

A Benchmark Study of Adversarial Attacks and Defenses in Telecom Retrieval Systems

Gonçalo Ferreira, Eva Maia, Isabel Praça

GECAD - Research Group on Intelligent Engineering and Computing for Advanced Innovation and Development

Instituto Superior de Engenharia do Porto (ISEP)

Porto, Portugal

{gdclf@isep.ipp.pt, egm@isep.ipp.pt, icp@isep.ipp.pt}

Abstract—Dense Hierarchical Retrieval (DHR) models are increasingly used to support information access and automation in telecom-domain applications, particularly for large-scale technical corpora such as 3GPP specifications. However, their robustness under adversarial input perturbations remains insufficiently understood in domain-specific settings. In this paper, we present a telecom-domain adversarial benchmarking study for DHR models under black-box evasion attacks. Utilizing the Telco-DPR dataset, a highly specialized technical corpus based on 3GPP standards, we evaluate the impact of adversarial text perturbations on retrieval performance across multiple model variants. We investigate a broad set of black-box attack strategies, including gradient-based, score-based, and decision-based methods, and analyze their transferability in telecom-domain retrieval scenarios. In addition, we evaluate a range of defense mechanisms, including lightweight input transformations and neural re-ranking models, under computationally constrained inference settings. Our results show that adversarial perturbations consistently degrade retrieval performance across all evaluated models, as well as significant trade-offs between robustness, accuracy, and computational efficiency across defense strategies. Therefore, this work provides a systematic and reproducible benchmark for understanding adversarial vulnerabilities in telecom-domain dense retrieval systems and offers insights for designing more robust information retrieval models.

Index Terms—Adversarial Machine Learning, Black-Box Attacks, Dense Retrieval, Real-Time Networks, System Security

I. INTRODUCTION

Modern telecommunication ecosystems increasingly rely on Artificial Intelligence (AI) to support information-intensive operations, including automated querying, diagnostics, and retrieval of large-scale technical documentation [1], [2]. In particular, telecom standards such as 3GPP specifications represent complex, evolving corpora that require efficient semantic retrieval mechanisms to enable rapid knowledge access [3].

Dense Hierarchical Retrieval (DHR)-style architectures, have emerged as effective solutions for encoding semantic relationships between queries and technical passages. These models enable scalable information retrieval by embedding queries and documents into a shared space, allowing fast similarity search over large knowledge bases. Despite their effectiveness, dense retrieval systems are known to be vulnerable to adversarial perturbations in input text [4], [5]. Prior work in adversarial machine learning has demonstrated that carefully crafted modifications to input queries can significantly degrade

model performance, even when perturbations are semantically subtle [6]. However, the majority of existing studies focus on general-domain datasets such as web text or open-domain question answering, leaving a gap in understanding how these vulnerabilities manifest in telecom-specific retrieval environments. This gap is particularly important because telecom-domain documents exhibit unique characteristics: they are highly structured, densely technical, and often governed by strict semantic conventions. As a result, the behavior of adversarial attacks and defense mechanisms in this domain may differ significantly from general Natural Language Processing (NLP) settings.

To address this limitation, this paper introduces a systematic benchmarking study of adversarial robustness in telecom-domain dense retrieval systems. Rather than focusing on deployment or system integration, our goal is to provide a controlled and reproducible evaluation framework that quantifies the impact of adversarial perturbations under a variety of attack and defense scenarios. Specifically, we make the following contributions:

- We construct an evaluation pipeline for studying adversarial robustness of dense retrieval models using Telco-DPR dataset;
- We benchmark multiple categories of adversarial attacks to assess their effectiveness in degrading retrieval performance;
- We evaluate lightweight and neural defense strategies under computational constraints relevant to efficient inference;
- We provide a detailed empirical study of the trade-offs between retrieval accuracy, adversarial robustness, and computational overhead across models and defenses.

The remainder of this paper is organized as follows. Section II provides the background on dense retrieval, adversarial threats, and existing defense mechanisms. Section III details the methodology, establishing the target retrieval models, the proxy architectures, and the black-box threat model used in our evaluation. Section IV reports the results of adversarial attacks. Section V evaluates the proposed defense strategies and their robustness–efficiency trade-offs. Finally, Section VI concludes the paper and outlines directions for future work.

II. STATE OF THE ART AND RELATED WORK

To support automated decision-making, next-generation telecom environments increasingly rely on large-scale processing of technical documentation, particularly 3GPP specifications [3]. These documents define complex protocol structures and operational procedures for mobile networks and therefore represent a challenging domain for information extraction and semantic retrieval. To enable machines to parse and extract actionable insights from this unstructured data, NLP-based dense retrieval systems have been widely adopted. In particular, Dense Information Retrieval (DIR) approaches and DHR architectures have been explored to map queries and documents into shared semantic embedding spaces for similarity-based search [7]. Within the current literature, the DHR architecture has emerged as the standard for this task, allowing automated diagnostics to extract relevant procedures by understanding semantic meaning [3]. However, the integration of these models into telecom-domain retrieval pipelines introduces new security considerations. Recent studies in adversarial machine learning have shown that dense retrieval models are vulnerable to carefully crafted input perturbations, which can significantly alter retrieval outcomes even when modifications are semantically minimal.

A. Adversarial Attacks in Retrieval Systems

While traditional telecommunications security often focuses on availability through resource exhaustion [8], Adversarial Machine Learning (AML) vectors represent emerging threats that directly target the integrity of the intelligence layer. Compromising an AI model's decision-making process inherently leads to service degradation. Prior work has demonstrated the vulnerability of machine learning systems in telecom-related domains including wireless signal classification [5] and control-plane optimization tasks [4]. These findings indicate that learned representations in communication systems can be sensitive to small input variations.

These vulnerabilities are magnified in Information Retrieval (IR) systems. Because dense retrievers map text into continuous vector spaces, minor coordinate shifts in embedding spaces drastically alter semantic interpretation. While early literature focused on white-box attacks, real-world adversaries in operational environments typically operate under strict black-box constraints, lacking access to model architecture or weights [9]. Attackers compromise these systems using three primary classes of methods. First, gradient-based methods, such as Fast Gradient Sign Method (FGSM) and Projected Gradient Descent (PGD), leverage gradient information from surrogate models to craft transferable adversarial examples [10], [11]. Second, score-based methods like ZOO [12] estimate vulnerabilities by analyzing the system's confidence outputs. Finally, decision-based methods, including HopSkipJump (HSJ) [13] and Boundary attacks, operate under strict black-box conditions by iteratively probing model outputs to construct adversarial examples [6].

B. Defenses and Latency Constraints

To mitigate adversarial vulnerabilities, a wide range of defense strategies has been proposed in the literature. One common approach is adversarial training, where models are trained on adversarially perturbed samples to improve robustness [6]. However, such methods require retraining and are computationally expensive, limiting their applicability in large-scale or frequently updated systems. Another category includes input transformation techniques, which aim to reduce adversarial noise prior to inference. These include feature squeezing, randomized smoothing, and dimensionality reduction techniques such as PCA-based filtering [14]. While efficient, these methods may introduce unintended distortion of semantic information.

More advanced approaches include neural re-ranking models such as Cross-Encoders, which refine retrieval results by performing fine-grained query-document interaction [15], and knowledge distillation frameworks, where compact student models are trained to approximate robust teacher models [16]. However, State of the Art deep-learning defenses introduce severe latency overheads and massive memory footprints, often requiring large containers (e.g., exceeding 11 GB) due to heavy model dependencies. This makes them unsuitable for real-time telecommunication environments. In contrast, lightweight linear methods based on geometric projections, such as Ridge regression, offer a superior alternative. By correcting embedding perturbations using closed-form transformations [17], these linear adapters outperform traditional neural re-rankers in terms of latency. Analyzing this robustness-efficiency trade-off is central to this study.

III. METHODOLOGY

In this section, we outline the methodology used to evaluate the adversarial robustness of DHR models in a telecom-domain benchmark setting. By focusing specifically on DHR architectures, our evaluation provides a targeted analysis rather than a generalized benchmark for all dense retrieval frameworks. Our objective is to systematically measure how retrieval performance degrades under different adversarial perturbation strategies and to compare the effectiveness of various defense mechanisms under controlled computational constraints.

A. Retrieval Architectures

To construct the evaluation setting, we adopt pre-trained retrieval models from the Telco-DPR dataset [3], a telecom-domain corpus derived from 3GPP technical specifications. We consider two retrieval architectures: the passage-level model (DHR-P), which relies on local semantic context, and the hybrid model (DHR-H), which combines passage-level relevance with document-level information through the integration of DHR-P and DHR-D scores. Both models have previously demonstrated strong performance on telecom-domain retrieval tasks and are used here as fixed evaluation targets to ensure a consistent and reproducible assessment of adversarial robustness.

B. Key Performance Indicators (KPIs)

We evaluate retrieval quality using standard Information Retrieval (IR) metrics: Accuracy at k ($\text{Acc}@k$, for $k \in \{1, 3, 5, 10\}$) measures the percentage of queries where the relevant document successfully appears within the top- k retrieved results; and Mean Reciprocal Rank at 10 ($\text{MRR}@10$) evaluates the overall ranking quality by calculating the average of the reciprocal ranks of the first relevant document. The clean baseline performance for these targeted models is summarized in Table I.

TABLE I
BASELINE RETRIEVAL PERFORMANCE ON THE COMPLETE TELCO-DPR DATASET [3].

Model	Acc@1	Acc@3	Acc@5	Acc@10	MRR@10
DHR-P	57.9	71.8	77.0	82.0	0.66
DHR-H	59.2	75.1	81.8	86.2	0.68

C. Perturbation Methods and Attacks

To evaluate robustness, we simulate adversarial perturbations using three widely studied categories of black-box attack methods: gradient-based (PGD, FGSM), score-based (ZOO), and decision-based (HSJ, Boundary). This categorization follows standard adversarial machine learning benchmarks and enables comparison across different levels of attacker knowledge and access. To execute these attacks realistically, we assume a strict black-box threat model, reflecting the operational constraints of real-world networks where adversaries lack access to the target model’s internal architecture, parameter weights, or training data. The attacker’s capabilities are strictly limited to querying the system and observing the final retrieved outputs.

Since the target retrieval models do not expose gradients or confidence scores, adversarial examples are generated using surrogate models that approximate their behavior. To support decision-based attacks such as HSJ and Boundary, we employ a K-Nearest Neighbors (KNN) surrogate, which approximates retrieval decisions using only query, output interactions and enables iterative estimation of decision boundaries from hard-label outputs [9]. For gradient-based (FGSM, PGD) and score-based (ZOO) attacks, we use a Multi-Layer Perceptron (MLP) surrogate trained to approximate the embedding space of the target DHR models. This differentiable surrogate provides the continuous representation required to optimize adversarial perturbations [11]. Adversarial queries generated on these surrogate models are subsequently transferred to the target retrieval systems to evaluate robustness under black-box attack conditions.

D. Defensive Countermeasures

Finally, to complete our evaluation pipeline, the last stage of our methodology assesses defensive countermeasures. Rather than retraining the underlying architecture, we evaluate distinct defensive paradigms. These include complex neural strategies,

such as post-retrieval verification (Cross-Encoders) and knowledge distillation (Teacher-Student), alongside lightweight input-level transformations, namely Feature Squeezing, Randomized Smoothing, and geometric projections. We focus on measuring the trade-off between robustness improvement and computational overhead introduced by each defense method, which is an important consideration in efficiency-aware retrieval benchmarking.

IV. ADVERSARIAL ATTACKS EVALUATION

The goal of this section is to present the results of our black-box attacks and evaluate the resilience of the target DHR architectures. To do this, we extracted queries from the Telco-DPR dataset and executed the attacks directly against our proxy models (KNN and MLP) to calculate the optimal adversarial perturbations. Once the queries were successfully corrupted in this surrogate environment, the noise was transferred and fed into the unseen target retrieval systems. By comparing the system’s performance on clean queries versus these transferred adversarial queries, we quantify the exact degradation in retrieval accuracy.

To ensure our findings are statistically stable and to comprehensively evaluate the true capabilities of all attack vectors, our evaluation is conducted on a robust subset of 300 queries sampled from the Telco-DPR dataset. Given the highly standardized vocabulary of 3GPP technical manuals, semantic variance within this domain is significantly lower than in open-domain literature. Consequently, this subset provides comprehensive coverage of the specialized semantic space, serving as a highly representative proxy for larger-scale operational behavior. This ensures statistically stable results and proves vulnerability without the need to generate thousands of redundant synthetic queries. As defined in our methodology, we target two configurations: the passage-only DHR-P and the hybrid DHR-H model. Table II summarizes the accuracy of the systems before and after the attacks.

TABLE II
IMPACT OF ADVERSARIAL ATTACKS ON RETRIEVAL ACCURACY (300 QUERIES)

Attack Method	Clean (DHR-P)	Attacked (DHR-P)	Clean (DHR-H)	Attacked (DHR-H)
HSJ	56.7%	8.7%	58.3%	44.3%
Boundary	56.7%	6.7%	58.3%	47.3%
PGD	56.7%	0.0%	58.3%	13.7%
FGSM	56.7%	0.0%	58.3%	11.7%
ZOO	56.7%	56.7%	58.3%	58.3%

The results reveal distinct behavioral patterns depending on the attack method and the target architecture. The decision-based attacks (HSJ and Boundary) were highly effective against the single-passage configuration, dropping the DHR-P accuracy from 56.7% to under 9%. However, the hybrid DHR-H model demonstrated significant natural resilience against these same threats, retaining over 44% accuracy. Conversely, the gradient-based attacks (PGD and FGSM) proved to be the most devastating threats. They successfully bypassed the hybrid model’s structural defenses, dropping the DHR-H accuracy to near 12%, while completely collapsing the DHR-P

accuracy to 0%. Finally, the score-based ZOO attack failed to alter the baseline accuracy at all.

Before detailing why these specific methods behave differently, it is crucial to rule out the possibility that the system’s degradation is simply due to natural instability or random data variance. To confirm this, we conducted a baseline stability evaluation by injecting Gaussian noise ($\sigma = 0.02$) into a batch of clean queries, a standard approach to simulate benign transmission errors or environmental variance. To ensure statistical reliability, this noise injection and retrieval process was repeated 50 times.

TABLE III
BASE MODEL STABILITY EVALUATION UNDER GAUSSIAN NOISE (50 INDEPENDENT TRIALS)

Evaluation Metric	Retrieval Accuracy@1
Original Baseline (No Noise)	60.00%
Mean Accuracy under Noise	60.04% ($\pm 1.12\%$)
Mean Performance Deviation	+0.04%

As Table III shows, the system’s performance under random noise remained completely stable (+0.04% deviation). This proves that the retrieval collapse is not a symptom of general fragility, but is specifically triggered by structured adversarial displacement. With this established, it is necessary to analyze the underlying dynamics of how these attacks transfer from the proxy to the real target.

A. Understanding the Attack Dynamics

In a black-box scenario, the success of an attack depends on its transferability, meaning the ability of an adversarial query generated on our proxy model to successfully fool the unseen target DHR models. Our evaluation highlights several key geometric and algorithmic factors that dictate this transferability.

First, the failure of the ZOO attack demonstrates the practical limits of certain methods. ZOO operates by making thousands of tiny modifications to the query and observing how the model’s confidence score changes. Given the large size of the 3GPP embedding vectors, this process requires an impractical number of query attempts to find a successful perturbation. Consequently, within realistic operational limits, the attack fails to converge and leaves the system unaffected.

Second, our results highlight the protective effect of hybrid architectures against decision-based attacks (HSJ and Boundary). These attacks rely on observing changes in the final output label to estimate where the error boundary lies. While they successfully confused the simpler DHR-P model, they struggled against the fully integrated DHR-H. By fusing passage-level scores with document-level context, the hybrid model creates a more complex decision space. Forcing the attacker to simultaneously trick both the passage and the document context requires significantly more query attempts than the algorithms could efficiently manage, turning structural depth into a natural first line of defense.

Lastly, the evaluation reveals a well-documented phenomenon known as the transferability paradox when comparing the two gradient-based methods (PGD and FGSM) [9]. Theoretically, PGD is a more advanced, multi-step attack designed to find the perfect perturbation. However, in our black-box setup, the simpler, single-step FGSM proved slightly more effective against the hybrid DHR-H (11.7% vs 13.7% accuracy). This occurs because PGD tends to overfit to the proxy model. It memorizes the specific parameters of the MLP surrogate so heavily that the resulting noise becomes too specific and fails to align perfectly when transferred to the actual DHR system. In contrast, FGSM applies a broad, single mathematical push. By not over-optimizing to the proxy, this generalized semantic shift translates more reliably across different architectures. This confirms that while hybrid retrieval structures can resist certain algorithms, they remain highly vulnerable to broad geometric displacements.

V. DEFENSE EVALUATION

The significant degradation of both retrieval architectures under adversarial perturbations highlights the need for robust defensive mechanisms. However, securing real-time telecommunication controllers introduces a strict operational constraint: any defense must neutralize adversarial noise without exceeding sub-second control-loop latencies. This section evaluates defensive approaches to identify a viable solution that balances computational efficiency with retrieval accuracy recovery, defined as the ability to successfully restore the correct document to the top-1 position.

A. Standard Input Filters

Guided by the defensive paradigms detailed in the State of the Art, our evaluation began with standard, low-complexity input transformations. Because real-time operations prioritize low latency, our initial goal was to test if these fast, non-parametric filters could efficiently neutralize adversarial queries before inference. We implemented three standard techniques: Feature Squeezing (reducing numerical precision), Randomized Smoothing (injecting controlled Gaussian noise), and PCA Denoising (projecting vectors into a lower-dimensional space).

We evaluated these filters using the 300 queries previously corrupted by the HSJ attack. We selected HSJ because decision-based attacks provide a standard benchmark for evaluating defense effectiveness without completely displacing the initial retrieval pool. Our evaluation showed that standard filters failed to provide meaningful recovery. The DHR-P accuracy remained consistently low (between 11.0% and 13.0%, slightly up from the 8.7% attacked state), and the hybrid DHR-H showed no significant improvement, reaching only 40.0% with Feature Squeezing, 43.0% with Randomized Smoothing, and 39.0% with PCA Denoising. These results represent negligible gains, or even slight degradation, compared to the 44.3% attacked baseline. Because dense retrieval requires highly precise vector representations, simplistic noise reduction either fails to remove the complex mathematical

perturbation, or it alters the embedding so severely that it destroys the query’s original technical meaning.

B. Advanced Defense Mechanisms and Generalization

The failure of standard filters required a shift toward the more complex defense mechanisms previously detailed in the State of the Art. To evaluate their practical viability under strict real-time constraints, we implemented three distinct solutions: Semantic Re-ranking (via a Cross-Encoder) acting as a post-retrieval defense, alongside two pre-retrieval filters: Neural Denoising (via a Teacher-Student model) and Geometric Alignment (via Ridge Adapter). Rather than re-examining their underlying theoretical principles, our focus in this section is to test their mitigation efficacy, generalization capabilities, and computational overhead. To ensure a rigorous and transparent evaluation, we structured our analysis around two distinct testing regimes. The practical viability of any defense must be measured in an Out-of-Sample (OOS) regime. By using a 300-query subset, the model is trained on a set of 200 queries and evaluated on 100 previously unseen ones. This ratio ensures sufficient data to learn the mitigation patterns without relying on synthetic data generation and measuring true generalization capability. Conversely, an In-Sample regime evaluates the defense on the exact same data used to train it. While unsuitable for real-world metrics, it serves as a necessary mathematical proof-of-concept to establish the absolute theoretical upper bound of a defense’s recovery potential.

Following this roadmap, we began our comparative Out-of-Sample evaluation against the HSJ attack. Having demonstrated the failure of standard filters across both models, we now exclusively target the hybrid DHR-H configuration for these advanced defenses. As established previously, its fused architecture provides superior baseline accuracy and natural structural robustness, making it the most viable candidate for this phase of testing. Table IV details this initial comparison.

TABLE IV
REALISTIC ACCURACY RECOVERY AGAINST HSJ (OUT-OF-SAMPLE, TARGET: DHR-H)

System State / Defense Mechanism	Accuracy@1 (Realistic OOS)
Clean Baseline	58.3%
Under Attack (HSJ)	40.0%
Semantic Re-ranking (Cross-Encoder)	51.0%
Geometric Alignment (Ridge Adapter)	49.0%
Neural Denoising (Teacher-Student)	45.0%

While the Cross-Encoder achieved the highest recovery (51.0%), it possesses a fundamental structural vulnerability. Semantic re-ranking performs well against decision-based attacks but is practically ineffective against severe perturbations like PGD and FGSM, which push the target document entirely out of the initial top-K retrieval pool, leaving the second-stage mechanism with no relevant candidates to evaluate. Consequently, first-stage pre-retrieval filters are strictly necessary. Guided by this limitation, we discarded the post-retrieval approach and expanded our OOS evaluation across all threat

vectors to directly compare the two pre-retrieval models (Table V).

TABLE V
OUT-OF-SAMPLE ACCURACY RECOVERY ACROSS ALL ATTACKS (TARGET: DHR-H)

Attack Mechanism	Teacher-Student	Ridge Adapter
HSJ	45.0%	49.0%
Boundary	24.0%	34.0%
PGD	23.7%	18.0%
FGSM	24.0%	15.0%

As detailed in Table V, the Ridge Adapter maintained superior recovery against decision-based attacks, demonstrating better generalization on small training batches. However, both architectures showed limited efficacy against continuous gradient perturbations (PGD and FGSM). Because gradient methods compute a mathematically optimized displacement for each query, they create a highly variable adversarial space that a training batch of 200 examples cannot fully capture. While increasing training data could improve results, the computational cost of generating gradient samples limits practical batch sizes, causing these pre-retrieval filters to achieve only partial OOS recovery. Despite this, we selected the Ridge Adapter as the primary candidate for further analysis due to its superior performance under low-data regimes and significantly lower computational footprint. In critical network environments, maintaining a large neural footprint is problematic, controllers require near-instantaneous execution, making lightweight linear projections far more sustainable. To determine if the observed degradation is a fundamental mathematical inability or strictly a training volume limitation, we evaluated the Ridge Adapter in an In-Sample regime (Table VI), establishing its theoretical upper bound.

TABLE VI
THEORETICAL UPPER BOUND RECOVERY VIA RIDGE ADAPTER (300 QUERIES, IN-SAMPLE)

Attack	DHR-P			DHR-H		
	Clean	Attacked	Defended	Clean	Attacked	Defended
HSJ	56.7%	7.0%	52.3%	58.3%	40.0%	53.0%
Boundary	56.7%	5.3%	54.0%	58.3%	45.3%	55.7%
PGD	56.7%	0.3%	56.7%	58.3%	12.7%	58.7%
FGSM	56.7%	0.0%	57.0%	58.3%	11.7%	58.3%

The In-Sample results unequivocally confirm that adversarial semantic displacement follows a consistent geometric pattern that is entirely reversible through linear algebra when sufficient data is available.

Finally, to determine its cross-threat flexibility, we evaluated a Specialist adapter (trained strictly on Boundary Attack noise) and a Generalist version (trained on HSJ noise) against a batch of 100 queries corrupted by the unseen Boundary Attack. The goal was to verify if a linear adapter can generalize its mitigation capabilities across the same threat family without specific retraining. While the Boundary attack heavily degraded the

clean baselines, dropping DHR-P accuracy from 60.0% to 8.0% and DHR-H from 59.0% to 49.0%, both defensive setups achieved exceptional recovery. The Specialist adapter restored accuracy back to 60.0% on DHR-P and 58.0% on DHR-H. Crucially, the Generalist adapter mitigated the unseen threat with near-identical efficacy, achieving 57.0% accuracy on both DHR-P and DHR-H. This minimal performance variance proves that the linear adapter effectively generalizes its geometric alignment capabilities across different variants of the same attack family.

C. Latency Analysis and Operational Viability

To validate these computational arguments, we quantify the inference speed of each defense, determining which solution is truly sustainable under strict real-time latency requirements (Table VII).

TABLE VII
COMPARISON OF INFERENCE LATENCY AND RECOVERY ACCURACY
(TARGET: DHR-H)

Defense Mechanism	OOS Accuracy@1 (HSJ)	Latency (ms)
Cross-Encoder	51.0%	7.424 ms
Teacher-Student	45.0%	0.291 ms
Ridge Adapter	49.0%	0.236 ms

Operating at just 0.236 ms per query, the Ridge Adapter is roughly 31 times faster than the Cross-Encoder and 20% faster than the Teacher-Student model. While the Teacher-Student model is fast, it requires loading neural weights and maintaining a larger memory footprint. In contrast, the Ridge Adapter relies entirely on linear algebra, eliminating the need for complex frameworks or specialized hardware acceleration. By providing a compelling balance between accuracy recovery and near-zero latency, it emerges as a highly viable security candidate for critical retrieval pipelines.

VI. CONCLUSION

DHR models are becoming increasingly important for telecom-domain information retrieval, but our results show that they remain vulnerable to adversarial perturbations. Although the hybrid architecture (DHR-H) exhibits greater robustness than passage-level retrieval alone, both models experience substantial performance degradation under strong black-box attacks. To mitigate these effects, we evaluated a diverse set of defense mechanisms spanning input transformations, neural re-ranking approaches, and lightweight linear adapters.

Our results highlight a clear trade-off between robustness improvement and computational cost. Furthermore, relying on complex deep learning architectures for defense can inadvertently introduce new layers of vulnerability. This is precisely where our proposed Ridge Adapter demonstrates its structural advantage, by utilizing a linear projection rather than a neural network, it breaks this potential vulnerability cycle. It achieves competitive recovery performance with minimal computational cost while avoiding the inherent security flaws

of deep-learning-based defenses. However, out-of-sample experiments reveal a noticeable generalization gap against complex gradient-based perturbations, indicating that robustness remains highly dependent on the diversity of adversarial training examples.

Overall, this work provides a systematic benchmark of adversarial attacks and defenses for telecom-domain dense retrieval systems and highlights the challenges of achieving robust retrieval under computational constraints. Future work will focus on improving defense generalization through larger adversarial datasets and evaluating robustness under adaptive attack scenarios.

ACKNOWLEDGEMENTS

This research was funded by the Celtic-Next Project 6G-SMART (C2023/2-20) and Project 6G-SMART (COMPETE2030-FEDER-01602800) co-funded by Portugal 2030. Furthermore, this work also received funding from the project UID/00760/2025.

REFERENCES

- [1] O-RAN Alliance, "O-ran: Towards an open and smart ran," 2018.
- [2] M. Polese, L. Bonati, S. D'Oro, S. Basagni, and T. Melodia, "Understanding o-ran: Architecture, interfaces, algorithms, security, and research challenges," *IEEE Commun. Surv. Tutor.*, 2023.
- [3] T. Saraiva, M. Sousa, P. Vieira, and A. Rodrigues, "Telco-dpr: A hybrid dataset for evaluating retrieval models of 3gpp technical specifications," in *Proc. IEEE WCNC*, 2025.
- [4] N. N. Sapavath, B. Kim, K. Chowdhury, and V. K. Shah, "Experimental study of adversarial attacks on ml-based xapps in o-ran," in *Proc. IEEE FNWF*, 2023.
- [5] B. Kim, Y. E. Sagduyu, K. Davaslioglu, T. Erpek, and S. Ulukus, "Channel-aware adversarial attacks against deep learning-based wireless signal classifiers," *IEEE Trans. Wireless Commun.*, 2021.
- [6] E. Habler, R. Bitton, D. Avraham *et al.*, "Adversarial machine learning threat analysis and remediation in open radio access network (o-ran)," *J. Netw. Comput. Appl.*, vol. 236, p. 104090, 2025.
- [7] V. Karpukhin *et al.*, "Dense passage retrieval for open-domain question answering," in *Proc. EMNLP*, 2020.
- [8] O-RAN ALLIANCE Working Group 11, "O-RAN Security Threat Modeling and Risk Assessment 1.0," O-RAN ALLIANCE, Technical Specification, 2023.
- [9] Y. Dong, F. Liao *et al.*, "Boosting adversarial attacks with momentum," in *Proc. CVPR*, 2018.
- [10] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *Proc. ICLR*, 2015.
- [11] A. Madry *et al.*, "Towards deep learning models resistant to adversarial attacks," in *Proc. ICLR*, 2018.
- [12] P.-Y. Chen *et al.*, "Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models," in *Proc. ACM AISec*, 2017.
- [13] J. Chen, M. I. Jordan, and M. J. Wainwright, "Hopskipjumpattack: A query-efficient decision-based attack," in *Proc. IEEE SP*, 2020.
- [14] C. Guo, M. Rana, M. Cisse, and L. van der Maaten, "Countering adversarial images using input transformations," in *Proc. ICLR*, 2018.
- [15] R. Nogueira and K. Cho, "Passage re-ranking with bert," *arXiv:1901.04085*, 2019.
- [16] A. Chiejina, B. Kim, K. Chowdhury, and V. K. Shah, "System-level analysis of adversarial attacks and defenses on intelligence in o-ran based cellular networks," in *Proc. ACM WiSec*, 2024.