DGSent: Domain-Grounded Sentence Embeddings with Geometric Space Restructuring for Airport Runway Maintenance Text Representation

Anonymous ACL submission

Abstract

Reliable evidence matching is essential for applying language models to airport runway maintenance, where decisions depend on dense standards and terminology-rich technical documents. However, general-purpose sentence encoders often fail to capture fine-grained domain semantics, suffering from anisotropic collapse and semantic drift in this specialized setting. We present DGSent, a domain-grounded sentence embedding model for runway maintenance. DGSent combines global geometry restructuring with retrieval-grounded knowledge distillation to improve inter-category separability and anchor sentence representations to authoritative domain evidence. We also construct and release an expert-curated runway maintenance knowledge dataset from standards and technical documents. Experiments show that DGSent consistently improves semantic matching and representation geometry, and further benefits downstream RAG when used as its embedding backbone. The code and data for DGSent are available[1].

1 Introduction

Airport runways are critical infrastructures in civil aviation, whose maintenance quality directly affects runway availability, operational safety, flight punctuality, and lifecycle costs (International Civil Aviation Organization, 2022; International Air Transport Association, 2024). Minor pavement distresses or delayed treatment may rapidly evolve into foreign object debris risks and large-scale flight disruptions. However, current maintenance still relies heavily on expert experience and manual interpretation of fragmented standards (Federal Aviation Administration, 2014). As airport traffic grows and inspection data become increasingly heterogeneous, intelligent and evidence-grounded decision support for runway maintenance is urgently needed.

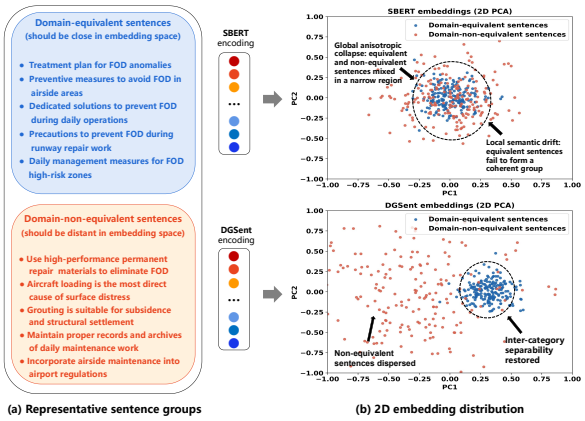

Figure 1: Geometric pathologies of SBERT on airport runway maintenance texts. SBERT collapses domain-equivalent (blue) and non-equivalent (orange) sentences into the same region, while DGSent aggregates equivalent sentences and disperses non-equivalent ones.

Large language models (LLMs) (OpenAI, 2023; Grattafiori et al., 2024; Minaee et al., 2024) provide a promising basis for intelligent runway maintenance systems, but their use in safety-critical domains requires outputs to be grounded in authoritative knowledge rather than parametric memory alone. In retrieval-augmented generation (RAG) (Gao et al., 2023), this grounding depends critically on sentence embeddings, which determine whether relevant evidence can be retrieved for downstream reasoning (Izacard et al., 2022). However, general-purpose sentence embedding models such as SBERT (Reimers and Gurevych, 2019) and BGE (Xiao et al., 2023) often fail to capture fine-grained semantics in specialized corpora dominated by dense terminology and structured technical knowledge (Tang and Yang, 2024).

Adapting sentence encoders to airport runway maintenance faces both data and representation challenges. Unlike biomedicine, finance, or law,

[1] https://anonymous.4open.science/r/anonymous-final-4B14

this domain lacks publicly available, expert-curated corpora and semantic matching benchmarks. More importantly, as shown in Figure 1, general-purpose encoders exhibit two geometric pathologies. First, global anisotropic collapse is amplified by the narrow topical distribution of runway texts, causing equivalent and non-equivalent sentences to occupy the same compact region (Su et al., 2021). Second, local semantic drift makes embeddings follow generic linguistic similarity rather than domain terminology, pulling together surface-similar but semantically different distress descriptions while separating paraphrases of the same maintenance fact (Anderson et al., 2024; Tang and Yang, 2024).

Existing domain adaptation methods, such as continued pretraining (Gururangan et al., 2020) and contrastive fine-tuning (Wei et al., 2025), mainly optimize data-driven objectives while leaving the geometry of the embedding space implicit. To address this limitation, we propose DGSent, a domain-grounded sentence embedding model for airport runway maintenance. DGSent combines global geometry restructuring with retrieval-grounded knowledge distillation based on BM25-guided sparse and dense scoring. The former improves isotropy and inter-category separability, while the latter anchors sentence representations to terminology-rich evidence from authoritative documents. Together, these two mechanisms produce embeddings that are both geometrically discriminative and domain-aware. Our contributions are summarized as follows:

- We release an expert-curated airport runway maintenance dataset from authoritative standards and technical documents, providing the first public benchmark for this vertical.
- We propose DGSent, a domain-grounded sentence embedding model that restructures the representation space and alleviates domain-induced anisotropic collapse.
- We introduce retrieval-grounded knowledge distillation to learn terminology-aware embeddings for evidence retrieval and LLM-based decision support.

2 Related Work

2.1 Domain Adaptation of Embedding Models

Sentence embedding models map text into dense vectors that underpin semantic retrieval and retrieval-augmented generation (Bengio et al., 2013; Lewis et al., 2020). The field has evolved from static word vectors (Mikolov et al., 2013; Pennington et al., 2014) to contextual encoders such as Sentence-BERT, and further to contrastively trained general-purpose models that perform competitively on open-domain benchmarks (Wang et al., 2022; Muennighoff et al., 2023). These encoders nevertheless deteriorate on specialized corpora dominated by terminology and structured technical knowledge (Beltagy et al., 2019; Lee et al., 2020), motivating domain adaptation strategies at varying granularities: corpus-level continued pretraining on in-domain text (Wang et al., 2021a), instance-level contrastive fine-tuning over mined pairs (Karpukhin et al., 2020; Izacard et al., 2022), and task-level retrieval-aligned post-training (Xiao et al., 2022). Such efforts focus on data and training objectives while leaving the encoder's internal representation behavior largely intact, motivating advances at the model and representation level.

2.2 Geometric Restructuring for Anisotropy Mitigation

The geometry of the embedding space itself critically affects representation quality. Pretrained encoders suffer from representation degeneration, in which contextual vectors concentrate in a narrow cone and pairwise cosine similarities are systematically inflated (Gao et al., 2019; Ethayarajh, 2019), with the anisotropy further traced to a handful of rogue dimensions that dominate the similarity geometry (Kovaleva et al., 2021; Timkey and van Schijndel, 2021). Considerable efforts have been devoted to restoring isotropy at the representation level. Mu and Viswanath (Mu and Viswanath, 2018) proposed *all-but-the-top*, a post-hoc method that removes the top principal components to flatten the spectrum. Li et al. (Li et al., 2020) introduced BERT-flow to map embeddings to a Gaussian latent space, and Su et al. (Su et al., 2021) showed that a simple whitening transformation achieves comparable gains at far lower cost. Wang and Isola (Wang et al., 2019) formalized the alignment-uniformity decomposition that explains why contrastive objectives such as SimCSE (Gao et al., 2021) implicitly enforce isotropy, and Wang et al. (Wang et al., 2020) further constrained the singular-value spectrum during training to suppress dominant directions. These approaches establish isotropy as a tractable property of the embedding space but are developed on open-domain corpora and mainly

redistribute generic linguistic variance, leaving domain terminology insufficiently grounded in the resulting representations.

2.3 Knowledge Distillation for Lexical-Aware Retrieval

Grounding dense representations to domain terminology calls for an explicit signal that anchors embeddings to lexical evidence. Sparse retrievers such as BM25 (Robertson and Zaragoza, 2009) remain competitive on terminology-rich corpora, and SPLADE (Formal et al., 2021) extends this principle by learning contextualized term importance within a transformer. To transfer such lexical evidence into dense encoders, knowledge distillation has become the dominant mechanism. Hofstätter et al. (Hofstätter et al., 2020) proposed Margin-MSE to distill cross-encoder relevance margins into a bi-encoder student, and Santhanam et al. (Santhanam et al., 2022) distilled a late-interaction teacher into a compact student in ColBERTv2. More recently, Lupart et al. (Lupart et al., 2025) introduced DiSCo, which distills query-document similarity scores rather than fixed representations, and Tamber et al. (Tamber et al., 2026) showed that cross-encoder listwise distillation consistently improves dense retrievers where contrastive fine-tuning degrades. These teachers, however, are themselves neural rankers operating on dense semantics, while hybrid sparse-dense frameworks (Chen et al., 2024) fuse the two channels only at inference, leaving the dense encoder unaware of lexical evidence during representation learning.

However, none jointly tackles the anisotropic geometry and the lack of lexical grounding inherent to specialized retrieval. Therefore we propose DGSent, which couples geometric restructuring of the embedding space with hybrid sparse-dense distillation to produce an encoder that is both geometrically well-behaved and terminology-aware.

3 DGSent: Domain-Grounded Sentence Embeddings

We propose DGSent, a sentence embedding framework tailored to the airport runway maintenance domain. DGSent introduces two coordinated components that operate at the representation level: a Geometry Reconstruction Layer that restructures the embedding space of the backbone encoder, and a Retrieval-Grounded Knowledge Distillation module that aligns the embedding with domain evidence during training. The former remains in the encoder at inference, while the latter is used only for supervision. Figure 2 shows the overall pipeline.

3.1 Overview

Each training instance consists of a query sentence q and a set of candidate sentences $\{r_1, \ldots, r_K\}$ with graded relevance labels, sampled from the runway maintenance corpus. Every sentence $x \in \{q, r_1, \ldots, r_K\}$ is encoded by a shared Transformer that produces token-level hidden states

$$H = [\mathbf{h}_1, \ldots, \mathbf{h}_N], \quad \mathbf{h}_n \in \mathbb{R}^d, \tag{1}$$

where N is the sequence length and d is the hidden size. The Geometry Reconstruction Layer transforms H into the sentence embedding $\mathbf{d}_{\text{sent}} \in \mathbb{R}^d$, which is the final output of DGSent.

To enrich $\mathbf{d}_{\text{sent}}$ with domain knowledge, the Retrieval-Grounded Knowledge Distillation module operates alongside the encoder during training. For each sentence x, it retrieves terminology-rich entries from a curated runway knowledge memory through hybrid sparse-dense scoring and forms a teacher distribution over the retrieved support. The encoder is trained to reproduce this distribution from $\mathbf{d}_{\text{sent}}$, transferring domain evidence as a relational signal rather than a point-wise geometric target. DGSent thereby couples a restructured embedding geometry with retrieval-grounded domain supervision, yielding a representation backbone well-suited to downstream retrieval and semantic matching.

3.2 Geometry Reconstruction Layer

The Geometry Reconstruction Layer maps the token sequence H_x into a sentence embedding whose geometry is reshaped for retrieval rather than for the original pretraining objective. It consists of a token-level projection bridge that re-parameterizes the feature space, followed by a multi-view aggregation that condenses the reshaped tokens into a single sentence vector.

GeometryBridge Transformer-based sentence encoders are known to concentrate their representations within a narrow anisotropic cone, where a few dominant directions absorb most of the variance and lexically distinct sentences become nearly collinear. This effect is amplified on terminology-dense domain corpora, in which limited surface

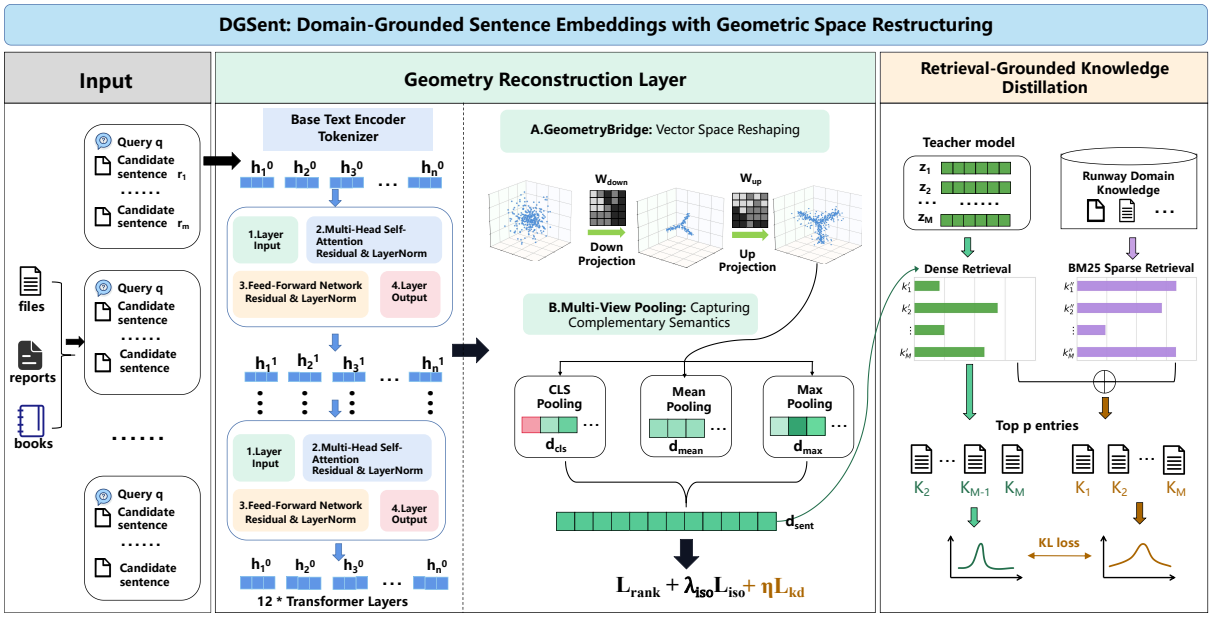

Figure 2: Overall architecture of **DGSent**. The Geometry Reconstruction Layer reshapes token representations via GeometryBridge ($\mathbf{W}_{\text{down}}$, $\mathbf{W}_{\text{up}}$) and aggregates them through multi-view pooling into the sentence embedding d_{sent}. At training time, the Retrieval-Grounded Knowledge Distillation module retrieves top-P evidence from a runway knowledge memory via hybrid BM25-dense scoring and distills the resulting relevance distribution into the encoder through a KL objective.

diversity further compresses semantically meaningful contrasts. GeometryBridge alleviates this issue with a learnable low-rank bottleneck inserted between the backbone output and the pooling stage. Each token representation $\mathbf{h}_n$ is projected into a rank-r subspace and then re-expanded to $\mathbb{R}^d$,

$$\tilde{\mathbf{h}}_n = \mathbf{W}_{\text{up}}\, \phi(\mathbf{W}_{\text{down}}\, \mathbf{h}_n)\,, \qquad (2)$$

where $\mathbf{W}_{\text{down}} \in \mathbb{R}^{r\times d}$ and $\mathbf{W}_{\text{up}} \in \mathbb{R}^{d\times r}$ with $r < d$, and $\phi(\cdot)$ denotes the GELU nonlinearity. The low-rank constraint encourages the encoder to retain directions informative for retrieval and suppress redundant dominant components, while the nonlinear re-expansion regenerates a d-dimensional space whose axes align with task-relevant variance rather than with the inherited anisotropic cone. Applying Eq. 2 at every position yields the reconstructed sequence $\tilde{H}_x = [\tilde{\mathbf{h}}_1, \ldots, \tilde{\mathbf{h}}_N]$.

Multi-View Pooling. No single pooling operator is universally optimal across the linguistic phenomena observed in a specialized domain. We therefore extract three complementary views from $\tilde{H}_x$: a CLS view $\mathbf{d}_{\text{cls}} = \tilde{\mathbf{h}}_{[\text{CLS}]}$ that captures sequence-level aggregation, a mean view $\mathbf{d}_{\text{mean}}$ obtained by averaging $\{\tilde{\mathbf{h}}_n\}_{n=1}^N$ to summarize content uniformly along the sequence, and a max view $\mathbf{d}_{\max}$ obtained by element-wise maximization over $\{\tilde{\mathbf{h}}_n\}_{n=1}^N$ that highlights salient per-dimension activations often associated with discriminative terminology. The three views are concatenated and projected back to $\mathbb{R}^d$,

$$\mathbf{d}_{\text{sent}} = \mathbf{W}_p\, [\mathbf{d}_{\text{cls}}; \mathbf{d}_{\text{mean}}; \mathbf{d}_{\max}] + \mathbf{b}_p, \qquad (3)$$

with $\mathbf{W}_p \in \mathbb{R}^{d\times 3d}$ and $\mathbf{b}_p \in \mathbb{R}^d$. The resulting $\mathbf{d}_{\text{sent}}$ inherits the restructured geometry of $\tilde{H}_x$ and serves as the sentence embedding output of DGSent.

3.3 Retrieval-Grounded Knowledge Distillation

This module anchors $\mathbf{d}_{\text{sent}}$ to the domain semantic manifold during training by retrieving terminology-rich evidence from a curated runway knowledge memory $\mathcal{M} = \{z_1, \ldots, z_M\}$ and distilling its relevance pattern into the encoder. Sparse and dense channels are combined to score memory entries, but here their fusion yields a teacher retrieval distribution that supervises the student, rather than a final retrieval result.

Hybrid Sparse-Dense Scoring For each memory entry $z_j \in \mathcal{M}$, BM25 (Robertson and Zaragoza, 2009) provides a lexical relevance score s_j^{bm25} between the surface form of x and z_j, while

the encoder provides a dense semantic score

$$s_j^{\text{dense}} \;=\; \cos\!\big(\mathbf{d}_{\text{sent}},\, \mathbf{e}_{z_j}\big), \tag{4}$$

where $\mathbf{e}_{z_j}$ is the precomputed embedding of z_j. The two channels are linearly fused as

$$\hat{s}_j \;=\; \lambda\, \bar{s}_j^{\text{bm25}} + (1-\lambda)\, s_j^{\text{dense}}, \tag{5}$$

where $\bar{s}_j^{\text{bm25}}$ is the lexical score min-max normalized to $[0,1]$ and $\lambda \in [0,1]$ trades off lexical exactness against distributional similarity. The lexical channel is decisive when terminology distinguishes otherwise paraphrastic sentences, while the dense channel preserves robustness to surface variation.

Top-P Distributional Distillation. The M entries are reranked by $\hat{s}_j$, and the top-P entries $\{z_{(1)}, \ldots, z_{(P)}\}$ are retained as the support set of x. The teacher distribution over this support is obtained by a temperature-scaled softmax over the hybrid scores,

$$p_T(j \mid x) \;=\; \frac{\exp(\hat{s}_{(j)}/\tau)}{\sum_{k=1}^{P} \exp(\hat{s}_{(k)}/\tau)}, \tag{6}$$

where τ controls the sharpness of the distribution. Compared with anchoring on a single top match, the distributional formulation absorbs complementary evidence from near-tie candidates and mitigates the influence of noisy retrievals. The student distribution is computed from $\mathbf{d}_{\text{sent}}$ over the same support using the dense channel alone,

$$p_S(j \mid x) \;=\; \frac{\exp(s_{(j)}^{\text{dense}}/\tau)}{\sum_{k=1}^{P} \exp(s_{(k)}^{\text{dense}}/\tau)}. \tag{7}$$

The teacher captures a BM25-enhanced consensus over authoritative entries surrounding x on the domain manifold, providing terminology-aware evidence that the dense channel alone cannot recover. Aligning p_S with p_T therefore internalizes lexical and terminological priors into the encoder, while imposing no point-wise geometric constraint on $\mathbf{d}_{\text{sent}}$ and leaving the representation space free to be reshaped by the Geometry Reconstruction Layer.

3.4 Training Objective

DGSent is trained end-to-end with a margin-based ranking loss, a distribution-level distillation loss that transfers retrieval-grounded domain knowledge into the encoder, and an isotropy regularizer that prevents the restructured space from collapsing back into a narrow cone. For triplets (x, x^+, x^-) mined from the domain corpus, where x^+ is a domain-relevant counterpart of x and x^- is a distractor, we abbreviate $s^+ = \cos(\mathbf{d}_{\text{sent}}, \mathbf{d}_{\text{sent}}^+)$ and $s^- = \cos(\mathbf{d}_{\text{sent}}, \mathbf{d}_{\text{sent}}^-)$. The ranking loss enforces a margin m between positive and negative pairs,

$$\mathcal{L}_{\text{rank}} \;=\; \max\!\big(0,\; m - s^+ + s^-\big). \tag{8}$$

The retrieval-grounded distillation loss aligns the student distribution in Eq. 7 with the teacher distribution in Eq. 6 via the Kullback-Leibler divergence,

$$\mathcal{L}_{\text{kd}} \;=\; \sum_{j=1}^{P} \tilde{p}_T^{\,j} \log \frac{\tilde{p}_T^{\,j}}{p_S^{\,j}}, \tag{9}$$

where $\tilde{p}_T^{\,j} = \text{sg}(p_T(j \mid x))$ and $p_S^{\,j} = p_S(j \mid x)$, and $\text{sg}(\cdot)$ denotes the stop-gradient operator applied to the teacher. Distributional alignment requires the student to reproduce the teacher ranking over memory entries rather than to match any absolute direction in the teacher embedding space, so that knowledge is transferred at a relational level and the anisotropy inherited from the teacher is not imprinted onto $\mathbf{d}_{\text{sent}}$.

To further counter representation collapse, let $\{\sigma_j\}_{j=1}^{d}$ be the singular values of the centred mini-batch embedding matrix in non-increasing order. The isotropy regularizer penalizes the energy of the leading K directions,

$$\mathcal{L}_{\text{iso}} \;=\; \sum_{j=1}^{K} \sigma_j^2 \Big/ \sum_{j=1}^{d} \sigma_j^2, \tag{10}$$

which is minimized when the spectrum is flat. The full training objective is

$$\mathcal{L} \;=\; \mathcal{L}_{\text{rank}} + \eta\, \mathcal{L}_{\text{kd}} + \lambda_{\text{iso}}\, \mathcal{L}_{\text{iso}}, \tag{11}$$

with $\eta, \lambda_{\text{iso}} \geq 0$ balancing distillation and isotropy, so that the Geometry Reconstruction Layer, multi-view pooling, and retrieval-grounded distillation are co-adapted so that the resulting $\mathbf{d}_{\text{sent}}$ jointly reflects discriminative geometry and grounded domain knowledge.

4 Experiments

We first describe the construction of the airport runway maintenance dataset and then evaluate DGSent from four complementary perspectives: (i) retrieval performance against representative sentence embedding models; (ii) embedding-space geometry, focusing on relevance separation and anisotropic

collapse; (iii) ablations of the Geometry Reconstruction Layer and retrieval-grounded knowledge distillation, with additional hyperparameter sensitivity analyses in the appendix; and (iv) downstream impact when DGSent is used as the embedding backbone in a runway-domain RAG pipeline. For all tables, the best results are marked in **bold**, and the second-best results are underlined.

4.1 Dataset Construction

We construct a graded retrieval benchmark for airport runway maintenance from three categories of authoritative materials: normative documents on runway maintenance, inspection and repair case reports from operational practice, and officially published textbooks covering pavement distress, inspection procedures, and operational response. Queries and candidate sentences are extracted verbatim from the source materials to preserve domain terminology and regulatory phrasing. The final dataset contains 766 queries, 1,352 candidate sentences, and 101,053 expert-annotated query-candidate pairs on a graded relevance scale, split into 80,843/10,105/10,105 for training, validation, and testing, with cross-split deduplication to avoid evaluation leakage. Corpus statistics, annotation protocols, quality control, and label distribution are reported in Appendix A.

4.2 Comparison with Sentence Encoders

We first evaluate whether DGSent improves sentence retrieval on the runway maintenance benchmark using MRR, NDCG@$\{3, 10\}$, and Recall@$\{3, 10\}$ as ranking metrics, with full definitions provided in Appendix B. We compare with five representative sentence encoders, all fine-tuned on the same training split: BERT (Devlin et al., 2019), a generic pretrained backbone; SBERT (Reimers and Gurevych, 2019), a siamese sentence encoder; MPNet (Song et al., 2020) and MiniLM (Wang et al., 2021b), widely used *sentence-transformers* encoders; and SimCSE (Gao et al., 2021), a contrastive encoder known to improve embedding uniformity. All systems are evaluated with the same candidate pool and retrieval protocol, so that performance differences primarily reflect representation quality rather than retrieval settings.

As shown in Table 1, the five baselines cluster within a narrow band, with MRR ranging from 0.51 to 0.58 and NDCG@10 from 0.23 to 0.27, and no single encoder dominates across metrics. This

Model	MRR↑	N@3↑	N@10↑	R@3↑	R@10↑
BERT	0.5103	0.2984	0.2286	0.4095	0.3148
SBERT	0.5486	0.3277	0.2694	0.4468	0.3617
MPNet	0.5508	0.3061	0.2385	0.4377	0.3381
MiniLM	0.5778	0.3022	0.2339	0.4415	0.3466
SimCSE	0.5375	0.3116	0.2588	0.4387	0.3492
DGSent	**0.8136**	**0.5440**	**0.5108**	**0.6552**	**0.6049**

Table 1: Retrieval performance on the airport runway maintenance benchmark. N and R denote NDCG and Recall, respectively.

near-tie among architecturally diverse encoders, all fine-tuned on the same in-domain data, reflects a known difficulty of terminology-dense verticals: standard similarity-based objectives tend to collapse fine-grained relevance distinctions into a narrow region of the embedding space, leaving little room for further ranking improvement regardless of the backbone. DGSent departs from this pattern by training the encoder to align its geometry with graded relevance levels directly, and improves over the strongest baseline by $+0.236$ MRR, $+0.216$ NDCG@3, $+0.241$ NDCG@10, $+0.208$ Recall@3, and $+0.243$ Recall@10. The gain is larger on NDCG than on MRR, indicating that DGSent produces a better-stratified ranking across relevance levels rather than merely promoting the top hit; the consistent Recall@$\{3, 10\}$ improvement further shows that more of the truly relevant evidence is concentrated within the top of the ranking, which is particularly valuable for downstream retrieval-augmented generation.

4.3 Anisotropy Analysis

We next test whether DGSent's retrieval gain is accompanied by a healthier embedding geometry. We report two diagnostics. The first is the mean cosine similarity at each relevance level, which reflects whether the encoder assigns distinguishable scores across relevance grades. The second is a uniformity score, defined as one minus the coefficient of variation of the eigenvalues of the embedding covariance matrix, with higher values indicating an isotropically used space rather than one collapsed along a few dominant directions. The probing data follow the label distribution in Table 5 of Appendix A, and formal definitions are provided in Appendix B.2.

Figure 3 makes the contrast immediately visible: every baseline curve is nearly flat across relevance levels, while DGSent's curve drops sharply and monotonically from level 5 to level 1. Ta-

Model	$L_5 \uparrow$	$L_4 \uparrow$	$L_2 \downarrow$	$L_1 \downarrow$	Unif $\uparrow$
BERT	0.9293	**0.9221**	0.9029	0.9078	0.2644
SBERT	**0.9331**	0.9062	0.8967	0.8843	0.3152
MPNet	0.6933	0.6494	0.6192	0.6019	0.3628
MiniLM	0.6767	0.6218	0.5865	0.5530	0.3987
SimCSE	0.3974	0.3038	0.2310	0.2163	0.4660
DGSent	0.7135	0.5548	**0.2036**	**0.0960**	**0.7513**

Table 2: Geometry diagnostics on the lexical-overlap-free probing set. L_ℓ denotes the mean cosine similarity at relevance level ℓ, and Unif is the uniformity score.

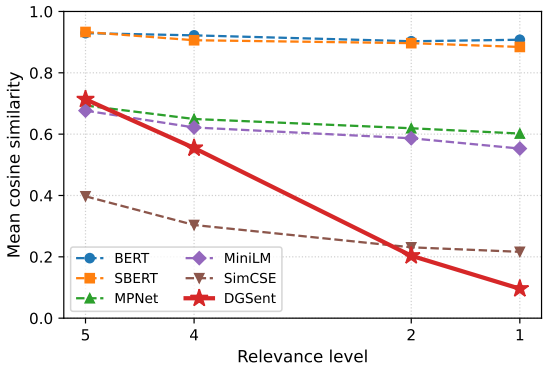

Figure 3: Mean cosine similarity across relevance levels. Unlike the nearly flat baseline curves, DGSent declines steeply and monotonically from level 5 to 1, indicating an embedding space organized by relevance.

ble 2 explains this contrast geometrically. BERT and SBERT compress all pairs into a narrow cone with cosine similarities concentrated around 0.88–0.93 regardless of relevance, and their uniformity scores remain low (0.2644, 0.3152), reproducing the cone-shaped collapse in Figure 1. Their seemingly high L_5 and L_4 values are an artefact of this collapse rather than evidence of semantic alignment, since irrelevant pairs receive almost equally high scores ($L_1 \approx 0.88$–0.91). MPNet and MiniLM partially relax this collapse, and SimCSE, whose contrastive pretraining targets isotropy, attains the highest baseline uniformity (0.4660); yet its cosine gap between level 5 and level 1 still stays below 0.20, indicating that a generic isotropy-promoting objective alone does not induce a relevance-aligned structure in this domain. DGSent simultaneously achieves the highest uniformity (0.7513) and by far the widest cosine spread (0.7135 vs. 0.0960, a gap of 0.62 that exceeds every baseline by more than 3×). The mild reduction in absolute similarity at the top relevance grades is the expected consequence of pushing irrelevant pairs apart in a non-collapsed space rather than packing everything into a narrow cone.

4.4 Ablation Study

We isolate the contributions of the two main components of DGSent: the Geometry Reconstruction Layer (GEOM) and Retrieval-Grounded Knowledge Distillation (KD). To disable GEOM, we remove the projection bridge and multi-view pooling, and revert to mean pooling over the backbone representations. To disable KD, we set $\eta = 0$, thereby removing the distillation objective. All variants are trained and evaluated under the same data split, training schedule, and remaining hyperparameters. Additional sensitivity analyses on the Geometry-Bridge bottleneck r and teacher support size P are provided in Appendix C. To jointly assess retrieval effectiveness and embedding geometry, we additionally compute the uniformity score defined in Section 4.3 on the same probing set used in the anisotropy analysis.

The ablation results are listed in Table 3 and visualized in Figure 4. We observe three patterns: (i) GEOM is the dominant lever. Removing it causes the largest collapse in both views: MRR drops by 0.2132 (a relative loss of 26.2%) and uniformity falls from 0.7513 to 0.4310, i.e., the embedding space reverts almost entirely to an anisotropic cone in which the four relevance grades are squeezed into a single dominant direction and distillation alone cannot pry them apart. (ii) KD is not a free add-on. Even when GEOM keeps the space globally well spread, disabling KD still removes 0.0972 MRR and 0.0969 NDCG@10 and lowers uniformity by 0.0829, indicating that an isotropic geometry without retrieval-grounded supervision fails to capture the fine-grained terminological evidence separating adjacent relevance grades. (iii) The two components are complementary, not redundant. The full model exceeds the sum of single-component gains on every metric, and MRR and Unif rise in lockstep across the three variants with no crossover, ruling out the alternative explanation that the gains stem from either channel alone or from incidental optimization effects. Overall, GEOM opens up the representation space

GEOM	KD	MRR$\uparrow$	N@3$\uparrow$	N@10$\uparrow$	R@3$\uparrow$	R@10$\uparrow$
✗	✓	0.6004	0.4096	0.3527	0.5276	0.4761
✓	✗	0.7164	0.4734	0.4139	0.6012	0.5204
✓	✓	**0.8136**	**0.5440**	**0.5108**	**0.6552**	**0.6049**

Table 3: Component-wise ablation. GEOM: Geometry Reconstruction Layer. KD: Retrieval-Grounded Knowledge Distillation.

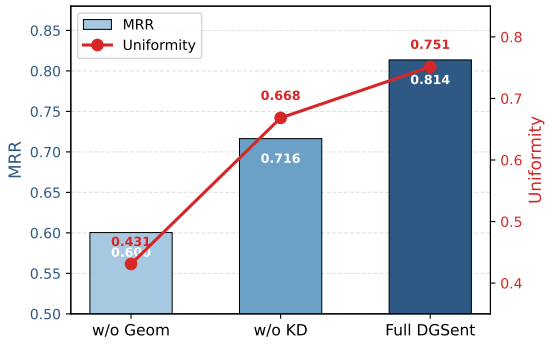

Figure 4: Joint retrieval-geometry view of the ablation. Bars denote MRR (left axis); the line denotes the uniformity score (right axis). Retrieval and geometry move together across variants, confirming that the two modules act on the same underlying failure mode.

to counter global anisotropic collapse, KD injects fine-grained retrieval-grounded signal to counter local semantic drift, and only their joint use produces the relevance-aligned geometry behind DGSent's retrieval gain.

4.5 Downstream Impact on RAG

We finally examine whether DGSent's representation-level gains propagate to the end-to-end RAG pipeline. DGSent is compared against the two representative baseline encoders, MiniLM and MPNet, with all other components fixed: each encoder embeds the same corpus, retrieves evidence under identical settings, and feeds it to the same generator. Evaluation is conducted on 4,000 expert-curated QA pairs covering regulation lookup, procedural reasoning, and defect-handling scenarios. We use Hit@1 to evaluate whether the most relevant evidence is ranked first, and report hallucination rate, completeness, and composite F1 to assess generation quality. Detailed descriptions of the QA dataset are provided in Appendix A.2, and metric definitions are provided in Appendix B.3.

The results are listed in Table 4. We observe three patterns. (i) DGSent substantially improves first-rank evidence retrieval. DGSent reaches 80.77% Hit@1, outperforming MiniLM by 9.23 points and MPNet by 26.15 points, showing that the improved encoder is better at prioritizing the exact regulation-relevant passage rather than merely retrieving broadly related text. (ii) Better retrieval translates into more reliable generation. Compared with MiniLM, DGSent reduces the hallucination rate from 30.53% to 22.59%, a 26.0% relative reduction, while improving F1 by 6.72 points. This indicates that higher-quality first-rank evidence gives the generator a stronger grounding signal and reduces reliance on parametric knowledge. (iii) The gains are consistent across the full pipeline. Hit@1, completeness, and F1 improve monotonically from MPNet to MiniLM to DGSent, while hallucination rate decreases in the same order. The relatively modest completeness gain of 2.94 points over MiniLM suggests that answer coverage is also constrained by the available corpus content, but the simultaneous improvements in retrieval precision, hallucination reduction, and F1 confirm that DGSent provides a cleaner evidence signal for downstream RAG.

Encoder	Hit@1↑	HR↓	Com↑	F1↑
MPNet	54.62%	52.76%	51.09%	44.31%
MiniLM	71.54%	30.53%	64.72%	64.99%
DGSent	**80.77%**	**22.59%**	**67.66%**	**71.71%**

Table 4: Downstream RAG performance under different sentence encoders, with all other pipeline components held fixed. Hit@1 evaluates first-rank evidence retrieval; HR, Com, and F1 denote hallucination rate, completeness, and the composite generation score, respectively.

5 Conclusion

In this paper, we presented DGSent, a domain-grounded sentence embedding model for airport runway maintenance. To support embedding research in this safety-critical vertical, we constructed an expert-curated knowledge dataset from authoritative standards and technical documents. Motivated by the observation that general-purpose encoders suffer from global anisotropic collapse and local semantic drift on runway maintenance texts, DGSent combines global geometry restructuring with retrieval-grounded knowledge distillation. The former expands the effective semantic space and improves inter-category separability, while the latter injects terminology-rich domain evidence into the encoder. Extensive experiments show that DGSent consistently improves semantic matching, representation geometry, and downstream RAG performance, leading to more accurate evidence retrieval and more reliable generation. These results suggest that explicitly modeling representation geometry and domain knowledge is a practical direction for adapting sentence encoders to safety-critical domain applications.

Limitations

This work has several limitations. First, the current evaluation is based on text-only standards, technical documents, and QA pairs; future work should incorporate multimodal inspection data such as runway images, sensor readings, and structured maintenance records. Second, the effectiveness of retrieval-grounded distillation depends on the quality and coverage of the domain knowledge memory. When authoritative documents are incomplete, outdated, or inconsistent, the teacher signal may also become biased.

Ethics Statement

This work studies domain-specific sentence embedding for airport runway maintenance. The experimental data are constructed from authoritative standards and technical documents and do not involve personal information, private user data, or sensitive individual-level records. We will ensure that any released resources comply with the licensing and usage requirements of the original sources. Since runway maintenance is safety-critical, the proposed model should be regarded as a representation learning component for evidence matching and decision support, rather than a substitute for certified expert judgment. Potential biases or incompleteness in source documents and pre-trained models may still affect the learned embeddings, as bias mitigation is not the focus of this work.

References

Peter Anderson, Mano Vikash Janardhanan, Jason He, Wei Cheng, and Charlie Flanagan. 2024. Greenback bears and fiscal hawks: Finance is a jungle and text embeddings must adapt. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 362–370, Miami, Florida, US. Association for Computational Linguistics.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 3615–3620, Hong Kong, China. Association for Computational Linguistics.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. 2013. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186. Association for Computational Linguistics.

Kawin Ethayarajh. 2019. How contextual are contextualized word representations? comparing the geometry of bert, elmo, and gpt-2 embeddings. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 55–65.

Federal Aviation Administration. 2014. AC 150/5380-6c: Guidelines and procedures for maintenance of airport pavements. Active; cancels AC 150/5380-6B.

Thibault Formal, Benjamin Piwowarski, and Stéphane Clinchant. 2021. SPLADE: Sparse lexical and expansion model for first stage ranking. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2288–2292, New York, NY, USA. Association for Computing Machinery.

Jun Gao, Di Li, Xu Wang, Tong Qin, and Tie-Yan Liu. 2019. Representation degeneration problem in training natural language generation models. In *ICLR*.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. In *Proceedings of EMNLP 2021*, pages 6894–6910. Association for Computational Linguistics.

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

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, and Vaughan. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360. Association for Computational Linguistics.

Sebastian Hofstätter, Sophia Althammer, Mete Schröder, Mete Sertkan, and Allan Hanbury. 2020. Improving efficient neural ranking models with cross-architecture knowledge distillation. *arXiv preprint arXiv:2010.02666*.

International Air Transport Association. 2024. Annual review 2024.

International Civil Aviation Organization. 2022. Annex 14 to the convention on international civil aviation: Aerodromes, volume i – aerodrome design and operations. Superseded previous editions on 3 November 2022.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2022. Unsupervised dense information retrieval with contrastive learning. *Transactions on Machine Learning Research*.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 6769–6781, Online. Association for Computational Linguistics.

Olga Kovaleva, Saurabh Kulshreshtha, Anna Rogers, and Anna Rumshisky. 2021. BERT busters: Outlier dimensions that disrupt transformers. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3392–3405.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474.

Bohan Li, Hao Zhou, Junxian He, Mingxuan Wang, Yiming Yang, and Lei Li. 2020. On the sentence embeddings from pre-trained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 9119–9130, Online. Association for Computational Linguistics.

Simon Lupart, Mohammad Aliannejadi, and Evangelos Kanoulas. 2025. DiSCo: LLM knowledge distillation for efficient sparse retrieval in conversational search. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*. Association for Computing Machinery.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. 2024. Large language models: A survey. *arXiv preprint arXiv:2402.06196*.

Jiaqi Mu and Pramod Viswanath. 2018. All-but-the-top: Simple and effective postprocessing for word representations. In *International Conference on Learning Representations (ICLR)*.

Niklas Muennighoff, Nouamane Tazi, Loic Magne, and Nils Reimers. 2023. MTEB: Massive text embedding benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2014–2037, Dubrovnik, Croatia. Association for Computational Linguistics.

OpenAI. 2023. GPT-4 technical report. *Preprint*, arXiv:2303.08774.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, pages 3982–3992. Association for Computational Linguistics.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389.

Keshav Santhanam, Omar Khattab, Jon Saad-Falcon, Christopher Potts, and Matei Zaharia. 2022. ColBERTv2: Effective and efficient retrieval via lightweight late interaction. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3715–3734, Seattle, United States. Association for Computational Linguistics.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. MPNet: Masked and permuted pre-training for language understanding. In *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*. Curran Associates Inc.

Jianlin Su, Jiarun Cao, Weijie Liu, and Yangyiwen Ou. 2021. Whitening sentence representations for better semantics and faster retrieval. *arXiv preprint arXiv:2103.15316*.

Manveer Singh Tamber, Suleman Kazi, Vivek Sourabh, and Jimmy Lin. 2026. Contrastive learning falls short: Improving dense retrieval with cross-encoder listwise distillation and synthetic data. In *Advances in Information Retrieval: 48th European Conference on Information Retrieval*, pages 104–121. Springer Nature Switzerland.

Yixuan Tang and Yi Yang. 2024. Do we need domain-specific embedding models? an empirical investigation. *arXiv preprint arXiv:2409.18511*.

William Timkey and Marten van Schijndel. 2021. All bark and no bite: Rogue dimensions in transformer language models undermine monotonicity. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4527–4546.

Kexin Wang, Nils Reimers, and Iryna Gurevych. 2021a. TSDAE: Using transformer-based sequential denoising auto-encoder for unsupervised sentence embedding learning. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 671–688, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Li Wang, Rui Zhao, and Ying Liu. 2019. Characterizing the impact of geometric properties of word embeddings on task performance. In *Proceedings of the Third Workshop on Evaluating Vector Space Representations for NLP*, pages 8–13, Minneapolis, USA. Association for Computational Linguistics.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2022. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*.

Lingxiao Wang, Jing Huang, Kevin Huang, Ziniu Hu, Guangtao Wang, and Quanquan Gu. 2020. Improving neural language generation with spectrum control. In *International Conference on Learning Representations (ICLR)*.

Wenhui Wang, Hangbo Bao, Shaohan Huang, Li Dong, and Furu Wei. 2021b. MiniLMv2: Multi-head self-attention relation distillation for compressing pretrained transformers. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3146–3160, Online. Association for Computational Linguistics.

Yubai Wei, Jiale Han, and Yi Yang. 2025. Adapting general-purpose embedding models to private datasets using keyword-based retrieval. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 6856–6870, Vienna, Austria. Association for Computational Linguistics.

Shitao Xiao, Zheng Liu, Yingxia Shao, and Zhao Cao. 2022. RetroMAE: Pre-training retrieval-oriented language models via masked auto-encoder. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 538–548, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. 2023. C-Pack: Packaged resources to advance general chinese embedding. *Preprint*, arXiv:2309.07597.

A Dataset Details

A.1 Runway Maintenance Retrieval Dataset

Corpus sources and preprocessing. The retrieval dataset is built from three categories of authoritative materials on runway and airside maintenance: (i) official normative documents on airport runway maintenance, (ii) inspection and repair case reports produced in practical maintenance operations, and (iii) officially published textbooks covering pavement distress, inspection procedures, and operational response. The vast majority of these source materials are written in Chinese, as they originate from Chinese civil aviation regulations, industrial standards, and operational practice; the dataset is therefore predominantly Chinese in its original form, and all examples shown in this paper are translated into English for presentation only, while training, evaluation, and released annotations are performed on the original Chinese text. Queries and candidate sentences are extracted directly from these sources to preserve domain terminology and regulatory phrasing; we apply only format unification, deduplication, and split organization, without any paraphrastic rewriting.

Query and candidate construction. Queries reflect realistic information needs in runway maintenance, including hazard identification, abnormality diagnosis, inspection requirements, disposal procedures, and emergency response. Candidate sentences are extracted from the same source materials as concise knowledge units that can serve as retrievable response targets. The dataset is released as two text files (queries and candidate sentences) and three tabular files (train, validation, test), with each record stored as (query_id, candidate_id, relevance_score).

Annotation protocol. Annotation is performed by domain experts with operational experience in runway maintenance, following a five-level graded relevance scale: L5, the candidate sentence provides explicit response content, actionable guidance, or authoritative regulatory support; L4, substantial but less direct support; L3, useful back-

ground or supporting information; L2, weakly relevant with limited operational value; L1, marginal topical overlap without meaningful information. Ambiguous samples are reconciled under unified guidelines, and overlapping or duplicated records across splits are removed to avoid evaluation leakage.

Label distribution. Table 5 reports the label distribution across splits. The dataset exhibits a non-uniform but well-covered relevance distribution, with both weakly and highly relevant pairs represented in every split.

A.2 RAG Question-Answering Dataset

Table 6 presents three representative samples drawn from a professional dataset of 4,000 high-standard question–answer records focused on airport runway management and maintenance. As with the retrieval dataset, the original records are predominantly in Chinese, and the samples in Table 6 are translated into English for display purposes only. The three samples concern, respectively, applicable pavement types for non-stop runway construction, potential defects of asphalt pavement used on airport aprons, and material-based classification of airport pavements together with their applicable service scenarios. All question and answer contents are strictly sourced from official industrial specifications and authoritative technical documents, with no semantic alteration, subjective revision, or paraphrasing, ensuring credibility and professional accuracy. The dataset offers comprehensive topic coverage over the core subdomains of runway maintenance, including pavement damage repair, underground facility management, and material selection across practical application scenarios.

A.3 Dataset Quality Assurance

All textual content in both the retrieval dataset and the RAG question-answering dataset originates from authoritative sources: official airport runway maintenance specifications, industrial technical standards, peer-reviewed textbooks, and inspection and repair case reports collected from operational practice. Since the source materials are predominantly Chinese, the released datasets and all annotation work are conducted on the original Chinese text, and English versions appearing in this paper are provided for illustration only and do not affect the underlying Chinese data or labels. No content is generated or rewritten by large language models, so terminology, regulatory phrasing, and procedural descriptions remain faithful to the original references. Annotation and validation are carried out by domain experts with operational experience in runway maintenance: each relevance label is first assigned by one annotator and then independently verified by a second, with disagreements reconciled under a unified annotation guideline, and a random subsample is re-examined by a senior reviewer to monitor labeling consistency. For the question-answering dataset, all 4,000 question–answer pairs are fully reviewed against the source documents to ensure factual correctness and the absence of hallucinated content. Across both datasets, exact and near-duplicate records are removed, and any overlap between the training, validation, and test splits is explicitly filtered out to prevent evaluation leakage.

Split	Total	L1	L2	L3	L4	L5
Train	80,843	31,163	20,744	13,464	10,044	5,428
Dev	10,105	3,997	2,575	1,666	1,224	643
Test	10,105	3,337	2,581	1,718	1,482	987
All	101,053	38,497	25,900	16,848	12,750	7,058

Table 5: Label distribution of the runway maintenance retrieval dataset over query-candidate pairs.

B Evaluation Metrics

This appendix gives the precise definitions of the evaluation metrics used in the main experiments. Let $\mathcal{Q}$ denote the query set, and for a query $q \in \mathcal{Q}$ let π_q be the ranking returned over the candidate pool, with $\pi_q(i)$ its i-th ranked candidate. Let $\mathrm{rel}(q, d) \in \{1, \ldots, 5\}$ be the expert-annotated graded relevance label of candidate d with respect to q, and let $\mathbb{I}(\cdot)$ denote the indicator function. A candidate is treated as relevant when $\mathrm{rel}(q, d) \geq 3$, consistent with the graded scale defined in Appendix A.1.

B.1 Retrieval Metrics

MRR Mean Reciprocal Rank rewards how early the first relevant candidate appears. Let r_q be the rank of the first relevant candidate in π_q,

$$\mathrm{MRR} = \frac{1}{|\mathcal{Q}|} \sum_{q \in \mathcal{Q}} \frac{1}{r_q}. \tag{12}$$

NDCG@k Normalized Discounted Cumulative Gain at cutoff k integrates graded relevance with a

Query	Answer
During airport pavement maintenance, if uneven settlement or loose surface is found in the subgrade, what materials or construction processes might be causing the problem?	Uneven settlement and surface loosening commonly have two causes. First, there are problems with the subgrade itself: for example, using poorly graded gravel mixtures to fill the subgrade, resulting in excessive voids that lead to water accumulation and softening; or insufficient compaction of sandy soil. Second, improper construction of the rubble base: if the rubble stones are irregularly shaped (e.g., flat, pointed) or not compacted properly, the interlocking effect will be insufficient, also causing loosening. Furthermore, if easily weathered rock is used directly for the rubble base, long-term use will lead to cracking and settlement. Therefore, during maintenance, it is necessary to check not only the compaction and drainage of the subgrade, but also the quality of the rubble base stones and the degree of compaction.
What problems may arise if asphalt pavement is adopted for airport aprons?	Asphalt pavements on airport aprons mainly face two issues. First, prolonged airflow from aircraft engines impairs pavement strength, especially when temperatures exceed 60°C and asphalt softens. Second, aviation oil corrodes the asphalt surface, dissolving asphalt binder, loosening aggregate and forming potholes, which eventually damages the pavement. Accordingly, asphalt pavement is rarely used at runway ends and aprons.
What are the material-based classifications of airport pavements? What scenarios are they respectively applicable to?	Airport pavements are mainly categorized by materials as follows: 1. Cement concrete pavement. It features high strength and good service performance yet requires high initial investment and long curing time, widely applied to main runways. 2. Asphalt pavement. It boasts smooth surface and immediate traffic accessibility, suitable for non-stop construction. Caution is needed in high-temperature and oil-contaminated areas. 3. Sand and gravel pavement. It has low bearing capacity and tends to produce dust and mud, seeing limited application nowadays. 4. Earth pavement. It is low-cost and easy to construct, primarily used for light aircraft airports, military emergency runways and emergency landing strips of large airports.5. Water aerodrome, exclusively for seaplanes.

Table 6: Display of question-answer samples in question answering dataset.

logarithmic position discount,

$$\mathrm{DCG}_q@k \;=\; \sum_{i=1}^{k} \frac{2^{\mathrm{rel}(q,\pi_q(i))}-1}{\log_2(i+1)}, \qquad (13)$$

$$\mathrm{NDCG}_q@k \;=\; \mathrm{DCG}_q@k \,/\, \mathrm{IDCG}_q@k, \qquad (14)$$

where $\mathrm{IDCG}_q@k$ is the DCG of the ideal ranking sorted by ground-truth relevance.

Recall@k The fraction of relevant candidates that are recovered within the top-k ranking. Let $\mathcal{R}_q = \{d : \mathrm{rel}(q,d) \geq 3\}$ denote the set of relevant candidates for query q,

$$\mathrm{R}@k \;=\; \frac{1}{|\mathcal{Q}|} \sum_{q\in\mathcal{Q}} \frac{\big|\{\pi_q(1),\ldots,\pi_q(k)\} \cap \mathcal{R}_q\big|}{|\mathcal{R}_q|}. \qquad (15)$$

B.2 Geometry Diagnostics

Let $\{\mathbf{v}_i\}_{i=1}^{N} \subset \mathbb{R}^d$ be the embeddings produced by a given encoder on the probing set, all ℓ_2-normalized before any cosine-based computation.

Relevance-conditioned cosine similarity For each relevance level ℓ, the mean cosine similarity over the corresponding pair subset $\mathcal{P}_\ell$ is

$$\bar{s}_\ell \;=\; \frac{1}{|\mathcal{P}_\ell|} \sum_{(q,d)\in\mathcal{P}_\ell} \cos\!\big(\mathbf{v}_q, \mathbf{v}_d\big). \qquad (16)$$

A discriminative encoder is expected to yield $\bar{s}_\ell$ that increases monotonically with ℓ.

Uniformity To quantify whether embeddings spread evenly across directions of the representation space rather than concentrate along a few dominant axes, we use a spectrum-based uniformity score. Stacking the embeddings into $\mathbf{X} \in \mathbb{R}^{N\times d}$ with mean $\boldsymbol{\mu}$, the centred covariance is

$$\boldsymbol{\Sigma} \;=\; \tfrac{1}{N-1}(\mathbf{X} - \mathbf{1}\boldsymbol{\mu}^\top)^\top(\mathbf{X} - \mathbf{1}\boldsymbol{\mu}^\top), \qquad (17)$$

with eigenvalues $\{\lambda_j\}_{j=1}^{d}$. Let

$$\bar{\lambda} = \frac{1}{d}\sum_{j=1}^{d}\lambda_j, \qquad \sigma(\lambda) = \sqrt{\frac{1}{d}\sum_{j=1}^{d}(\lambda_j - \bar{\lambda})^2}, \qquad (18)$$

denote the mean and standard deviation of the spectrum. The uniformity score is

$$\mathcal{U} \;=\; 1 - \frac{\sigma(\lambda)}{\bar{\lambda}}. \qquad (19)$$

$\mathcal{U} \to 1$ when the spectrum is flat and variance is shared across all directions; smaller values indicate that the embedding space collapses onto a few dominant axes.

B.3 Downstream RAG Metrics

To evaluate the impact of the retriever on downstream generation, we measure both the quality of generated answers and the quality of retrieved evidence.

Answer-level metrics Let N_t denote the total number of evaluation questions, N_h the number of responses containing unsupported content, and N_c the number of factually correct responses. The hallucination rate and completeness are defined as

$$\mathrm{HR} = \frac{N_h}{N_t}, \qquad \mathrm{Acc} = \frac{N_c}{N_t}, \tag{20}$$

$$\mathrm{Comp} = \frac{1}{N_t} \sum_{i=1}^{N_t} C_i, \tag{21}$$

where $C_i \in [0, 1]$ measures the coverage of key information against the reference answer, scored by domain experts. The composite score is the weighted harmonic mean of Acc, Comp, and $1 - \mathrm{HR}$,

$$\mathrm{F1} = \frac{2}{\frac{0.5}{\mathrm{Acc}} + \frac{0.5}{\mathrm{Comp}} + \frac{1}{1 - \mathrm{HR}}}, \tag{22}$$

which jointly rewards factual correctness, informational coverage, and faithfulness to the retrieved evidence.

Retrieval-side metrics Let $\mathcal{D}_q$ be the set of ground-truth relevant documents for query q and $\mathcal{R}_q(k)$ the top-k retrieved documents. Hit@k measures whether any relevant document appears in the top-k list,

$$\mathrm{Hit@}k = \frac{1}{|\mathcal{Q}|} \sum_{q \in \mathcal{Q}} \mathbb{I}\big(\mathcal{R}_q(k) \cap \mathcal{D}_q \neq \emptyset\big). \tag{23}$$

MAP@k further evaluates the overall ranking quality. Let $P_i(q) = \frac{1}{i} \sum_{j=1}^{i} \mathrm{rel}(d_j)$ be the precision at position i. Then

$$\mathrm{AP@}k(q) = \frac{1}{\min(|\mathcal{D}_q|, k)} \sum_{i=1}^{k} \mathrm{rel}(d_i)\, P_i(q), \tag{24}$$

$$\mathrm{MAP@}k = \frac{1}{|\mathcal{Q}|} \sum_{q \in \mathcal{Q}} \mathrm{AP@}k(q), \tag{25}$$

where $\mathrm{rel}(d_i) \in \{0, 1\}$ indicates whether the i-th retrieved document is relevant.

C Hyperparameter Analysis

We analyze two key hyperparameters: the bottleneck dimensionality r of the Geometry Reconstruction Layer, and the hybrid retrieval configuration (λ, P) of the retrieval-grounded distillation module. All other components and training settings are held fixed.

r	MRR↑	NDCG@10↑	R@10↑
128	0.5846	0.3291	0.4467
192	**0.7164**	**0.4139**	**0.5204**
384	0.6972	0.3981	0.5063
512	0.6036	0.3437	0.4728

Table 7: Effect of the bottleneck dimensionality r of the Geometry Reconstruction Layer.

λ	P	MRR↑	NDCG@10↑	R@10↑
0.2	4	0.7994	0.4978	0.5931
0.3	**4**	**0.8136**	**0.5108**	**0.6049**
0.4	4	0.7433	0.4267	0.5396
0.3	2	0.7822	0.4880	0.5843
0.3	6	0.7520	0.4308	0.5514

Table 8: Effect of the sparse-dense fusion weight λ and the top-P support size in retrieval-grounded distillation.

C.1 Bottleneck Dimensionality

Table 7 reports retrieval performance as r varies in $\{128, 192, 384, 512\}$. Performance is non-monotonic in r: r=192 attains the best results on all three metrics, while r=512 leaves the geometric regularization too weak and r=128 over-compresses the semantic content. The gain of the Geometry Reconstruction Layer therefore comes from a balanced trade-off between information preservation and structural constraint, rather than from dimensionality reduction itself.

C.2 Hybrid Scoring Weight and Support Size

Table 8 reports the effect of the sparse-dense fusion weight λ in Eq. 5 and the top-P support size used to construct the teacher distribution in Eq. 6. A moderate weight λ=0.3 together with P=4 yields the best performance: weaker BM25 weighting under-utilizes terminological evidence, whereas stronger BM25 weighting over-emphasizes lexical overlap. Similarly, P=2 provides insufficient support while P=6 admits noisy candidates.