KaLM-Embedding-V2: Superior Training Techniques and Data Inspire A Versatile Embedding Model

Xinping Zhao, Xinshuo Hu, Zifei Shan, Shouzheng Huang, Yao Zhou, Xin Zhang, Zetian Sun, Zhenyu Liu, Dongfang Li, Xinyuan Wei, Youcheng Pan, Yang Xiang, Meishan Zhang, Haofen Wang, Jun Yu, Baotian Hu ✉, Min Zhang

[1]Shenzhen Loop Area Institute (SLAI), [2]Tencent

深圳河套學院 Shenzhen Loop Area Institute

Tencent 腾讯

Superior Training Techniques and High-quality Data

Focal-style Reweighting Mechanism

- It emphasizes difficult samples, where the more difficult the sample, the larger the weight.

$$w_i = (1 - \frac{e^{s(\mathbf{q}_i,\mathbf{p}_i^+)/\tau}}{Z_i})^\gamma, \quad \mathcal{L} = \mathop{\mathbb{E}}_{i\in N}\left[-w_i \log \frac{e^{s(\mathbf{q}_i,\mathbf{p}_i^+)/\tau}}{Z_i}\right]$$

Online Hard Negative Mixing Strategy

- It synthesizes new informative hard negatives via pair-wise or list-wise mixing.

$$\mathbf{h}_i^- = \frac{\tilde{\mathbf{h}}_i^-}{\|\tilde{\mathbf{h}}_i^-\|_2}, \quad \tilde{\mathbf{h}}_i^- = \lambda \mathbf{p}_{i,j}^- + (1-\lambda)\mathbf{p}_{i,k}^-, \quad j \neq k, \quad j,k \in [1,M]$$

$$\mathbf{s}_i^- = \frac{\tilde{\mathbf{s}}_i^-}{\|\tilde{\mathbf{s}}_i^-\|_2}, \quad \tilde{\mathbf{s}}_i^- = \sum_{m=1}^{M} \lambda_m \mathbf{p}_{i,m}^-, \quad \text{s.t.} \sum_{m=1}^{M} \lambda_m = 1,$$

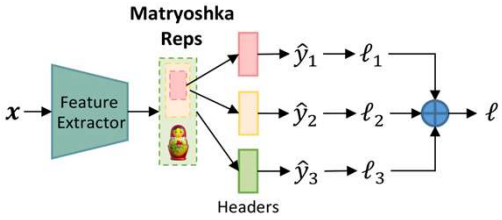
Contrastive Distillation

- It distills fine-grained soft signals from the teacher model that capture nuanced differences.

$$\mathcal{L}_{KL} = D_{KL}(P_t \| P_s) = \sum_i P_t(i) \log \frac{P_t(i)}{P_s(i)}, \quad P_t(i) = \frac{e^{z_{t,i}/\tau}}{\sum_j e^{z_{t,j}/\tau}}, \quad P_s(i) = \frac{e^{z_{s,i}/\tau}}{\sum_j e^{z_{s,j}/\tau}}$$

Matryoshka Representation Learning

- It enables flexible-dimensional embeddings.



Comprehensive Data Curation Recipe

- Around 20 categories of weakly supervised data as well as 100 categories of supervised data
- Hard Negative Mining
- Task-specific instructions
- Persona-based Data Synthesis
- Example-based multi-class labeling

Lychee-KaLM-Embedding-Data

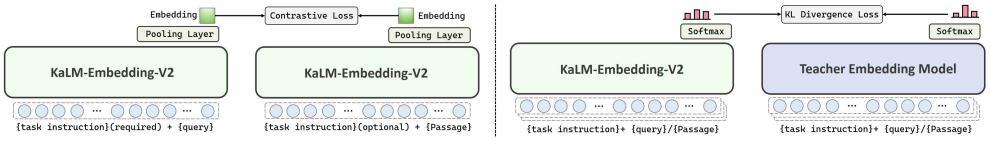
KaLM-Embedding/KaLM-embedding-finetuning-data
Viewer · Updated Nov 27, 2025 · 6.34M · 852 · 25

HIT-TMG/KaLM-embedding-pretrain-data
Viewer · Updated Nov 27, 2025 · 23.7M · 1.82k · 21

Training Framework and Recipe

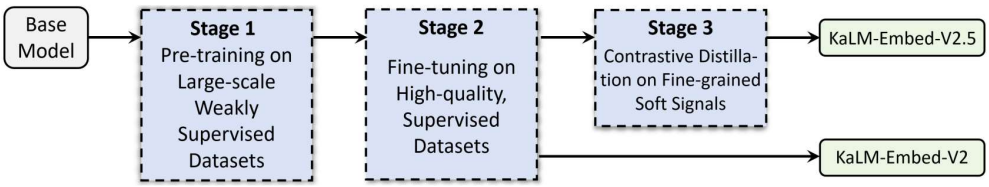
Overall Training Framework

- The left illustrates the workflow of contrastive learning, while the right shows that of contrastive distillation.



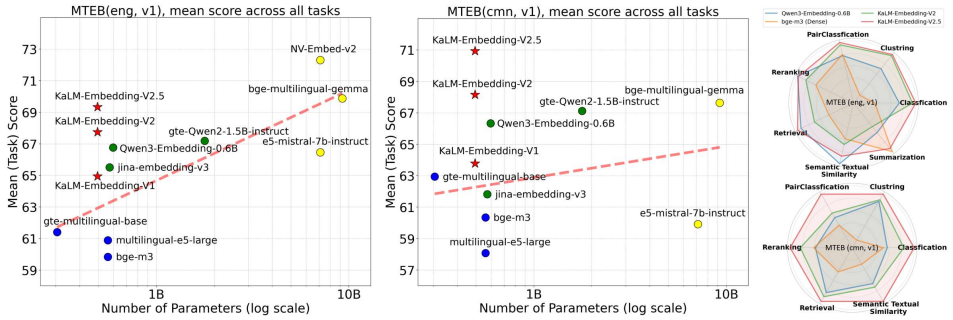
Progressive Multi-stage Training Recipe

- Pre-training on large-scale, weakly supervised datasets over 20 categories to endow the model with strong generalization.
- Fine-tuning on over 100 categories of high-quality supervised datasets to further improve the overall model performance.
- Contrastive Distillation with fine-grained soft knowledge from a stronger teacher model to capture nuanced semantic differences.

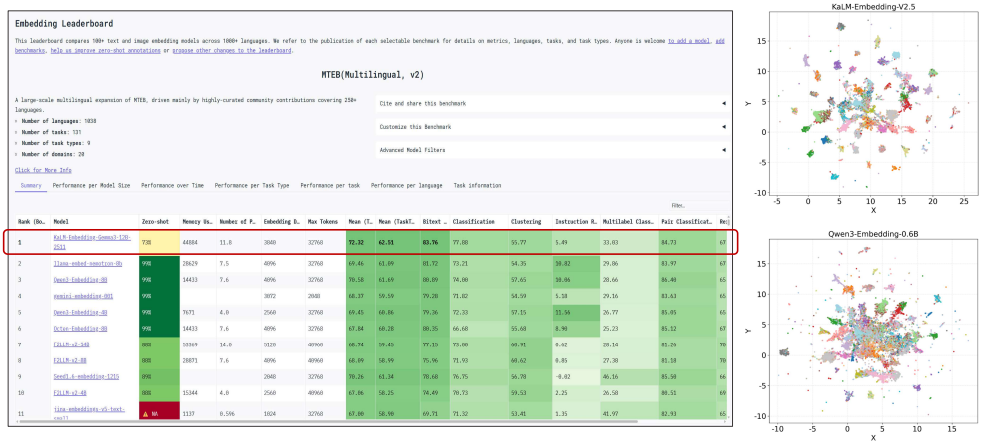


Experimental Results

- Our KaLM-Embedding-V2 series significantly outperforms models of comparable size, rivaling models 3–26x larger.



- KaLM-Embed-Gemma3 further advances SOTA on MMEB. KaLM-Embed-V2.5 shows superior intra-class compactness and inter-class separability.



ICLR 2026

If you are interested in our work, please contact us!

Email: xinpingzhao@slai.edu.cn, baotianhu@slai.edu.cn