

Zen3-Embedding: A Packaging of Qwen3-Embedding for Zen Retrieval

Technical Whitepaper v2025.06

Zach Kelling
Zen LM Research Team
research@zenlm.org
papers.zenlm.org

June 2025

Abstract

Zen3-Embedding is a repackaging of Alibaba’s **Qwen3-Embedding** series, used as the first-stage retrieval encoder in Zen retrieval pipelines. It is *not* a from-scratch model: it is the openly released, Apache-2.0 licensed **Qwen/Qwen3-Embedding** family (0.6B, 4B, and 8B), produced by the Qwen team and built on the Qwen3 dense foundation models [5]. Qwen3-Embedding is a *causal* (decoder-only) LLM adapted for embeddings: an [EOS] token is appended to the input and the final-layer hidden state at that position is taken as the embedding. The models accept task instructions on the query side, support Matryoshka-style user-defined output dimensions (from 32 up to the model’s native dimension), span 100+ languages including code, and handle a 32K context. Native embedding dimensions are 1024 (0.6B), 2560 (4B), and 4096 (8B). On the Qwen team’s evaluation, the flagship Qwen3-Embedding-8B reports a Mean(Task) of **70.58** on the MTEB Multilingual (MMTEB) leaderboard — ranked No. 1 as of 5 June 2025 — with MTEB English v2 Mean(Task) 75.22, C-MTEB Mean(Task) 73.83, and MTEB-Code 80.68 [5]. This document describes how the upstream model is wired into Zen retrieval; all weights, training, and reported numbers are the Qwen team’s and are cited as such. Claims in earlier revisions of a bespoke “7680-dimensional” bidirectional “Zen MoDE” encoder trained for “Decentralized Semantic Optimization,” with BitDelta index compression, were fabricated and have been removed.

Contents

1	Introduction	2
2	Architecture	2
2.1	Causal Decoder Adapted for Embeddings	2
2.2	Embedding Extraction (Last-Token Pooling)	3
2.3	Matryoshka Output Dimensions	3
2.4	Use in the Zen Retrieval Stack	3
2.5	Instruction-Aware Embedding	3
3	Training	3
3.1	Provenance, Not Training	3
4	Evaluation	4
4.1	Flagship (8B) Headline Numbers	4
4.2	Across Model Sizes	4
4.3	A Note on Earlier Revisions	5

5	Related Work	5
5.1	Dense Retrieval	5
5.2	Contrastive Learning for Embeddings	5
5.3	Matryoshka Representations	5
6	Conclusion	5

1 Introduction

Dense vector embeddings are the foundation of modern semantic search, retrieval-augmented generation (RAG), and knowledge-base indexing. The quality of embeddings determines the precision of document retrieval, which directly bounds the quality of RAG-powered applications. Yet most deployed embedding models face a trilemma:

- **Quality vs. size:** High-quality embeddings require large models, but API latency and memory constraints favor small models.
- **Generality vs. specialization:** General-purpose embeddings perform adequately on all tasks but rarely excel on any specific domain.
- **Monolingual vs. multilingual:** Multilingual models traditionally underperform language-specific models on high-resource languages.

Rather than train an embedding model from scratch, Zen3-Embedding adopts the Qwen3-Embedding series [5], which targets these trade-offs directly. Our contribution is integration and packaging, not modeling. The upstream model provides:

1. **A spectrum of sizes** — 0.6B, 4B, and 8B, letting a deployment trade quality against latency and memory.
2. **Instruction-aware embeddings** — The query may be prefixed with a task instruction that steers the representation toward the task, without fine-tuning.
3. **Matryoshka output dimensions** — A single model supports user-defined output dimensions from 32 up to its native dimension, so index storage can be traded against retrieval quality without a separate model.
4. **Multilingual and code coverage** — 100+ languages, including programming languages, with a 32K context.

We do not modify the Qwen3-Embedding weights; Zen3-Embedding is the upstream model wired into the Zen retrieval stack.

2 Architecture

2.1 Causal Decoder Adapted for Embeddings

Qwen3-Embedding is *not* a bidirectional BERT-style encoder and has no mixture-of-experts “Zen MoDE” backbone. Per the Qwen team [5], each model is a fine-tune of the corresponding Qwen3-*Base dense, causal* (decoder-only) LLM (e.g. the 8B from Qwen3-8B-Base, the 0.6B from Qwen3-0.6B-Base). The “7.1B-parameter, 48-layer, 64-expert top-4 bidirectional” specification in earlier revisions did not correspond to any deployed weights and has been removed.

Released sizes and native embedding dimensions [5]:

- Qwen3-Embedding-0.6B: 28 layers, native embedding dimension 1024.
- Qwen3-Embedding-4B: 36 layers, native embedding dimension 2560.
- Qwen3-Embedding-8B: 36 layers, native embedding dimension 4096.
- Context length: 32,768 tokens (all sizes).

2.2 Embedding Extraction (Last-Token Pooling)

The embedding is the final-layer hidden state at an appended `[EOS]` token [5]. Because attention is causal, that last position has attended over the entire input, so its hidden state serves as the sequence representation:

$$\mathbf{e} = \mathbf{h}_{[\text{EOS}]}^{(L)} \in \mathbb{R}^{d_{\text{model}}}, \quad (1)$$

where d_{model} is the model’s native dimension (1024 / 2560 / 4096). There is no `[CLS]` token, no mean+CLS dual pooling, and no learned projection to a 7680-dim space; those constructions in earlier revisions were fabricated.

2.3 Matryoshka Output Dimensions

Qwen3-Embedding supports user-defined output dimensions from 32 up to the native dimension [5, 4], so a shorter prefix of the embedding can be used to shrink index storage at a modest quality cost. The supported range is continuous (32 to d_{model}), not the fixed ladder $\{256, \dots, 7680\}$ asserted in earlier revisions.

2.4 Use in the Zen Retrieval Stack

Zen indexing computes Qwen3-Embedding vectors for documents (no instruction) and builds a standard vector index (e.g. HNSW) over them; at query time the query embedding (optionally with a task instruction) is used for similarity search, and the Qwen3-Reranker stage re-scores the shortlist. The “DSO-native 7680-dim format,” “BitDelta 1.58-bit index compression,” and “peer-to-peer node” descriptions of earlier revisions did not describe this model and have been removed.

2.5 Instruction-Aware Embedding

Task instructions are prepended to queries using a structured template:

Instruction: `<task_description>`

Query: `<query_text>`

The instruction shifts the attention pattern toward task-relevant dimensions of the embedding space. Evaluated instruction types include:

- `Retrieve relevant documents`: general retrieval
- `Find semantically similar sentences`: semantic similarity
- `Classify the following text`: classification
- `Find duplicates`: deduplication

Instructions are applied to queries only (not documents), preserving document index reusability.

3 Training

3.1 Provenance, Not Training

Zen3-Embedding performs no training of its own. The models were trained by the Qwen team; we summarize their published approach at a high level and attribute it accordingly [5].

Per the Qwen team, Qwen3-Embedding starts from the Qwen3-Base dense LLMs and is adapted to embeddings with a contrastive objective (the standard last-token / [EOS] representation with an InfoNCE-style loss over positive (query, document) pairs and negatives), with the Qwen3 generative models used to synthesize a large, multilingual portion of the training pairs [5]. The contrastive loss has the usual form

$$\mathcal{L}_{\text{InfoNCE}} = -\log \frac{\exp(\text{sim}(\mathbf{q}, \mathbf{d}^+)/\tau)}{\exp(\text{sim}(\mathbf{q}, \mathbf{d}^+)/\tau) + \sum_j \exp(\text{sim}(\mathbf{q}, \mathbf{d}_j^-)/\tau)}, \quad (2)$$

with $\text{sim}(\cdot, \cdot)$ cosine similarity. For the authoritative dataset composition, synthetic-data pipeline, and hyperparameters, see the Qwen team’s report [5].

The fabricated training details in earlier revisions — “bidirectional MLM adaptation,” a 2.1B-pair corpus with specific per-source counts, a synthetic generator named “Zen3-Medium,” and a 32,768 GradCache batch — did not describe any real training run and have been removed.

4 Evaluation

All numbers in this section are the Qwen team’s published results for Qwen3-Embedding [5], reproduced here with attribution. We ran no evaluations of our own and make no independent benchmark claims.

4.1 Flagship (8B) Headline Numbers

Table 1: Qwen3-Embedding-8B, as reported by the Qwen team [5]. The MMTEB Mean(Task) of 70.58 ranked No. 1 on the MTEB Multilingual leaderboard as of 5 June 2025.

Benchmark	Score
MTEB Multilingual (MMTEB), Mean(Task)	70.58
MTEB Multilingual (MMTEB), Mean(Type)	61.69
MTEB English v2, Mean(Task)	75.22
MTEB English v2, Mean(Type)	68.70
C-MTEB (Chinese), Mean(Task)	73.83
MTEB-Code (nDCG@10)	80.68

4.2 Across Model Sizes

Table 2: Released sizes and native embedding dimensions [5]. The 8B is the flagship; the Qwen team’s report gives the full MMTEB breakdown for each size, with quality increasing with size.

Model	Layers	Native dim
Qwen3-Embedding-0.6B	28	1024
Qwen3-Embedding-4B	36	2560
Qwen3-Embedding-8B	36	4096

The 8B MMTEB Mean(Task) of 70.58 is given above; per-size MMTEB/MTEB scores for the 0.6B and 4B models are reported in the Qwen team’s paper and we refer the reader there rather than restating numbers we have not independently verified [5].

4.3 A Note on Earlier Revisions

Earlier revisions of this document reported a fabricated “MTEB average 74.3,” a “Zen2-Embedding” comparison column, BEIR/MIRACL tables, and a Matryoshka/BitDelta storage table at 7680 dimensions. None of those described this model or any real measurement; they were fabricated and have been removed. For per-task and per-dataset breakdowns of the genuine numbers, see the Qwen team’s report and the HuggingFace model cards [5].

5 Related Work

5.1 Dense Retrieval

Dense passage retrieval [1] established bi-encoder architectures as practical alternatives to BM25 for open-domain question answering. Qwen3-Embedding sits in the more recent line of repurposing decoder-only LLMs as embedding models via last-token pooling and contrastive fine-tuning [5].

5.2 Contrastive Learning for Embeddings

SimCSE [2] demonstrated that contrastive learning with dropout as data augmentation substantially improves semantic textual similarity. E5 [3] and subsequent work showed that web-scale weakly supervised contrastive training with curated hard negatives matches or exceeds expensive human-annotated pairs. Qwen3-Embedding follows this contrastive-fine-tuning paradigm on top of the Qwen3 base LLMs [5].

5.3 Matryoshka Representations

Matryoshka Representation Learning [4] enables flexible dimension truncation with minimal quality loss; Qwen3-Embedding exposes this as user-selectable output dimensions from 32 up to each model’s native dimension [5].

6 Conclusion

Zen3-Embedding is the first-stage retrieval encoder of Zen retrieval pipelines, implemented as a direct deployment of Alibaba’s Apache-2.0 Qwen3-Embedding models (0.6B/4B/8B) [5]. Qwen3-Embedding is a causal decoder LLM adapted for embeddings via last-token ([EOS]) pooling, with instruction-aware queries, Matryoshka output dimensions, 100+ language coverage, and a 32K context. The flagship 8B reports an MMTEB Mean(Task) of 70.58 (No. 1 as of 5 June 2025) per the Qwen team. We contribute integration into the Zen retrieval stack, not a new model, and we report the Qwen team’s published numbers with attribution rather than benchmarks of our own. The “DSO-native 7680-dimensional Zen MoDE encoder” and “BitDelta index compression” claims of earlier revisions have been removed as fabrications.

Acknowledgments and Attribution

Zen3-Embedding is built entirely on Qwen3-Embedding (Qwen/Qwen3-Embedding, Apache-2.0), developed by the Qwen team at Alibaba; all model weights, training, and reported metrics are theirs. We thank the Qwen team for releasing the models openly and the MTEB/MMTEB maintainers for the evaluation suites.

References

- [1] V. Karpukhin et al., “Dense Passage Retrieval for Open-Domain Question Answering,” *EMNLP*, 2020.
- [2] T. Gao, X. Yao, D. Chen, “SimCSE: Simple Contrastive Learning of Sentence Embeddings,” *EMNLP*, 2021.
- [3] L. Wang et al., “Text Embeddings by Weakly-Supervised Contrastive Pre-Training,” *arXiv:2212.03533*, 2022.
- [4] A. Kusupati et al., “Matryoshka Representation Learning,” *NeurIPS*, 2022.
- [5] Y. Zhang, M. Li, D. Long, X. Zhang, H. Lin, B. Yang, P. Xie, A. Yang, D. Liu, J. Lin, F. Huang, J. Zhou (Qwen Team, Alibaba), “Qwen3 Embedding: Advancing Text Embedding and Reranking Through Foundation Models,” *arXiv:2506.05176*, 2025. Models: `Qwen/Qwen3-Embedding-{0.6B,4B,8B}` (Apache-2.0).