Encoder-Aware Sequence-Level Knowledge Distillation for Low-Resource Neural Machine Translation

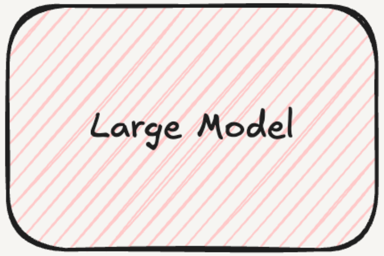
Menan Velayuthan[1], Nisansa de Silva[1], Surangika Ranathunga[2]
[1]Dept. of Computer Science & Engineering, University of Moratuwa, Sri Lanka
[2]School of Mathematical and Computational Sciences, Massey University, Auckland, New Zealand
Presented by : Menan Velayuthan

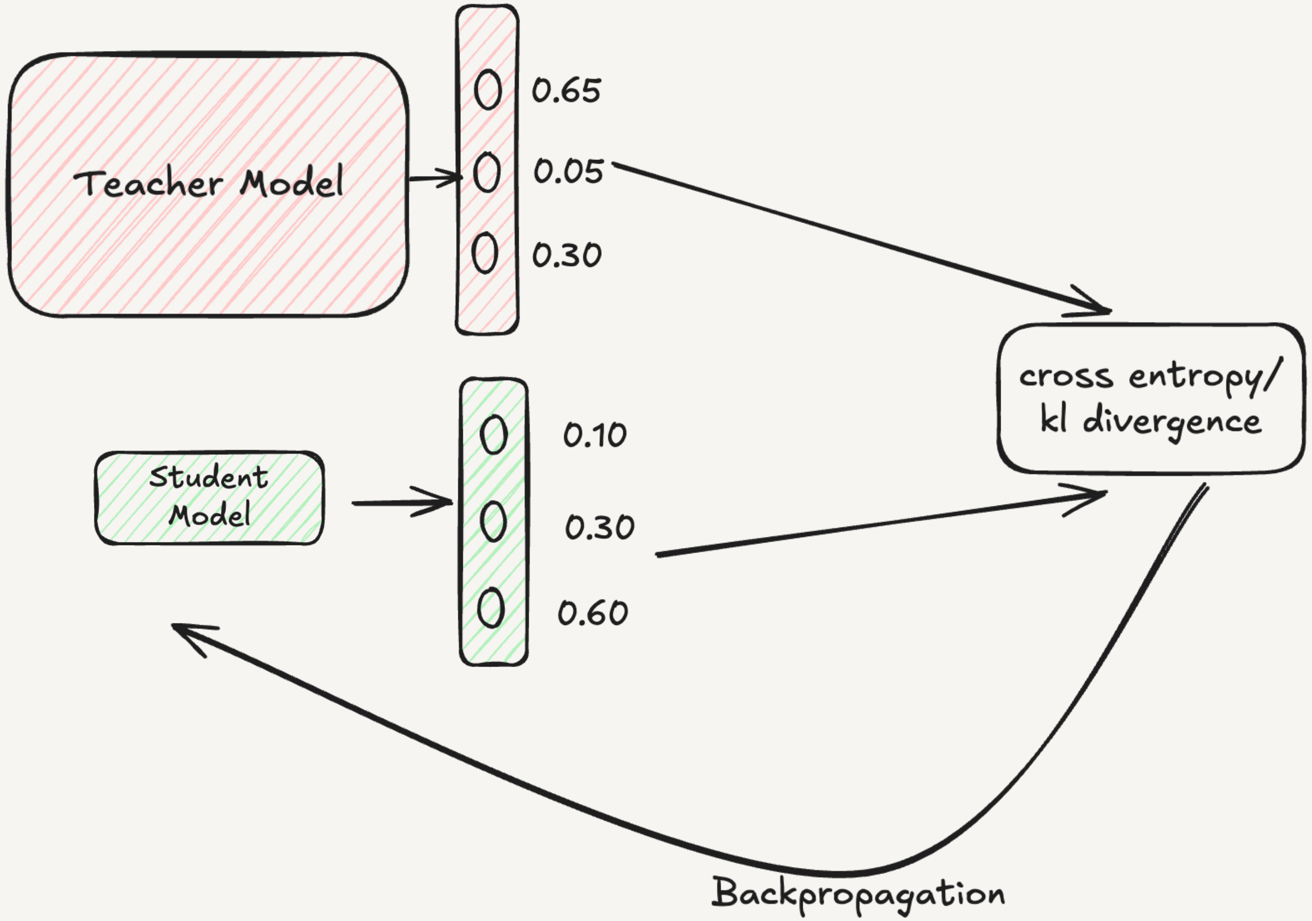
LoResMT@NAACL

Motivation

- Domain specific Neural Machine Translation (NMT) systems are of high demand as general NMT systems have limited applications (Saunders,2022).
- Sequence-Level Distillation (SLD) (Kim and Rush, 2016) enabled Knowledge Distillation (KD) (Hinton, 2015), to be applied for Sequence-to-Sequence (Seq2Seq) problem.
- Currey et al.(2020) utilized SLD to successfully perform multi-domain adaptation for NMT in high resource language setting.
- While LLMs excel in high-resource translation, encoder-decoder models outperform them in low-resource settings, making them still relevant and worth studying (Zhu et al., 2024).
- We hypothesize that SLD in encoder-decoder models **primarily distills the decoder**, resulting in limited encoder learning and weaker domain adaptation in low-resource settings; to address this, we introduce **encoder alignment to enhance knowledge transfer and adaptability**.

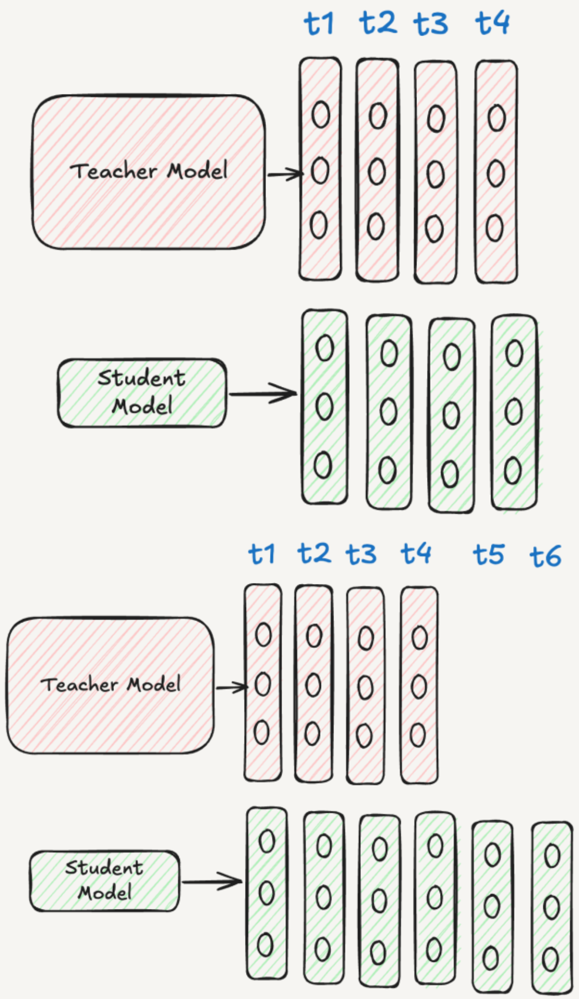
Knowledge Distillation

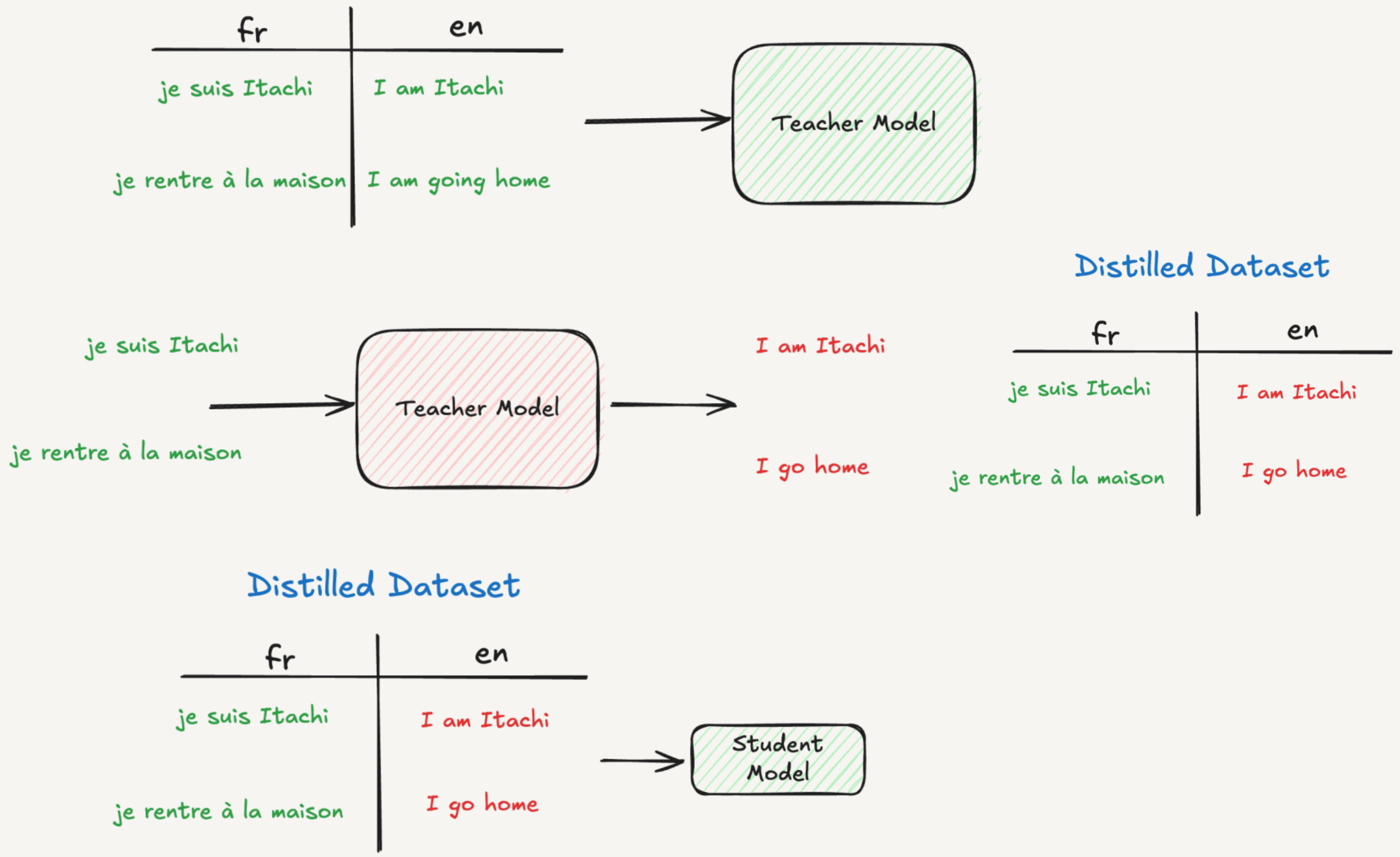




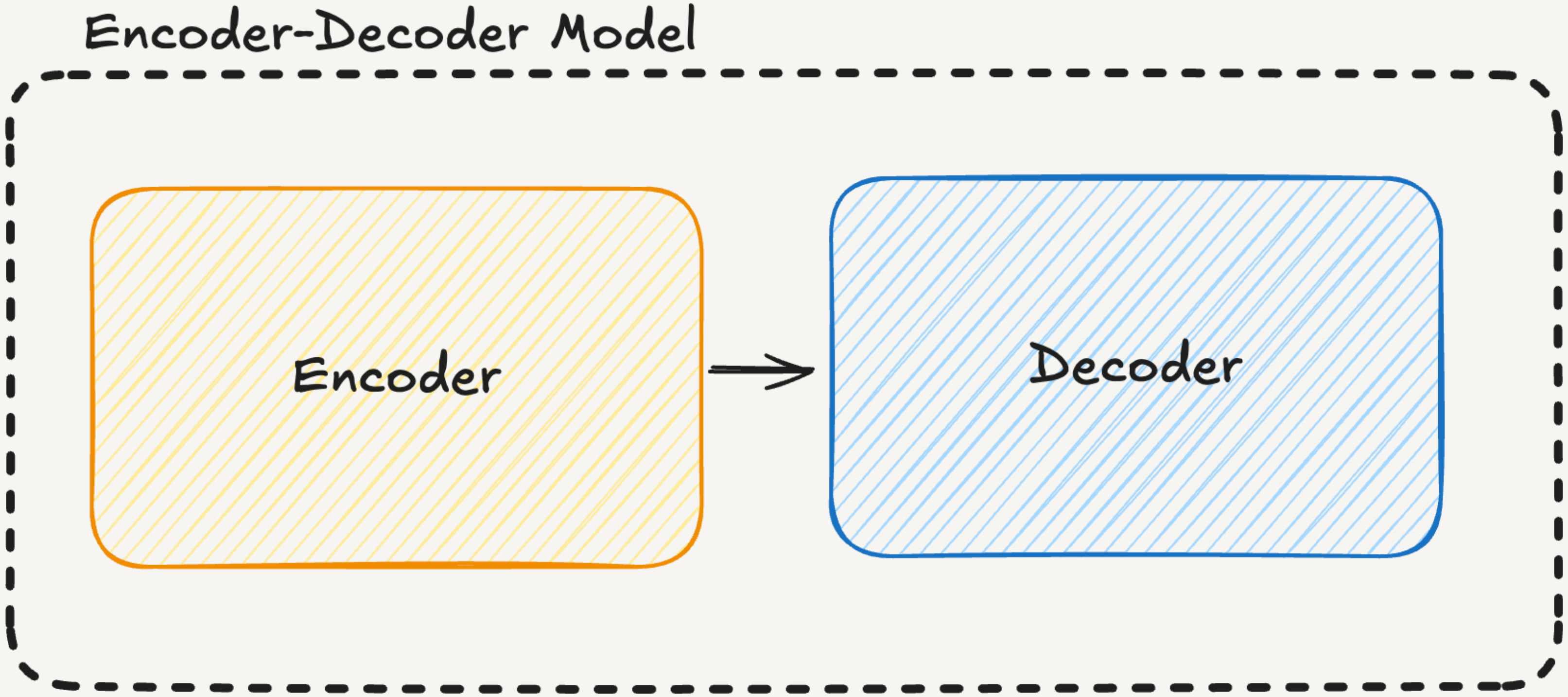
Sequence Level Distillation

Seq2Seq Task





Our Problem

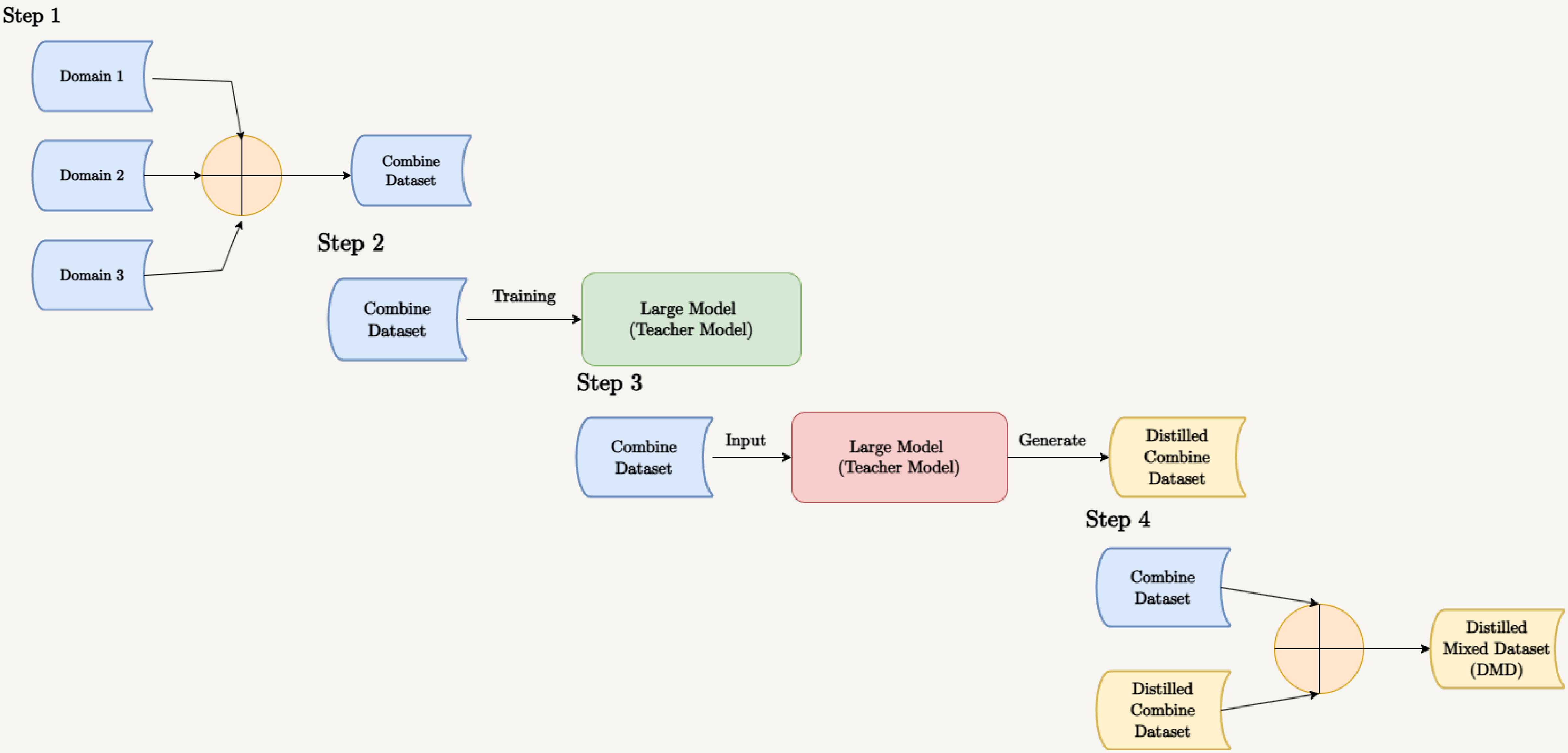


Research Questions

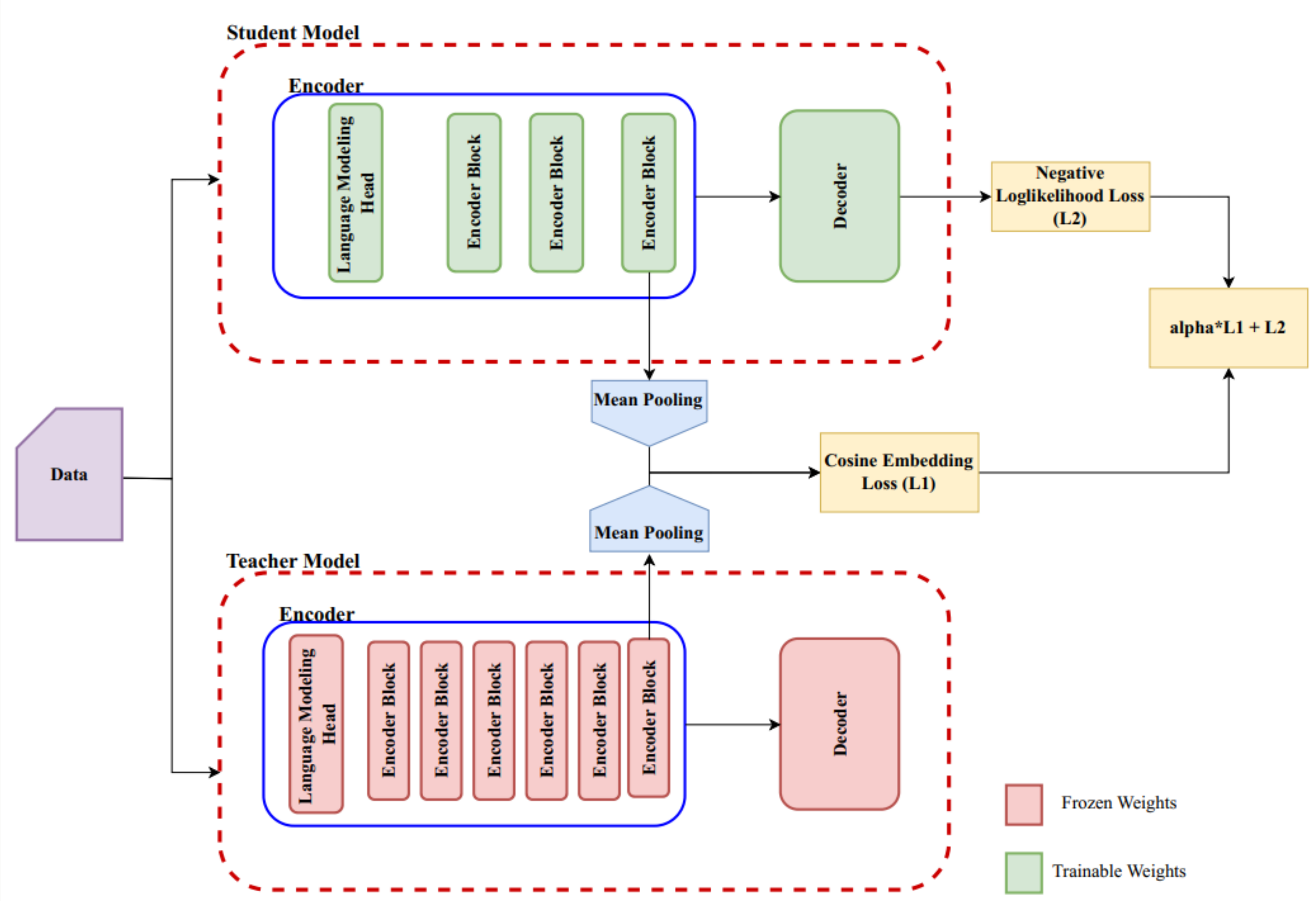
- **RQ1**: Is sequence-level knowledge distillation inherently decoder-focused in encoder-decoder NMT architectures?
- **RQ2**: Can encoder alignment improve knowledge transfer and domain adaptability in low-resource settings?

Methodology

Distilled Mixed Dataset(DMD) Creation



Proposed Teacher-Student Encoder Alignment



$$\mathrm{L}_{\mathrm{total}} = \alpha \cdot L_1 + L_2$$

- Here α is the attenuation factor used to control the contribution of the cosine embedding loss.
- We chose mean pooling based on BehnamGhader et al. (2024).
- A cosine-based loss function (Barz and Denzler,2019) was used for encoder alignment between teacher and student.

Experimental Setup

Two Studies

- **German-English**: simulated low resource setting, due to large number of domains (European Parliament, Law, Medical and News Commentary, Ted Talks, Open Subtitles).
- **Sinhala-English**: bonafide low resource setting.

German-English

- **Stage 1**: train on 4 domains (European Parliament, Law, Medical and News Commentary) and test on their respective test sets and one out of domain test set Flores200.
- **Stage 2**: further finetune the models from **Stage 1** on unseen domains (Ted talks, Open Subtitles).

Experimental Setup Contd.

Sinhala-English

- Study the impact of alpha.
- Train on 3 domains (CC Align, Open Subtitles, Sri Lankan Government) and test on their respective test sets and one out of domain test set Flores200.

Naming Conventions

- **L-ADO**: Large model trained on the All-Domain Original dataset.
- **S-ADO**: Small model trained on the All-Domain Original dataset.
- **S-ADD**: Small model trained on the All-Domain Distilled dataset (vanilla sequence-level distillation (Kim and Rush, 2016)).
- **S-DMD-NoAlign**: Small model trained on the Distilled Mixed Dataset (DMD) without teacher-student encoder alignment (as followed in (Currey et al., 2020)).
- **S-DMD-Align**: Small model trained on the DMD with teacher-student encoder alignment(using the proposed methodology).

Simulated Low Resource Setting (German-English)

Model	med	parl	law	news	Flores
L-ADO	63.27	56.33	63.73	53.80	50.89
S-ADO	62.31	55.66	62.39	53.28	50.23
S-ADD	62.36	56.08	62.92	53.87	50.90
S-DMD-NoAlign	61.38	55.49	61.89	52.91	49.88
S-DMD-Align	**63.43**	**56.92**	**64.08**	**54.88**	**52.90**

Table 2: ChrF scores of models trained with various configurations, evaluated on in-domain test sets (med, parl, law, news) and the out-of-domain Flores200 development-test set.

Model	opensub	ted
L-ADO	39.94	51.32
S-ADO	39.21	51.03
S-ADD	39.59	50.51
S-DMD-NoAlign	39.13	50.41
S-DMD-Align	**40.43**	**51.94**

Table 3: ChrF scores for Stage 1 models fine-tuned on single domains (Open Subtitles and Ted2020) to evaluate domain adaptation. Each model is fine-tuned on an individual domain and evaluated on its corresponding test set.

Real Low Resource Setting (English-Sinhala)

α	ccalign	opensub	gov	Flores
1.0	38.91	28.71	44.25	**28.11**
2.0	39.06	28.88	44.35	28.04
3.0	38.79	28.21	43.80	27.43
4.0	**39.54**	**28.91**	**44.66**	27.54
5.0	37.96	28.43	43.65	27.73
6.0	36.27	27.89	41.85	25.41
7.0	38.59	28.86	43.91	27.71

Table 4: ChrF scores of our model trained on the English–Sinhala language pair with different α values using the distilled dataset, evaluated on three in-domain test sets and the out-of-domain Flores200 development-test set.

Model	alpha	ccalign	opensub	gov	Flores
L-ADO	–	41.95	28.88	**48.44**	29.81
S-ADO	–	39.23	28.58	45.69	28.34
S-ADD	–	38.41	28.67	43.46	27.15
S-DMD-NoAlign	–	42.34	30.11	47.62	30.47
S-DMD-Align	1.0	42.78	30.36	47.25	30.54
S-DMD-Align	4.0	**43.11**	**30.42**	48.20	**31.03**

Table 5: ChrF scores of models trained with various configurations for the English–Sinhala translation direction, evaluated on three in-domain test sets and the out-of-domain Flores200 development-test set.

Conclusion

- RQ1 Answered: We confirmed that sequence-level distillation mainly transfers decoder knowledge, limiting encoder learning and leading to suboptimal performance.
- RQ2 Answered: Introducing encoder alignment improves knowledge transfer, resulting in better generalization and domain adaptability, especially in low-resource settings.
- Practical Impact: Our approach is effective even in compute-poor environments, making it a viable solution for multi-domain NMT under real-world constraints.

Acknowledgement

- This work was funded by the Google Award for Inclusion Research (AIR) 2022 received by Dr. Surangika Ranathunga and Dr. Nisansa de Silva.
- We would like to thank the National Languages Processing (NLP) Centre, at the University of Moratuwa for providing the GPU to execute the experiments related to the research.
- We would like to thank everyone involved in LoResMT and NAACL in making this possible.

References

Danielle Saunders. 2022. Domain adaptation and multi-domain adaptation for neural machine translation: A survey. J. Artif. Int. Res., 75.

Yoon Kim and Alexander M. Rush. 2016. Sequence-level knowledge distillation. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 1317–1327, Austin,Texas. Association for Computational Linguistics.

Geoffrey Hinton. 2015. Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531.

Anna Currey, Prashant Mathur, and Georgiana Dinu.2020. Distilling multiple domains for neural machine translation. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 4500–4511, Online. Association for Computational Linguistics.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2024. Multilingual machine translation with large language models: Empirical results and analysis. In Findings of the Association for Computational Linguistics: NAACL 2024, pages 2765–2781,Mexico City, Mexico. Association for Computational Linguistics.

Parishad BehnamGhader, Vaibhav Adlakha, Marius Mosbach, Dzmitry Bahdanau, Nicolas Chapados, and Siva Reddy. 2024. LLM2Vec: Large Language Models Are Secretly Powerful Text Encoders. In First Conference on Language Modeling.

Björn Barz and Joachim Denzler. 2019. Deep learning on small datasets without pre-training using cosine loss. 2020 IEEE Winter Conference on Applications of Computer Vision (WACV), pages 1360–1369.

Thank you!