neighbor zero-shot inference. In *Proceedings of Conference on Empirical Methods in Natural Language Processing*, pages 3254–3265.

Yixuan Tang and Yi Yang. 2024. MultiHop-RAG: Benchmarking retrieval-augmented generation for multi-hop queries. *arXiv preprint arXiv:2401.15391*.

Milena Trajanoska, Riste Stojanov, and Dimitar Trajanov. 2023. Enhancing knowledge graph construction using large language models. *arXiv preprint arXiv:2305.04676*.

Robin Wagner, Emanuel Kitzelmann, and Ingo Boersch. 2025. Mitigating hallucination by integrating knowledge graphs into LLM inference – a systematic literature review. In *Proceedings of Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 795–805.

Jiaan Wang, Yunlong Liang, Fandong Meng, Zengkui Sun, Haoxiang Shi, Zhixu Li, Jinan Xu, Jianfeng Qu, and Jie Zhou. 2023. Is ChatGPT a good NLG evaluator? a preliminary study. *arXiv preprint arXiv:2303.04048*.

Shuting Wang, Jiejun Tan, Zhicheng Dou, and Ji-Rong Wen. 2024a. OmniEval: An omnidirectional and automatic RAG evaluation benchmark in financial domain. *arXiv preprint arXiv:2412.13018*.

Yu Wang, Nedim Lipka, Ryan A Rossi, Alexa Siu, Ruiyi Zhang, and Tyler Derr. 2024b. Knowledge graph prompting for multi-document question answering. In *Proceedings of AAAI Conference on Artificial Intelligence*, pages 19206–19214.

Hui Wei, Shenghua He, Tian Xia, Fei Liu, Andy Wong, Jingyang Lin, and Mei Han. 2024. Systematic evaluation of LLM-as-a-judge in LLM alignment tasks: Explainable metrics and diverse prompt templates. *arXiv preprint arXiv:2408.13006*.

Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong Zhang. 2024. Benchmarking retrieval-augmented generation for medicine. In *Proceedings of Findings of the Association for Computational Linguistics: ACL*, pages 6233–6251.

Zhentao Xu, Mark Jerome Cruz, Matthew Guevara, Tie Wang, Manasi Deshpande, Xiaofeng Wang, and Zheng Li. 2024. Retrieval-augmented generation with knowledge graphs for customer service question answering. In *Proceedings of ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2905–2909.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380.

Liang Yao, Jiazhen Peng, Chengsheng Mao, and Yuan Luo. 2025. Exploring large language models for knowledge graph completion. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1–5.

Xiaohan Yuan, Jinfeng Li, Dongxia Wang, Yuefeng Chen, Xiaofeng Mao, Longtao Huang, Hui Xue, Wenhai Wang, Kui Ren, and Jingyi Wang. 2024. S-Eval: Automatic and adaptive test generation for benchmarking safety evaluation of large language models. *arXiv preprint arXiv:2405.14191*.

Feng Zhang, Chengjie Pang, Yuehan Zhang, and Chenyu Luo. 2025. CAMB: A comprehensive industrial LLM benchmark on civil aviation maintenance. *arXiv preprint arXiv:2508.20420*.

Yuzhe Zhang, Yipeng Zhang, Yidong Gan, Lina Yao, and Chen Wang. 2024. Causal graph discovery with retrieval-augmented generation based large language models. *arXiv preprint arXiv:2402.15301*.

Xuejiao Zhao, Siyan Liu, Su-Yin Yang, and Chunyan Miao. 2025. MedRAG: Enhancing retrieval-augmented generation with knowledge graph-elicited reasoning for healthcare copilot. In *Proceedings of the ACM on Web Conference*, page 4442–4457.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *Proceedings of Advances in Neural Information Processing Systems*, pages 46595–46623.

Yigeng Zhou, Wu Li, Yifan Lu, Jing Li, Fangming Liu, Meishan Zhang, Yequan Wang, Daojing He, Honghai Liu, and Min Zhang. 2025. Reflection on knowledge graph for large language models reasoning. In *Proceeding of Findings of the Association for Computational Linguistics: ACL*, pages 23840–23857.

Xiangrong Zhu, Yuexiang Xie, Yi Liu, Yaliang Li, and Wei Hu. 2025. Knowledge graph-guided retrieval augmented generation. In *Proceedings of Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8912–8924.

A Additional Results

A.1 Ablation on KG Size and KG Generation LLM

We assess the impact of KG construction parameters by varying the number of records (from 100 to 500) and the LLM used (GPT-4o vs. Phi-4-mini). Tables 5 and 7 report GPT-4o evaluations on global sensemaking questions. KEO consistently outperforms baselines, with optimal performance typically achieved when the KG is constructed from 200 to 300 records. While KGs built by Phi-4-mini result in slightly lower scores, KEO still surpasses vanilla and text-chunk RAG methods, highlighting the robustness of our dynamic prompting strategy for high-quality KG construction.

On knowledge-to-action questions (Tables 6 and 10), varying KG size and generation LLM does not significantly alter the performance trend—text-chunk RAG remains more effective. This reinforces the observation that specific action-oriented tasks favor localized retrieval over structured abstraction.

A.2 Ablation on LLM Evaluators

To evaluate robustness against evaluator bias, we conduct comparative assessments using both GPT-4o and Llama-3.3-70B-Instruct. As shown in Figures 2 and 3, both evaluators prefer KEO over baselines in global sensemaking tasks, with Llama-3.3 showing even stronger preference for KEO. The trend of stronger LLMs amplifying the performance benefits of KEO remains consistent. Tables 8 and 9 confirm this, with optimal KG size again ranging between 200 and 300 records.

For knowledge-to-action questions, Llama evaluations (Tables 11 and 12) mirror those of GPT-4o, favoring text-chunk RAG slightly over KEO. This consistency across evaluators increases confidence in the observed performance tradeoffs between tasks that benefit from structured reasoning and those that require precise, localized retrieval.

B KG Creation Prompt

Here we provide the prompt used to extract knowledge graph triplets from aviation maintenance text, enabling structured knowledge representation for the KEO framework.

Role— You are extracting knowledge graph triplets from aviation maintenance text.

Task— Extract informative triplets directly from the text following the examples. Format each triplet as: `<entity1, relation, entity2>`. Do not add any extra words, line

Table 7: Evaluation results on 83 global sensemaking questions using GPT-4o as the evaluator. The knowledge graph employed by the KEO method is constructed using Phi-4-mini with varying numbers of aviation records.

Model \ KG type	Phi-4-mini-100			Phi-4-mini-200			Phi-4-mini-300			Phi-4-mini-400			Phi-4-mini-500		
	TC	**VN**	**KG**	**TC**	**VN**	**KG**	**TC**	**VN**	**KG**	**TC**	**VN**	**KG**	**TC**	**VN**	**KG**
gemma-3-it	4.05 ± 0.56	3.68 ± 0.62	4.31 ± 0.29	4.11 ± 0.44	3.72 ± 0.53	$\mathbf{4.33 \pm 0.29}$	4.15 ± 0.41	3.71 ± 0.54	4.29 ± 0.30	4.11 ± 0.52	3.70 ± 0.50	4.27 ± 0.37	4.14 ± 0.42	3.74 ± 0.50	4.26 ± 0.35
phi-4	3.90 ± 0.47	3.71 ± 0.51	4.01 ± 0.42	3.86 ± 0.53	3.64 ± 0.54	4.06 ± 0.42	3.86 ± 0.51	3.68 ± 0.55	$\mathbf{4.13 \pm 0.40}$	3.89 ± 0.50	3.68 ± 0.53	4.09 ± 0.41	3.95 ± 0.47	3.69 ± 0.49	4.01 ± 0.42
mistral-nemo-it	3.82 ± 0.57	3.46 ± 0.54	3.81 ± 0.50	3.77 ± 0.60	3.43 ± 0.72	$\mathbf{3.85 \pm 0.49}$	3.78 ± 0.59	3.47 ± 0.58	3.80 ± 0.52	3.82 ± 0.56	3.46 ± 0.54	3.76 ± 0.50	3.74 ± 0.67	3.49 ± 0.60	3.75 ± 0.46

Table 8: Evaluation results on 83 global sensemaking questions using Llama-3.3-70B-Instruct as the evaluator. The knowledge graph employed by the KEO method is constructed using GPT-4o with varying numbers of aviation records.

Model \ KG type	GPT-4o-100			GPT-4o-200			GPT-4o-300			GPT-4o-400			GPT-4o-500		
	TC	**VN**	**KG**	**TC**	**VN**	**KG**	**TC**	**VN**	**KG**	**TC**	**VN**	**KG**	**TC**	**VN**	**KG**
gemma-3-it	4.47 ± 0.45	4.20 ± 0.62	4.88 ± 0.25	4.52 ± 0.28	4.31 ± 0.23	$\mathbf{4.93 \pm 0.20}$	4.49 ± 0.27	4.32 ± 0.21	4.90 ± 0.22	4.47 ± 0.23	4.30 ± 0.28	4.91 ± 0.22	4.49 ± 0.23	4.24 ± 0.47	4.93 ± 0.21
phi-4	4.34 ± 0.13	4.31 ± 0.21	4.48 ± 0.21	4.32 ± 0.21	4.29 ± 0.25	4.47 ± 0.20	4.29 ± 0.42	4.30 ± 0.23	$\mathbf{4.50 \pm 0.23}$	4.29 ± 0.40	4.28 ± 0.32	4.42 ± 0.42	4.32 ± 0.20	4.29 ± 0.25	4.45 ± 0.45
mistral-nemo-it	4.31 ± 0.26	4.07 ± 0.61	4.34 ± 0.45	4.28 ± 0.31	4.13 ± 0.43	$\mathbf{4.42 \pm 0.17}$	4.20 ± 0.39	4.09 ± 0.49	4.40 ± 0.13	4.28 ± 0.28	4.12 ± 0.50	4.37 ± 0.16	4.27 ± 0.33	4.11 ± 0.49	4.39 ± 0.22

Table 9: Evaluation results on 83 global sensemaking questions using Llama-3.3-70B-Instruct as the evaluator. The knowledge graph employed by the KEO method is constructed using Phi-4-mini with varying numbers of aviation records.

Model \ KG type	Phi-4-mini-100			Phi-4-mini-200			Phi-4-mini-300			Phi-4-mini-400			Phi-4-mini-500		
	TC	**VN**	**KG**	**TC**	**VN**	**KG**	**TC**	**VN**	**KG**	**TC**	**VN**	**KG**	**TC**	**VN**	**KG**
gemma-3-it	4.45 ± 0.44	4.25 ± 0.46	4.85 ± 0.27	4.48 ± 0.22	4.33 ± 0.23	4.81 ± 0.28	4.50 ± 0.26	4.28 ± 0.28	$\mathbf{4.88 \pm 0.24}$	4.43 ± 0.44	4.27 ± 0.30	4.84 ± 0.26	4.50 ± 0.26	4.31 ± 0.28	4.84 ± 0.27
phi-4	4.33 ± 0.21	4.32 ± 0.21	4.47 ± 0.20	4.33 ± 0.17	4.27 ± 0.31	4.48 ± 0.19	4.33 ± 0.18	4.30 ± 0.24	$\mathbf{4.50 \pm 0.22}$	4.32 ± 0.21	4.34 ± 0.19	4.46 ± 0.18	4.33 ± 0.18	4.32 ± 0.19	4.45 ± 0.17
mistral-nemo-it	4.26 ± 0.30	4.11 ± 0.44	4.39 ± 0.12	4.31 ± 0.27	4.00 ± 0.67	$\mathbf{4.42 \pm 0.14}$	4.32 ± 0.28	4.11 ± 0.48	4.39 ± 0.19	4.26 ± 0.34	4.15 ± 0.44	4.37 ± 0.19	4.26 ± 0.38	4.10 ± 0.52	4.41 ± 0.18

Table 10: Evaluation results on 50 generalizable knowledge-to-action questions using GPT-4o as the evaluator. The knowledge graph employed by the KEO method is constructed using Phi-4-mini with varying numbers of aviation records.

Model \ KG type	Phi-4-mini-100			Phi-4-mini-200			Phi-4-mini-300			Phi-4-mini-400			Phi-4-mini-500		
	TC	**VN**	**KG**	**TC**	**VN**	**KG**	**TC**	**VN**	**KG**	**TC**	**VN**	**KG**	**TC**	**VN**	**KG**
gemma-3-it	$\mathbf{3.89 \pm 0.66}$	3.84 ± 0.63	3.77 ± 0.69	3.82 ± 0.68	3.83 ± 0.64	3.77 ± 0.72	3.86 ± 0.65	3.77 ± 0.67	3.72 ± 0.74	3.86 ± 0.63	3.78 ± 0.68	3.78 ± 0.64	3.85 ± 0.66	3.84 ± 0.65	3.81 ± 0.69
phi-4	4.14 ± 0.45	4.00 ± 0.55	3.97 ± 0.53	$\mathbf{4.21 \pm 0.45}$	4.02 ± 0.59	3.90 ± 0.53	4.14 ± 0.47	4.04 ± 0.52	3.86 ± 0.67	4.14 ± 0.57	4.04 ± 0.50	3.91 ± 0.60	4.15 ± 0.45	4.05 ± 0.48	3.92 ± 0.54
mistral-nemo-it	3.71 ± 0.67	$\mathbf{3.75 \pm 0.79}$	3.70 ± 0.78	3.68 ± 0.76	3.73 ± 0.79	3.68 ± 0.86	3.65 ± 0.74	3.62 ± 0.85	3.51 ± 0.89	3.70 ± 0.69	3.66 ± 0.82	3.58 ± 0.87	3.66 ± 0.73	3.70 ± 0.74	3.49 ± 0.91

Table 11: Evaluation results on 50 generalizable knowledge-to-action questions using Llama-3.3-70B-Instruct as the evaluator. The knowledge graph employed by the KEO method is constructed using GPT-4o with varying numbers of aviation records.

Model \ KG type	GPT-4o-100			GPT-4o-200			GPT-4o-300			GPT-4o-400			GPT-4o-500		
	TC	**VN**	**KG**	**TC**	**VN**	**KG**	**TC**	**VN**	**KG**	**TC**	**VN**	**KG**	**TC**	**VN**	**KG**
gemma-3-it	3.98 ± 0.60	3.96 ± 0.61	3.77 ± 0.78	4.00 ± 0.59	3.96 ± 0.62	3.85 ± 0.68	$\mathbf{4.00 \pm 0.54}$	3.97 ± 0.61	3.81 ± 0.74	3.99 ± 0.60	3.87 ± 0.71	3.87 ± 0.69	4.00 ± 0.60	3.94 ± 0.59	3.78 ± 0.74
phi-4	4.32 ± 0.37	4.27 ± 0.39	4.19 ± 0.41	4.33 ± 0.36	4.24 ± 0.43	4.17 ± 0.44	$\mathbf{4.36 \pm 0.33}$	4.24 ± 0.43	4.23 ± 0.46	4.32 ± 0.36	4.29 ± 0.36	4.15 ± 0.47	4.32 ± 0.33	4.26 ± 0.41	4.14 ± 0.47
mistral-nemo-it	$\mathbf{4.03 \pm 0.53}$	3.87 ± 0.73	3.75 ± 0.77	3.96 ± 0.63	3.81 ± 0.78	3.84 ± 0.68	3.92 ± 0.62	3.80 ± 0.80	3.78 ± 0.79	4.00 ± 0.60	3.76 ± 0.80	3.74 ± 0.78	3.83 ± 0.71	3.82 ± 0.74	3.82 ± 0.76

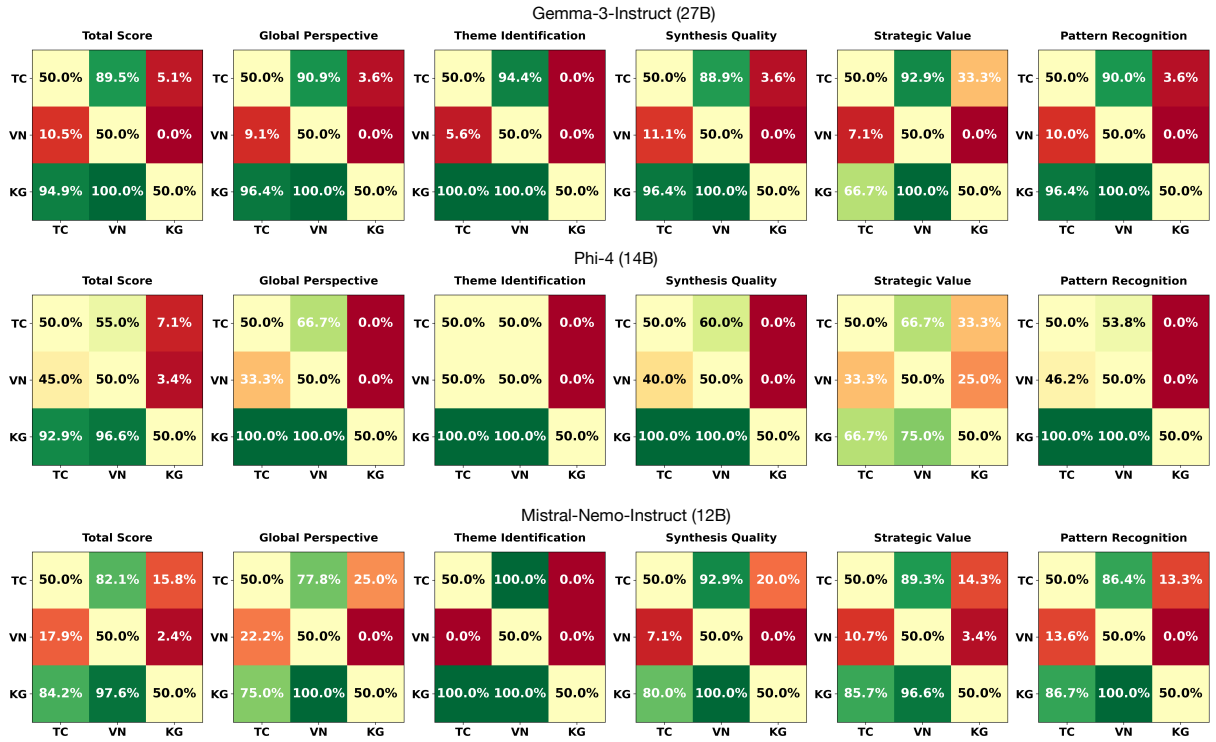

Figure 3: Head-to-head win rate matrix of row method over column method (TC: text-chunk RAG, VN: vanilla LLM, KG: our method KEO) on the same 83 global sensemaking questions, evaluated by Llama-3.3-70B-Instruct. Compared to GPT-4o evaluation (Figure 2), Llama shows a higher preference for answers generated using RAG-based methods. Nonetheless, the same trend persists: stronger LLMs tend to amplify the advantage of the KEO approach.

Table 12: Evaluation results on 50 generalizable knowledge-to-action questions using Llama-3.3-70B-Instruct as the evaluator. The knowledge graph employed by the KEO method is constructed using Phi-4-mini with varying numbers of aviation records.

Model \ KG type	Phi-4-mini-100			Phi-4-mini-200			Phi-4-mini-300			Phi-4-mini-400			Phi-4-mini-500		
	TC	VN	KG	TC	VN	KG	TC	VN	KG	TC	VN	KG	TC	VN	KG
gemma-3-it	**4.00 ± 0.59**	3.93 ± 0.62	3.85 ± 0.71	**4.00 ± 0.59**	3.97 ± 0.60	3.83 ± 0.72	3.99 ± 0.59	3.94 ± 0.61	3.80 ± 0.76	4.00 ± 0.62	3.95 ± 0.62	3.86 ± 0.64	4.00 ± 0.62	3.95 ± 0.62	3.85 ± 0.69
phi-4	**4.35 ± 0.31**	4.20 ± 0.45	4.18 ± 0.46	4.33 ± 0.37	4.26 ± 0.45	4.19 ± 0.43	4.32 ± 0.37	4.22 ± 0.43	4.11 ± 0.62	4.26 ± 0.58	4.28 ± 0.39	4.17 ± 0.50	4.32 ± 0.36	4.27 ± 0.44	4.15 ± 0.46
mistral-nemo-it	**3.92 ± 0.56**	3.88 ± 0.71	3.87 ± 0.67	3.92 ± 0.68	3.83 ± 0.76	3.80 ± 0.76	3.88 ± 0.62	3.76 ± 0.80	3.64 ± 0.86	3.92 ± 0.60	3.73 ± 0.79	3.70 ± 0.78	3.83 ± 0.65	3.76 ± 0.75	3.61 ± 0.85

breaks, or explanatory notes. Focus on extracting factual relationships from the text.

Goal— Generate structured entity-relation-entity triplets that capture factual relationships from aviation maintenance records while maintaining consistency across the knowledge graph.

Guidelines— Use only these relation types: `OWNED BY`, `INSTANCE OF`, `FOLLOWED BY`, `HAS CAUSE`, `FOLLOWS`, `EVENT DISTANCE`, `HAS EFFECT`, `LOCATION`, `USED BY`, `INFLUENCED BY`, `TIME PERIOD`, `PART OF`, `MAINTAINED BY`, `DESIGNED BY`. When extracting triplets, prefer to use existing nodes from the knowledge graph if possible, rather than inventing new entity mentions.

Style— Extract triplets in the format `<entity1, relation, entity2>`. Each triplet should appear on a new line with no numbering or bullets. Focus on factual relationships that can be directly inferred from the text.

Example— `TEXT: THE WRIGHT BROTHERS DESIGNED THE FIRST SUCCESSFUL AIRPLANE IN 1903 IN KITTY HAWK.`
`Triplets:`
`<FIRST SUCCESSFUL AIRPLANE, DESIGNED BY, WRIGHT BROTHERS>`
`<FIRST SUCCESSFUL AIRPLANE, TIME PERIOD, 1903>`
`<FIRST SUCCESSFUL AIRPLANE, LOCATION, KITTY HAWK>`

Target Text: `{text}`
Triplets:

C Global Sensemaking Question Prompts

Here we provide the prompts used to generate global sensemaking questions for the QA benchmark, covering three subtypes: comprehensive, context-specific, and category-specific.

C.1 Comprehensive Question Prompt

Role— You are an expert in aviation safety and maintenance analysis.

Task— Generate high-level, global sensemaking questions based on a summary of an aviation maintenance dataset. These questions should require a comprehensive understanding of patterns and relationships that span the entire dataset.
Goal— Generate exactly `{num_questions}` questions that encourage holistic insight into overarching themes, systemic issues, and strategic implications present in the dataset. The dataset summary is provided below: `{data_summary}`.
Guidelines— The questions must: (1) require a holistic understanding of the dataset, not just individual records; (2) focus on patterns, trends, causal relationships, and systemic insights; (3) enable strategic reasoning or decision-making at an organizational level; and (4) avoid questions that can be answered through simple statistics or localized analysis.
Style— Generate exactly `{num_questions}` questions. Each question should appear on a new line, with no numbering or bullets. Use formulations such as: "What are the...", "How do...", "Which factors...", "What patterns...".
Examples— What are the most common systemic maintenance challenges reflected across the dataset? How do maintenance trends vary by aircraft type, and what are the implications? Which recurring issues suggest gaps in standard operating procedures? What are the long-term safety patterns indicated by the data?

C.2 Context-Specific Question Prompt

Role— You are an expert in aviation safety and maintenance analysis.
Task— Generate context-specific sensemaking questions based on representative aviation maintenance records. These questions should help uncover causes, patterns, and interactions relevant to the particular maintenance context.
Goal— Given representative maintenance records from the `{context_type}` context: `{sample_records}`, generate exactly `{num_questions}` questions that promote analysis across multiple records and inform actionable understanding.
Guidelines— The questions should: (1) Focus on understanding *why* failures occur. (2) Identify *what* can be done to prevent similar issues. (3) Explore *how* different factors interact. (4) Reveal patterns across similar incidents. (5) Support decision-making for maintenance and safety.
Style— Generate exactly `{num_questions}` questions. Each question should appear on a new line, with no numbering or bullets. Focus on questions that require synthesis of multiple incidents or factors, rather than simple summaries or isolated examples.

C.3 Category-Specific Question Prompt

Role— You are an expert in aviation safety and maintenance analysis.
Task— Generate high-quality sensemaking questions tailored to a specific analytical category in aviation maintenance.
Goal— Generate exactly `{num_questions}` questions that align with the category: `{category}`. These questions should reflect deep insight into aviation maintenance practices and support strategic reasoning.
Category— `{config['description']}`
Context— `{context_prompt}`
Guidelines— The questions should: (1) Require synthesis across multiple data points to answer. (2) Focus on actionable insights for aviation safety. (3) Be specific to the aviation maintenance domain. (4) Support strategic decision-making. (5) Reveal patterns and relationships in the data.
Style— Use the following starter patterns as inspiration, but generate varied, comprehensive questions: `{config['template_starters']}` Generate exactly `{num_questions}` questions. Each question should appear on a new line, with no numbering or bullets.

D Task-Specific Evaluation Metrics

Global Sensemaking Questions. Since these questions have no predefined gold answers, we use an LLM-based evaluator for both absolute and comparative scoring. For absolute scoring, the evaluator rates each answer on a 1–5 scale according to task-specific criteria, providing detailed explanations. For comparative scoring, pairwise comparisons are conducted between answers from different methods (Vanilla LLM, Text-chunk RAG, and KEO). The evaluator selects the better answer for each pair, and we report the win rate of each method across the full set of comparisons.

The evaluation is guided by these criteria:

- **Global Perspective:** Does the answer reflect dataset-wide insights?
- **Theme Identification:** Are key recurring patterns or topics clearly identified?
- **Synthesis Quality:** Does the response integrate information across different records?
- **Strategic Value:** Are the insights useful for high-level decision-making?
- **Pattern Recognition:** Are underlying trends or systemic relationships revealed?

Knowledge-to-Action Questions. We apply both automatic and LLM-based evaluation. For automatic evaluation, we compute ROUGE-F1 scores between predicted and gold-standard actions. For human-aligned evaluation, we prompt an LLM with the question, predicted answer, and gold-standard answer, and ask it to rate the response using the following criteria (1–5 scale):

- **Correctness:** Is the predicted action factually accurate?
- **Completeness:** Does it cover all necessary action steps?
- **Practicality:** Is the recommended action feasible in real-world maintenance?
- **Safety:** Does the response preserve or enhance operational safety?
- **Clarity:** Is the action clearly and precisely articulated?

E LLM Evaluator Prompts

Here we provide the prompts used by the LLM evaluator to assess both global sensemaking and knowledge-to-action questions.

E.1 Global Sensemaking Evaluation Prompt

Role— You are evaluating an LLM-generated answer to a **global sensemaking** question in the domain of aviation maintenance.
Input— Question: `{question}`; LLM Answer: `{answer}`
Definition— Global sensemaking questions require synthesis across entire datasets to identify overarching themes, systemic patterns, and strategic insights.
Evaluation Criteria— Rate the answer on a 1–5 scale for each of the following:

- **Global Perspective:** Does the answer demonstrate understanding of dataset-wide patterns?
- **Theme Identification:** Are major themes and patterns clearly identified?
- **Synthesis Quality:** How well does it synthesize information across multiple sources?
- **Strategic Value:** Does it provide insights useful for high-level decision making?
- **Pattern Recognition:** Are complex relationships and dependencies identified?

Output Format—

```
Global Perspective: [score] - [explanation]
Theme Identification: [score] - [explanation]
Synthesis Quality: [score] - [explanation]
Strategic Value: [score] - [explanation]
Pattern Recognition: [score] - [explanation]
Global Sensemaking Assessment: [overall score]
```

E.2 Knowledge-to-Action Evaluation Prompt

Role— You are evaluating a predicted answer to a maintenance-action question, assessing its correctness and safety based on the gold-standard response.
Input— `Question: {question}` `Ground Truth Answer: {ground_truth}` `Predicted Answer: {predicted}`
Evaluation Criteria— Rate on a 1–5 scale:

- **Correctness:** How factually accurate is the predicted answer?
- **Completeness:** Does it cover all necessary action steps?
- **Practicality:** Are the actions feasible and implementable?
- **Safety:** Would the actions maintain or improve operational safety?
- **Clarity:** Is the answer easy to understand and follow?

Output Format—

```
Correctness: [score] - [explanation]
Completeness: [score] - [explanation]
Practicality: [score] - [explanation]
Safety: [score] - [explanation]
Clarity: [score] - [explanation]
Overall Score: [average score] - [summary of quality]
```

F Relation Extraction

F.1 Lack of a Relation Extraction Gold Standard

The OMIn dataset has previously been processed through a knowledge extraction pipeline covering NER, CR, NEL, and RE (Mealey et al., 2025). A random subset of 100 records was human-annotated to provide gold standards for NER, CR, and NEL, with the same subset manually reviewed for RE.

Table 13: Distribution of relation types in the gold standard for RE.

Relation	Strict Count	Loose Count
OWNED BY	1	2
INSTANCE OF	5	9
FOLLOWED BY	0	20
HAS CAUSE	13	93
FOLLOWS	1	8
EVENT DISTANCE	0	1
HAS EFFECT	13	94
LOCATION	12	23
USED BY	4	5
INFLUENCED BY	0	1
TIME PERIOD	6	29
PART OF	21	30
MAINTAINED BY	8	9
DESIGNED BY	0	1

The original FAA dataset contains sentences such as: "After takeoff, engine quit. Wing fuel tank sumps were not drained during preflight because they were frozen." (Federal Aviation Administration, 2024). Although prior work demonstrated that a KG could in principle be derived from such records (Mealey et al., 2025), the construction of a gold standard for RE was deferred. This omission stemmed from the fact that each RE tool under evaluation employed a distinct relation schema, making a unified global GS impractical. However, since our KG–RAG pipeline requires structured relations, we address this gap by creating a dedicated RE gold standard.

F.2 Creation of a Relation Extraction Gold Standard

In their evaluation of RE tools, prior studies compared four different systems, each adopting its own set of relations (Mealey et al., 2025). To establish consistent gold standards, we defined a relation schema that combines the REBEL subset of Wikidata relations (Huguet Cabot and Navigli, 2021) with additional domain-specific relations relevant to aviation maintenance. Specifically, we included relations such as `MAINTAINED BY` and `DESIGNED BY` to capture critical maintenance-specific dependencies (Table 13).

This schema was selected because the REBEL tool outperformed others in prior experiments on the OMIn dataset, identifying 220 relations and providing a comprehensive basis for KG construction. Using this relation set, we created 100 gold-standard annotations through a standard human

annotation process.

RE Strict Gold Standard - In order to create a relationship triple (head, relationship, tail), the head and tail entities must already exist as nodes from the NER or CR processes in order for the relationship to be added.

RE Loose Gold Standard - If a head or tail node does not already exist, a new node will be created when the relationship is added.

The loose count was introduced to mitigate the limitations of generating the KG via a single, complex prompt per record.

This metric served as a relaxed evaluation by only requiring that the LLM to extract a plausible relation type from the correct record ID. This less strict measure allowed for a fairer assessment of the LLM's core ability to identify relevant relational concepts within the complex input text, despite the difficulty posed by the single-prompt extraction methodology, as noted in Appendix B.

G Study Scope and Constraints

KEO framework's evaluation is constrained to ensure the integrity of the knowledge-transfer assessment and the viability of real-world application. The research limits the evaluation to locally deployable LLMs, such as Gemma-3, Phi-4, and Mistral-Nemo, as this proves the framework's suitability for secure, in-house deployment in safety-critical domains where large, proprietary cloud APIs are unsuitable. Stronger models, such as GPT-4o and Llama-3.3, are employed solely as judges for a robust, objective evaluation. To rigorously test the transferability of structured knowledge, the target answer corpus, MaintNet records (Akhbardeh et al., 2020), were excluded from the retrieval corpus; instead, both RAG and KEO are strictly restricted to the OMIn dataset corpus only to force the systems to answer by leveraging transferred maintenance knowledge successfully. This domain focus highlights the inherent challenge of working in this specialized field, as there is a significant lack of data from the U.S. military and defense sectors in the public domain, with the few available records. Others include NASA's Prognostics Center of Excellence (PCOE) (NASA Prognostics Center of Excellence, 2023), which offers no free-response natural language text, and NASA's Aviation Safety Reporting System (ASRS) (NASA Aviation Safety Reporting System), which provides only a limited selection of 50 records regarding 30 topics.