ICLR 2025
The Thirteenth International Conference on Learning Representations

☕ ESE: Espresso Sentence Embeddings

Xianming Li[1], Zongxi Li[2] (presenter), Jing Li[1,3*], Haoran Xie[2], Qing Li[1,3]

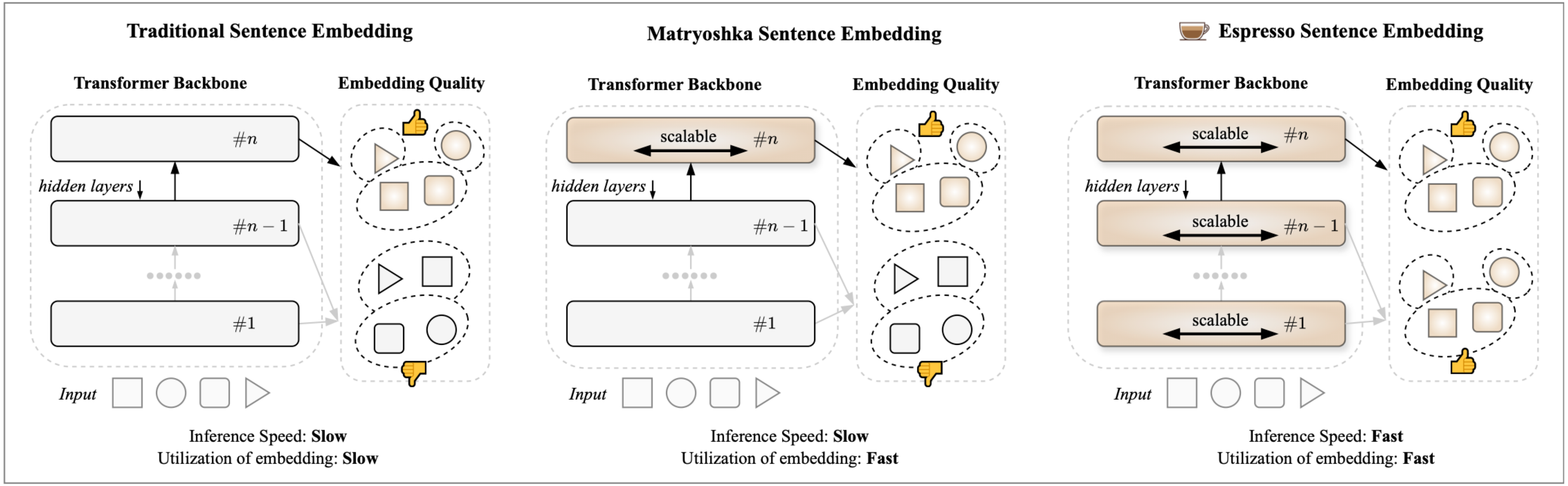
[1] Department of Computing,
[3] Research Centre on Data Science & Artificial Intelligence,
The Hong Kong Polytechnic University, Hong Kong SAR
[2] School of Data Science,
Lingnan University, Hong Kong SAR

xianming.li@connect.polyu.hk,
{zongxili,hrxie}@ln.edu.hk,
{jing-amelia.li,qing-prof.li}@polyu.edu.hk

Introduction

- High-quality sentence embeddings are essential for NLP tasks (STS, RAG, etc.)
- Existing methods lack scalability across **model depth** and **embedding size**
- Compromise inference efficiency

Introduction



- The comparison of traditional (left), MRL (middle) (Kusupati et al., 2022), and the proposed ESE (right) sentence embedding models.
- The gray blocks represent Transformer layers that are fine-tuned in the full model setting, while the coffee-colored blocks indicate layers used in scalable settings.

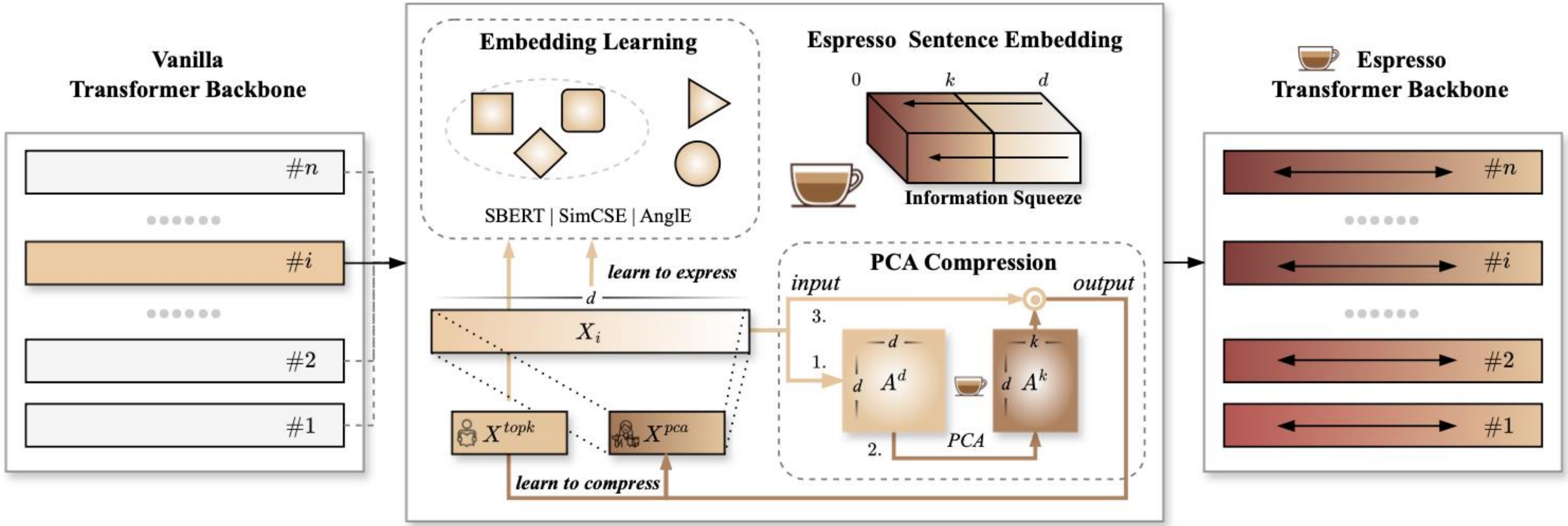
Introduction

- ESE (Espresso Sentence Embeddings) addresses this via:
  - **Learn-to-express**: Allocates crucial features to shallow layers
  - **Learn-to-compress**: Uses PCA for compacting embedding dimensions

- To the best of our knowledge, *we are the first to learn sentence embedding with information compression, presenting scalable embedding inference to both model depths and embedding dimensions*.

ESE Framework

- **Learn-to-express**: Enhances shallow layer representations
- **Learn-to-compress**: Applies PCA for dimension reduction
- Joint training optimizes both processes



(left) Vanilla Transformer backbone, where each layer is not scalable

(center) ESE training with the learn-to-express (to scale model depths) and learn-to-compress (to scale embedding sizes) processes.

(right) Trained Espresso Transformer backbone, where each layer is scalable

Learn to Express

- **Objective**: Improve embedding quality in shallow layers to allow scalable model depths

- We cache each layer's sentence embeddings for $i \in [1, n-1]$ and then jointly train their first-k-dimension sub-embeddings by a weighted loss.

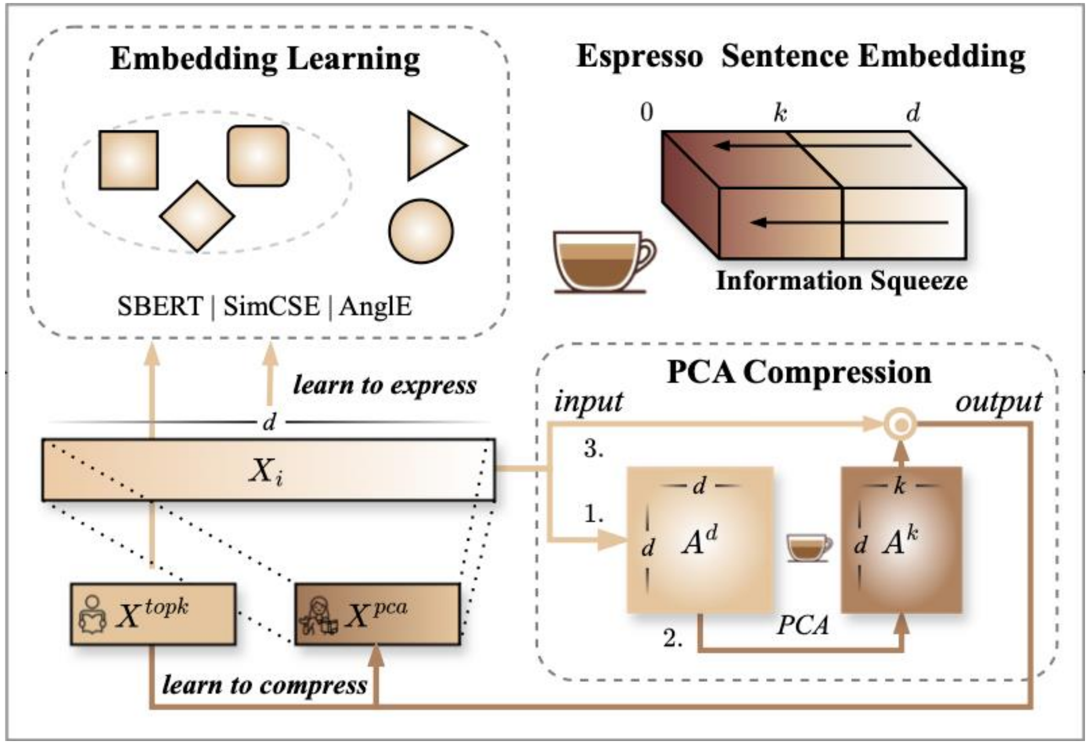
$$\mathcal{L}_{le} = \sum_{i=1}^{n-1} w_i * \mathrm{loss}(\mathbf{X}_i^k, \mathcal{G}) + \mathrm{loss}(\mathbf{X}_n^k, \mathcal{G}).$$

- The loss(·) can be any loss function for sentence embedding learning, e.g., contrastive loss (Gao et al., 2021) or AnglE loss (Li & Li, 2024a). We use the latter one by default.

- Question: we don't know if the first-k-dimension sub-embedding has all the essential information.

Learn to Compress

- **Objective**: Reduce embedding dimensionality while preserving key features
- Uses PCA on embedding dependencies instead of direct compression
- Enables inference with only the first k-dimensions



STS Experiments

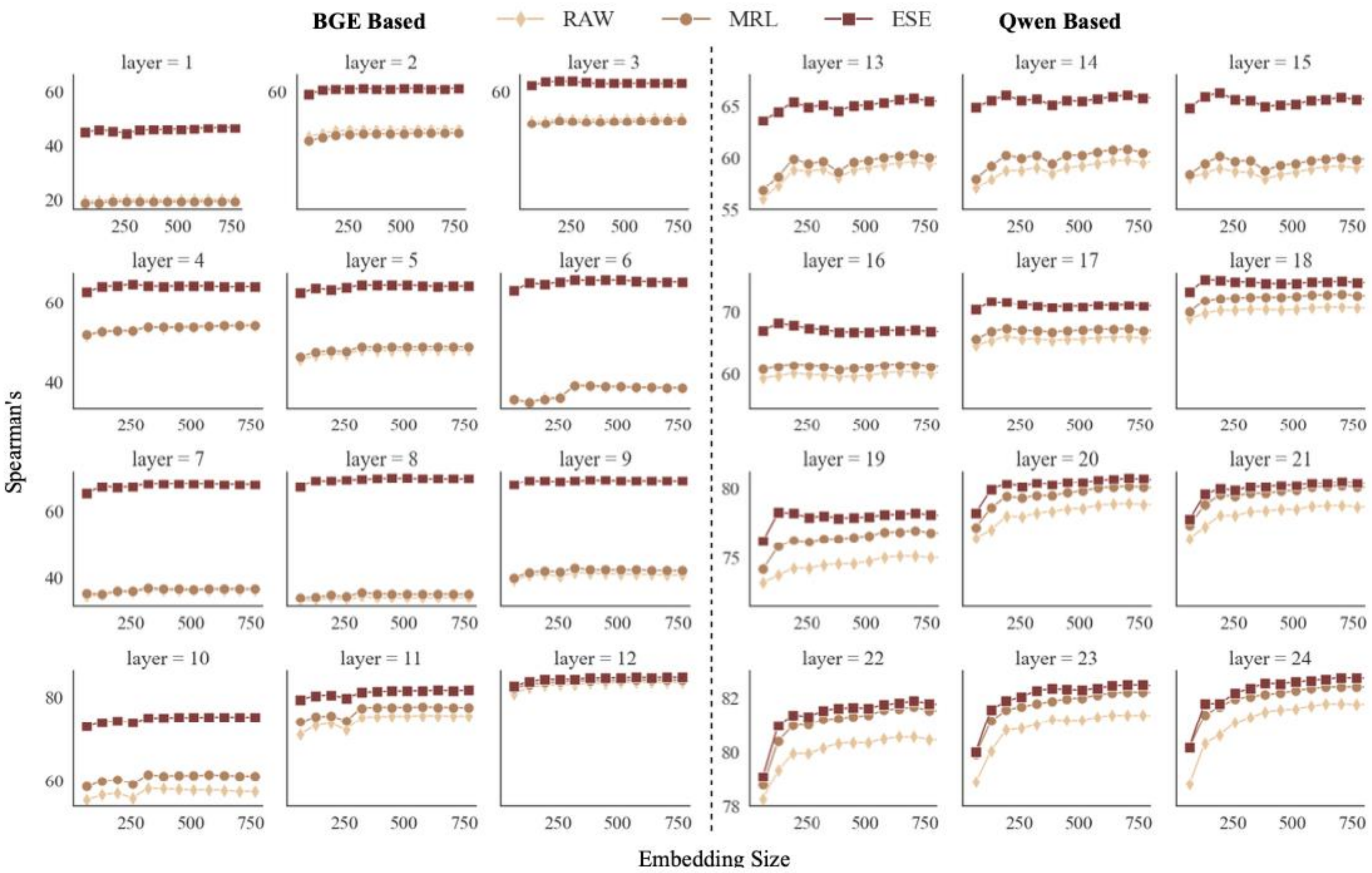
- Evaluated on standard STS benchmark datasets based on different backbones
- Compared with RAW (baseline), MRL, and ESE
- Results:
  - ESE performs better than baselines at shallow layers
  - Embeddings maintain quality with fewer parameters

Table 1: STS benchmark results. The last column ($\prec$ Avg.) is the average results of shallow layers (except the last one), while the remaining correspond to the last-layer results. Avg.: average results over varying benchmark datasets. RAW: the original model; MRL: Kusupati et al. (2022). The coffee-colored cells: the best results for each backbone model; boldfaced numbers: the overall best results. For $\prec$ Avg., ESE performs significantly better than baselines: p-value $< 5\%$ (paired t-test).

Model	STS12	STS13	STS14	STS15	STS16	STS-B	SICK-R	Avg.	$\prec$ Avg.
bge-base-en-v1.5 (Xiao et al., 2023)									
RAW	78.03	84.18	82.27	87.96	85.47	86.41	79.88	83.46	45.60
+ MRL	75.90	87.87	83.97	88.92	85.07	87.17	79.18	84.01	46.18
+ ESE	77.70	86.97	83.57	89.43	86.16	87.27	80.32	84.49	**66.27**
UAE-Large-V1 (Li & Li, 2024a)									
RAW	79.09	89.62	85.02	89.51	86.61	**89.06**	**82.10**	85.86	44.80
+ MRL	78.26	90.19	84.91	89.48	86.17	88.49	79.28	85.25	44.97
+ ESE	**79.64**	**90.40**	**85.76**	**90.33**	**86.64**	88.54	81.09	**86.06**	59.12
Qwen1.5-0.5B (Bai et al., 2023)									
RAW	75.91	83.77	80.04	86.05	82.91	85.32	78.98	81.85	56.59
+ MRL	76.30	85.04	80.68	86.15	83.12	85.65	79.45	82.34	58.22
+ ESE	76.43	85.70	81.75	86.30	83.67	85.76	80.16	82.82	59.99

STS Experiments

- Results of the STS benchmark for each of the last 12 layers of BGE-based backbone (left part) and Qwen-based backbone (right part).
- For each layer's result, the x-axis shows the embedding size, and the y-axis shows the average Spearman's correlation over varying benchmark datasets.



RAG Experiment

- Evaluated on HotpotQA for retrieval-augmented generation (RAG)
- Compared different embedding sizes (64-768 dimensions)
- Results:
  - ESE outperforms baselines across all embedding sizes
  - Larger performance gain at smaller embedding sizes

embedding size	**Model**		
	RAW	+ MRL	+ ESE
64	29.86	30.48	**32.28**
128	38.85	38.90	**39.28**
256	42.05	42.20	**42.50**
512	44.16	44.20	**44.44**
768	45.06	45.09	**45.31**

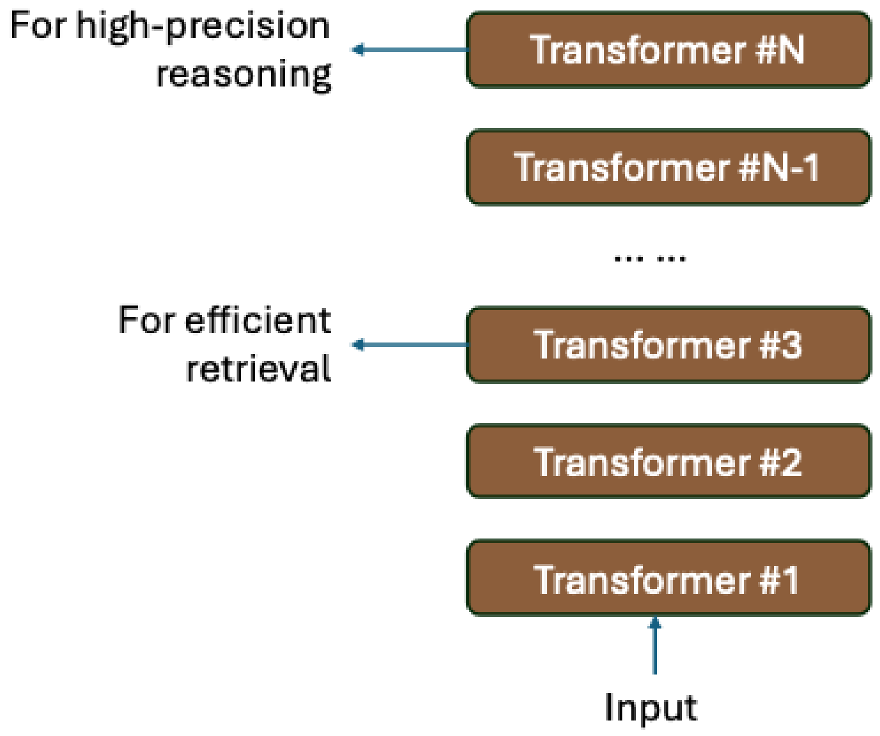
Application Scenarios

Different enterprise applications have diverse requirements for response speed and accuracy:

- Real-time content retrieval →
  requires small, efficient models
- Chatbots and reasoning applications →
  require deeper, more complex models

Diverse resource requirements due to device limitations:

- Server deployment →
  relatively small computing power limitations
- PC/mobile →
  strict computing power limitations



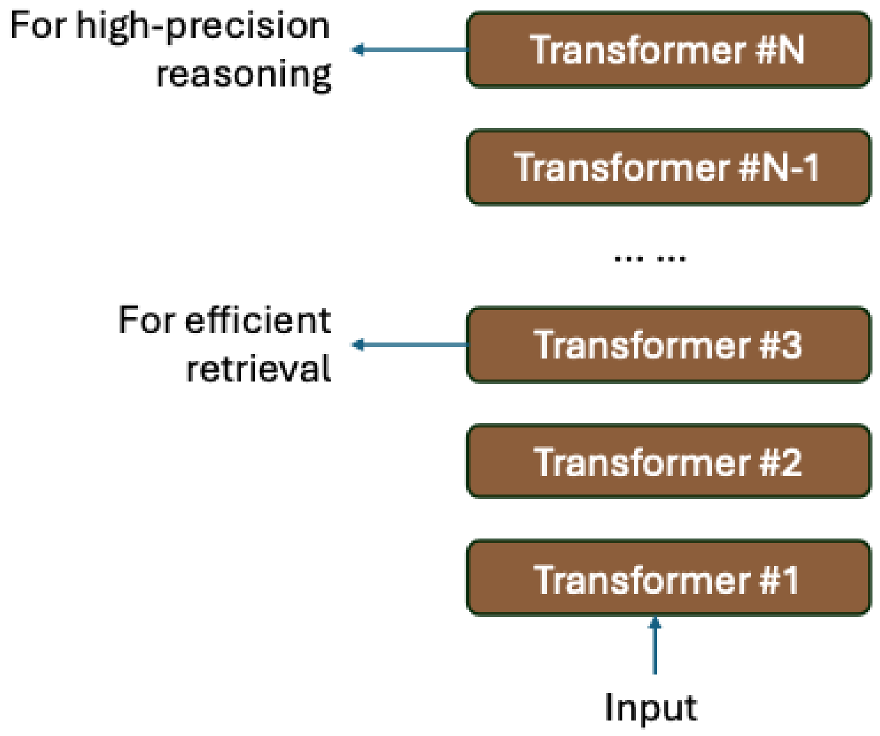
Application Scenarios

Traditional approach:

- Models with multiple scales need to be trained and maintained for different applications
- Updating the model requires retraining all versions

ESE solution:

- A single trained model can dynamically scale to meet various computing needs
- Reduces storage and operational overhead
- Only need to train one model for once when updating information



THE HONG KONG POLYTECHNIC UNIVERSITY
香港理工大學

Lingnan University 嶺南大學 香港 Hong Kong

ICLR

Thank you!

Openreview

Code Repo

WeChat of Xianming

WeChat of Zongxi

ESE: Espresso Sentence Embeddings

Xianming Li, **Zongxi Li**, Jing Li, Haoran Xie, Qing Li

`xianming.li@connect.polyu.hk, zongxili@ln.edu.hk`