# Bilingual Tabular Inference: A Case Study on Indic Languages

https://enindicinfotabs.github.io


Chaitanya Agarwal[1*], Vivek Gupta[2*†],
Anoop Kunchukuttan[3], Manish Shrivastava[1]

[1]LTRC, IIIT Hyderabad; [2]University of Utah;
[3]AI4Bharat; [3]Microsoft Research



[†]On Academic Job Market

Slides borrowed from Chaitanya

# TABULAR INFERENCE (TNLI)

- The **tabular natural language inference** problem is similar to standard NLI.
- But here, the **premises are tabular data**.
- Task: Given a premise table, decide whether a hypothesis is **true (entailment)**, **false (contradiction)** or **undetermined (neutral)**.
- **Can be converted to the standard NLI task by linearizing tables.**

| Joe Strummer | |
|---|---|
| **Birth Name** | John Graham Mellor |
| **Born** | 1952-08-21 Ankara, Turkey |
| **Died** | 2002-12-22 Broomfield, Somerset, England |
| **Genres** | Punk Rock, Post Punk |
| **Occupation(s)** | Musician, Songwriter, Radio Host, Actor |
| **Instruments** | Vocals, Guitar, Piano |
| **Years Active** | 1970-2002 |
| **Labels** | CBS, Sony, Hellcat, Mercury |
| **Associated Acts** | The 101ers, The Clash |

**H1: John Graham Mellor plays less instruments than the number of labels he has worked for.**

**H2: Joe Strummer changed his surname after he became a guitar player.**

**H3: Joe Strummer was active in the sports industry for over three decades.**

Check out INFOTABS (Gupta et al., 2020)
https://infotabs.github.io

# MOTIVATION

- **Knowledge Issues:** Tabular Data / Knowledge Base is mostly present in High Resource Languages like English. High disparity across language in Wikipedia.

- **Model Issues:** Translating tabular data while maintaining the intent, context, and the same quality of the source English language is difficult. Latency and Throughput Issues.

- **Research Aspects:** However, making inferences in a low resource language over the English tabular source is a practical requirement.

# MOTIVATION

- Tabular Data / Knowledge Base is mostly present in High Resource Languages like English.
- Translating tabular data while maintaining the intent, context, and the same quality of the source English language is difficult.
- However, making inferences in a low resource language over the English tabular source is a practical requirement.
- **Can we construct a TNLI task in a bilingual setting?**

# BILINGUAL TABULAR INFERENCE (bTNLI)

- A novel TNLI task wherein, the **tabular premise** is in a **high resource language**, while the **textual hypothesis** is in a **low resource language**.

**PREMISE**

| Joe Strummer | |
|---|---|
| **Birth Name** | John Graham Mellor |
| **Born** | 1952-08-21 Ankara, Turkey |
| **Died** | 2002-12-22 Broomfield, Somerset, England |
| **Genres** | Punk Rock, Post Punk |
| **Occupation(s)** | Musician, Songwriter, Radio Host, Actor |
| **Instruments** | Vocals, Guitar, Piano |
| **Years Active** | 1970-2002 |
| **Labels** | CBS, Sony, Hellcat, Mercury |
| **Associated Acts** | The 101ers, The Clash |

H1: John Graham Mellor plays less instruments than the number of labels he has worked for.

**HYPOTHESIS**

H1_hindi: जॉन ग्राहम मेलर उन लेबलों की संख्या की तुलना में कम वाद्य बजाते हैं जिनके लिए उन्होंने काम किया है।

# BILINGUAL TABULAR INFERENCE (bTNLI)

- A novel TNLI task wherein, the **tabular premise is in a high resource language**, while the textual hypothesis is in a **low resource language**.

**QUESTIONS?**

- How can we create a dataset that can evaluate bTNLI?

PREMISE

| Joe Strummer | |
|---|---|
| **Birth Name** | John Graham Mellor |
| **Born** | 1952-08-21 Ankara, Turkey |
| **Died** | 2002-12-22 Broomfield, Somerset, England |
| **Genres** | Punk Rock, Post Punk |
| **Occupation(s)** | Musician, Songwriter, Radio Host, Actor |
| **Instruments** | Vocals, Guitar, Piano |
| **Years Active** | 1970-2002 |
| **Labels** | CBS, Sony, Hellcat, Mercury |
| **Associated Acts** | The 101ers, The Clash |

H1: John Graham Mellor plays less instruments than the number of labels he has worked for.

HYPOTHESIS

H1_hindi: जॉन ग्राहम मेलर उन लेबलों की संख्या की तुलना में कम वाद्य बजाते हैं जिनके लिए उन्होंने काम किया है।

# BILINGUAL TABULAR INFERENCE (bTNLI)

- A novel TNLI task wherein, the **tabular premise is in a high resource language**, while the **textual hypothesis is in a low resource language**.

**QUESTIONS?**

- How can we create a dataset that can evaluate bTNLI?
- How well can multilingual models (for example, mBERT) reason about bTNLI?

PREMISE

| Joe Strummer | |
|---|---|
| **Birth Name** | John Graham Mellor |
| **Born** | 1952-08-21 Ankara, Turkey |
| **Died** | 2002-12-22 Broomfield, Somerset, England |
| **Genres** | Punk Rock, Post Punk |
| **Occupation(s)** | Musician, Songwriter, Radio Host, Actor |
| **Instruments** | Vocals, Guitar, Piano |
| **Years Active** | 1970-2002 |
| **Labels** | CBS, Sony, Hellcat, Mercury |
| **Associated Acts** | The 101ers, The Clash |

H1: John Graham Mellor plays less instruments than the number of labels he has worked for.

HYPOTHESIS

H1_hindi: जॉन ग्राहम मेलर उन लेबलों की संख्या की तुलना में कम वाद्य बजाते हैं जिनके लिए उन्होंने काम किया है।

# OUR CONTRIBUTIONS

- We introduce the novel task of bilingual tabular inference (bTNLI).
- EI-INFOTABS, a dataset for bilingual tabular inference which contains hypotheses in 11 Indian languages, while retaining the English tabular premises from the INFOTABS dataset.
- To create EI-INFOTABS, we leverage cutting-edge machine translation models which provide high-quality translations of the hypotheses.
- We access reasoning ability of state-of-the-art multilingual models trained with varying strategies over EI-INFOTABS.

# EXAMPLE FROM EI-INFOTABS

| Joe Strummer | |
|---|---|
| **Birth Name** | John Graham Mellor |
| **Born** | 1952-08-21 Ankara, Turkey |
| **Died** | 2002-12-22 Broomfield, Somerset, England |
| **Genres** | Punk Rock, Post Punk |
| **Occupation(s)** | Musician, Songwriter, Radio Host, Actor |
| **Instruments** | Vocals, Guitar, Piano |
| **Years Active** | 1970-2002 |
| **Labels** | CBS, Sony, Hellcat, Mercury |
| **Associated Acts** | The 101ers, The Clash |

English Tabular Premise

| Language | Hypothesis | Label |
|---|---|---|
| **Hindi** | जॉन ग्राहम मेलर उन लेबलों की संख्या की तुलना में कम वाद्य बजाते हैं जिनके लिए उन्होंने काम किया है। | ENTAIL |
| **Telugu** | జోయ్ స్ట్రమ్మర్ తాను జన్మించిన దేశంలోనే మరణించాడు. | CONTRADICT |
| **Assamese** | জো ষ্ট্ৰামাৰে তেওঁৰ জীৱনকালত বহুতো পুৰস্কাৰ লাভ কৰিছিল। | NEUTRAL |
| **Oriya** | ଜୋ ଷ୍ଟ୍ରୁମର ଜଣେ ଏକକ ଗୌଳୀର ରକ ସଙ୍ଗୀତକାର ଥିଲେ। | CONTRADICT |
| **Tamil** | ஜோ ஸ்ட்ரம்மர் 2002 முதல் தீவிரமாக செயல்பட்டு வந்தார். | ENTAIL |
| **Punjabi** | ਜੋ ਸੋਨੀ ਨਾਲ ਕੰਮ ਕਰਨ ਤੋਂ ਪਹਿਲਾਂ ਹੈਲਕੈਟ ਨਾਲ ਕੰਮ ਕਰਦਾ ਸੀ । | NEUTRAL |
| **Gujarati** | જ્હોન ગ્રેહામ મેલરે તેમના જીવનની શરૂઆતમાં તેમનું નામ બદલીને જો સ્ટ્રુમર રાખ્યું હતું. | NEUTRAL |

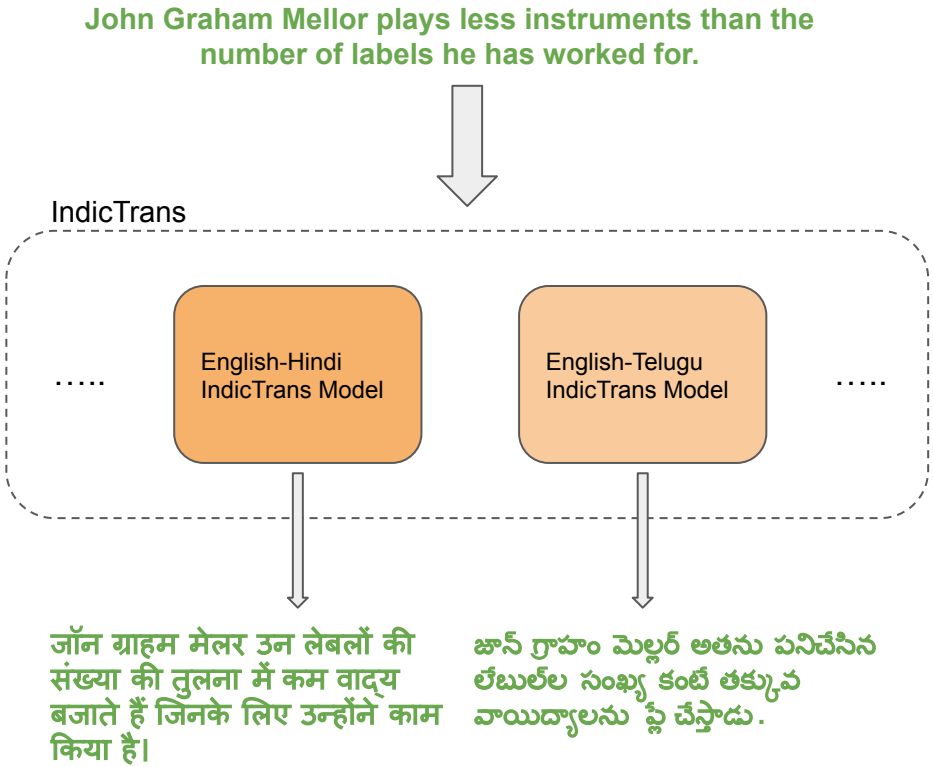
# CHALLENGES

- How does one **estimate quality of machine translation in an unsupervised manner and at scale**?

# TRANSLATING HYPOTHESES

We **machine translate** the English hypotheses provided in INFOTABS **to 11 major Indian languages**.

We use **IndicTrans**, an **open-sourced state-of-the-art Indic NMT model** which **outperforms commercial MT systems** like Google Translate, Bing Translate, etc.




Check out IndicTrans (Ramesh et al., 2022)
https://indicnlp.ai4bharat.org/indic-trans/

# MEASURING THE QUALITY OF TRANSLATIONS

## Automatic Evaluation

We use **BERTScore** (Zhang* et al., 2020)

An automatic scoring metric for sentence similarity, **between the source and back-translated English sentences**.

We use IndicTrans to generate Indic to English back-translated data.

# MEASURING THE QUALITY OF TRANSLATIONS

## Automatic Evaluation

We use **BERTScore** (Zhang* et al., 2020)

An automatic scoring metric for sentence similarity, **between the source and back-translated English sentences**.

We use IndicTrans to generate Indic to English back-translated data.

## Human Evaluation

We follow the guidelines recommended in (Agirre et al., 2016) to conduct human evaluation.

We **(a.) diversely sample** source-translation pairs in each language, (b.) **prepare a common Direct Assessment** (Graham et al., 2013) **scoring strategy**, and (c.) **get the sampled data evaluated** by language experts.

# MEASURING THE QUALITY OF TRANSLATIONS

## Automatic Evaluation

We use **BERTScore** (Zhang* et al., 2020)

An automatic scoring metric for sentence similarity, **between the source and back-translated English sentences**.

We use IndicTrans to generate Indic to English back-translated data.

## Human Evaluation

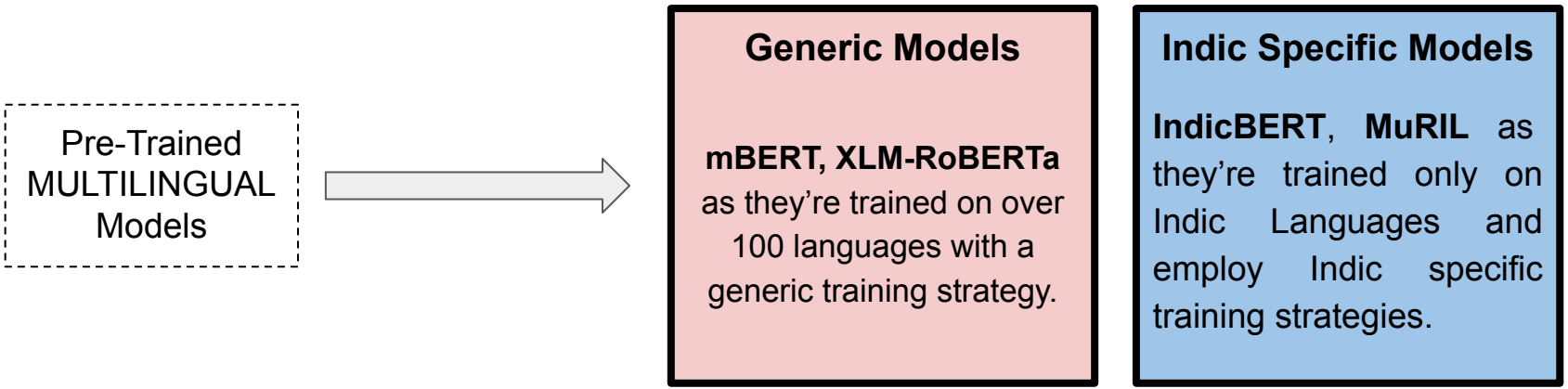
We follow the guidelines recommended in (Agirre et al., 2016) to conduct human evaluation.

We **(a.) diversely sample** source-translation pairs in each language, (b.) **prepare a common Direct Assessment** (Graham et al., 2013) **scoring strategy**, and (c.) **get the sampled data evaluated** by language experts.

We observe high quality scores and note that our machine translated dataset closely preserves the semantics of the source and is fluent.

# EXPERIMENTS

*How well do existing pre-trained multilingual language models perform on the bTNLI task?**

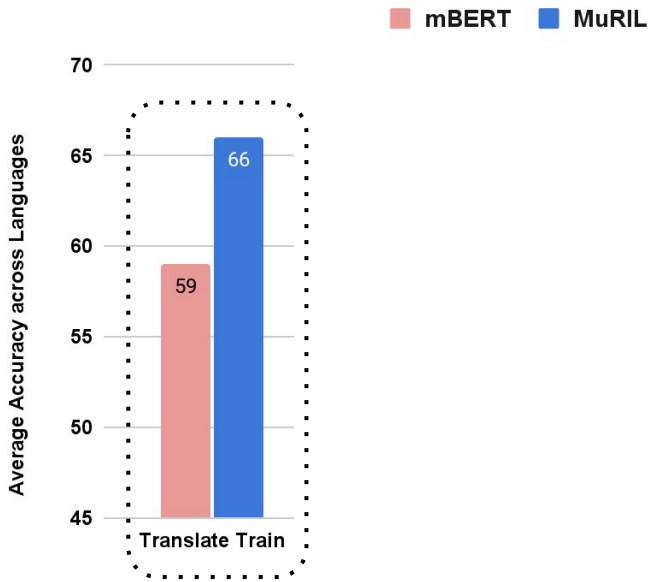



Check out IndicBERT (Kakwani et al., 2020)
https://github.com/AI4Bharat/indic-bert

Check out MuRIL (Khanuja et al., 2021)
https://arxiv.org/abs/2103.10730

*We linearize the tables into paragraphic representations

# RESULTS & ANALYSIS




Results on α1 test split

**Translate** **Train**

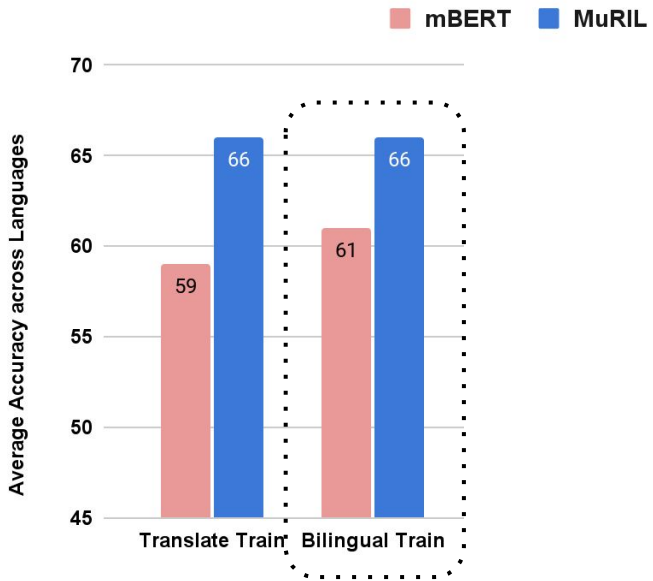
We fine-tune and evaluate the multilingual models on EN-$IN_i$ premise-hypothesis pairs where $IN_i$ is one of the 11 Indic languages.

Training and evaluation done on each language separately.

Each fine-tuned model is evaluated on the same Indic language it was fine-tuned on.

**We also conduct cross-lingual evaluation of these models*

# RESULTS & ANALYSIS




Results on α1 test split

## Bilingual Train

We fine-tune the multilingual models on both EN-EN and EN-$IN_i$ and evaluate on EN-$IN_i$ premise-hypothesis pairs.

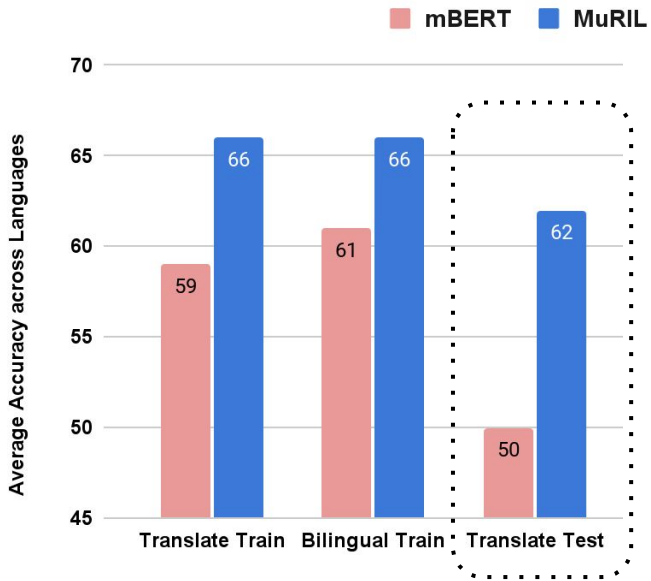
Training and evaluation done on each language separately.

Each fine-tuned model is evaluated on the same Indic language it was fine-tuned on.

Addition of English training data aids the performance of mBERT.

**We also conduct cross-lingual evaluation of these models*

# RESULTS & ANALYSIS




Results on α1 test split

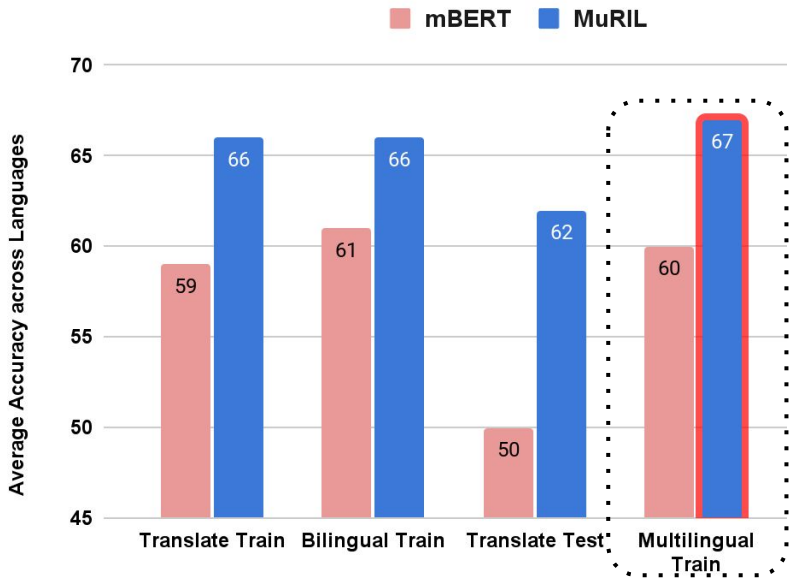
## Translate Test

We fine-tune the multilingual models on EN-EN and evaluate on EN-$IN_i$ premise-hypothesis pairs.

The fine-tuned model is then evaluated on each Indic language.

**Poor Zero Shot Cross Lingual Transfer from INFOTABS (English) to EI-INFOTABS (Indic).**

# RESULTS & ANALYSIS




Results on α1 test split

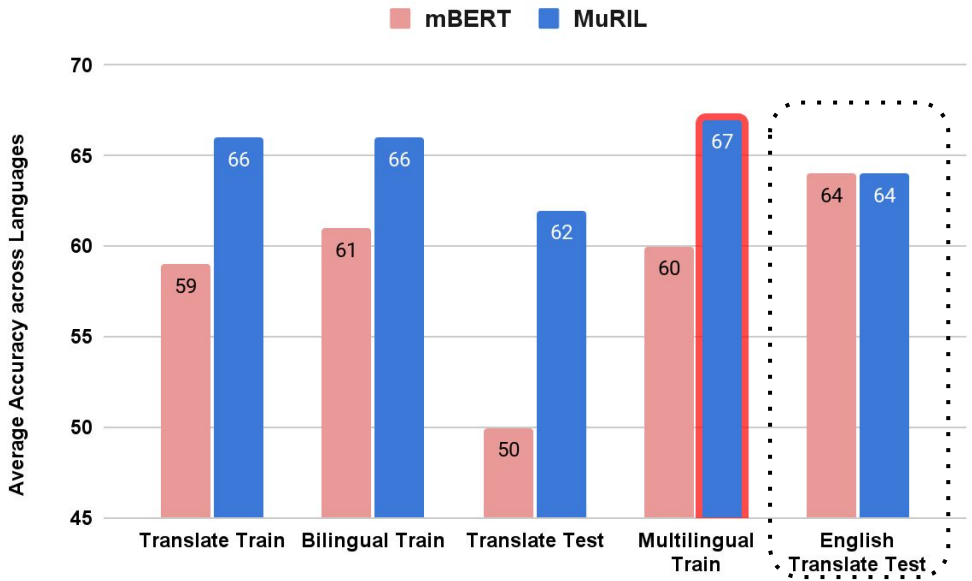
## Multilingual Train (Unified Model)

We fine-tune the multilingual models on all languages including English.

The fine-tuned model is then evaluated on each Indic language.

MuRIL performs best in this setting and forms the benchmark for this task.

# RESULTS & ANALYSIS




Results on α1 test split

## English Translate Test

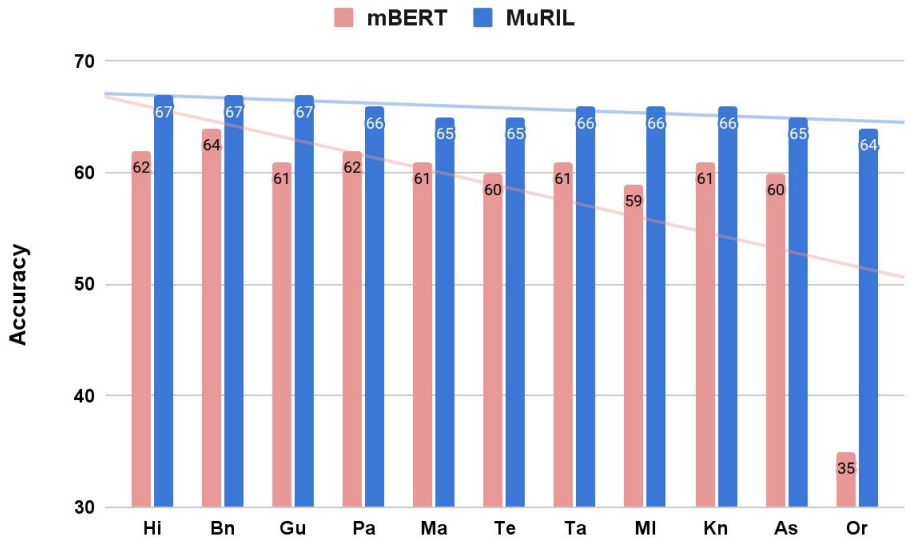
We fine-tune the multilingual models on EN-$IN_i$ premise-hypothesis pairs and evaluate on EN-$ENIN_i$ premise-hypothesis pairs where $ENIN_i$ represents $IN_i$ to EN back-translated hypotheses.

Training and evaluation done on each language separately.

Each fine-tuned model is evaluated on the same Indic language it was fine-tuned on.

Forms a strong Translate then Test baseline on EI-INFOTABS.

# LANGUAGE SPECIFIC PERFORMANCE



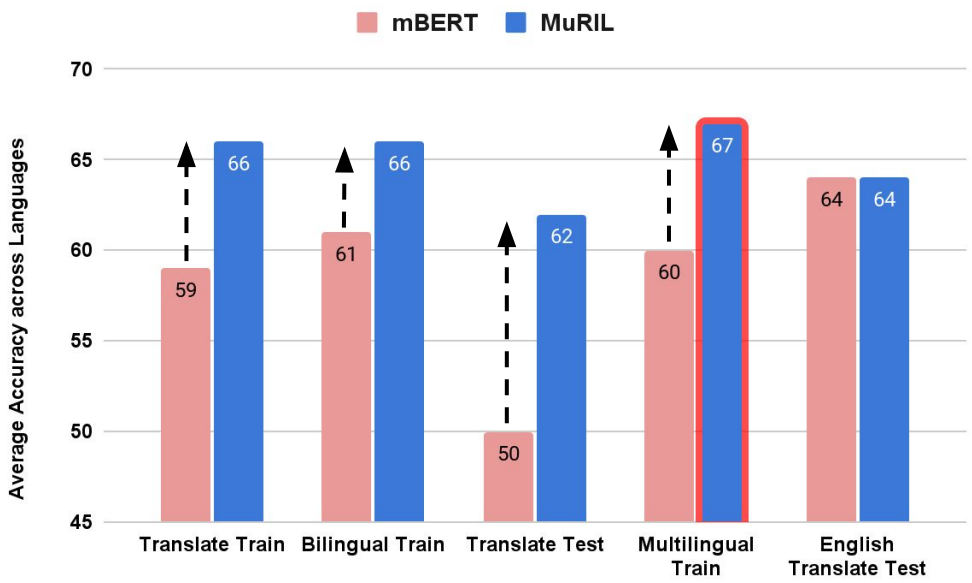

MuRIL's (Unified Model) results on α1 test split

Models perform best on Hindi(*hi*) and Bengali(*bn*). This is expected as they are high resource languages in the Indic context.

Assamese(*as*) and Oriya(*or*) are extremely low resource in the Indic context.

**Pre-training or fine-tuning on Bengali aids the performance on Assamese due to their high degree of relatedness.**

# MULTILINGUAL MODEL COMPARISON



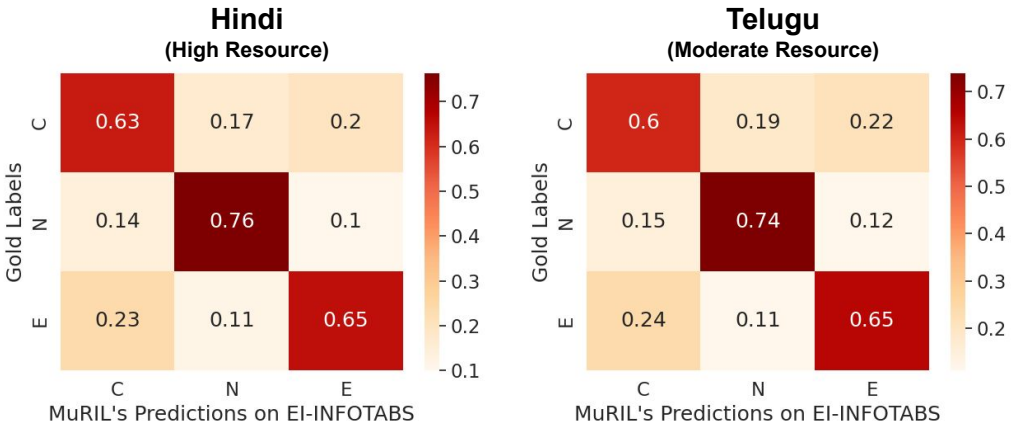

Results on α1 test split

MuRIL performs best across all languages and experiments beating IndicBERT and mBERT due to

- large parameter count
- Indic specific pre-training data
- Indic specific pre-training objectives (Khanuja et al., 2021)

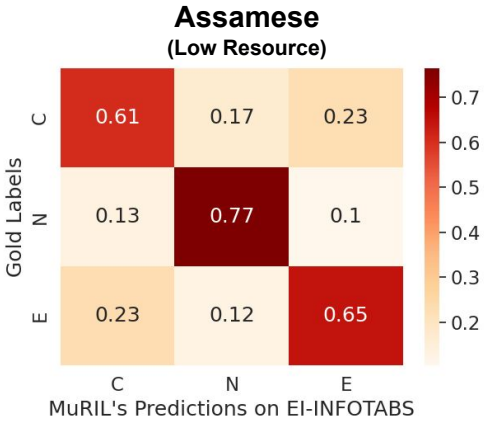
MuRIL is multilingually robust as it performs similarly across all experiments and languages.

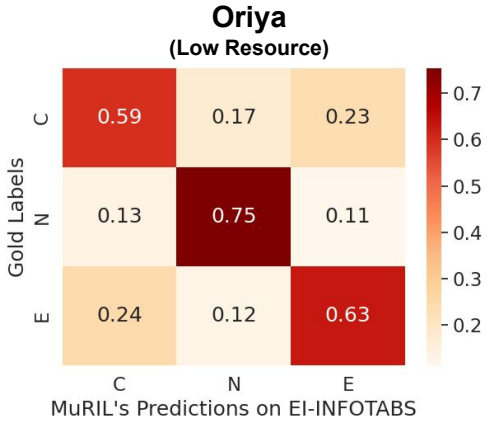
# CONFUSION MATRICES










MuRIL performs similarly across all languages irrespective of their degree of low resource-*edness* in the Indic context.

Neutral is predicted with the highest confidence.

MuRIL confuses Entailment with Contradiction and vice-versa.

As noted earlier, MuRIL is multilingually robust in the Indic context.

# EI-INFOTABS v/s INFOTABS

| Model (Rep) | Dev | $\alpha_1$ | $\alpha_2$ | $\alpha_3$ |
|---|---|---|---|---|
| $\text{BERT}_B$ (BPR) | 63.00 | 63.54 | 52.57 | 48.17 |
| $\text{RoBERTa}_B$ (TabFact) | 68.06 | 66.7 | 56.87 | 55.26 |
| $\text{RoBERTa}_L$ (BPR) | 76.42 | 75.29 | 66.50 | 64.26 |
| $\text{RoBERTa}_L$ (TabFact) | 77.61 | 75.06 | 69.02 | 64.61 |
| Human | **79.78** | **84.04** | **83.88** | **79.33** |

*

Table 4: The human benchmarks and several baselines on evaluation set of INFOTABS as reported in Gupta et al. (2020) (TabFact) and Neeraja et al. (2021) (BPR). Here subscript $X_L$ and $X_B$ represent X model L: Large and B: Base versions respectively.

**RoBERTa$_{LARGE}$ forms the benchmark on INFOTABS*

Baselines on EI-INFOTABS are within an **absolute margin of 10%** when compared to those on INFOTABS.

But EI-INFOTABS is a **more challenging dataset** than INFOTABS due to

1. bilinguality within the premise-hypothesis pair

2. low resource nature of Indic languages

Thus, baselines on EI-INFOTABS are strong and promising.

# RoBERTa v/s MuRIL

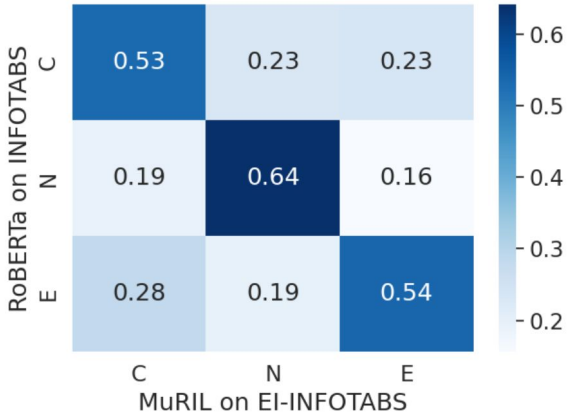



Figure 3: Consistency Matrix which measures the deviation of our best performing model, MuRIL (Multilingual-Train)'s predictions on the $\alpha_1$ test set for Hindi as compared to that of RoBERTa$_{\text{LARGE}}$ on the $\alpha_1$ test set of INFOTABS.

MuRIL disagrees with RoBERTa$_{\text{LARGE}}$ **on 47% of examples** with the Contradiction and Entailment labels.

However, for Neutral labels, it **only disagrees on around 36%** of the examples.

However, the models fine-tuned on EI- INFOTABS broadly mimic the performance of RoBERTa$_{\text{LARGE}}$ on INFOTABS.

Both models **predict Neutral hypotheses** with the **highest confidence.**

Both models **confuse Entailment with Contradiction** inference label and vice-versa.

# TAKEAWAYS

- EI-INFOTABS, a dataset for bilingual tabular inference which contains hypotheses in 11 Indian languages, while retaining the English tabular premises from the INFOTABS dataset.
- The dataset offers immense potential as it opens up avenues in (a) multilingual tabular NLI, (b) bilingual claim verification, (c) and evaluation of multilingual models.
- To create EI-INFOTABS,, we leverage cutting-edge machine translation models which provide high-quality translations of the hypotheses.
- We access reasoning ability of state-of-the-art multilingual models trained with varying strategies over EI-INFOTABS,.

enindicinfotabs.github.io