IMPROVING ZERO SHOT CROSS-LINGUAL TRANSFER FOR LOW RESOURCE NLP

TEAM MINDLESS

TEAM MEMBERS

K.H.T.Chathumina - 200088D

P.L.I.D Puranegedara - 200487B

N. D. Ranathunga - 200517U

SUPERVISORS

Dr. Surangika Ranathunga

Dr. Nisansa de Silva

Dr. Mokanarangan Thayaparan

ZERO SHOT CROSS-LINGUAL TRANSFER

- When tasks specific data is not available for a low resource language, a multilingual model can be trained or fine-tuned with available task data in a high resource language.
- Then directly use the model for inference in low resource language task.
- Cross lingual representation of the multilingual models helps to achieve this.

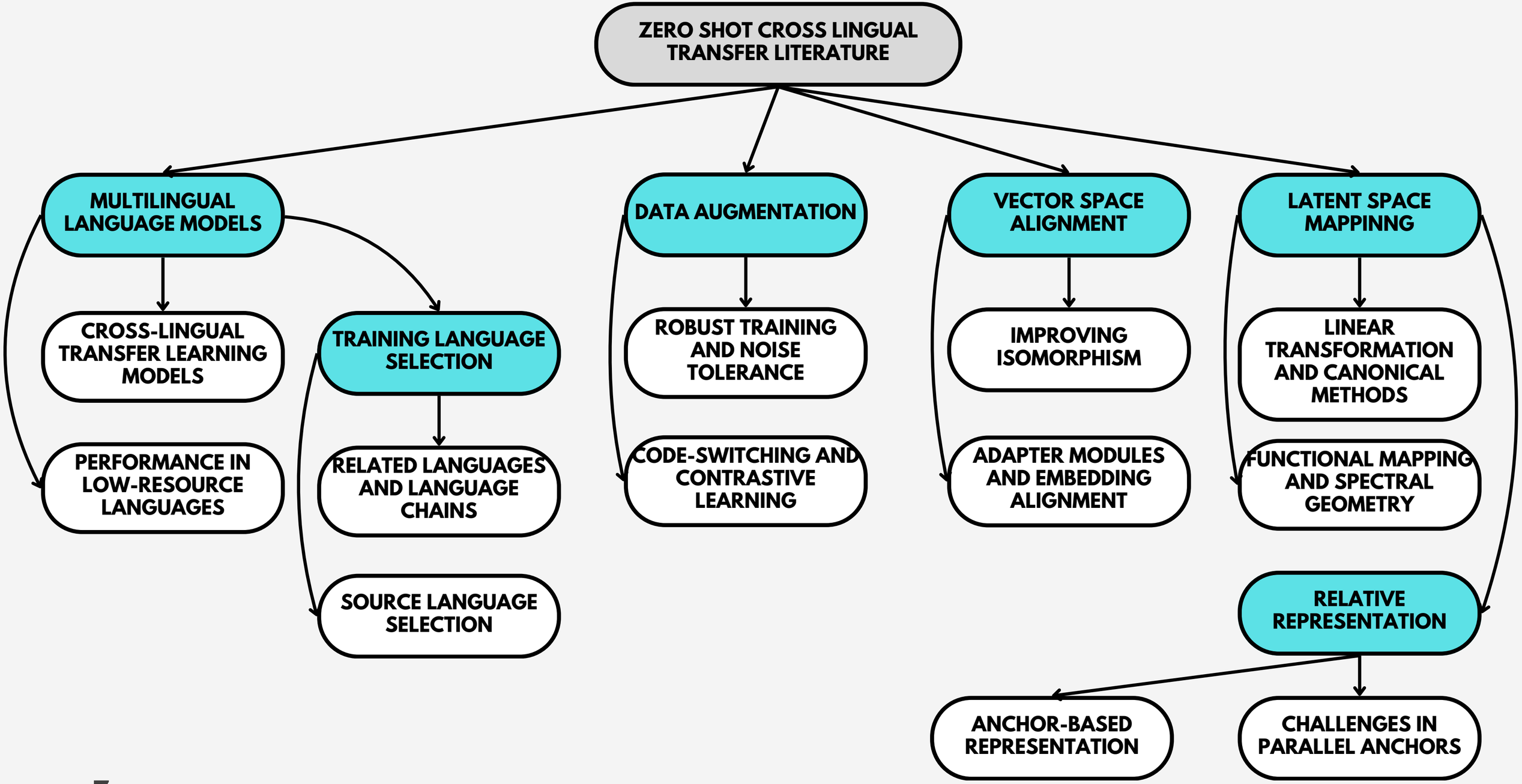
BACKGROUND

- **Data Scarcity**: Low-resource languages lack sufficient annotated and raw data, making model fine-tuning difficult.
- **Ineffective Models**: XLM-R and mBERT show poor zero-shot performance for low-resource languages, needing task-specific data that is often unavailable.
- **Suboptimal Fine-tuning**: Fine-tuning on high-resource languages or translating data shows limited success due to unreliable translation systems.
- **Non-Isomorphic Embedding Spaces**: Misaligned vector spaces across languages hinder cross-lingual transferability, complicating direct mapping between languages.

PROBLEM STATEMENT

- How to improve the cross lingual representation of the already pretrained mutilingual models for downstream tasks of low resource languages when task specific data is not there?

LITERATURE REVIEW


ZERO SHOT CROSS LINGUAL TRANSFER LITERATURE
MULTILINGUAL LANGUAGE MODELS
CROSS-LINGUAL TRANSFER LEARNING MODELS
PERFORMANCE IN LOW-RESOURCE LANGUAGES
TRAINING LANGUAGE SELECTION
RELATED LANGUAGES AND LANGUAGE CHAINS
SOURCE LANGUAGE SELECTION
DATA AUGMENTATION
ROBUST TRAINING AND NOISE TOLERANCE
CODE-SWITCHING AND CONTRASTIVE LEARNING
VECTOR SPACE ALIGNMENT
IMPROVING ISOMORPHISM
ADAPTER MODULES AND EMBEDDING ALIGNMENT
LATENT SPACE MAPPINNG
LINEAR TRANSFORMATION AND CANONICAL METHODS
FUNCTIONAL MAPPING AND SPECTRAL GEOMETRY
RELATIVE REPRESENTATION
ANCHOR-BASED REPRESENTATION
CHALLENGES IN PARALLEL ANCHORS

MULTILINGUAL MODELS

PERFORMANCE IN LOW-RESOURCE LANGUAGES

- Ebrahimi et al. (2022) [1] introduced AmericasNLI, an extension of XNLI, to include 10 Indigenous languages from the Americas.
- **XLM-R's (A. Conneau et al [2]) zero-shot performance on these languages was poor, with an average accuracy of 38.48%.**
- The challenges in cross-lingual transfer, especially with distant languages, emphasize the need for further research in this area.

[1] A. Ebrahimi et al., "AmericasNLI: Evaluating zero-shot natural language understanding of pretrained multilingual models in truly low-resource languages," Proc. 60th Annu. Meeting Assoc. Comput. Linguistics (ACL), Dublin, Ireland, 2022, pp. 6279-6299.

[2] A. Conneau et al., "Unsupervised cross-lingual representation learning at scale," Proc. 58th Annu. Meeting Assoc. Comput. Linguistics (ACL), Jul. 2020, pp. 8440-8451.

TRAINING LANGUAGE SELECTION

SOURCE LANGUAGE SELECTION

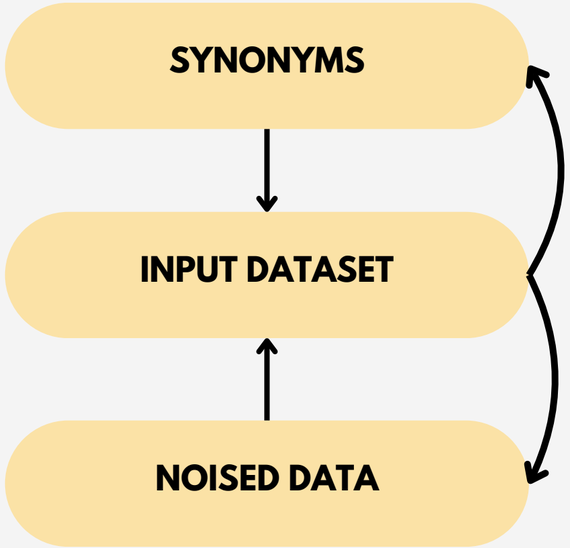
- Eronen (2023) [3] examined the impact of source language selection on NLP tasks like sentiment analysis, named entity recognition, and dependency parsing.
- The study proposes selecting the source training language based on linguistic similarity to the target language by measuring the distance between languages.
- **Significant improvements were achieved by selecting an optimal source language for training, rather than defaulting to English.**

[3] J. Eronen, M. Ptaszynski, and F. Masui, "Zero-shot cross-lingual transfer language selection using linguistic similarity," Information Processing & Management, vol. 60, no. 3, p. 103250, 2023.

DATA AUGMENTATION

ROBUST TRAINING

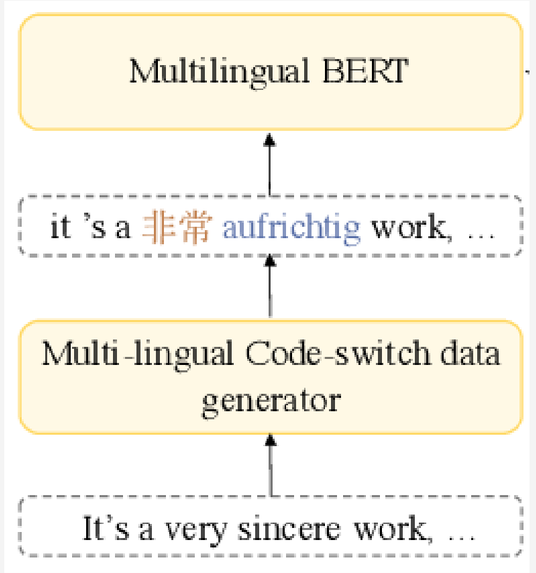
- Aepli (2022) [4] proposes augmenting data from high-resource languages with character-level noise.



[4] N. Aeppli and R. Sennrich, "Improving zero-shot cross-lingual transfer between closely related languages by injecting character-level noise," 2022.

CODE-SWITCHING

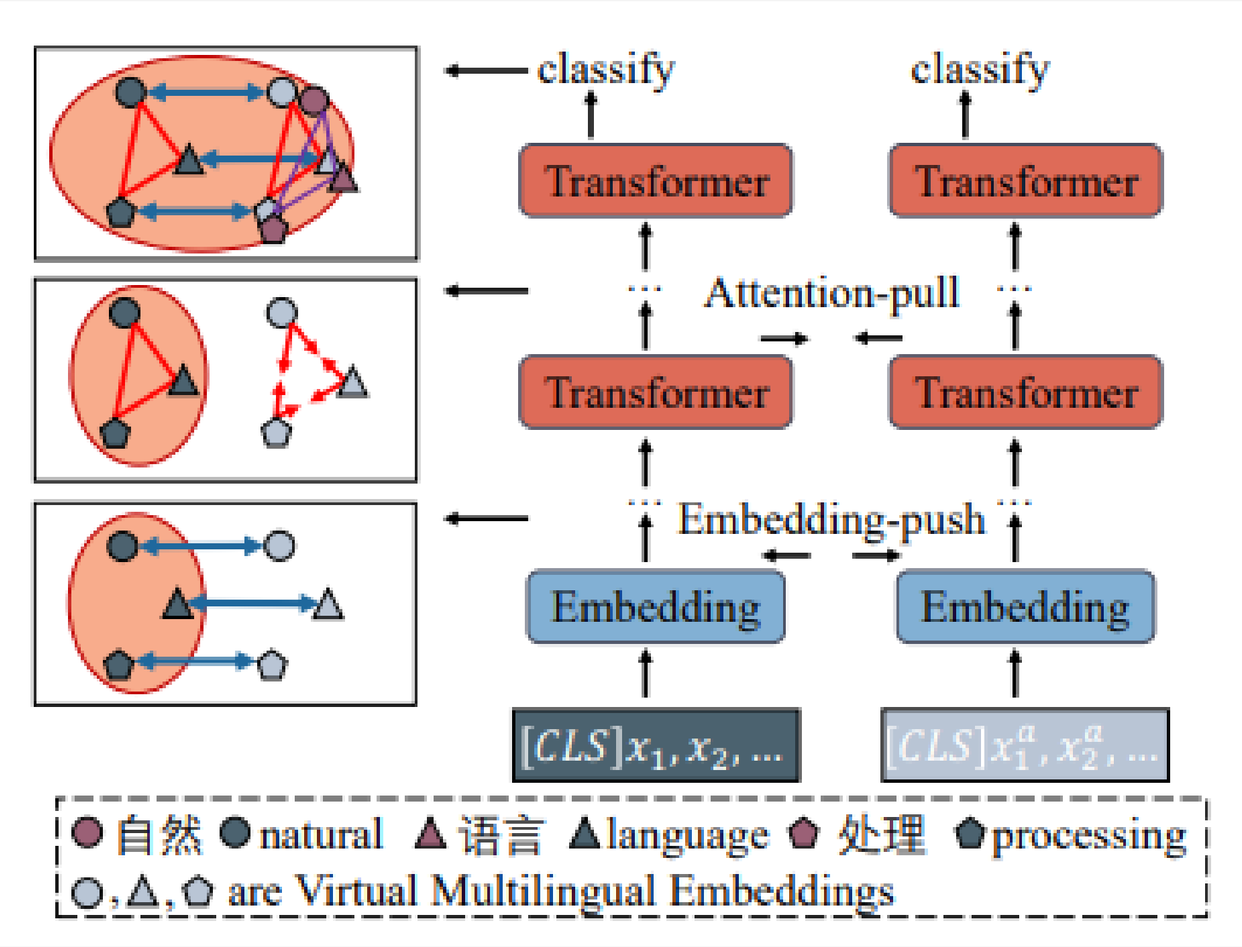
- Li (2024) [5] introduces a Code-Switching (CS) method to improve zero-shot cross-lingual transfer.



[5] Z. Li, C. Hu, J. Chen, Z. Chen, X. Guo, and R. Zhang, "Improving zero-shot cross-lingual transfer via progressive code-switching," 2024.

VECTOR SPACE ALIGNMENT

EMBEDDING ALIGNMENT



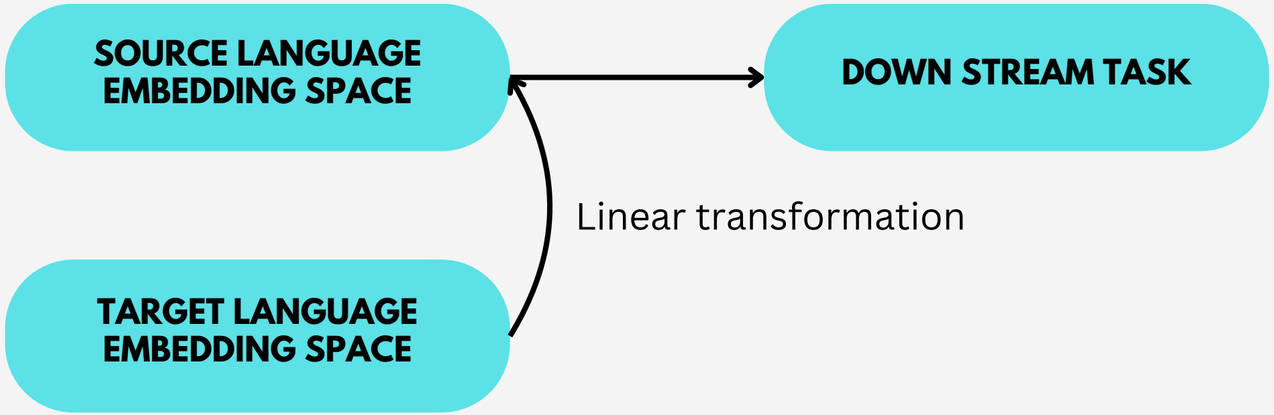
- Ding et al. (2022) [6] address the limitations of language embedding cluster distances in cross-lingual transfer with three approaches: **robust targets, attention-pull, and embedding-push**.
- Despite being evaluated on the XNLI dataset, these methods show low performance gains, particularly on the PAWS-X dataset, making them less significant for practical use.

[6] K. Ding et al., "A simple and effective method to improve zero-shot cross-lingual transfer learning," Proc. 29th Int. Conf. Comput. Linguistics, Gyeongju, Republic of Korea, Oct. 2022, pp. 4372-4380.

LATENT SPACE MAPPING

LINEAR TRANSFORMATION

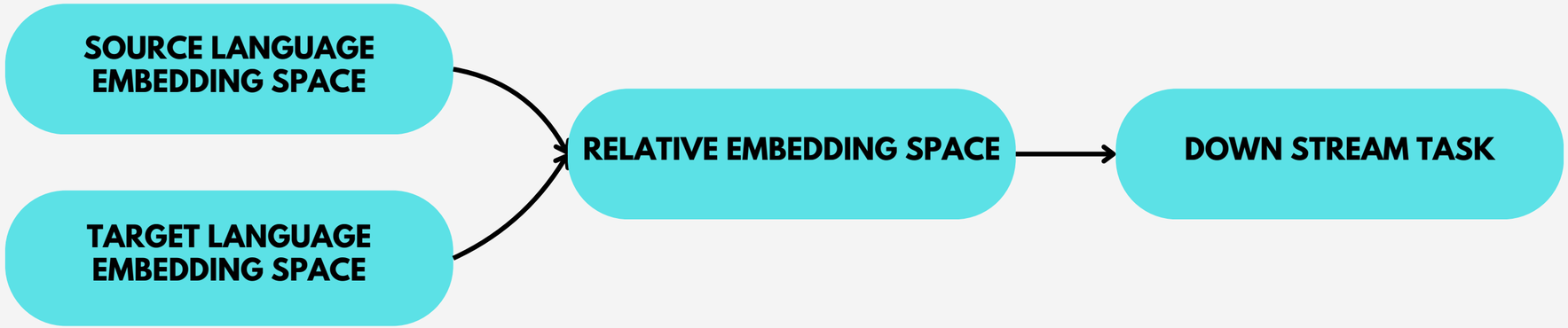
- Mikolov et al. (2013) [7] introduced the linear transformation method, a highly influential approach in aligning monolingual embedding spaces across languages.
- The method is based on the discovery that words and their translations exhibit similar geometric structures in monolingual embedding spaces.
- This only works in isomorphic vector spaces. But in practice, embedding spaces are not isomorphic



[7] T. Mikolov, Q. V. Le, and I. Sutskever, "Exploiting similarities among languages for machine translation," 2013.

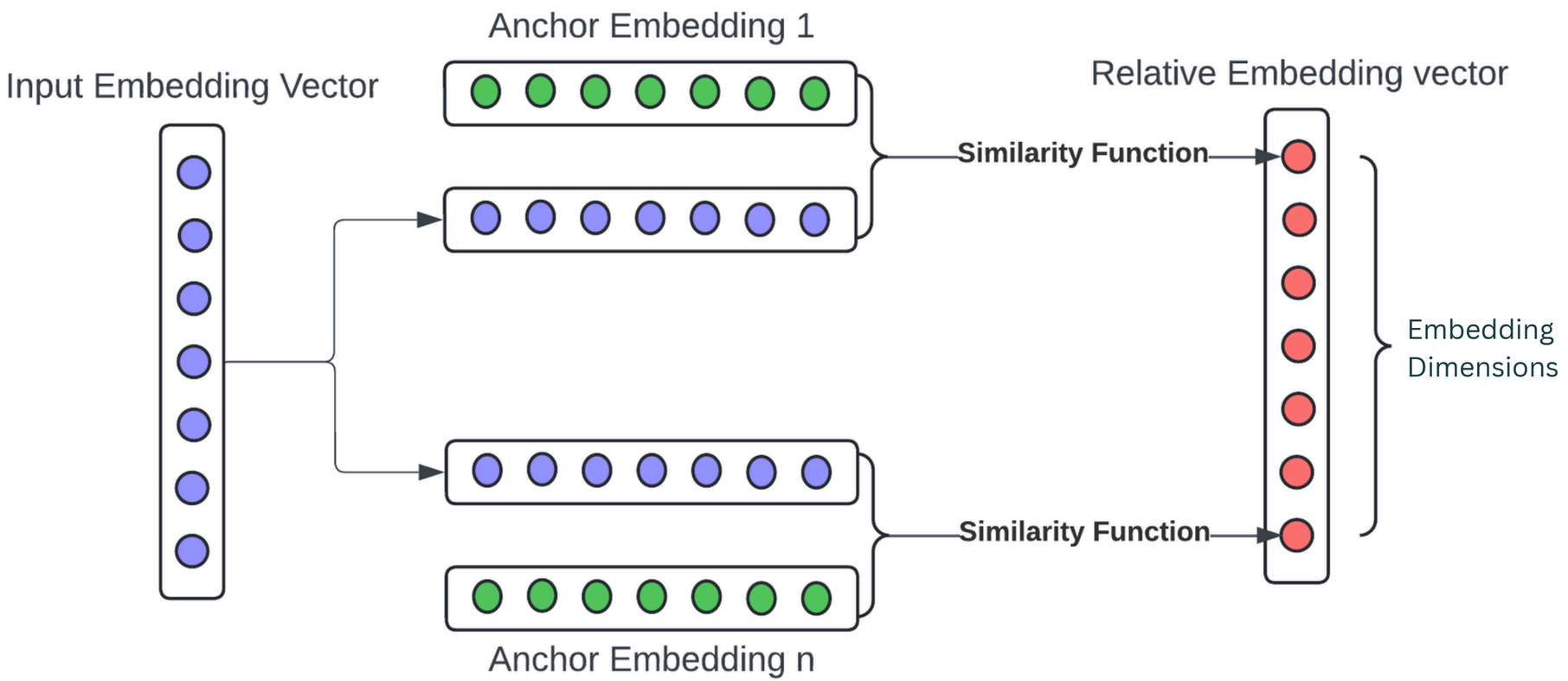
RELATIVE REPRESENTATION

- Moschella et al. (2022) [8] introduce the concept of "relative representations" to create invariant data representations in neural networks, resilient to transformations like random initialization and data shuffling.
- **Relative representations remain consistent despite transformations in the latent space, as they rely on angles (cosine similarity) rather than absolute positions.**
- This approach is applicable across various models, including image processing, monolingual language models, autoencoders, and graph neural networks, transforming original embeddings into a common, shareable space.



[8] L. Moschella et al., "Relative representations enable zero-shot latent space communication," arXiv preprint, arXiv:2209.15430, 2022.

RELATIVE REPRESENTATION

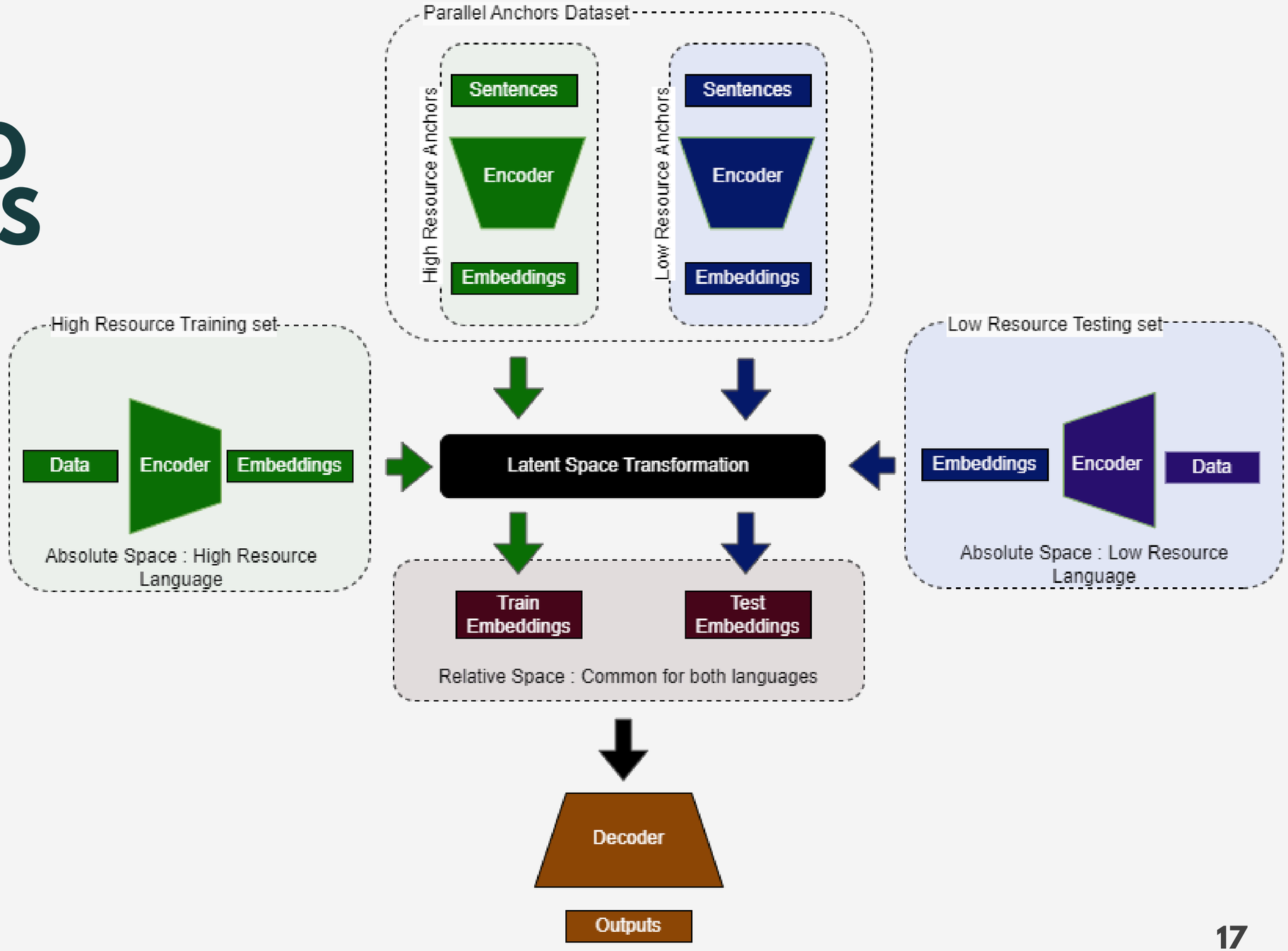


RESEARCH OBJECTIVES

- **Adapting Anchor-Based Zero-Shot Cross-Lingual Transfer**
  - Adapt anchor-based zero-shot cross-lingual transfer from high-resource to low-resource languages.
  - Transform latent embedding spaces into a common space for effective cross-lingual transfer.
- **Exploring and Adapting Architectural and Cross-Lingual Approaches to Improve the Performance for low resource languages**
  - Explore and adapt architectural and cross-lingual changes to enhance anchor-based zero-shot cross-lingual transfer.
- **Adapting Identified Approaches for Multiple Tasks**
- **Generalizing Approaches for Multiple Low-Resource Languages**

PROPOSED METHODOLOGY

ADAPTING ANCHOR BASED ZERO SHOT CROSS LINGUAL TRANSFER



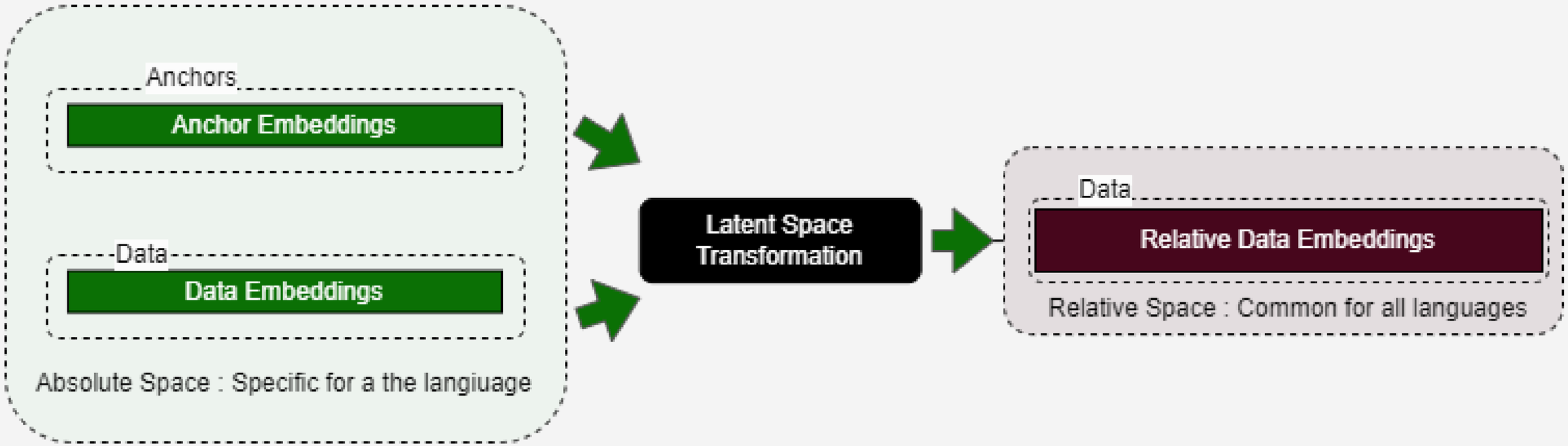
LATENT SPACE TRANSFORMATION

Latent Space Transformation

Maps embeddings from absolute space into a common relative space

Methods of Latent Space Transformation

- Relative Representation
- Affaine Transformation



Methods of Latent Space Transformation

RELATIVE REPRESENTATION

Relative representation transforms latent space by calculating the similarity between an absolute embedding vector and each anchor embedding vector.

$$Similarity\_Function(data\_embeddings, anchor\_embeddings) = relative\_data\_embeddings$$

- **Similarity Function**
  - Cosine Similarity
  - Soft Cosine Similarity
  - L1 Similarity

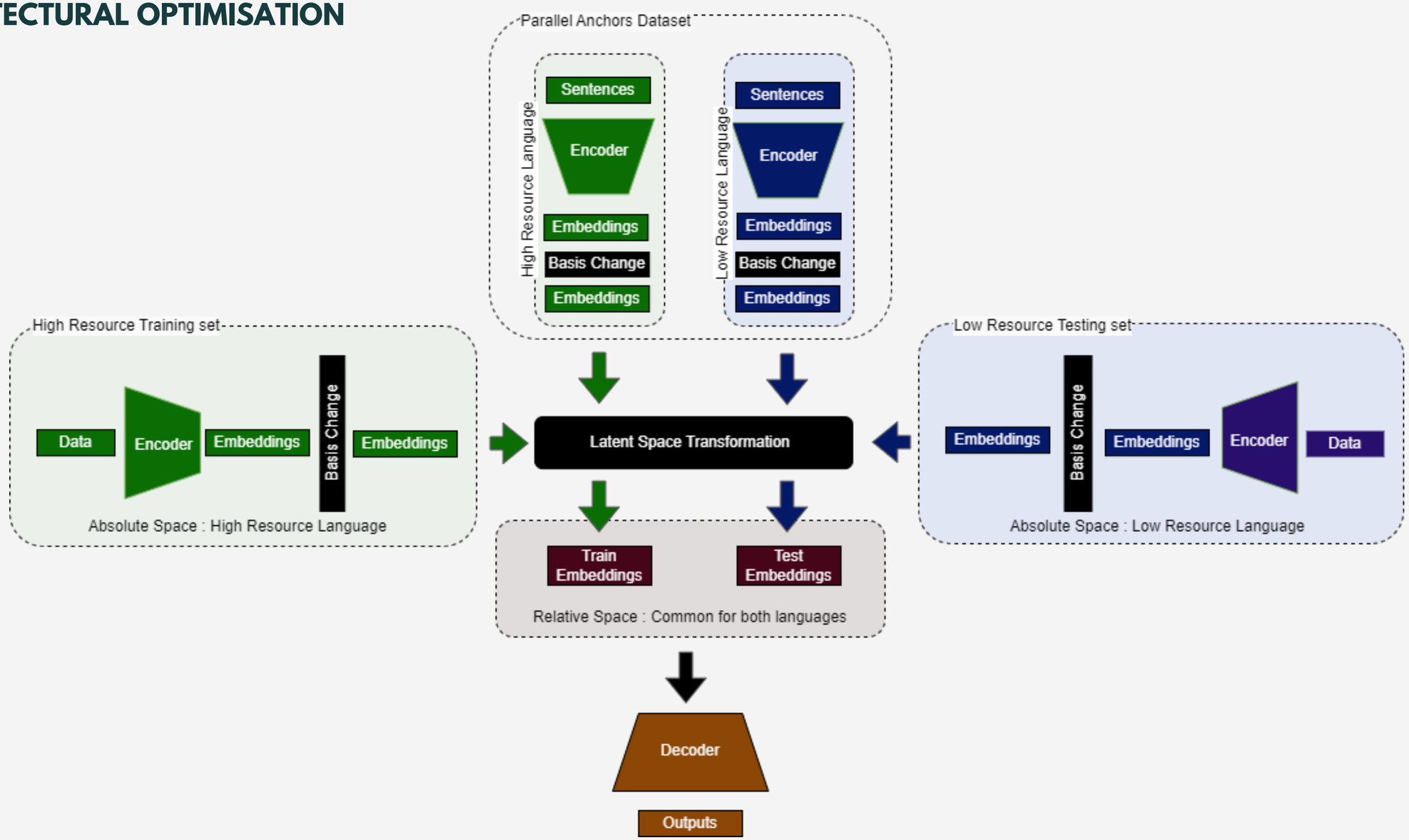
ENHANCING THE PERFORMANCE

WE PROPOSE THREE METHODS TO ENHANCE THE PERFORMANCE

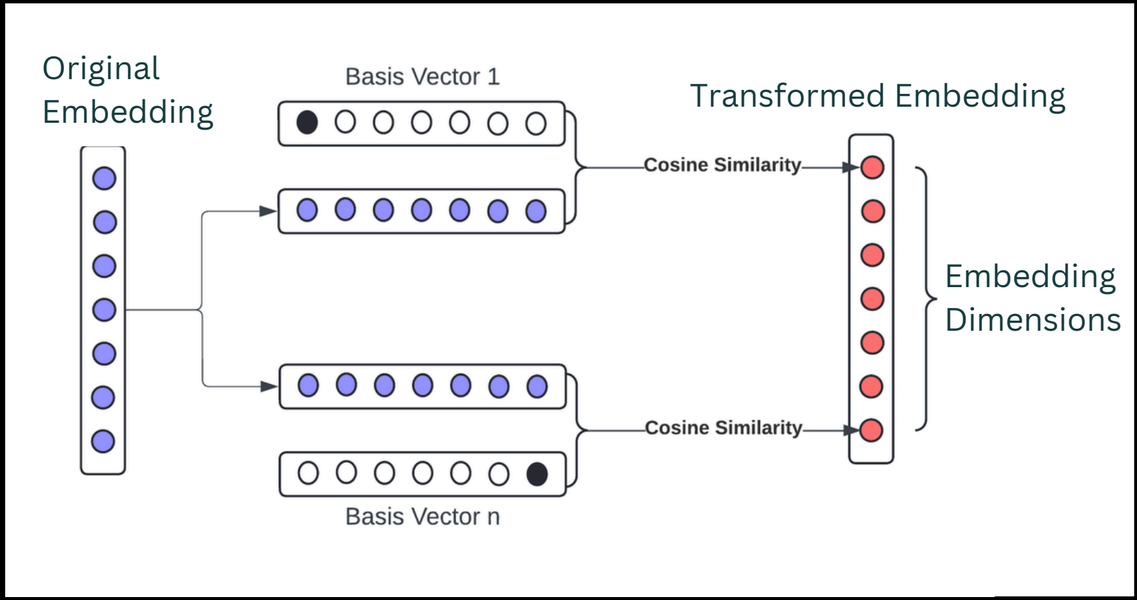
- Architectural Optimisation
- Adapting Cross-Lingual data augmentation
- Anchor optimisation

Enhancing Performance

ARCHITECTURAL OPTIMISATION



Enhancing Performance



Architectural Optimisation

Latent Space Transformation with Cosine Similarity:

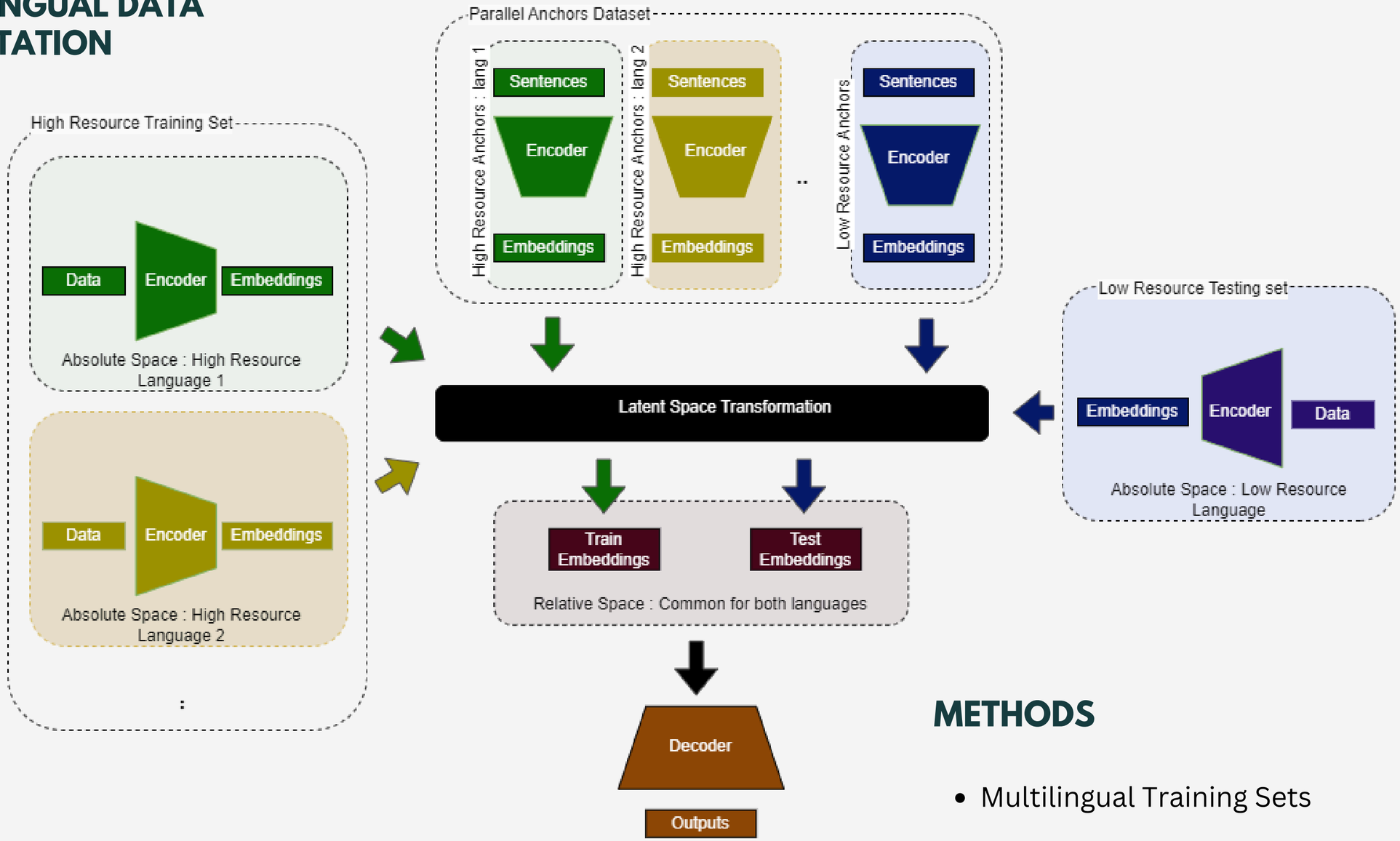
- Transforms vectors into angles relative to a base axis using cosine similarity.
- Enhances vector representation in the latent space.
- The cosine similarity between the vector and the basis vector of that dimension represents each vector dimension.
- Captures more semantic meaning across different languages.

Basis Transformation with Anchor Vectors:

- Represents embedding vectors relative to anchor vectors.
- Aligns and optimizes embeddings to improve the capture of semantic relationships.

Enhancing Performance

CROSS LINGUAL DATA AUGMENTATION



METHODS

- Multilingual Training Sets

Enhancing Performance

ANCHOR OPTIMISATION

Explore the relationship between anchors and embedding space:

- Find the how the selection of anchors affects the accuracy.
- Find methods to extract best anchor points from a given set of anchors.

Using In-Domain Anchors:

- Incorporate domain-specific anchors to enhance alignment within the target domain.
- Leads to improved performance and accuracy in the transformed space.

.

Closest ‘n’ Anchors to Training Set:

- Identify and use the closest set of n anchors from a given larger set to the training set.

Most Different ‘n’ Anchors from Anchor Set:

- Identify and use the most distinct n anchors from a larger set using methods like Fixed-Size Determinantal Point Processes (DPP).

Baseline and Evaluation Metrics

- **Baseline**
  - Concept: Train a model using only a high-resource language. No data or adaptations from the target (low-resource) language are used during training.
  - Evaluation: Post-training, the model is directly applied to the low-resource language.
- **Evaluation Metrics**
  - Accuracy: Percentage of correct predictions out of all instances (Used in Classification, NER, SA).
  - Precision: Percentage of correct positive predictions out of all positive predictions (Used in Classification, NER, SA).
  - Recall: Percentage of correct positive predictions out of all actual positive instances (Used in Classification tasks).
  - F1-Score: Harmonic mean of Precision and Recall (Used in Classification, NER, SA).
  - BLEU Score: Measures similarity between generated text and reference translations (Used in Text Generation).

Initial Results

Baseline: fine tuning the encoder on English dataset and testing on target language dataset

RR: the relative representation concept without any additional improvements

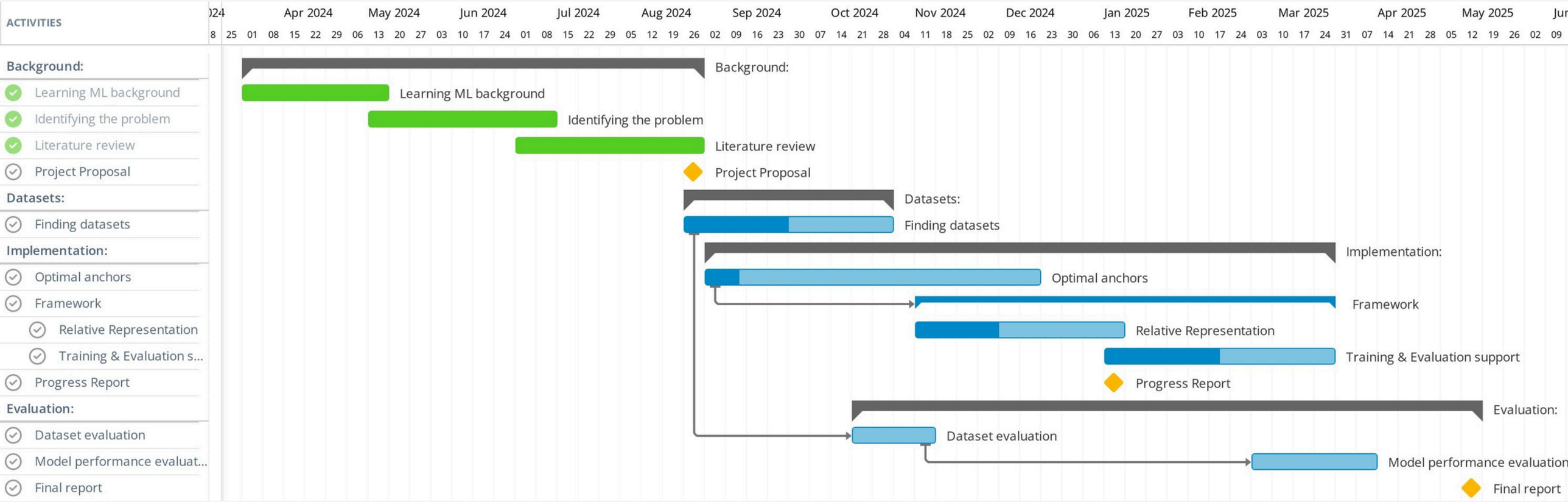
Improved RR: the relative representation concept with basis transformations

Experiment	News(F1)	Emotion(F1)	NER(F1)
Baseline	74.16	48.84	71.47
RR	74.75	51.43	71.6
Improved RR	**75.05**	**52.12**	**71.9**

COLLECTED DATASETS FOR VARIOUS TASKS AND LANGUAGES

Task	Language	Details
Offensive Language Detection	Sinhala	SOLD
	English	Hate Speech and Offensive Language Dataset
		TweetEval
		Hate Speech Data
News Category	Sinhala	NSINA
	English	News Category Dataset
		News Article Categories
	Bengali	Bangla Newspaper Dataset
	Tamil	Tamil News Dataset
Emotion Classification	Multi	XED: A Multilingual Dataset for Sentiment Analysis and Emotion Detection
	Indonesian	Indonesian Emotion Classification
	English	TweetEval
Word in Context Classification	Multi	XL-WiC: A Multilingual Benchmark for Evaluating Semantic Contextualization
Hate Speech Detection	English	Hate Speech and Offensive Language Dataset
	Sinhala	Sinhala Unicode Hate Speech Dataset
	Bengali	Bengali Hate Speech Dataset
Movie Review	Tamil	Tamil Movie Reviews

TIMELINE


ACTIVITIES
Apr 2024
May 2024
Jun 2024
Jul 2024
Aug 2024
Sep 2024
Oct 2024
Nov 2024
Dec 2024
Jan 2025
Feb 2025
Mar 2025
Apr 2025
May 2025
Background:
Learning ML background
Identifying the problem
Literature review
Project Proposal
Datasets:
Finding datasets
Implementation:
Optimal anchors
Framework
Relative Representation
Training & Evaluation s...
Progress Report
Evaluation:
Dataset evaluation
Model performance evaluat...
Final report
Training & Evaluation support
Model performance evaluation

CONCLUSION

- **Prior Work**: Focused on improving isomorphism in multilingual vector spaces but didn't explore a shared, language-agnostic space.
- **Proposed Approach**: Build on latent space transformation to create a common language-agnostic space.
- **Initial Results**: Promising, with ongoing work to optimize anchor selection for the best representation.

THANK YOU

Q & A