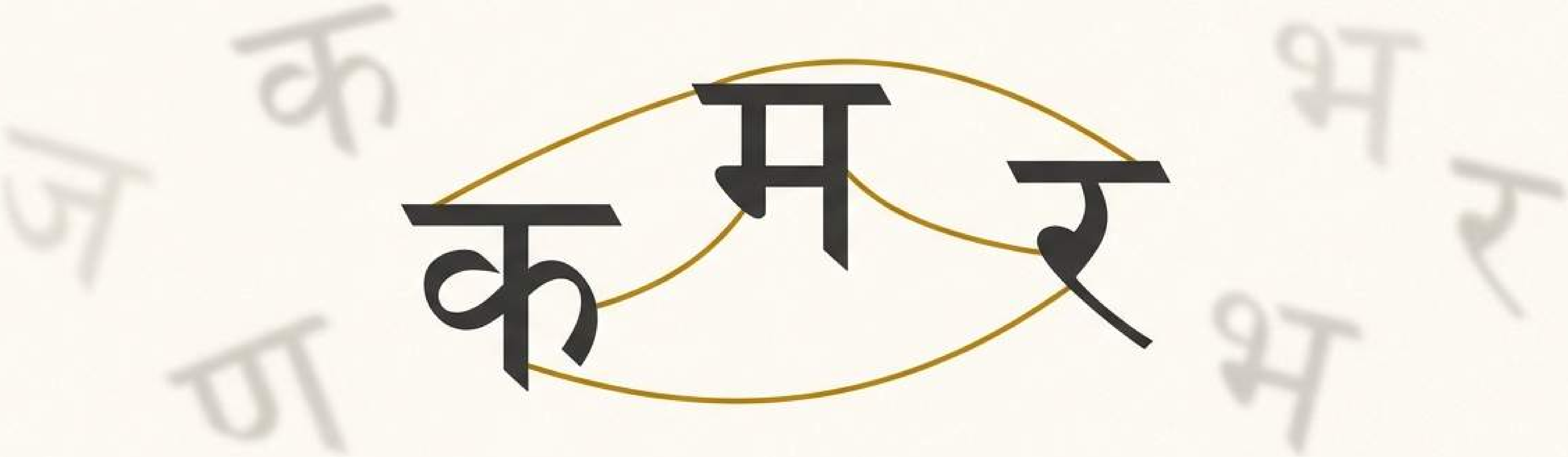
BHASHA SHARED TASK 2: INDIC WORD GROUPING

Team Horizon's 1st Place Approach at BHASHA Task 2



Manav Dhamecha, Gaurav Damor, Sunil Choudhary, Pruthwik Mishra

The Challenge: Identifying Semantic Units in Free Word Order Languages

1. **What are Local Word Groups (LWGs)?**

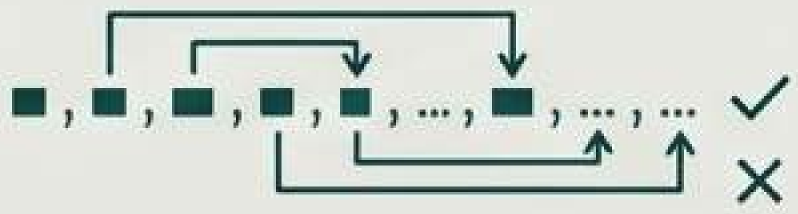
   LWGs are "semantically cohesive units consisting of a sequence of words that convey a single and complete meaning." Deeply rooted in the Indian grammatical tradition (Panini), they include noun compounds, verb groups with auxiliaries, and light verb constructions.

2. **Why is it Difficult?**

   The core problem stems from the free word order nature of most Indian languages. While the order of words *within* a group is fixed, the groups themselves can move freely in a sentence, creating significant ambiguity for computational models.

3. **The Competitive Benchmark**

   The BHASHA Task 2 requires systems to join tokens into correct word groupings. The evaluation metric is a demanding Exact Match Accuracy, where a prediction is correct only if the *entire* grouped sentence perfectly matches the gold standard.



From a Sentence to Grouped Semantic Units

INPUT SENTENCE

भारतीय भाषाओं में स्थानीय शब्द समूह मिलते हैं

The goal is to insert '_' separators between words that belong to the same Local Word Group, creating a single token for each semantic unit.

DESIRED OUTPUT

[भारतीय__भाषाओं__में] [स्थानीय__शब्द__समूह] [मिलते__हैं]

Our Approach: Reframing Word Grouping as Token Classification

Instead of complex rule-based systems, we modeled the task as a sequence labeling problem. Each token in a sentence is classified into one of three categories.

[word1]	[word2]	[word3]	[word4]
B - Begin	I - Inside	O - Outside	B - Begin
Marks the beginning of a multi-word group.	Marks a token inside a multi-word group.	Marks a token that is not part of any group (a delimiter).	

This simple annotation allows powerful Transformer models to learn the grouping patterns directly from data.

A Simple and Reproducible Fine-tuning Pipeline

Models Evaluated

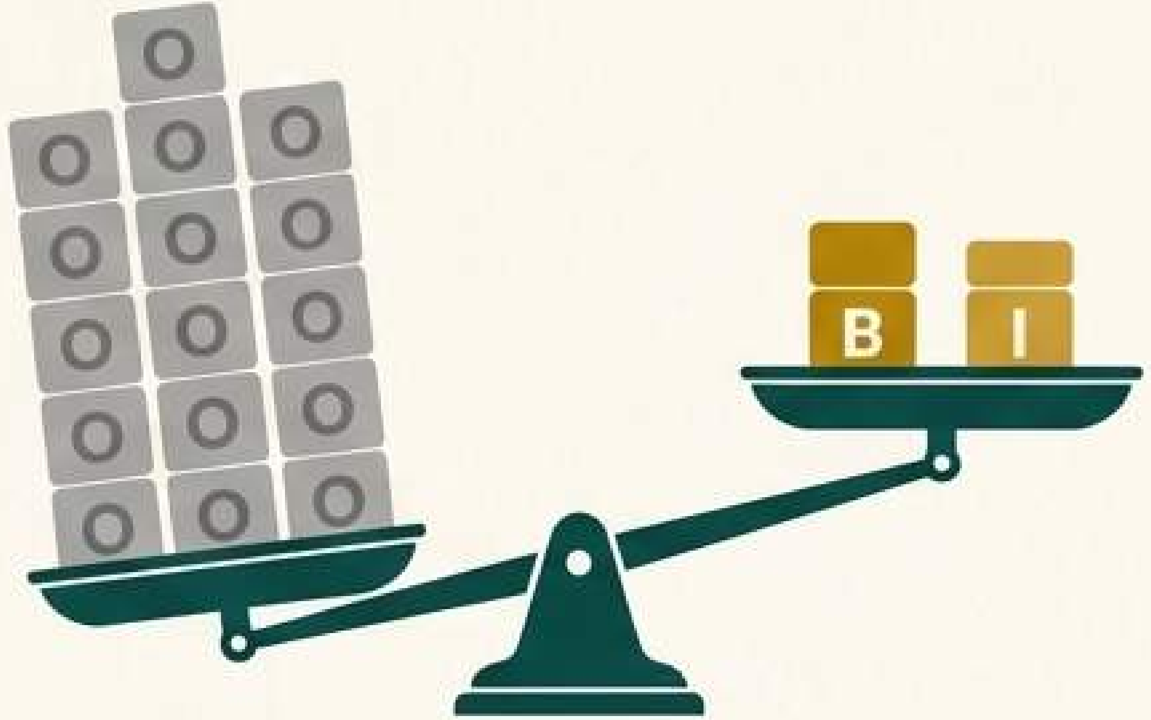
- **MuRIL**: Strong coverage for Indian languages. (Our eventual champion)
- **XLM-Roberta**: A general-purpose multilingual encoder.
- **IndicBERT v2**: An Indic-specific pretrained model.

The entire pipeline was built using the HuggingFace `AutoModelForTokenClassification` framework, ensuring easy replication.

The Key Refinement: Using Weighted Loss to Address Class Imbalance

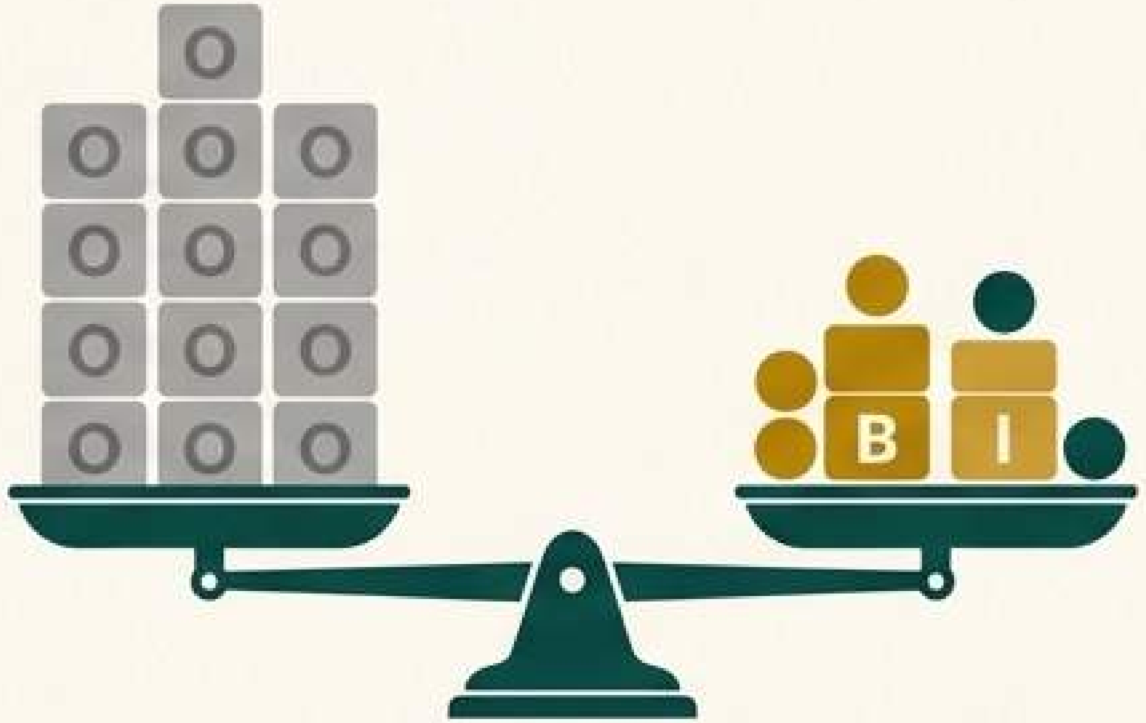
The Problem: The ‘O’ Label Bias

In word-grouping datasets, most tokens are delimiters, corresponding to the ‘O” (Outside) label. This creates a significant class imbalance, biasing the model towards predicting ‘O’ for all tokens.



The Solution: Class-Weighted Cross-Entropy

We calculated simple **inverse-frequency weights** from the training data. This technique slightly **upweights** the loss for the minority ‘B’ and ‘I’ labels during training, forcing the model to pay more attention to them.



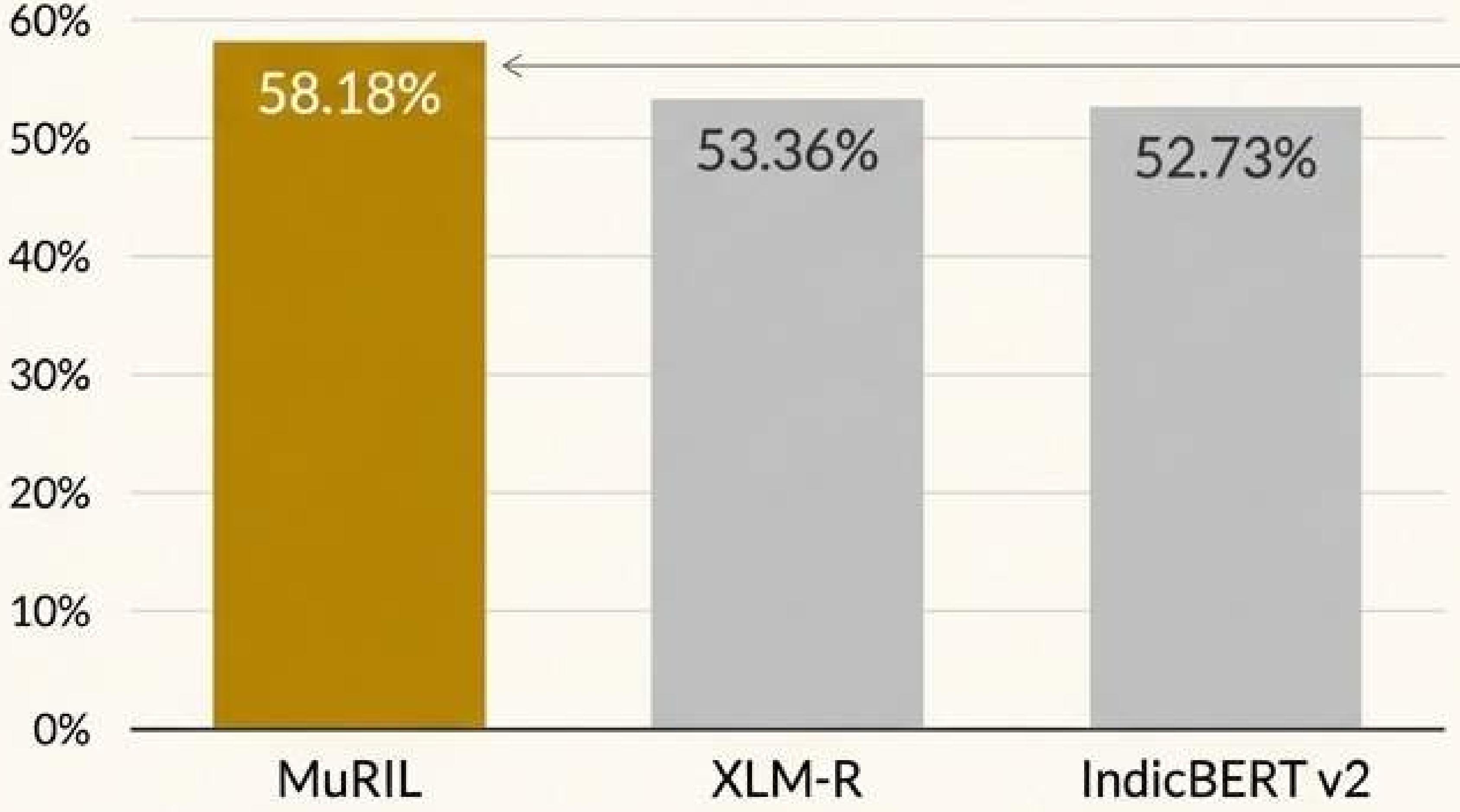
This simple change empirically improved token recall for B/I labels and delivered a **1-2% absolute increase** in Exact Match accuracy.

Impact: A 1st Place Finish with 58.18% Exact Match Accuracy

1st Place and 58.18%

among all participating teams in BHASHA Task 2.

Final Model Performance (Test Set Exact Match %)



MuRIL's superior performance is likely due to its targeted pretraining on Indian languages and a cased vocabulary that better preserves morphemic cues.

The Power of Refinement: From Official Submission to Final Result

Official Challenge Submission

45.13%

+13% Improvement

Post-Challenge Refined Model

58.18%

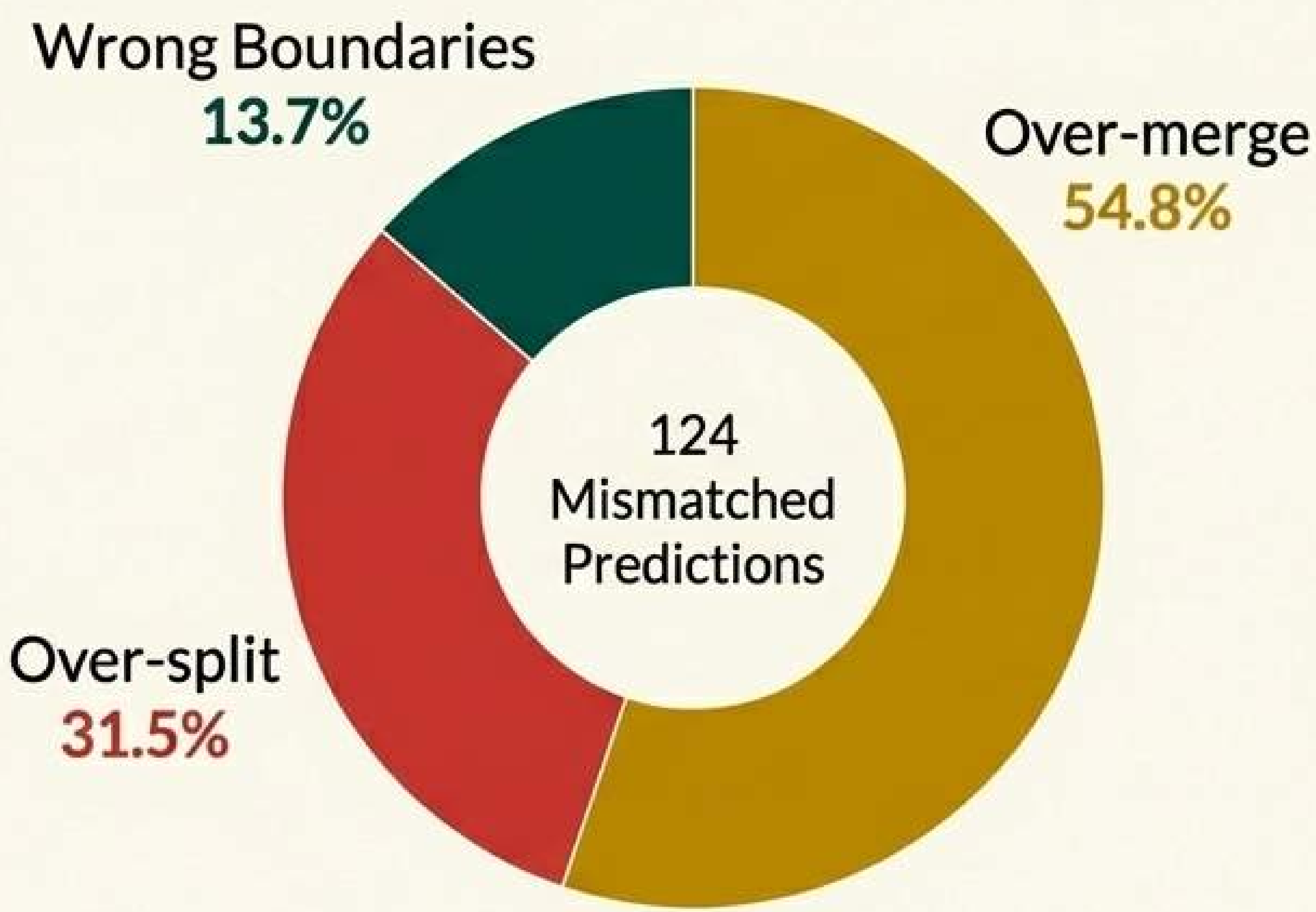
What Drove the Improvement?

1. Systematic implementation of the **class-weighted loss function.**
2. Improved logic for **boundary reconstruction and cleanup** during decoding.

This demonstrates the impact of meticulous tuning and analysis.

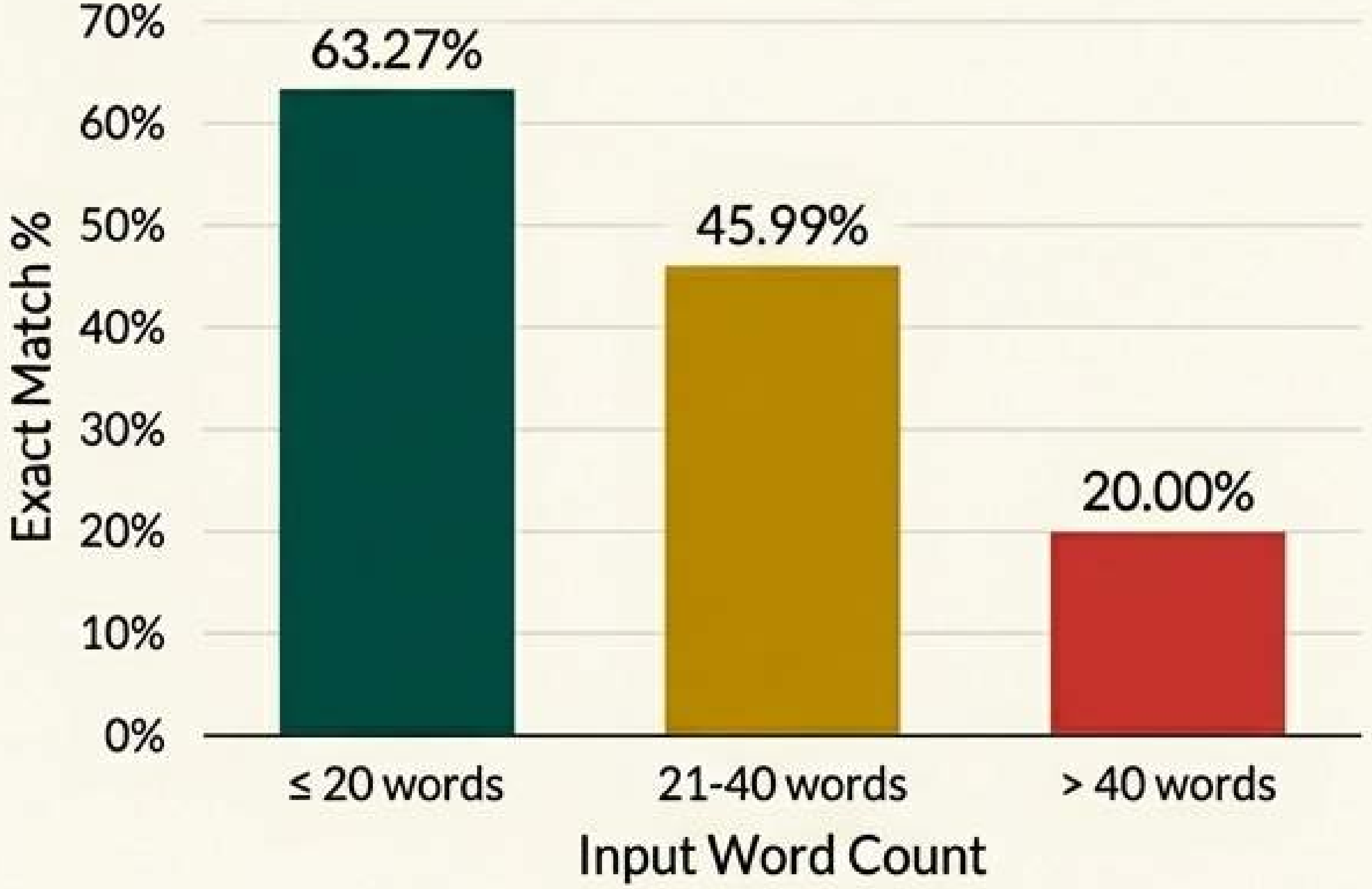
Where the Model Excels and Where It Struggles: An Error Analysis

Common Error Types (Test Set)



Sensitivity to Sentence Length

The model's accuracy degrades significantly as sentences get longer.



The model also struggles with long compounds, multi-word expressions, and annotation inconsistencies present in the gold data.

Contributions and Future Directions

Key Takeaways / Contributions

- **A Simple Pipeline Works**
  A straightforward BIO token-classification framework is highly effective for this task.
- **Class Weighting is Critical**
  Mitigating the 'O' label bias is essential for achieving top performance.
- **MuRIL is Superior**
  The model's targeted pretraining on Indic languages provides a distinct advantage.

Future Work

- **Ensembles**
  Combine predictions from MuRIL and XLM-R through voting or reranking.
- **Sequence-to-Sequence Models**
  Explore architectures like mT5 that directly generate grouped output, potentially avoiding subword alignment issues.
- **Larger Models / Adapters**
  Improve generalization on long sentences and rare compounds.