Project Outline: Training and Analyzing Models for Cross-Language Bias and Knowledge Transfer

1. Model Training Setup

- **Model 1: Canadian Hansard Model (Excluding Nunavut Proceedings)**
  - **Training Data**: Full corpus of the **Canadian Hansard**, excluding any content related to Nunavut, covering a similar time span to the **Nunavut Hansard** for direct comparison.
  - **Objective**: Establish the baseline for English-language embeddings and analyze how Canadian legislative language reflects cultural and political norms.
- **Model 2: Nunavut Hansard Model**
  - **Training Data**: Full corpus of the **Nunavut Hansard** to create embeddings reflective of legislative discussions that include Indigenous perspectives and cultural nuances.
  - **Objective**: Understand how the Inuktitut language represents cultural and political terms when trained independently.
- **Model 3: Multilingual Model (Canadian Hansard + Underrepresented Nunavut Hansard)**
  - **Training Data**: Combine the **Canadian Hansard** (overrepresented) with the **Nunavut Hansard** (underrepresented), sampling aligned spans of time to maintain temporal consistency.
  - **Objective**: Investigate cross-linguistic influences, particularly how biases in the overrepresented English dataset might affect embeddings and language representation in the underrepresented Inuktitut corpus.

2. Model Architecture and Hyperparameters

- **Consistent Architecture**: Use a transformer architecture like **BERT** or a custom implementation with the same number of layers, attention heads, and hidden dimensions for all models.
- **Hyperparameters**:
  - **Learning Rate**: Start with a common rate, such as `2e-5`.
  - **Batch Size**: Maintain consistency (e.g., `32`).
  - **Training Epochs**: Train each model for an equivalent number of epochs (e.g., `10`), with checkpoints to monitor overfitting.
- **Tokenizer Customization**:
  - Ensure the tokenizer can handle the unique morphological structure of **Inuktitut** alongside English.

3. Data Preprocessing

- **Text Cleaning and Tokenization**:
  - Uniformly clean and tokenize text from all corpora to standardize training inputs.
  - Handle unique characters and structures for Inuktitut, especially if syllabic writing is present.
- **Time Span Sampling**:
  - Sample data from equivalent periods across both corpora to match the temporal scope and ensure alignment in content and context.

4. Training Process

- **Train Each Model**:
  - Train each model separately using the same infrastructure (e.g., GPUs/TPUs) for consistency.
  - Implement **early stopping** or validation checks to avoid overfitting.
- **Multilingual Training Strategy**:
  - For the **multilingual model**, set proportions to ensure that the **Canadian Hansard** is overrepresented (e.g., 80%) while maintaining a 20% representation of the **Nunavut Hansard**.

5. Embedding Analysis

- **Word and Phrase Selection**:
  - Select culturally significant words and phrases (e.g., “leader,” “community,” “tradition,” “elder”) to extract embeddings.
- **Embedding Space Comparison**:
  - Use **cosine similarity** to compare embeddings of the selected words across the models.
  - Apply **t-SNE** or **PCA** for visualization to identify clustering patterns and semantic shifts.

6. Cross-Language Bias and Knowledge Transfer Evaluation

- **Bias Transfer Indicators**:

  - Analyze whether words associated with biases in the **Canadian Hansard model** shift in representation when compared to the **Nunavut Hansard model**.
  - Check if embeddings for terms that may carry Western-centric or gendered biases in English maintain those associations in Inuktitut within the **multilingual model**.

- **Semantic Integrity**:

  - Assess whether culturally specific terms in the **Nunavut Hansard model** maintain their positions in the **multilingual model** or shift closer to English-biased terms.

7. Visualization and Quantitative Analysis

- **Clustering and Embedding Overlaps**:
  - Visualize embeddings to show relationships between culturally important terms in each model.
- **Embedding Distance Metrics**:
  - Quantify shifts using **cosine distance** or **Euclidean distance** between word embeddings to evaluate how much influence the overrepresented corpus exerts on the underrepresented one.

8. Interpretation and Reporting

- **Highlight Key Findings**:
  - Identify terms that exhibit significant shifts, reflecting possible bias transfer or loss of cultural nuance.
- **Contextual Analysis**:
  - Discuss the implications of these findings in terms of how LLMs trained on dominant languages can distort or align with underrepresented language representations.

9. Recommendations for Mitigation

- **Balanced Training Suggestions**:
  - Provide recommendations for how to better balance training data to minimize bias transfer, such as through **data augmentation** or **fine-tuning** on culturally enriched corpora.
- **Cultural Preservation Strategies**:
  - Suggest ways to preserve cultural integrity, such as training with a focus on **culturally relevant data** and **context-aware learning**.

Timeline (6-7 Weeks)

- **Week 1**: Preprocess and align data; set up training environment.
- **Weeks 2-3**: Train each model (use parallel training where possible).
- **Week 4**: Extract embeddings and conduct initial comparisons.
- **Weeks 5-6**: Perform in-depth analysis, visualization, and quantitative evaluations.
- **Week 7**: Finalize analysis, document findings, and prepare for presentation.