TRACK 1 · CORE AI/ML | EFFICIENT ML & OPTIMIZATION: PEFT/LoRA, DISTILLATION, INFERENCE EFFICIENCY

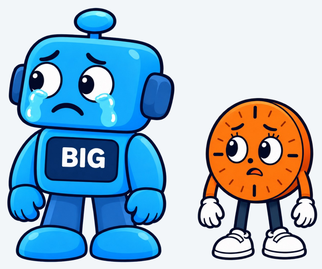
ElasticSkip: An Any-Depth Query Encoder for Efficient Dense Retrieval

Deepak Saini, **Suchith Chidananda Prabhu**, Aamod Khatiwada

Yardi School of Artificial Intelligence, IIT Delhi · Advisors: Sumeet Agarwal, Manik Varma

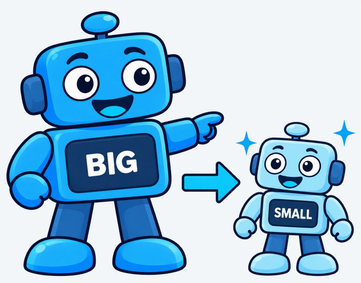
Where you cut matters more than how you heal the cut — and one base can serve every depth.

The problem, in four frames

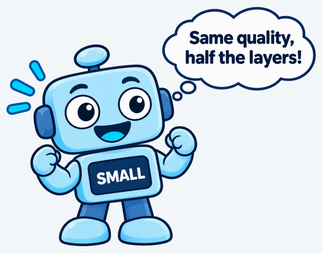


A big encoder ranks best — but it is slow on every query.

Documents are embedded once, offline. Time does not matter here.



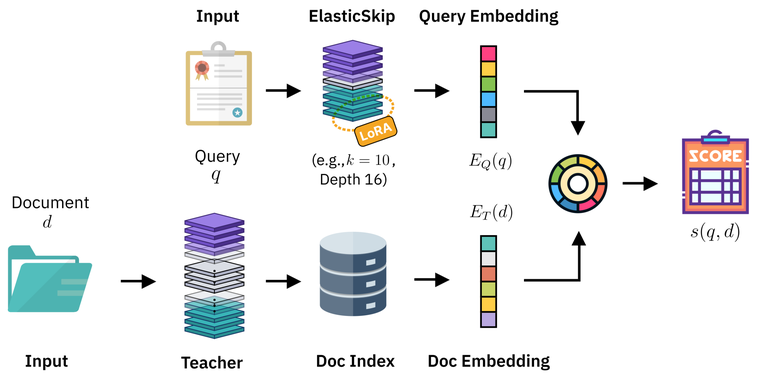
So the big model clones a small one from itself.



The clone answers queries fast. Everyone is happy.

Why this works

- Keep the big encoder on the **document** side — the index is built once and frozen.
- Compress only the **query** tower. It is the entire online cost.



One dot product against a frozen index. Only the query side is ever on the clock.

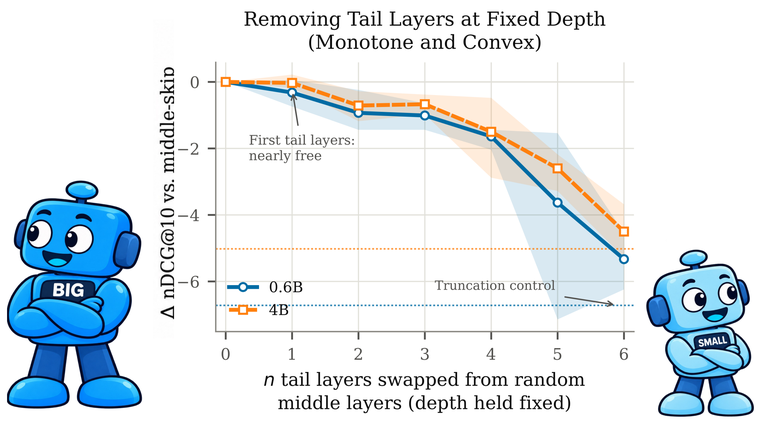
Our thesis

- Cut the **middle**. Keep the **prefix** and the **tail**.
- One base serves **every depth** — clones on demand, no retraining.

Limitations of prior SOTA

- Prior work clones badly: it prunes by **importance**, a rule borrowed from generative LLMs.
- That deletes **4–5 of the last 6 layers** — the ones that build the query vector.
- And every latency target needs its **own** clone, trained from scratch.

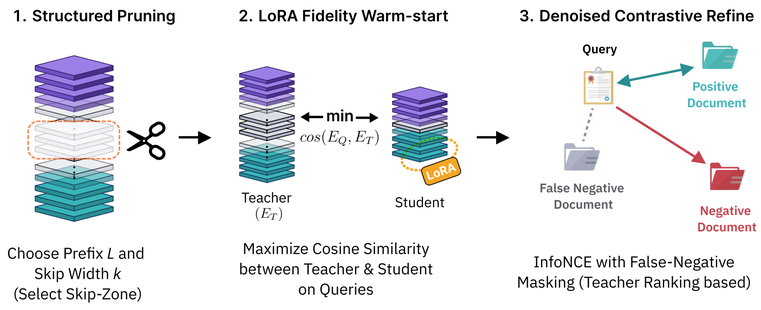
The tail is causally necessary



Swap tail layers for middle layers at fixed depth: monotone, convex collapse. What pruning gives up is the tail, not the depth.

How we clone

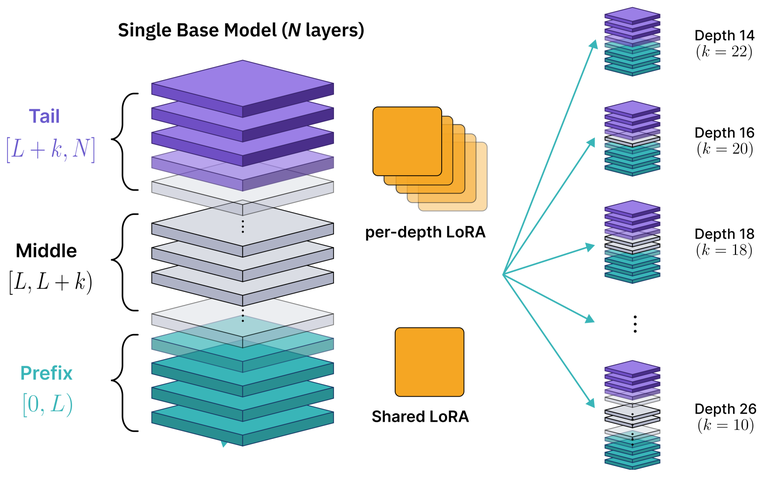
- Cut the **middle**, keep the **prefix** and the **tail** — a fixed structural cut, no calibration data.
- Align the clone into the frozen index space, then sharpen it on real queries.



Skip the middle → align into the index space → denoised contrastive refinement.

One base, many clones

- A shared prefix adapter plus a **per-depth tail adapter**: every depth on the frontier comes from one set of weights.
- A new operating point is a cheap adapter, not a retrained model — and adapters **merge at deploy** for zero runtime overhead.



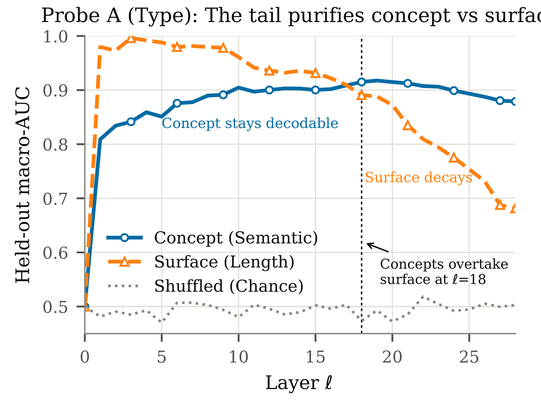
The executed depth is chosen at deployment, not fixed at training time.

Novelty

- **Contiguous middle-skip.** A fixed cut that needs no calibration data, so it transfers.
- **Elastic depth.** One base spans the whole frontier through cheap per-depth adapters.
- **Index-space training.** Align, then refine with false-negative masking.

What the probes show

- Across depth the tail trades surface form for **concepts** — and the readout is where query and document finally agree on meaning.



Concept decodability holds near 0.91; surface decays 0.99 → 0.68. The curves cross at layer 18.

Formulation

- **Score.** $s(q, d) = \langle E_Q(q) | E_T(q) \rangle$ — one dot product.
- **Skip.** Run $[0, L) \cup [L + k, N)$; depth $= N - k$. Fix L, sweep k.
- **Objective.** Denoised InfoNCE — mask the teacher's top-M as likely false negatives.

Where you cut decides everything

Selection · MS MARCO dev	Final-6	0.6B	4B
Middle-skip (ours)	**6 / 6**	**37.44**	**40.78**
Norm-ratio (HARNESS-LM)	2 / 2	35.74	39.56
Block-Influence (ShortGPT)	2 / 1	35.94	36.68
Random (control)	—	31.90	39.75
Tail (truncation)	0 / 0	30.72	35.76
Prefix (drop-early)	6 / 6	30.85	31.36
span (max – min)		**6.72**	**9.42**

MS MARCO dev, nDCG@10, keep 14/28 and 16/36. *Same recipe every row; only the cut moves — it spans 6.72 / 9.42 nDCG, the objective just 2.03 / 1.00.*

What it buys you

9.4× more nDCG from the cut than from the objective

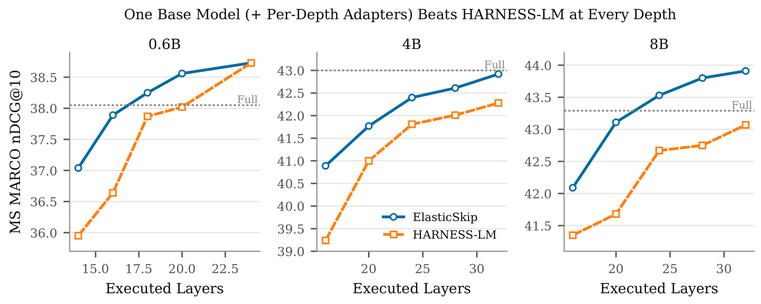
50% of query-tower layers removed (0.6B)

+2.82 nDCG gained out of domain (BEIR, 4B)

Against SOTA, at three scales

Depth · MS MARCO dev	Ours	HARNESS-LM	Δ
0.6B — 24 of 28	38.73	38.73	+0.00
0.6B — 14 of 28	**37.04**	35.95	**+1.09**
4B — 32 of 36	**42.92**	42.28	**+0.64**
4B — 16 of 36	**40.89**	39.24	**+1.65**
8B — 32 of 36	**43.91**	43.07	**+0.84**
8B — 16 of 36	**42.09**	41.35	**+0.74**

MS MARCO dev, nDCG@10. *Ahead at 14 of 15 operating points, tied at the 15th. At 8B / 32 the clone beats its own teacher (43.29).*



One elastic base tracks or beats a per-depth specialist family at every scale.

And out of domain

BEIR · 11 collections	0.6B	4B
Uncompressed	52.1	58.4
HARNESS-LM	41.8	46.1
ElasticSkip	**43.7**	**49.0**
Δ out of domain	**+1.91**	**+2.82**
Δ in domain	+1.09	+1.65

Zero-shot BEIR, 11 collections, nDCG@10, most aggressive depth. *Only the query tower is compressed; a fixed cut needs no in-domain calibration.*

Trade-offs & impact

- Keeping the tail is necessary but **not sufficient** — prefix-drop keeps all six and is still worst.
- Main runs are **single-seed**; across eight other encoder families the result is still open.
- **Half the layers** on the online path, index never rebuilt — one artifact instead of a family.

Code, paper & full results