ADAPT
Engaging Content
Engaging People

Iterative Layer Pruning for Efficient Translation Inference

A World Leading SFI Research Centre

Science Foundation Ireland
SFI For what's next

Yasmin Moslem
ADAPT Centre
School of Computer Science and Statistics
Trinity College Dublin

Muhammad Hazim Al Farouq
Kreasof AI
Research Labs

John D. Kelleher
ADAPT Centre
School of Computer Science and Statistics
Trinity College Dublin

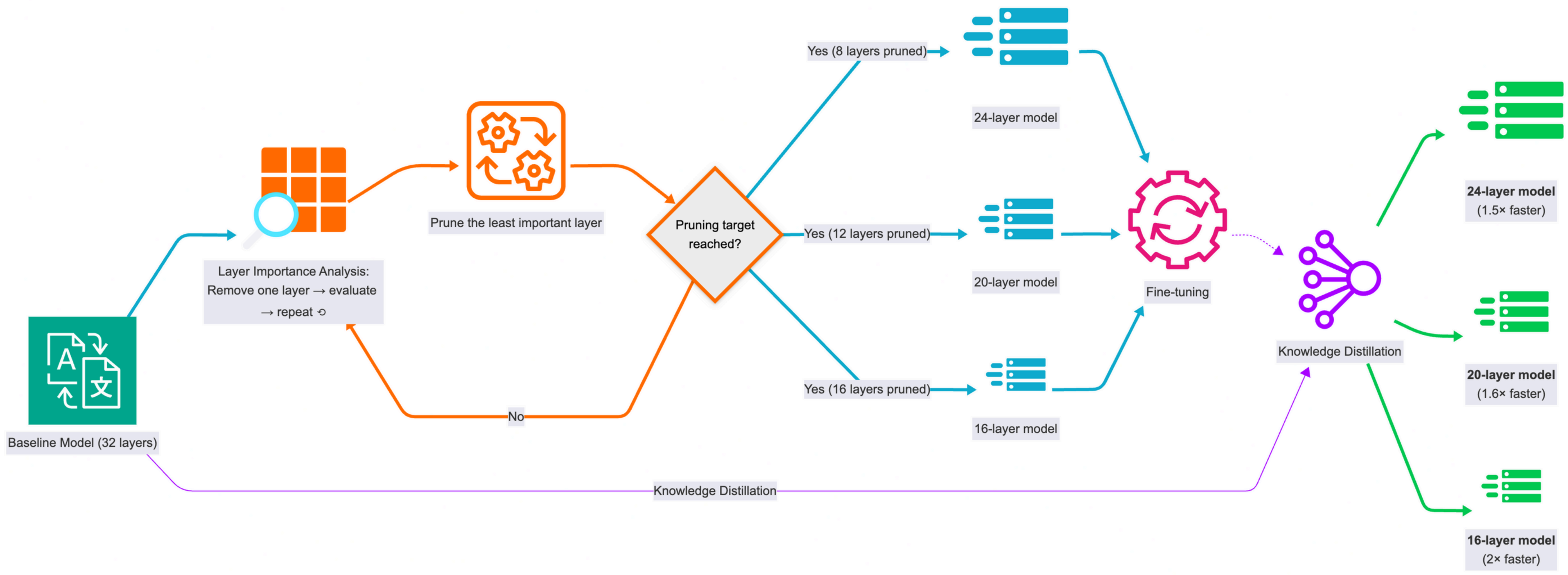
Iterative Layer Pruning

- Incrementally identifying and removing layers with minimal contribution to translation quality, one layer at a time.

- Fine-tuning the pruned models on the training dataset to restore the translation quality.
- Knowledge distillation from the teacher baseline.

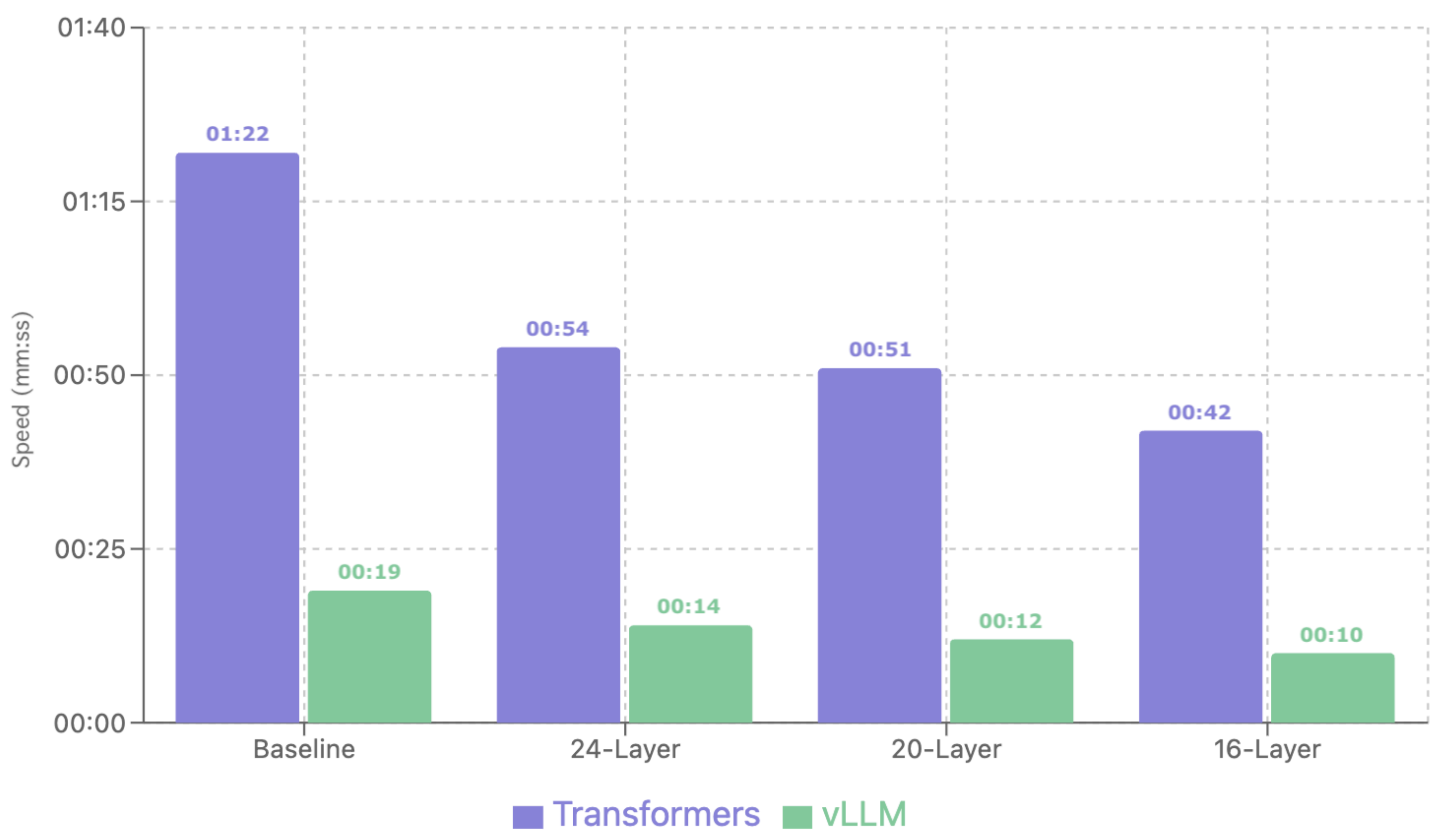
Layer Importance Evaluation

1. Remove one layer of the model.
2. Evaluate the model (chrF++).
3. Repeat for the rest of the layers.
4. Prune the least important layer (best chrF++ without it).
5. Repeat #1 to #4 until reaching the pruning target.


Baseline Model (32 layers)
Layer Importance Analysis: Remove one layer → evaluate → repeat ⟲
Prune the least important layer
Pruning target reached?
No
Yes (8 layers pruned)
Yes (12 layers pruned)
Yes (16 layers pruned)
24-layer model
20-layer model
16-layer model
Fine-tuning
Knowledge Distillation
Knowledge Distillation
24-layer model (1.5× faster)
20-layer model (1.6× faster)
16-layer model (2× faster)

Language	Model	Layers	chrF++ ↑	COMET ↑	Params (B) ↓	Speed (mm:ss) ↓
CES-DEU	Baseline	32	**52.79**	**87.18**	8.03	00:47
	Pruned + FT	24	51.35	85.70	6.28	00:34
		20	49.45	83.95	5.41	**00:27**
		16	45.79	79.39	**4.54**	**00:27**
ENG-ARZ	Baseline	32	42.03	81.45	8.03	01:22
	Pruned + FT	24	**58.38**	**85.74**	6.28	00:54
		20	55.69	84.50	5.41	00:51
		16	51.17	82.10	**4.54**	**00:42**

Evaluation of layer pruning experiments. For translation from Czech to German (CES-DEU), pruning 8 layers and then fine-tuning the resulting model retains 98% of the translation quality (as measured by COMET). Interestingly, for translation from English to Egyptian Arabic (ENG-ARZ), the model resulting from pruning up to 16 layers and then fine-tuning outperforms the Aya-Expanse-8B baseline for this language pair.



⚡ Inference Speed Comparison: Pruned models achieve up to ~2× speedup.

Knowledge Distillation

- Transferring knowledge from a larger model (teacher) to a smaller one (student).
- In “sequence-level” knowledge distillation, the student model is trained to generate sequences that match the teacher's sequence outputs.

Model	Layers	KD	chrF++ ↑	COMET ↑
Baseline 32B	40	-	54.57	87.76
Baseline 8B	32	-	52.79	87.18
Pruned + FT	24	☒	51.35	85.70
		✓	52.68	86.50
	20	☒	49.45	83.95
		✓	51.25	85.19
	16	☒	45.79	79.39
		✓	48.60	81.39

Evaluation of knowledge distillation (KD). Fine-tuning the pruned models on a combination of authentic and synthetic data (from Aya-Expanse-32B) improved the Czech to German (CES-DEU) translation quality, with the 24-layer pruned model nearly matching the performance of the Aya-Expanse-8B baseline.

Iterative Layer Pruning for
Efficient Translation Inference

Efficient Speech Translation
through Model Compression and
Knowledge Distillation

Code (GitHub repository)
ymoslem/Model-Compression

Trinity College Dublin
The University of Dublin