



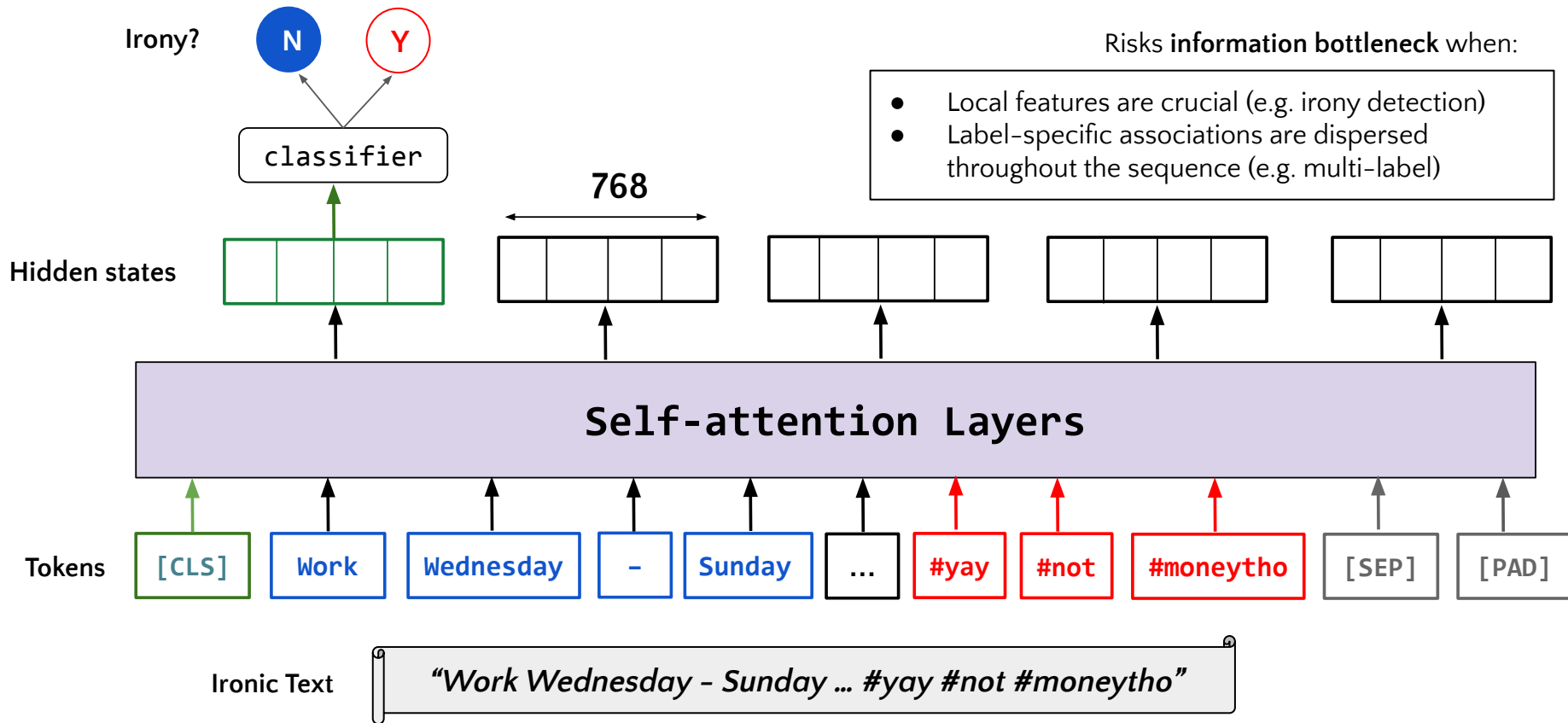
Inceptive Transformers: Enhancing Contextual Representations through Multi-Scale Feature Learning Across Domains and Languages

Asif Shahriar^{1,2}, Rifat Shahriyar¹, M Saifur Rahman¹

¹ Bangladesh University of Engineering and Technology, ² BRAC University
Proceedings of EMNLP 2025 (Long paper #2943)



Encoder and the [CLS] Token



A Step towards Solution

Problem

[CLS] representation is insufficient

Idea

Identify the **IMPORTANT** local features

Question

WHICH local features are **IMPORTANT**?

Inspiration

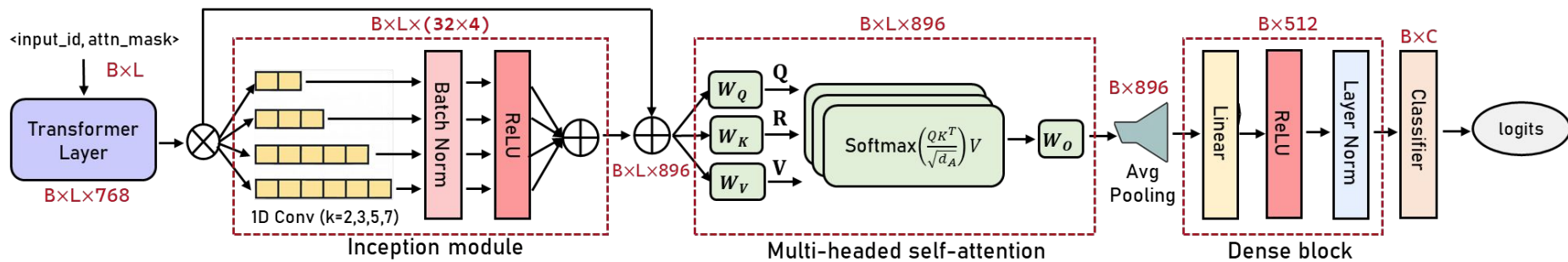
Going Deeper with Convolutions ^[1]

Approach

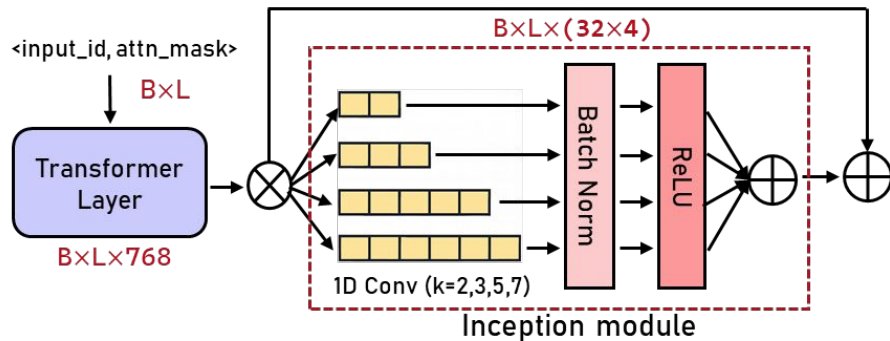
Capture features at **MULTIPLE** scales simultaneously

[1] Going Deeper with Convolutions, Szegedy et al., CVPR 2015

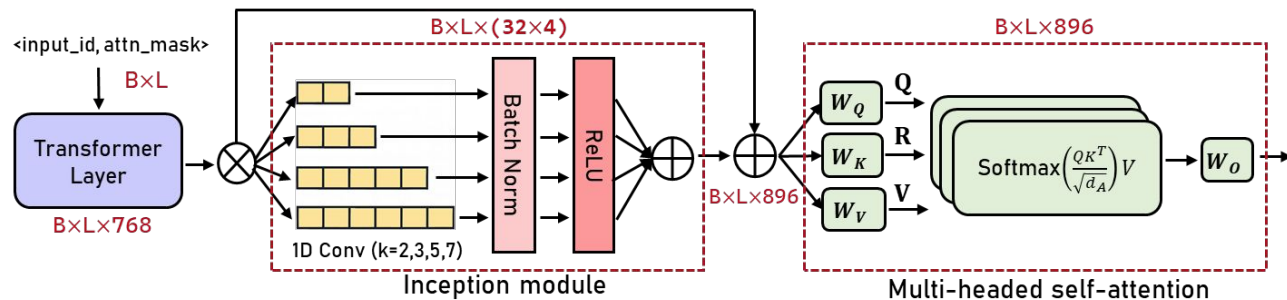
Inceptive Transformer Architecture



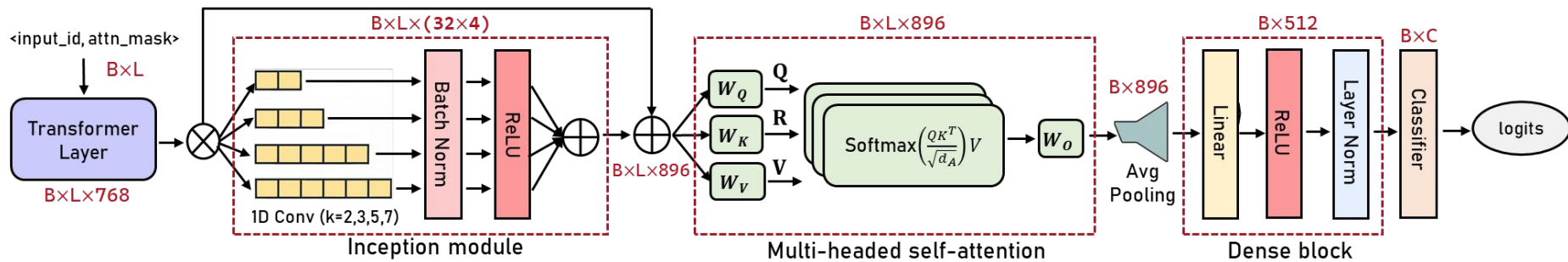
Component I: Inception Module



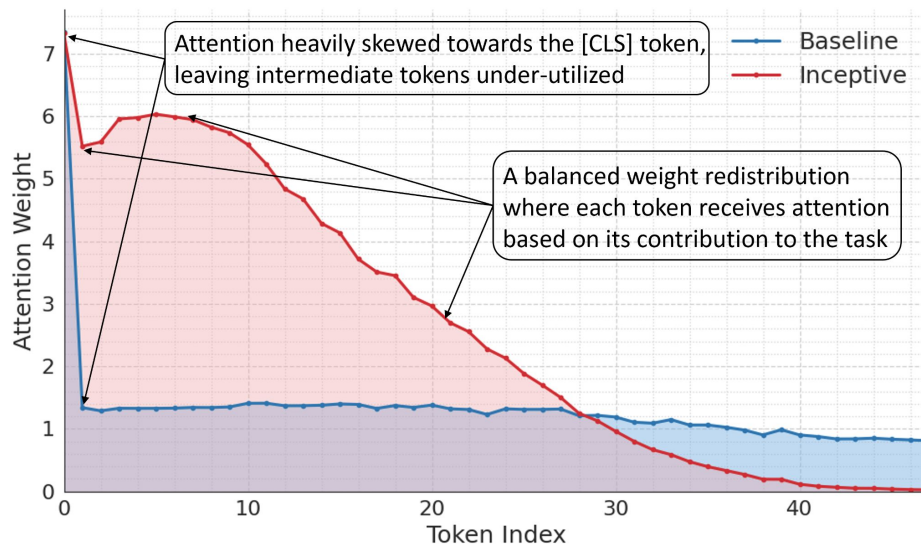
Component II: Self-attention Layer



Component III: Dense Block



Okay ... But HOW does Inception Help?



[CLS] Work Wednesday - Sunday ... #yay #not #moneytho [SEP] [PAD]

Datasets, Baselines & Evaluation

Task	Dataset	#Texts	#Classes	Avg Token Length	Metric
Irony Detection	TweetEval	4,061	2	22	Accuracy
Emotion Recognition	TweetEval	5,052	4	24	Accuracy
Bangla Emotion Recognition	Bangla-Banglish	80,098	6	18	Accuracy
Disease Identification	OHSUMED	13,929	23	290	F1-score
Anti-Covid Vaccine Concerns Classification	CAVES	9,921	11	58	F1-score

General Models

- RoBERTa
- DeBERTa v3
- ModernBERT
- XLM-RoBERTa

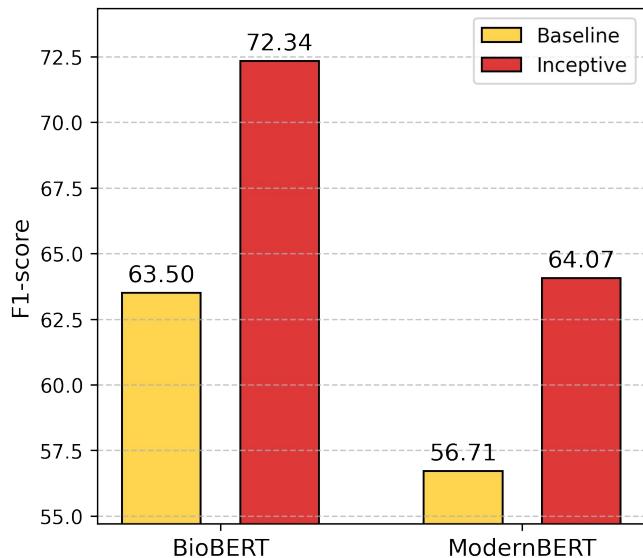
Domain-Specific Models

- BERTweet
- BioBERT
- CT-BERT
- BanglaBERT

Experimental Rigor

- Avg across 10 runs
- 10-fold CV
- Wilcoxon Rank Test

Results I: Disease Identification



Wilcoxon: $P = 0.00195$
(significant)

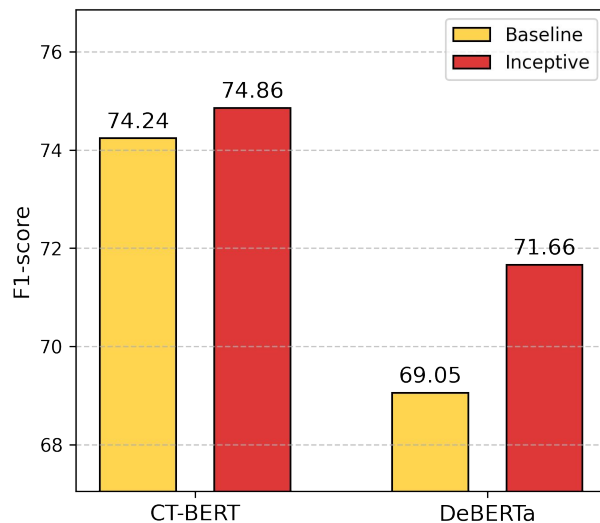
10-fold CV Std Dev:
Baseline: 1.35
Inceptive: 0.62

Avg across all models

Performance gain: **10.81%**

Inference Overhead: **7.35%**

Results II: Anti-Vaccine Concern Classification



Wilcoxon: $P = 0.00195$
(significant)

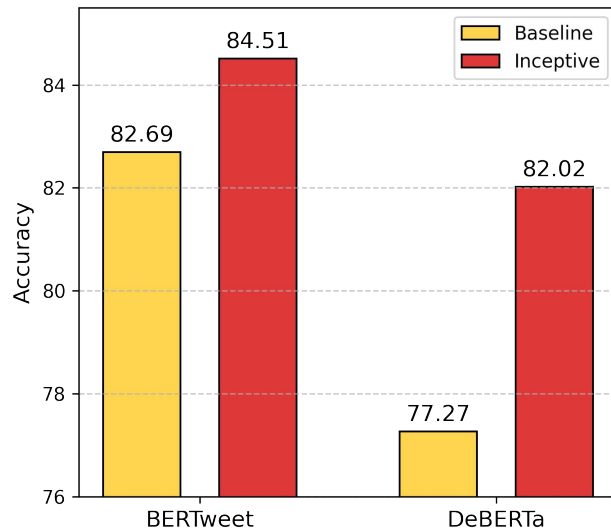
10-fold CV Std Dev:
Baseline: 0.94
Inceptive: 0.88

Avg across all models

Performance gain: **2.55%**

Inference Overhead: **2.23%**

Results III: Irony Detection



Wilcoxon: $P = 0.00585$
(significant)

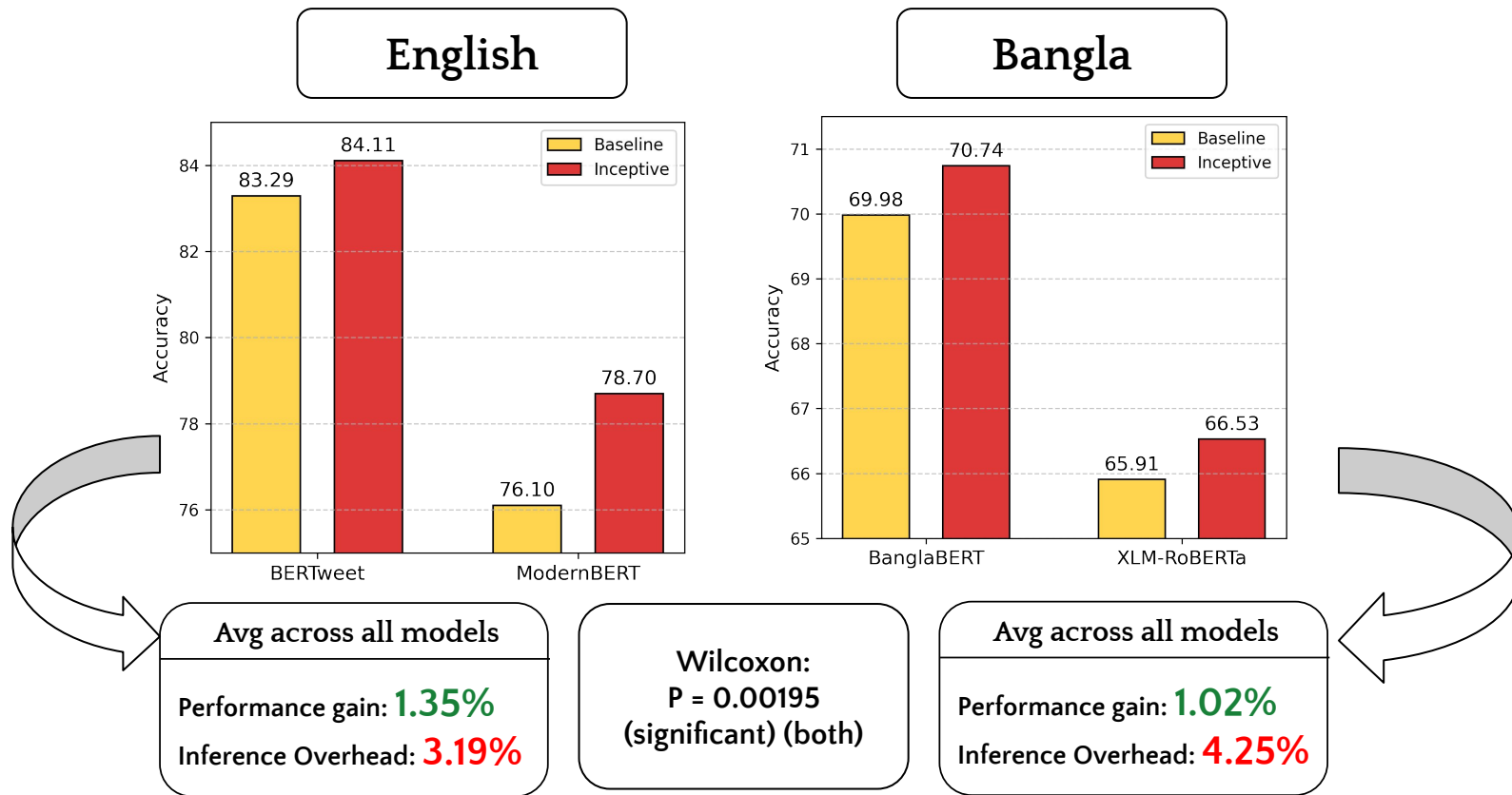
10-fold CV Std Dev:
Baseline: 1.20
Inceptive: 1.27

Avg across all models

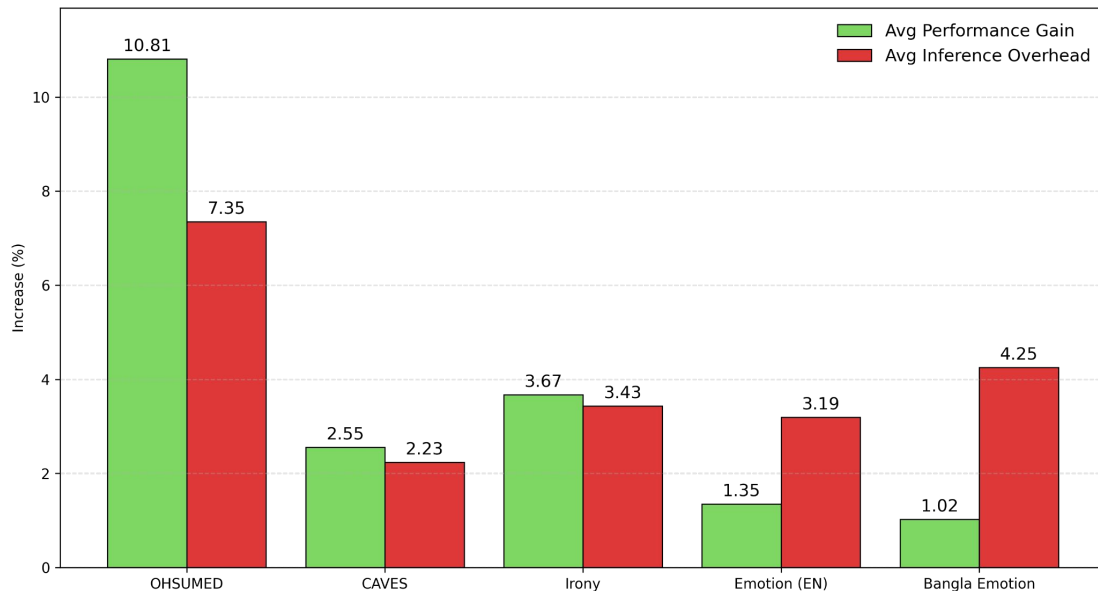
Performance gain: **3.67%**

Inference Overhead: **3.47%**

Results IV: Emotion Recognition



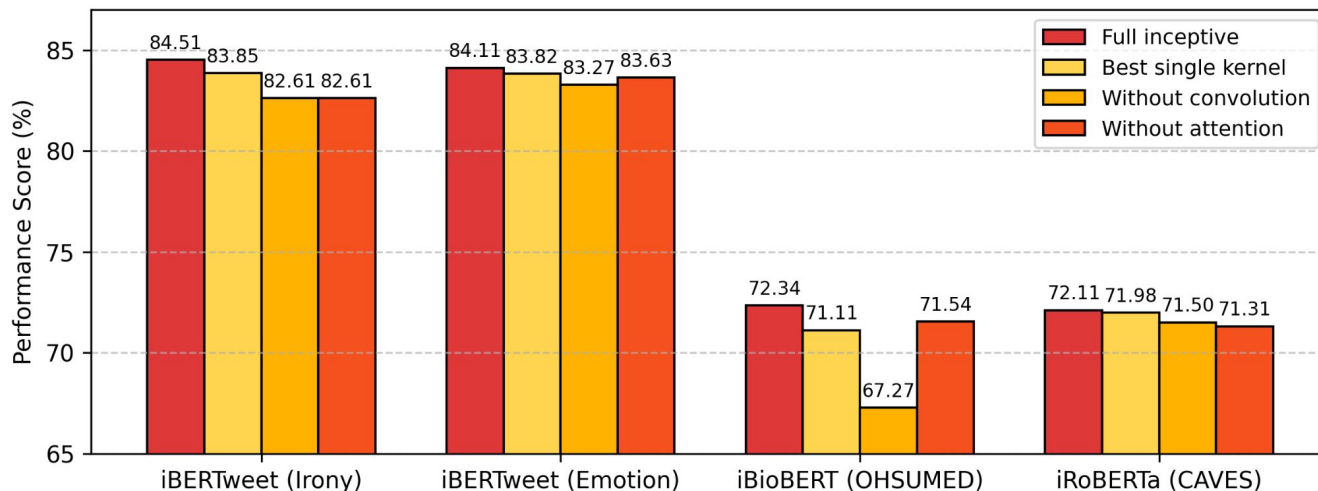
Results At A Glance



**Wilcoxon: $P < 0.05$
(significant)**

**10-fold CV Std Dev:
inceptive < baseline**

Ablation Studies



Why Should We Care?

Improves performance by increasing token coverage

Modular and plug-and-play by nature

Maintains inference and memory efficiency

Generalizable across models, domains and languages

Asterix Out Loud

Best performance depends on the number of output channels in conv branches

However, all configurations perform better than baselines

Focus on encoder models only

How will it perform on encoder-decoder models?

Impact on long-range dependencies not tested

Exciting future work!

Thank you for your attention!