T-VEC: A Telecom-Specific Vectorization Model with Enhanced Semantic Understanding via Deep Triplet Loss Fine-Tuning

Vignesh Ethiraj* **Ashwath David*** **Sidhanth Menon***
Divya Vijay* **Vidhyakshaya Kannan***

{vignesh.e, ashwath.d, sidhanth.m, divya.v, vidhyakshaya.k}@netoai.ai

NetoAI

🤗 Model 🤗 Data

Abstract

The specialized vocabulary and nuanced concepts of the telecommunications industry pose persistent challenges for standard Natural Language Processing (NLP) models. Generic embedding models often struggle to represent telecom-specific semantics, limiting their utility in retrieval and downstream tasks. We present T-VEC (Telecom Vectorization Model), a domain-adapted embedding model fine-tuned from the `gte-Qwen2-1.5B-instruct` backbone using a triplet loss objective. Fine-tuning was performed on T-Embed, a high-quality, large-scale dataset covering diverse telecom concepts, standards, and operational scenarios. Although T-Embed contains some proprietary material and cannot be fully released, we open source 75% of the dataset to support continued research in domain-specific representation learning. On a custom benchmark comprising 1500 query-passage pairs from IETF RFCs and vendor manuals, T-VEC surpasses MPNet, BGE, Jina and E5, demonstrating superior domain grounding and semantic precision in telecom-specific retrieval. Embedding visualizations further showcase tight clustering of telecom-relevant concepts. We release T-VEC and its tokenizer to support semantically faithful NLP applications within the telecom domain.

1 Introduction

Text embeddings—dense vector representations of text—serve as the backbone for many modern NLP applications, including semantic search, dialogue systems, and information retrieval (Reimers and Gurevych, 2019). While general-purpose models such as BERT (Devlin et al., 2019) and GPT-2 (Radford et al., 2019) have shown strong performance on broad benchmarks, their effectiveness often degrades in specialized technical domains characterized by domain-specific jargon, overloaded terminology, and structural ambiguity (Gururangan et al., 2020).

Telecommunications exemplifies such a domain. It features an unusually dense mix of acronyms (e.g., *MME*, *SMF*, *gNB*), technical jargon (*handover*, *QoS parameters*), and ambiguous terms (*cell*, *sector*, *core*), many of which carry very different meanings in general contexts. This linguistic complexity is further amplified by rapidly evolving standards (e.g., *5G*, *LTE*, *NFV*) and layered architectures (e.g., *RAN*, *core*, and *transport networks*).

Despite its real-world importance, telecommunications remains underserved in NLP research. Existing models struggle to accurately interpret telecom language, limiting performance in tasks like fault log analysis, technical document retrieval, customer intent classification, and regulatory compliance. Addressing this domain-language gap is vital for deploying effective AI solutions in operational telecom environments.

To bridge this gap, we introduce **T-VEC (Telecom Vectorization Model)**, a domain-adapted sentence embedding model trained via deep triplet loss fine-tuning. Our core contributions are threefold:

1. **T-Embed.** We construct a high-quality, large-scale telecom dataset, T-Embed, covering diverse telecom concepts, standards, and operational contexts. Although the dataset contains proprietary information and the full dataset cannot be publicly released, we open source 75% of it (MIT license).

2. **Open-Source Domain-Specific Embedding Model.** We release T-VEC, the first open-source embedding model specialized for the telecommunications domain. T-VEC is obtained via full-model fine-tuning of `gte-Qwen2-1.5B-instruct` using a triplet loss objective, with updates across all **338 transformer layers**. This yields domain-

*Equal contribution

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track, pages 2449–2460
November 4-9, 2025

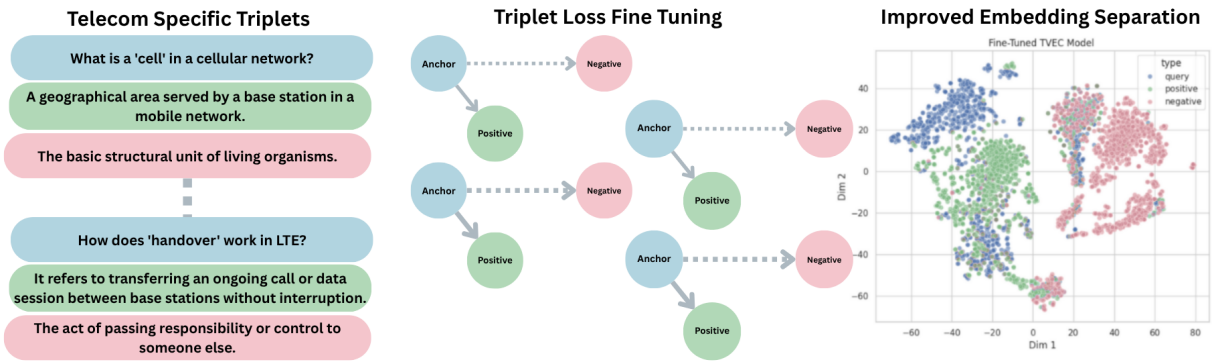


Figure 1: **From noisy telecom jargon to meaningful machine understanding.** T-VEC learns telecom semantics by training on curated triplets: a domain-specific query (anchor), a true paraphrase (positive), and a deceptive distractor (negative). Through triplet loss fine-tuning, the model learns to pull related meanings closer while pushing apart unrelated ones, resulting in clear, telecom-aware clusters in embedding space.

aligned representations for retrieval and semantic matching tasks in telecom. The trained model is publicly available to support reproducibility and real-world applications.

3. **Open-Source Telecom Tokenizer.** We release the first tokenizer tailored to telecom corpora and integrate it into T-VEC. Built by extending the `gte-Qwen2-1.5B-instruct` tokenizer with domain-specific vocabulary, it improves token segmentation and subword coverage for telecom acronyms, protocol names, and configuration terms. Shared tokens retain their original embeddings, while new tokens are randomly initialized and updated during fine-tuning. This approach enables T-VEC to represent telecom jargon more accurately without distrupting the pretrained model.

Our comprehensive evaluations demonstrate that T-VEC achieves state-of-the-art performance on standard benchmarks (leading MTEB average score) while exhibiting superior understanding of telecom semantics compared to its base model and other strong general-purpose models on our domain-specific benchmarks.

2 Related Work

Generating effective text representations is a fundamental challenge in NLP. Sentence-Transformers (Reimers and Gurevych, 2019) popularized the use of siamese network structures with pre-trained models like BERT (Devlin et al., 2019) to create semantically meaningful sentence embeddings. Subsequent research has produced numerous powerful general-purpose embedding models, including MPNet-based models (`all-mpnet-base-v2`) (Song et al., 2020), E5 (`e5-base-v2`) (Wang et al., 2024), BGE (`bge-base-en-v1.5`) (Chen et al., 2024), GTE (`gte-base`, now including Qwen2-based variants like our base model) (Li et al., 2023), Jina Embeddings (Jina AI Team, 2023), and instruction-tuned models like Instructor (Su et al., 2023). These models excel on general language tasks due to training on vast, diverse web corpora.

However, the limitations of general models in specialized domains are well-documented (Gururangan et al., 2020; Tang and Yang, 2025). While domain-specific embeddings have been extensively studied in healthcare (Alsentzer et al., 2019; Lee et al., 2019), finance (Anderson et al., 2024), accelerator physics (Hellert et al., 2024) engineering (Braun et al., 2021), cybersecurity (Roy et al., 2017) and law (Chalkidis et al., 2020), there has been little progress in creating or evaluating telecom-specific text embeddings. Previous work (Roychowdhury et al., 2024) includes a detailed study of domain-adapted sentence embeddings in the telecom sector, emphasizing the challenges and methods for effective document retrieval. Despite this work, public telecom datasets and standardized benchmarks comparable to those in other domains remain scarce. Standard evaluation suites such as MTEB (Muennighoff et al., 2022) do not include telecom standards, network logs, or regulatory filings, and telecom-oriented benchmarks are largely absent. As a result, research and development of domain-adapted embedding models for telecommunications has lagged behind other high-impact verticals, motivating our release of T-VEC and its

supporting telecom-specific evaluation resources.

Fine-tuning using objectives like triplet loss (Schroff et al., 2015) is particularly effective for learning fine-grained semantic similarity relevant to tasks like search and retrieval within a specific domain. While some domain adaptations might only involve fine-tuning the final layers or adding small adapter modules, our work pursues deep fine-tuning, modifying a significant portion of the base model's weights to fundamentally reshape its representational space for the target domain.

3 Methodology

3.1 Base Model

Our base model is `gte-Qwen2-1.5B-instruct` from Alibaba-NLP[1], a 1.5B-parameter transformer producing 1536-dimensional embeddings. It supports sequences up to 32K tokens and belongs to the Qwen2 family (Bai et al., 2023), known for strong multilingual and instruction-following capabilities. We fine-tune this model to specialize its representations for telecom-specific tasks.

3.2 Curating T-Embed: A Telecom Triplet Embedding Dataset

The fine-tuning dataset, T-Embed, was curated by a team of experienced telecommunications professionals. This process was designed to ensure that the resulting dataset would not only be large in scale, but also exhibit the depth, breadth, and nuance required to capture the full complexity of telecom language, operations, and standards.

The final dataset comprises **100,000 triplets**, capturing a wide spectrum of telecom knowledge, including thousands of unique concepts, procedures, and system artifacts. To support open research in domain-adapted representation learning, we publicly release **75,000 triplets** (75% of the dataset).

Further details on T-Embed, including token statistics, topic-wise query distribution, and vocabulary characteristics, are provided in Appendix A.

3.2.1 Topic and Subdomain Coverage

The curation process began with an exhaustive mapping of the telecommunications knowledge landscape. Experts systematically identified and catalogued all major and minor subdomains relevant to the industry, including but not limited to:

- **Wireless Technologies.** 3G, 4G/LTE, 5G NR, Wi-Fi, NB-IoT, and legacy standards.
- **Network Domains.** Radio Access Network (RAN), Core Network (EPC, 5GC), Transport (IP/MPLS, optical), Access (FTTx, DSL), and OSS/BSS.
- **Network Functions.** Coverage of both traditional (e.g., HSS, MME, SGW, PGW) and next-generation (e.g., AMF, SMF, UPF, AUSF, NRF, gNB, eNB) network elements.
- **Operational Procedures.** Fault management, alarm correlation, performance monitoring (KPI/KQI), configuration management, software upgrades, and network slicing.
- **Technical Documentation.** Vendor-specific manuals, 3GPP technical specifications, RFCs, ITU-T recommendations, and regulatory filings.
- **Emerging Topics.** O-RAN, virtualization (NFV, SDN), edge computing, private networks, and AI/ML for telecom.

This comprehensive taxonomy guided the balanced sampling of source materials, ensuring that both foundational and cutting-edge topics were sufficiently represented.

3.2.2 Domain Vocabulary and Semantic Ambiguity

Telecom language is characterized by dense layers of acronyms, abbreviations, and polysemous terms. The curation team placed special emphasis on vocabulary diversity and contextual disambiguation. For each subdomain, domain experts compiled extensive lists of:

- **Acronyms and Abbreviations.** e.g., "MME" (Mobility Management Entity), "SMF" (Session Management Function), "gNB" (next-gen NodeB), "O-RAN" (Open RAN).
- **Jargon and Technical Terms.** e.g., "handover," "RRC state," "bearer," "QoS parameter," "paging," "cell reselection," "sector," "slice."
- **Ambiguous Terms.** Words with multiple meanings in telecom and general English (e.g., "cell," "core," "sector," "handover").

Triplet construction explicitly targeted these terms to ensure the model would learn to resolve ambiguity based on context.

3.2.3 Triplet Generation Methodology

Let $\mathcal{D} = \{(a_i, p_i, n_i)\}_{i=1}^{N}$ denote our curated corpus of triplets, where each triplet is defined as fol-

[1] https://huggingface.co/Alibaba-NLP/gte-Qwen2-1.5B-instruct

lows:

$$(a, p, n) \in \mathcal{A} \times \mathcal{P} \times \mathcal{N},$$

with

$a \in \mathcal{A}$ **(Anchor):** A telecom-specific input (e.g., a query, log entry, or protocol message) sampled from our domain corpus.

$p \in \mathcal{P}$ **(Positive):** A text unit that is *semantically equivalent* or *contextually aligned* with a, drawn from the same technical subdomain.

$n \in \mathcal{N}$ **(Negative):** A text unit that is *lexically or topically plausible* relative to a but *semantically incorrect*, *irrelevant*, or *subtly misleading*.

The triplet generation process was iterative and multi-layered:

1. **Seed Collection.** Anchors were curated using a diverse corpus, including technical manuals, standards, incident tickets, and regulatory documents, as reference.
2. **Positive Selection.** For each anchor, positives were manually paraphrased or retrieved using expert knowledge to ensure semantic closeness, often reflecting real-world telecom paraphrase phenomena (e.g., different vendor terminology for the same concept).
3. **Negative Mining.** Negatives were not chosen at random; instead, "hard negatives" were prioritized. These are texts that are lexically or topically similar to the anchor but diverge in subtle, domain-relevant ways (e.g., confusing "handover" with "cell reselection," or "core" with "RAN").
4. **Quality Assurance.** Each triplet underwent review by at least two domain experts. Disagreements were resolved through discussion or further research. Ambiguous or low-quality triplets were iteratively refined or discarded.

3.3 Fine-Tuning with Triplet Loss

We fine-tune the `gte-Qwen2-1.5B-instruct` model using the triplet loss objective (Schroff et al., 2015) to encourage semantically meaningful embeddings. Given a triplet $(a, p, n) \in \mathcal{A} \times \mathcal{P} \times \mathcal{N}$, the model minimizes the following loss:

$$L(a, p, n) = \max\big(0,\ d(E(a), E(p)) - d(E(a), E(n)) + \alpha\big) \quad (1)$$

where $E(\cdot)$ denotes the embedding function, and $d(\cdot, \cdot)$ is the cosine distance, defined as

$$d(x, y) = 1 - \cos(\theta_{x,y}),$$

with $\cos(\theta_{x,y})$ being the cosine similarity between x and y, and α is a margin hyperparameter. The objective ensures that the distance between anchor and negative exceeds that of the anchor and positive by at least α. We construct triplets with query-like anchors (e.g., user questions or descriptions) to bias the model toward high retrieval performance on telecom-specific tasks.

3.3.1 Deep Model Architecture Modifications

A key differentiator of our approach lies in the depth of fine-tuning. Rather than constraining updates to lightweight adapters or a small subset of final layers, we perform end-to-end fine-tuning across **338 layers** of the `gte-Qwen2-1.5B-instruct` architecture. This enables substantial adaptation of internal representations, effectively reconfiguring a large fraction of the model's parameters to align with domain-specific semantics.

3.3.2 Magnitude and Distribution of Weight Adaptation

To quantify the depth of fine-tuning, we compute for each updated parameter tensor W the L2 norm

$$\Delta(W) = \|W_{\text{fine}} - W_{\text{base}}\|_2.$$

Across all modified tensors, the mean L2 change is

$$\overline{\Delta} = \frac{1}{M} \sum_{m=1}^{M} \Delta(W_m) = 0.7735,$$

indicating substantial redistribution of model capacity toward telecom-specific features.

Figure 5a visualizes the top-20 tensors with the largest $\Delta(W)$. These tensors, spanning MLP gate, up- and down-projection weights across layers 0–8, exhibit a broadly distributed adaptation pattern. The pervasiveness of these changes confirms that T-VEC's domain specialization arises from deep, architecture-wide weight modifications rather than superficial surface tuning.

4 Evaluation

We evaluate T-VEC's domain specialization primarily within the telecommunications domain. Our evaluation framework comprises three components: (1) a held-out test set of telecom triplets, (2) a domain-specific retrieval benchmark, and (3) embedding space analysis via similarity distributions and t-SNE projections. For completeness, we report T-VEC's performance on standard embedding

benchmarks (e.g., STS, classification, MTEB tasks) in Appendix D.

4.1 Telecom Triplet Evaluation

To assess the model's ability to capture fine-grained semantic distinctions, we construct a held-out set of telecom triplets (a, p, n) with no overlap with the training distribution. Each triplet consists of an anchor a, a semantically related positive p, and a plausible but incorrect negative n. The model is evaluated based on its ability to satisfy the triplet constraint:

$$d(E(a), E(p)) < d(E(a), E(n)) \quad (2)$$

where $E(\cdot)$ is the embedding function and $d(\cdot, \cdot)$ denotes cosine distance. Triplet accuracy is defined as the proportion of test triplets for which the constraint holds.

Table 1: **Telecom-Specific Triplet Evaluation.** Accuracy on a held-out test set of telecom triplets measuring semantic discrimination. Each model is evaluated on its ability to embed the anchor closer to the positive than the negative in cosine space.

Model	Triplet Accuracy
T-VEC	**0.9380**
GTE-Qwen2-1.5B-instruct	0.0135
all-mpnet-base-v2	0.0685
bge-base-en-v1.5	0.0414
e5-base-v2	0.0168
jina-embeddings-v2-base-en	0.0290
instructor-xl	0.0321
gte-base	0.0169
multilingual-e5-base	0.0120
all-MiniLM-L6-v2	0.0637

T-VEC achieves a triplet accuracy of 0.9380, substantially outperforming both its base model and leading general-purpose embedding models. This indicates a robust ability to disambiguate nuanced telecom semantics.

4.2 Telecom Retrieval Evaluation

To assess T-VEC's effectiveness in domain-specific retrieval, we constructed a custom benchmark consisting of 1500 query-passage pairs derived from telecommunications documentation. The benchmark corpus comprises IETF RFCs[2] and vendor technical manuals[3]. These documents were chosen for their authoritative status and technical specificity, making them ideal for evaluating retrieval systems that require precise semantic understanding in specialized domains.

RFCs (Requests for Comments) are public-domain specifications that define key protocols, architectures, and operational guidelines for the Internet and telecom infrastructure. They offer a rich source of structured, formal, and jargon-heavy content, which presents a meaningful challenge for semantic retrieval models. Similarly, vendor manuals often describe configuration parameters, troubleshooting workflows, and protocol extensions.

Preprocessing. We removed artifacts (e.g., ASCII drawings, null characters), deduplicated near-identical passages, and filtered out non-informative boilerplate content to ensure semantic quality and relevance. This benchmark provides a realistic and challenging testbed for evaluating retrieval in specialized technical domains, where understanding precise semantics and domain terminology is essential. Each document was segmented into semantically coherent chunks using structural markers such as section headers and paragraph boundaries. These chunks served as the unit of retrieval. To simulate realistic information needs, we used a large language model (LLM) to generate one query per chunk. Each chunk was passed as input to the LLM, which returned a corresponding query that is topically and semantically aligned with the content of the chunk. The exact prompting strategy is detailed in Appendix A. Each query is paired with its originating (ground-truth) passage, along with several hard negatives sampled from the same corpus to encourage fine-grained semantic discrimination. We evaluated a range of publicly available embedding models on this benchmark using cosine similarity-based retrieval. As shown in Table 2, T-VEC outperformed all other models across all metrics, including CosineSim@1, Recall@5, and top-1 match rate.

4.3 Embedding Space Analysis

To analyze the semantic geometry of T-VEC's learned embedding space, we visualize cosine similarity distributions and t-SNE plots (Figure 2).Positives are tightly grouped near high cosine similarity values, while negatives remain well separated. This

[2] https://www.rfc-editor.org/

[3] Scraped from publicly available documentation hosted on official vendor sites such as Cisco, Juniper, and Huawei.

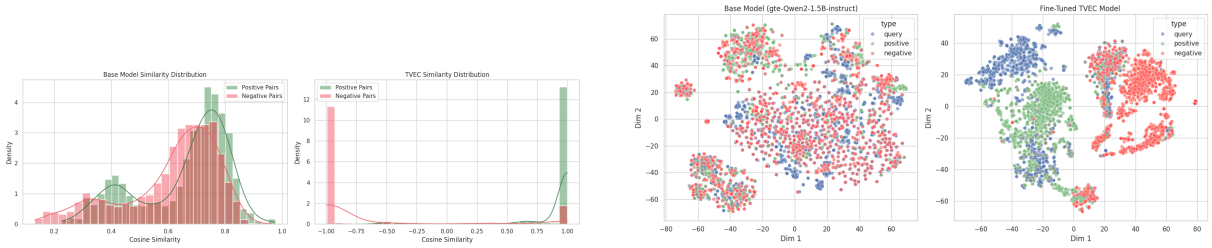


Figure 2: **Embedding space analysis.** Left: Cosine similarity distributions for positive (green) and negative (red) telecom pairs. T-VEC (right) demonstrates clearer separation than the base model. Right: t-SNE visualization of embeddings. T-VEC embeddings form tighter clusters with improved separation between anchor, positive, and negative samples.

Table 2: Comparison of cosine similarity-based evaluation metrics across embedding models.

Metric	T-VEC	Qwen2	MPNet	BGE	E5	Jina	Instr.	GTE	mE5	MiniLM
CosineSim@1	**0.78**	0.72	0.70	0.69	0.69	0.67	0.71	0.68	0.69	0.65
Avg_CosineSim@5	**0.74**	0.70	0.67	0.66	0.67	0.64	0.68	0.63	0.65	0.61
Top1_CosineMatch	**0.80**	0.75	0.72	0.71	0.72	0.69	0.74	0.70	0.72	0.67
Recall@5_cosine	**0.83**	0.76	0.74	0.73	0.74	0.70	0.76	0.71	0.72	0.68

confirms that the model has effectively internalized domain-specific semantics.

5 Real-World Deployment

Our domain-specific embedding model has been integrated into a production-grade platform that supports chat-based interaction with a growing corpus of internal and external documents. It enables users to retrieve precise, context-aware answers from a collection of organizational knowledge bases, forum discussions, and technical documentation.

The chatbot uses our custom-trained embedding model to improve retrieval performance for specialized terminology and nuanced queries that general-purpose models often struggle with. The model powers dense retrieval over a hybrid index (dense + sparse), ensuring high recall and semantic fidelity. It has been optimized for performance in noisy, real-world environments with domain-specific jargon, abbreviations, and informal user queries.

The chatbot has indexed over 10,000 documents across various formats (e.g., PDFs, Markdown), and supports multi-document reasoning via chunked embedding aggregation. The chatbot interface has handled over 50,000 queries in pilot deployments, with human evaluation suggesting significant improvement in answer relevance over baseline models such as `text-embedding-ada-002`.

User insights are being fed back into a continuous retraining loop, allowing the embedding model and retrieval logic to co-evolve with real user interactions.

Deployment challenges included latency optimization, query disambiguation, and integrating user feedback into the model improvement cycle. We addressed these via efficient vector search infrastructure (FAISS with GPU support), prompt-tuning pipelines, and lightweight feedback interfaces embedded into the chat UI.

Overall, this deployment demonstrates the viability and impact of domain-adapted embeddings in augmenting enterprise productivity.

6 Conclusion

We introduced T-VEC, a 1.5B parameter telecom-specific text embedding model derived from `gte-Qwen2-1.5B-instruct`. Through extensive and deep fine-tuning on a large, manually curated telecom dataset using triplet loss, T-VEC achieves state-of-the-art performance on telecom-specific semantic understanding tasks, significantly outperforming general models and its own base model.

We release T-VEC and its telecom-specific tokenizer under the MIT license to support transparency, reproducibility, and broader adoption across telecom and NLP communities.

Future work includes expanding the T-VEC fine-tuning dataset with even more diverse telecom data, exploring architectural enhancements specifically for telecom NLP tasks, and deploying T-VEC in

real-world telecom applications to further validate and refine its capabilities.

Limitations

While T-VEC demonstrates strong performance on telecom-specific tasks, it exhibits notable limitations in terms of generalization to broader natural language understanding. As shown in Table 6, T-VEC underperforms significantly on the `sentence-transformers/all-nli` triplet benchmark, achieving an average triplet score of only 0.6150—far below general-purpose sentence embedding models like `all-mpnet-base-v2` or `bge-base-en-v1.5`. This degradation reflects a well-documented trade-off in domain-adaptive fine-tuning (Gururangan et al., 2020; Howard and Ruder, 2018; Lee et al., 2020): models optimized for in-domain semantic distinctions often lose representational flexibility when applied to out-of-distribution (OOD) contexts.

This over-specialization is likely driven by T-VEC's intensive fine-tuning on telecom triplets, which may have narrowed its semantic space to focus exclusively on patterns, terminology, and structure relevant to telecommunications. While this narrowing enables precise disambiguation and ranking within the target domain, it reduces the model's ability to encode more abstract, general-purpose semantic relationships that are crucial in tasks like natural language inference, paraphrase detection, and commonsense reasoning.

Additionally, our fine-tuning did not employ strategies such as multi-domain pretraining, continual learning, or regularization techniques (e.g., feature drift control or contrastive mixing) to explicitly preserve generalization. Exploring such methods is a promising direction for future work, especially for applications that demand both high in-domain precision and robust zero-shot generalization.

While T-VEC demonstrates strong retrieval gains, retrieval evaluation was conducted on a domain-specific benchmark of limited size (1500 query passage pairs). The test queries were LLM-generated from telecom texts; broader evaluation on human-authored queries and larger retrieval pools remains future work. Quantitative analysis in real-world deployment scenarios, at scale, or on human-authored queries has not been performed. Broader evaluation is a key avenue for future work.

Finally, there are practical integration constraints. T-VEC's embedding dimensionality must match the downstream system (e.g., 768 dimensions for many Transformer-based chatbots). Mismatched dimensions require projection layers or model retraining, introducing latency and deployment complexity. Another promising direction is to explore smaller, more efficient models for deployment, which could reduce memory footprint and inference time while retaining strong in-domain performance. Implementing parameter-efficient approaches such as Low-Rank Adaptation (LoRA) could enable more scalable experimentation while reducing overfitting risks. Exploring such methods is a promising direction for future work, especially for applications that demand both high in-domain precision and robust zero-shot generalization.

References

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

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.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, and 1 others. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Daniel Braun, Oleksandra Klymenko, Tim Schopf, Yusuf Kaan Akan, and Florian Matthes. 2021. The language of engineering: Training a domain-specific word embedding model for engineering. In *MSIE*.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14, Vancouver, Canada. Association for Computational Linguistics.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. LEGAL-BERT: The muppets straight out of law school. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2898–2904, Online. Association for Computational Linguistics.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *arXiv preprint arXiv:2402.03216*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

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, Online. Association for Computational Linguistics.

Thorsten Hellert, João Montenegro, and Andrea Pollastro. 2024. Physbert: A text embedding model for physics scientific literature. *APL Machine Learning*, 2(4):046105.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339, Melbourne, Australia. Association for Computational Linguistics.

Jina AI Team. 2023. Jina Embeddings 2: World's First Open Source 8K Text Embedding Models That Rival OpenAI. `https://jina.ai/news/jina-embeddings-2-worlds-first-open-source-8k-text-embedding-model/`. Jina AI Blog, August 15, 2023.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*.

Zhen Li and 1 others. 2023. Towards general text embeddings with mixture of experts. *Preprint*, arXiv:2311.05723. `https://arxiv.org/abs/2311.05723`.

Niklas Muennighoff and 1 others. 2022. Mteb: Massive text embedding benchmark. *Preprint*, arXiv:2210.07316. `https://arxiv.org/abs/2210.07316`.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. OpenAI Blog.

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 and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Arpita Roy, Youngja Park, and Shimei Pan. 2017. Learning domain-specific word embeddings from sparse cybersecurity texts. *arXiv preprint arXiv:1709.07470*.

Sujoy Roychowdhury, Sumit Soman, Ranjani Hosakere Gireesha, Vansh Chhabra, Neeraj Gunda, Subhadip Bandyopadhyay, and Sai Krishna Bala. 2024. Investigating distributions of telecom adapted sentence embeddings for document retrieval. *arXiv preprint arXiv:2406.12336*.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

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*, volume 33, pages 16857–16867. Curran Associates, Inc.

Hongjin Su, Weijia Shi, Jungo Kasai, Yizhong Wang, Yushi Hu, Mari Ostendorf, Wen-tau Yih, Noah A. Smith, Luke Zettlemoyer, and Tao Yu. 2023. One embedder, any task: Instruction-finetuned text embeddings. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1102–1121, Toronto, Canada. Association for Computational Linguistics.

Yixuan Tang and Yi Yang. 2025. Finmteb: Finance massive text embedding benchmark. *arXiv preprint arXiv:2502.10990*.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Multilingual e5 text embeddings: A technical report. *arXiv preprint arXiv:2402.05672*.

A Fine-Tuning Dataset Details

We present summary statistics of our fine-tuning dataset in Table 4, reporting token-level metrics across query, positive, and negative components. This includes average, minimum, and maximum token counts, as well as the overall vocabulary size computed across all fields. In Table 3, we show example triplets selected to target telecom-specific

Table 3: Example triplets targeting telecom-specific ambiguity and jargon. Each triplet illustrates disambiguation of acronyms, polysemous terms, or technical vocabulary.

Query	Positive Response	Negative Response
What is the function of the SMF in 5G core networks?	It handles session management and IP address allocation for user equipment.	It measures signal strength and adjusts beam direction in the RAN.
How is handover managed during inter-gNB transitions?	The RRC handles signaling for seamless user mobility between gNBs.	It encodes the user's location in the IP header for routing.
What is a bearer in the context of LTE?	A bearer is a virtual channel with specific QoS parameters assigned to data flows.	A bearer is the physical antenna that transmits radio signals.
How does paging work in idle RRC state?	The network sends paging messages to wake idle UEs when there's incoming data.	Paging refers to dynamic spectrum sharing between different frequency bands.
What role does the core play in 5G slicing?	The core ensures logical isolation and resource allocation across slices.	The core handles beamforming and antenna configuration at the edge.

challenges, including domain-specific acronyms, polysemous terminology (e.g., "core," "cell"), and industry jargon. These triplets are curated to explicitly test the model's ability to resolve ambiguity in context.

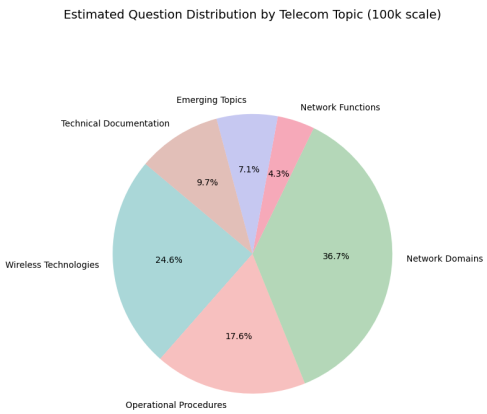


Figure 3: Estimated distribution of telecom-related queries across key annotated topic categories.

Additionally, we visualize dataset structure via three distribution plots: (1) Figure 3 illustrates the estimated question distribution across key telecom topics, (2) Figure 4a displays the token distribution in queries, and (3) Figures 4b and 4c show token count distributions for positive and negative responses respectively. These plots provide insights into both the linguistic complexity and topical balance of the dataset.

Table 4: Token statistics across dataset components.

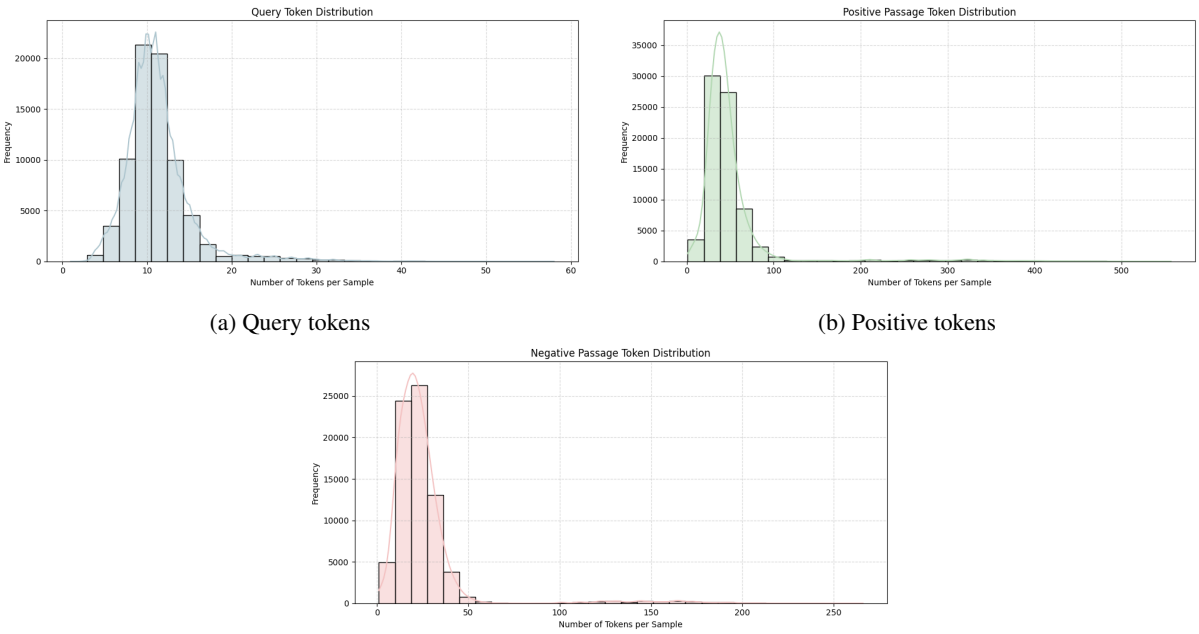
Field	Avg. Tokens	Min Tokens	Max Tokens
Query	11.26	8	58
Positive Response	50.26	14	557
Negative Response	25.00	11	266
Vocabulary Size		50,586	

B Retrieval Dataset Details

We constructed the **T-VEC Retrieval Dataset**, consisting of 1500 query-passage pairs, to evaluate telecom-specific knowledge retrieval capabilities. The dataset is derived from publicly available RFC documents and consists of natural-language queries paired with relevant technical passages. Each query is designed to retrieve a semantically appropriate excerpt, simulating realistic information-seeking scenarios in the telecom domain.

Query lengths average around 100 characters, while passage lengths vary more widely, with some exceeding 5,000 characters due to the inclusion of dense technical detail. Each passage includes provenance metadata such as the originating RFC ID, allowing traceability and potential reuse in downstream document-level tasks.

Each query is a natural-language question designed to retrieve a semantically relevant technical



(a) Query tokens

(b) Positive tokens

(c) Negative tokens

Figure 4: Token count distributions across query, positive, and negative responses in the fine-tuning dataset.

excerpt from an RFC document. Passages range in length from 80 to over 5,000 characters and include provenance metadata such as RFC source ID.

C Query Generation Prompt for Telecom Retrieval

To construct our domain-specific retrieval dataset, we formulated a prompt tailored to elicit high-quality, information-seeking queries grounded in telecommunications literature. Each prompt instance provides the model with a snippet from an RFC or related technical document and instructs it to generate a semantically relevant question that would ideally retrieve the given passage.

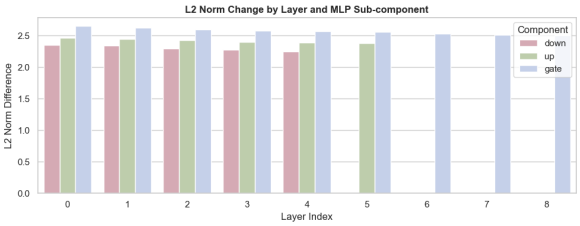
Prompt Template.

> You are a telecommunications expert assisting in the development of a technical search engine. Given a snippet from an RFC or vendor document, generate a question that would retrieve this snippet as a top-ranked result.
>
> **Guidelines:**
>
> - The question must be answerable from the content of the passage, though it need not cover the entire snippet.
> - Prefer domain-relevant, technical terminology, and use paraphrasing where possible instead of copying text verbatim.
> - Avoid overly broad or overly specific questions. Keep the focus on key technical concepts present in the passage.
> - Limit the query to a maximum of 20 words.
>
> **Output Formatting:** Return only the query on a single line with no quotation marks, metadata, or explanation.

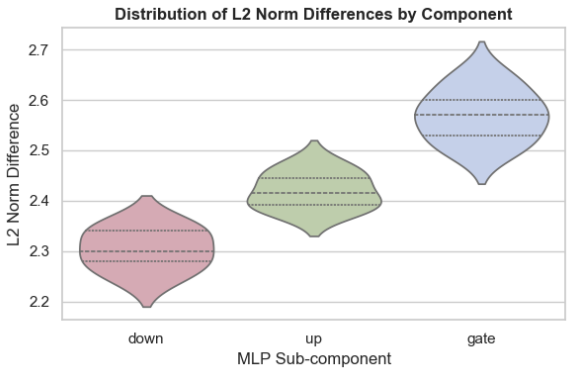
D Evaluation on Standard Benchmarks

Although standard evaluation suites such as MTEB (Muennighoff et al., 2022) do not include telecom-specific corpora, we evaluate T-VEC and several baselines on a range of general-purpose semantic tasks. These include classic Semantic Textual Similarity (STS) datasets (STS12–STS16, STS-Benchmark), Natural Language Inference (NLI) via AllNLI, and additional retrieval-oriented tasks from MTEB such as ArguAna and SciDocsRR.

Semantic Textual Similarity (STS). We report Spearman correlation ($\times 100$) across eight standard STS tasks in Table 7. Spearman's rank correlation coefficient ρ measures the strength of the monotonic relationship between two ranked variables.



(a) **Layer-wise Parameter Adaptation.** We plot the L_2-norm differences $\Delta(W) = \|W_{\text{fine}} - W_{\text{base}}\|_2$ for each transformer layer (0–19), grouped by MLP sub-component (gate, up-proj, down-proj). Gate projections exhibit the largest shifts across layers.



(b) **Distribution of L_2 Norm Changes.** Violin plot visualizing the spread of parameter shifts across all MLP weight tensors. The wide base and sharp tails reflect high variability and the presence of deeply adapted subspaces.

Figure 5: **Visualization of Weight Adaptation.** Left: Per-layer changes highlight systematic adaptation in MLP sub-components. Right: Distributional view emphasizes the extent and variance of fine-tuning across model weights.

As a non-parametric metric, Spearman correlation is robust to non-linear relationships and is widely adopted for evaluating semantic similarity (Cer et al., 2017). This makes it especially suitable for STS tasks where the goal is to capture relative semantic closeness rather than absolute distance. T-VEC consistently matches or outperforms its base model (GTE), achieving the highest average score across all datasets.

Table 5: Estimated average performance of various embedding models on the MTEB benchmark suite. Higher scores indicate better average task performance.

Model Name	Avg. Score
T-VEC	**0.825**
bge-base-en-v1.5	0.815
gte-base	0.805
gte-Qwen2-1.5B-instruct	0.795
instructor-xl	0.785
e5-base-v2	0.780
jina-embeddings-v2-base-en	0.775
all-mpnet-base-v2	0.770
multilingual-e5-base	0.765
all-MiniLM-L6-v2	0.760

Natural Language Inference. To assess generalization beyond the target domain, we evaluated all models on the AllNLI benchmark using the `sentence-transformers/all-nli` dataset. As shown in Table 6, T-VEC underperforms models trained specifically for general-domain NLI, reflecting a typical trade-off: domain specialization improves in-domain performance at the potential cost of generalization (Gururangan et al., 2020; Howard and Ruder, 2018; Lee et al., 2020). Prior work has shown that aggressive fine-tuning on domain-specific corpora can lead to feature overfitting, where models excel in narrow contexts but degrade on out-of-distribution (OOD) or general-domain tasks.

Table 6: Performance on the `sentence-transformers/all-nli` triplet evaluation. Higher scores indicate better general-domain NLI capabilities.

Model Name	Avg. Triplet Score
all-mpnet-base-v2	0.9620
bge-base-en-v1.5	0.9610
jina-embeddings-v2-base-en	0.9590
gte-base	0.9470
all-MiniLM-L6-v2	0.9380
e5-base-v2	0.9230
instructor-xl	0.9220
multilingual-e5-base	0.9210
gte-Qwen2-1.5B-instruct	0.8660
T-VEC	**0.6150**

Overall MTEB Performance. To provide a broader view of overall embedding quality, Table 5 reports the estimated average performance of each model across the MTEB benchmark suite. T-VEC ranks highest on this aggregate metric, suggesting strong generalization despite its domain-specific training. Table 5 reports averages computed over a selected subset of MTEB v1 tasks

Table 7: STS performance comparison of T-VEC against baseline and recent embedding models. Scores are Spearman correlation ($\times 100$).

Task	**T-VEC**	Qwen2	BGE	MPNet	GTE	E5	Instr.	Jina	MiniLM	MultiE5
ArguAna	61.15	62.34	**63.62**	46.52	57.15	51.60	54.88	44.15	50.17	47.83
SciDocsRR	83.97	81.56	87.49	**88.65**	87.08	82.83	79.54	83.11	87.12	80.39
STS12	**80.32**	72.81	78.03	72.63	75.71	73.49	74.08	74.28	72.37	77.93
STS13	**88.22**	84.70	84.18	83.48	85.73	83.00	85.05	84.18	80.60	76.89
STS14	**82.75**	78.80	82.27	78.00	81.51	80.45	80.32	78.81	75.59	77.53
STS15	88.26	87.45	87.96	85.66	**88.81**	88.18	88.36	87.55	85.39	88.37
STS16	84.78	84.94	**85.47**	80.03	83.82	83.66	83.78	85.35	78.99	82.70
STS-B	**88.05**	85.38	86.42	83.42	85.74	85.48	83.05	84.84	82.03	84.20
Average	**82.19**	79.75	81.93	77.30	80.69	78.59	78.63	77.78	76.53	76.98

(e.g., Arguana, SciDocs), rather than the full leaderboard set. The high average is largely driven by strong gains on retrieval-oriented tasks (e.g., SciDocsRR, STS13–15). Small variations also arise from the probabilistic nature of evaluation (e.g., stochasticity in negative mining and training).