Insights from Pixel Language Modeling

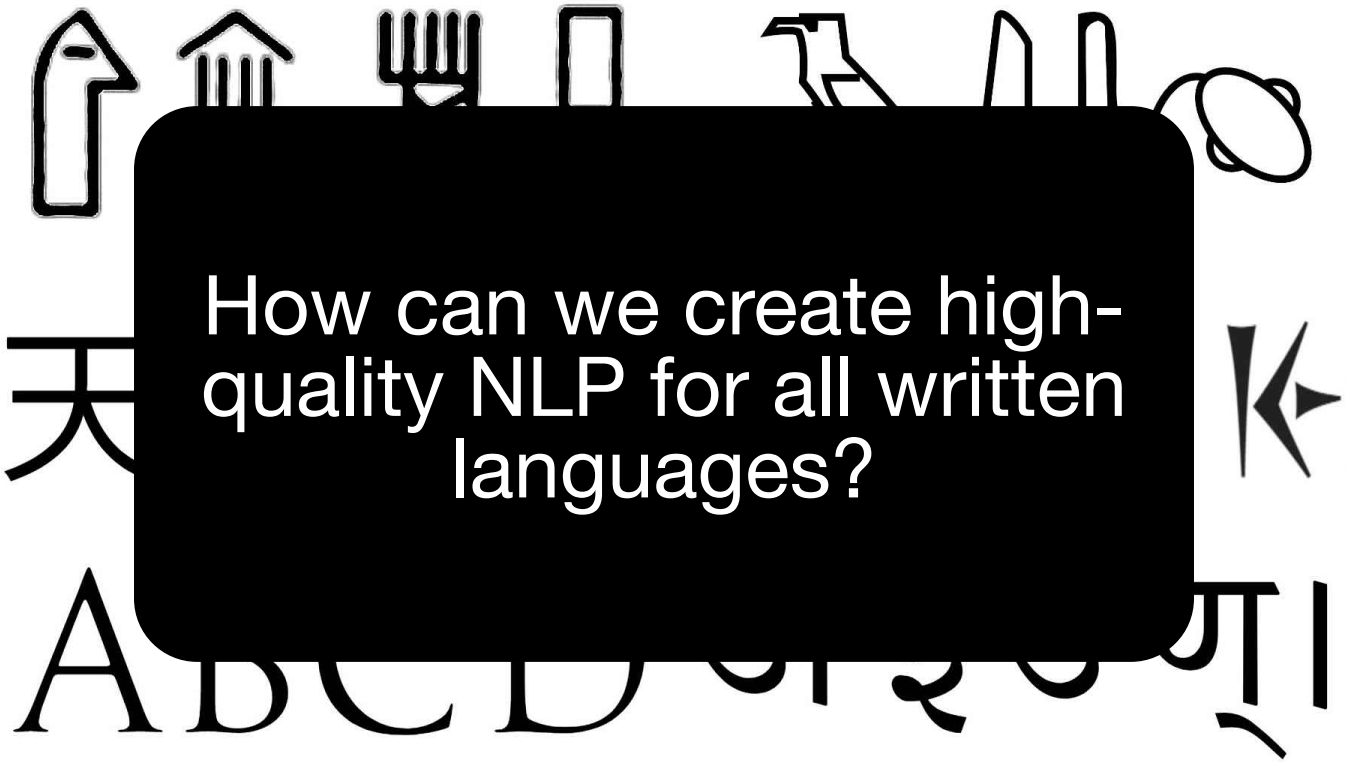
Desmond Elliott

Tokenization Workshop at ICML 2025

Warning: The final part of the talk contains dataset samples that are racist in nature.

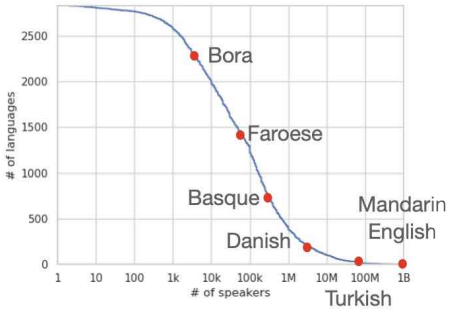
Why are you not at ICML?

The **African Computer Vision Summer School** (ACVSS) unites outstanding African students and researchers with leading computer vision and AI experts.


How can we create high-quality NLP for all written languages?

NLP is an Open Vocabulary Problem

- There are 3,000 written languages
  - 400 with >1M speakers
  - NLP usually covers 100 languages
  - Technological exclusion for billions



- NLP is an **open vocabulary problem** where the units are either
  - "Trained" over a corpus: Byte-Pair Encoding
    - Unseen tokens not in the vocabulary without a byte-level backoff
  - Corpus independent: characters / bytes
    - Need to deal with longer sequence lengths

van Esch et al. Writing System and Speaker Metadata for 2,800+ Language Varieties. LREC 2022.
Joshi et al. The State and Fate of Linguistic Diversity and Inclusion in the NLP World. ACL 2020.

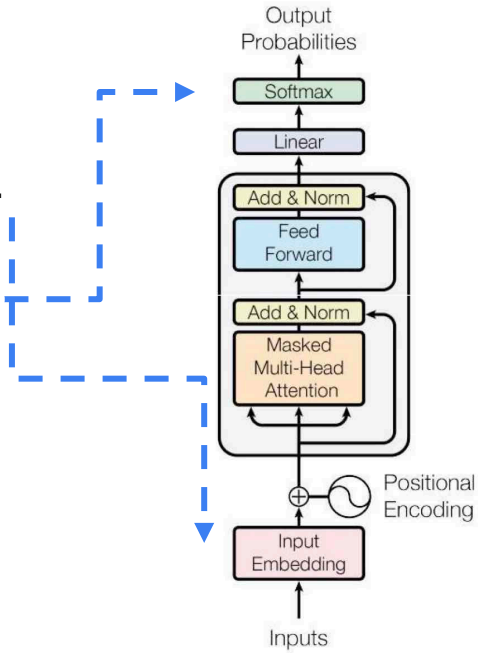
The Vocabulary Bottleneck

The vocabulary of language models creates a bottleneck in two locations:

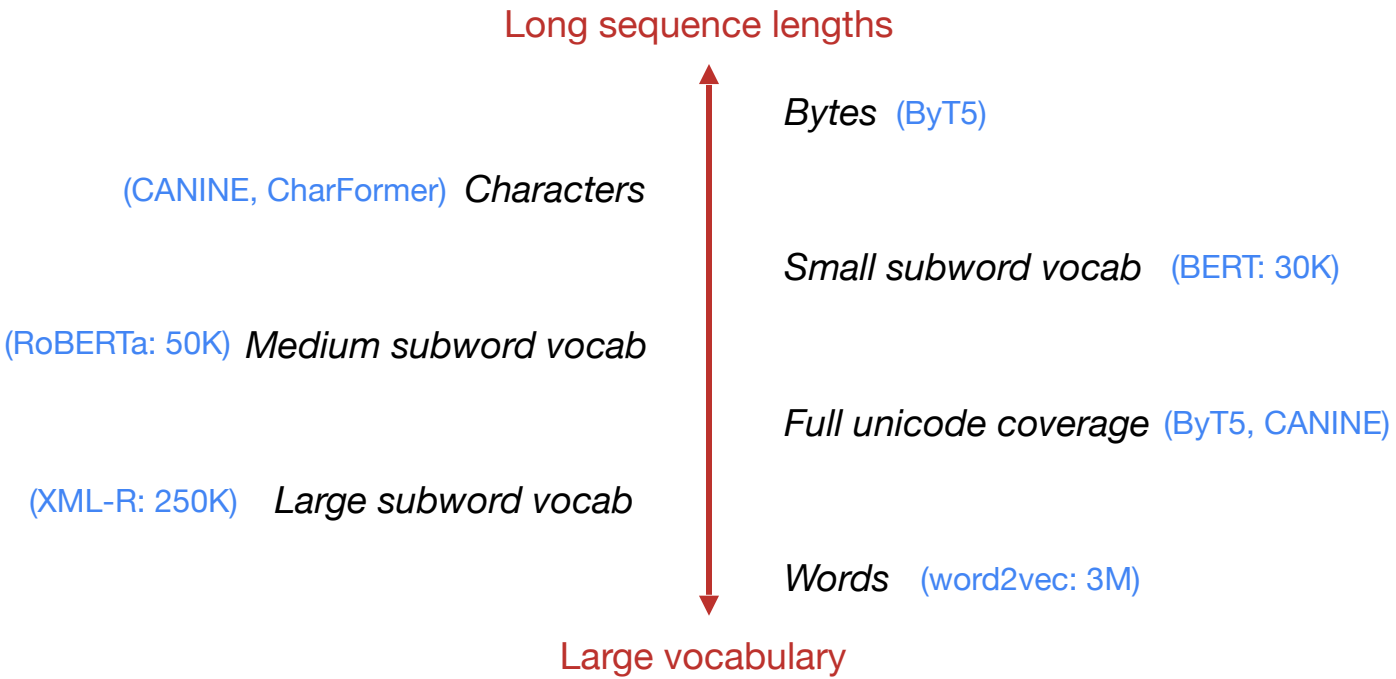
1. *Representational bottleneck* in the Embedding layer
2. *Computational bottleneck* in the token output layer

$$p_\theta(x_i|x_{<i}) := \frac{e^{z_i}}{\sum_j e^{z_j}}$$

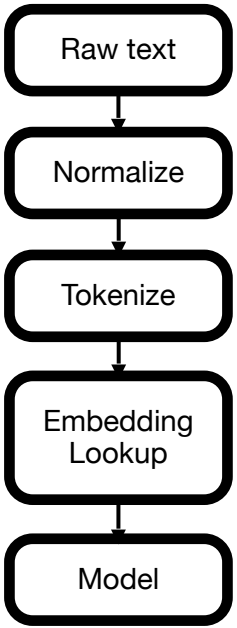
All words in vocabulary



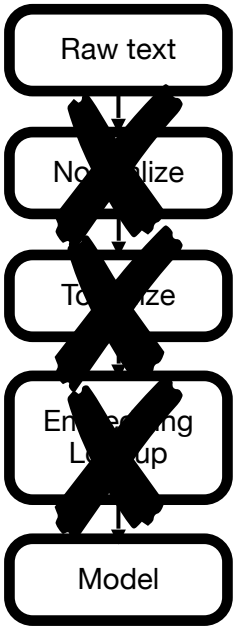
Where's the sweet spot?



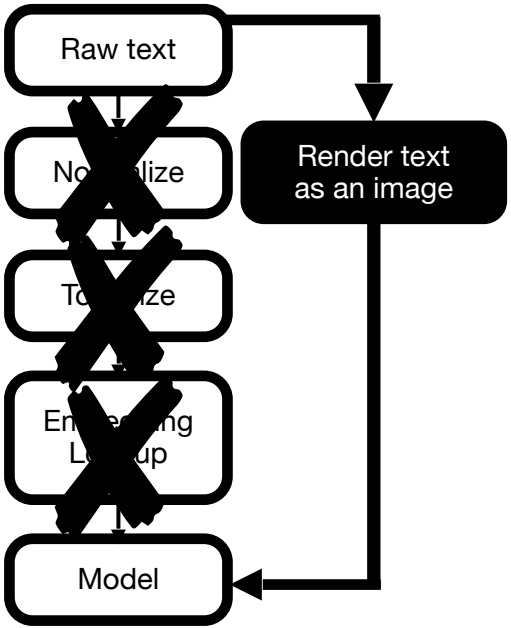
Alternative: treat language as vision



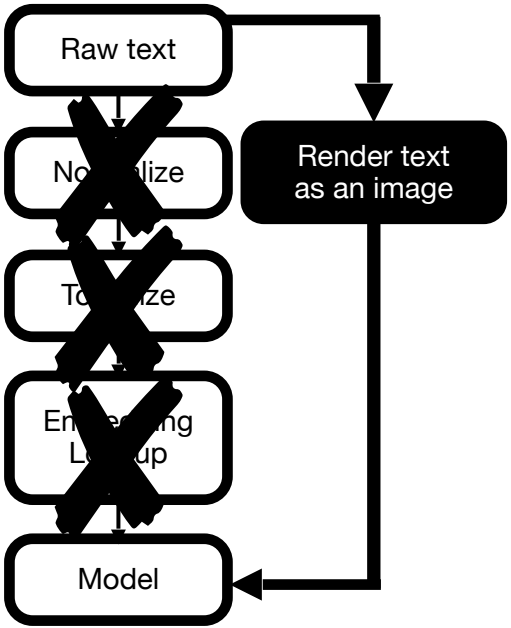
Alternative: treat language as vision



Alternative: treat language as vision

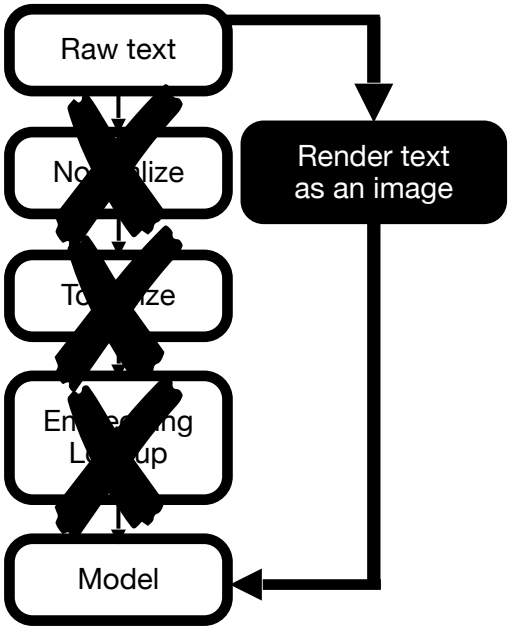


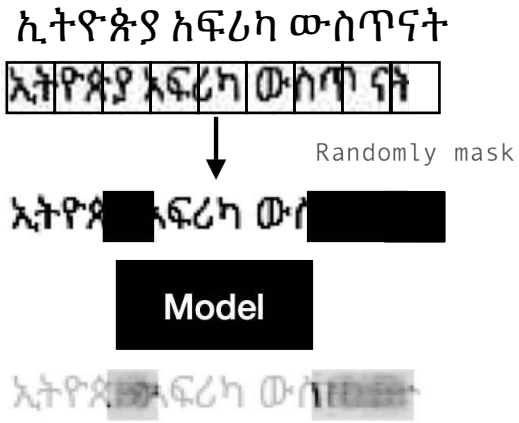
Alternative: treat language as vision



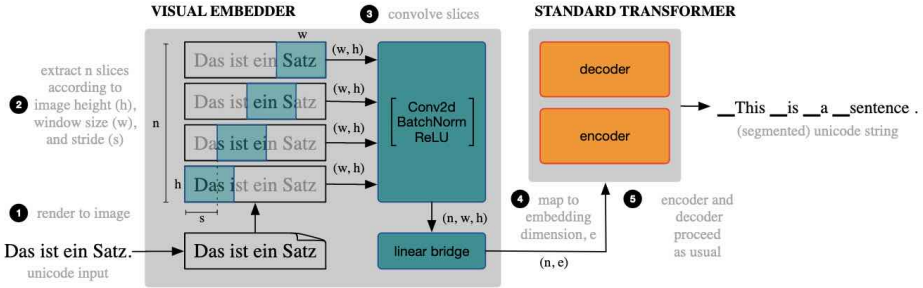
ኢትዮጵያ አፍሪካ ውስጥናት

Alternative: treat language as vision





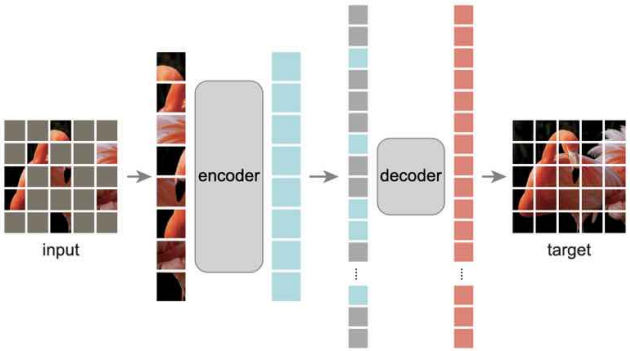
Inspiration



1. Robust Open Vocabulary Translation from Visual Text Representations
(Salesky et al. EMNLP 2021)

You can learn a translation model that transforms from visual representations to discrete tokens

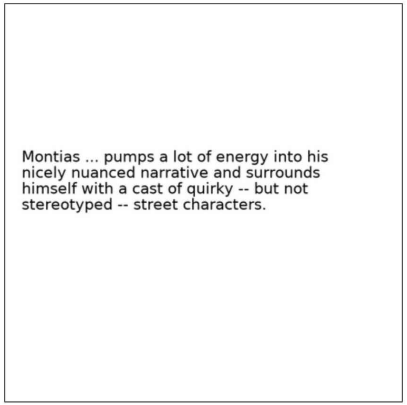
Inspiration



2. Masked Autoencoders are Scalable Visual Learners
(He et al. 2021)

You can learn a vision encoder without any label supervision, so why not a language encoder?

Inspiration



		SST
CLIP-ViT	B/32	70.8
	B/16	75.5
	L/14	**80.8**
	L/14-336px	**80.5**

3. Learning Transferable Visual Models From Natural Language Supervision
(Radford et al. ICML 2021)

You can learn a sentiment classification model using contrastive image–text supervision

Overview

1. The Pixel Language Model
2. Text Rendering Matters
3. Going Multilingual
4. Historical Document Processing

Language Modelling with Pixels

ICLR 2023


P. Rust


J. F. Lotz


E. Bugliarello


E. Salesky


M. de Lhoneux


D. Elliott

The Model

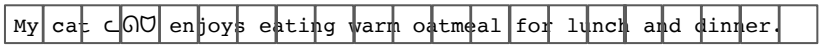
My cat ᓚᘏᗢ enjoys eating warm oatmeal for lunch and dinner.

The Model

16pixel x 16pixel patch

Google Noto Fonts

PyGame / PangoCairo



1 Render Text as Image

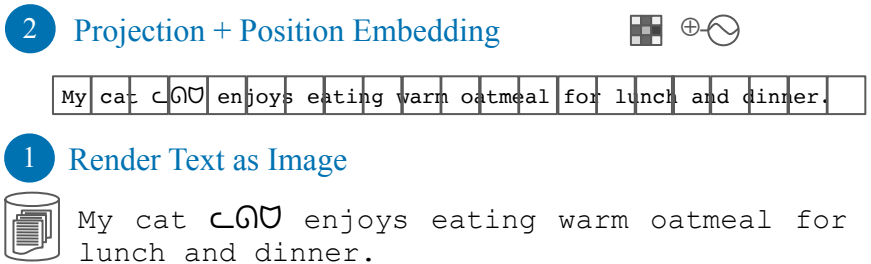
My cat ᓚᘏᗢ enjoys eating warm oatmeal for lunch and dinner.

The Model

16pixel x 16pixel patch

Google Noto Fonts

PyGame / PangoCairo

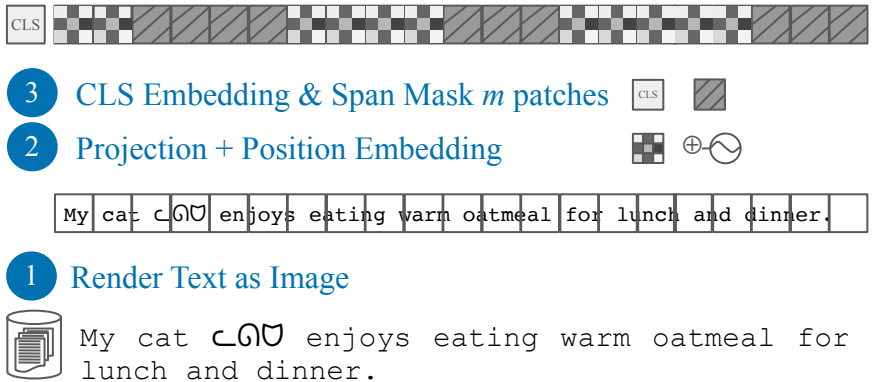


The Model

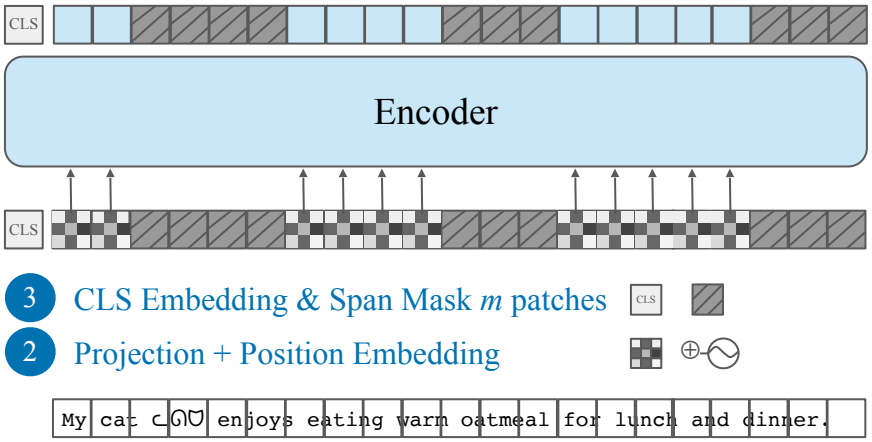
16pixel x 16pixel patch

Google Noto Fonts

PyGame / PangoCairo



The Model



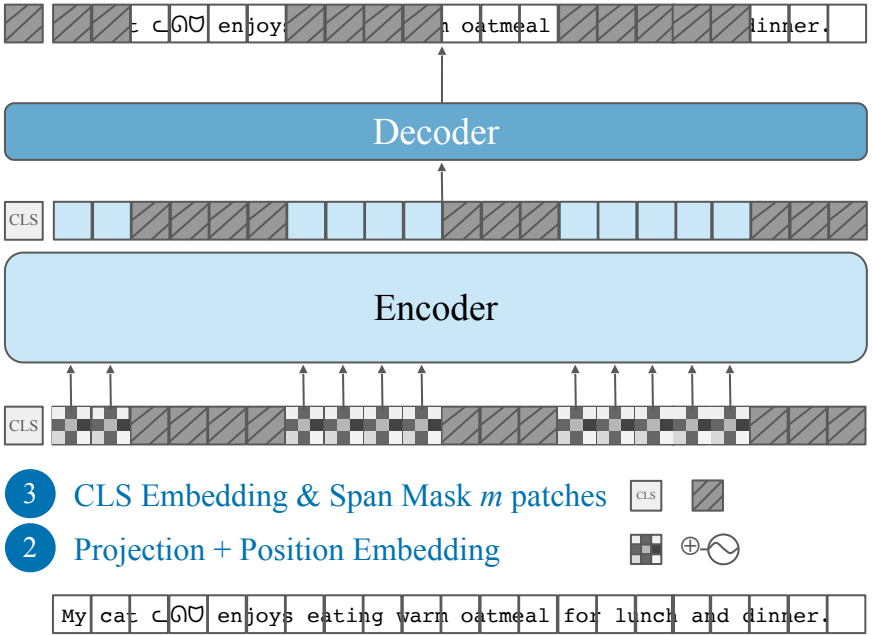
12 Layers

16pixel x 16pixel patch

Google Noto Fonts

PyGame / PangoCairo

The Model



8 Layers

12 Layers

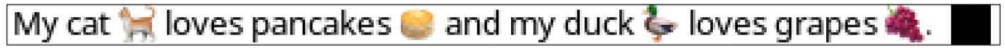
16pixel x 16pixel patch

Google Noto Fonts

PyGame / PangoCairo

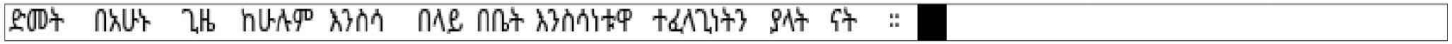
Flexible Text Renderer

- Emoji



- Left-to-right, right-to-left, and logosyllabic writing systems

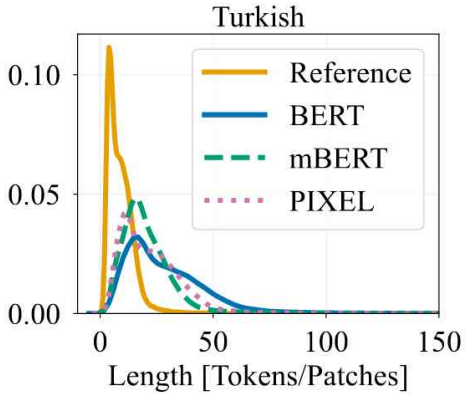
- Word-level rendering



Rendered Text is Compact

- PIXEL encoding produces sequence lengths that are at least as long as as BERT.
  - Universal Dependencies datasets with human reference segmentations
  - No length penalty for languages, unlike some LLMs (Ahia et al. 2023)

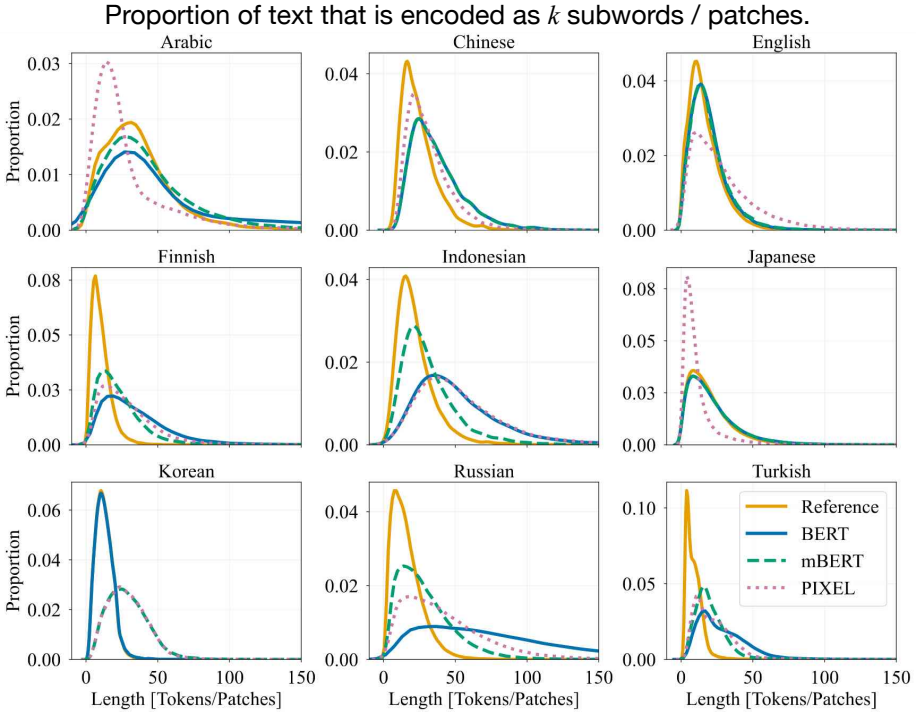
Proportion of text that is encoded as k subwords / patches.



Ahia et al. EMNLP 2023. Do All Languages Cost the Same? Tokenization in the Era of Commercial Language Models

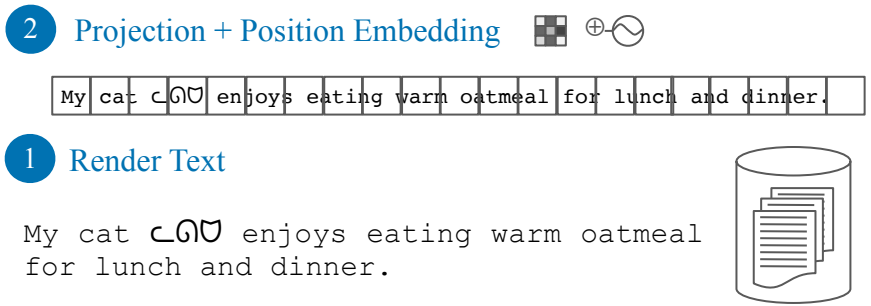
Rendered Text is Compact

- PIXEL encoding produces sequence lengths that are at least as long as as BERT.
  - Universal Dependencies datasets with human reference segmentations
  - No length penalty for languages, unlike some LLMs (Ahia et al. 2023)

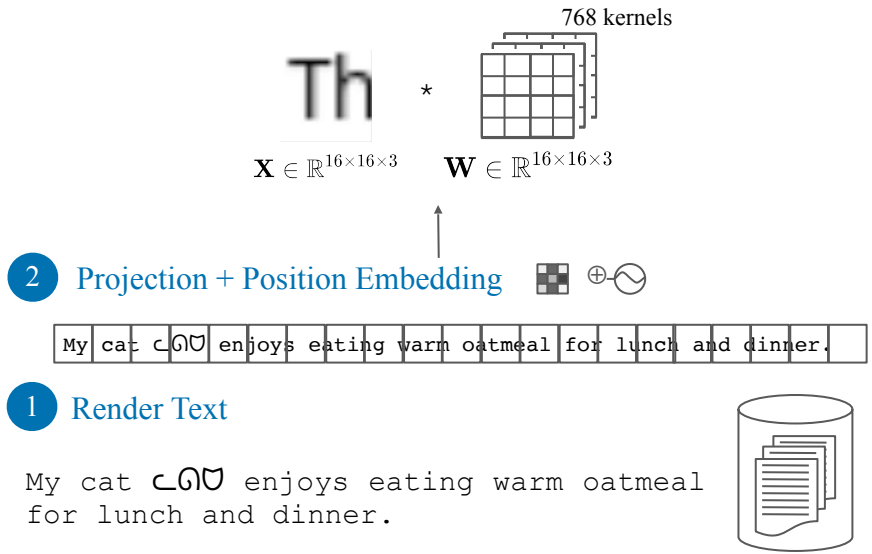


Ahia et al. EMNLP 2023. Do All Languages Cost the Same? Tokenization in the Era of Commercial Language Models

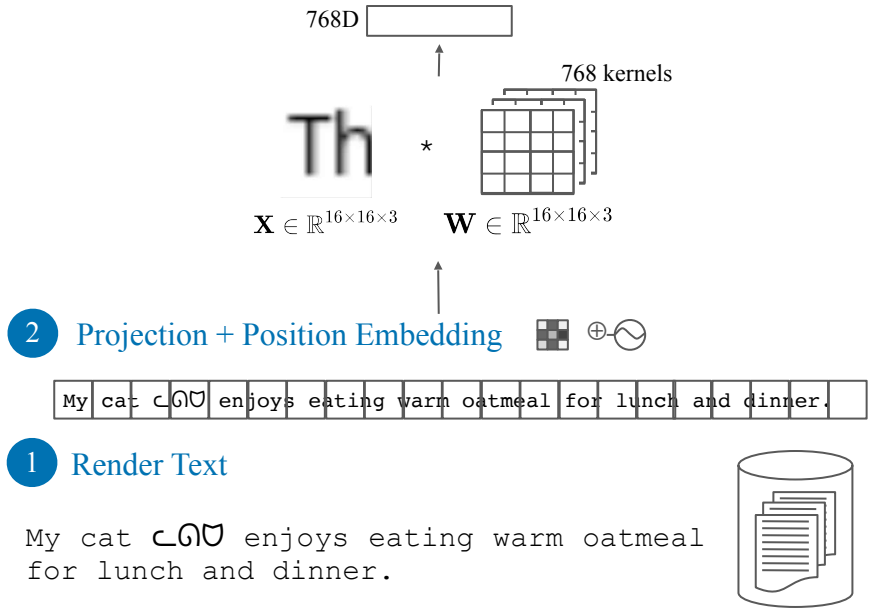
“Embedding” Layer



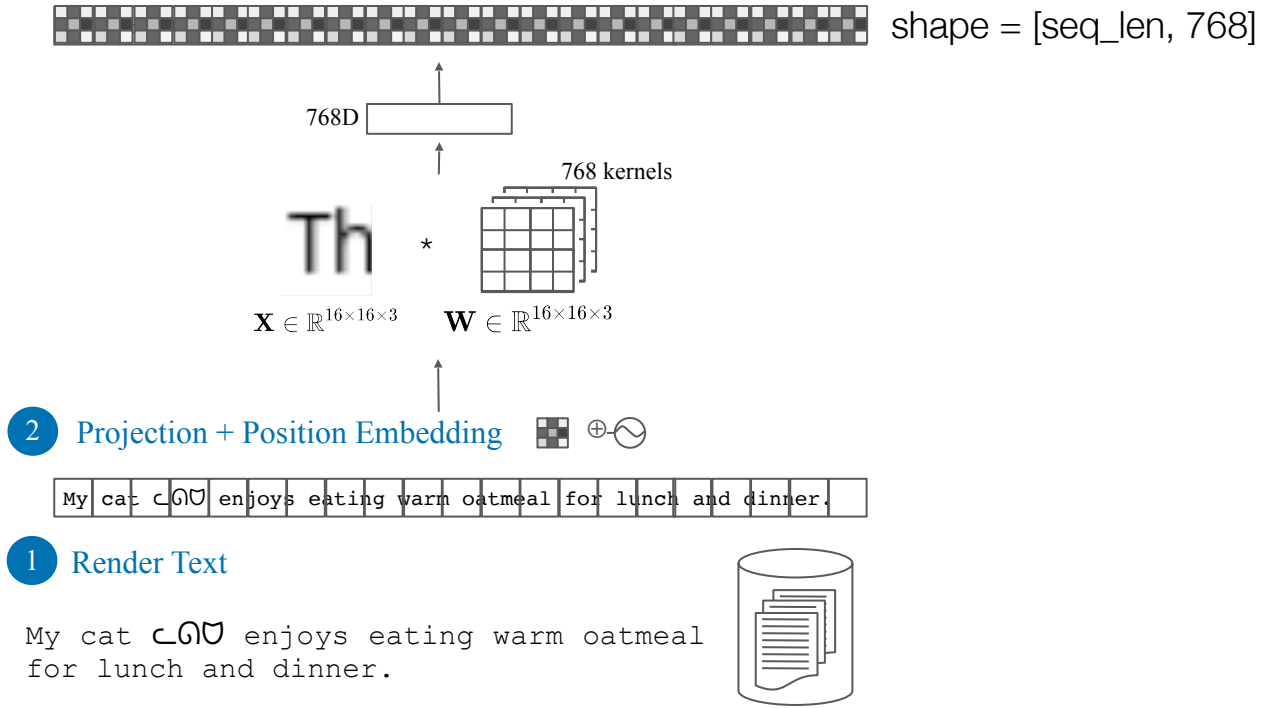
“Embedding” Layer



“Embedding” Layer



“Embedding” Layer

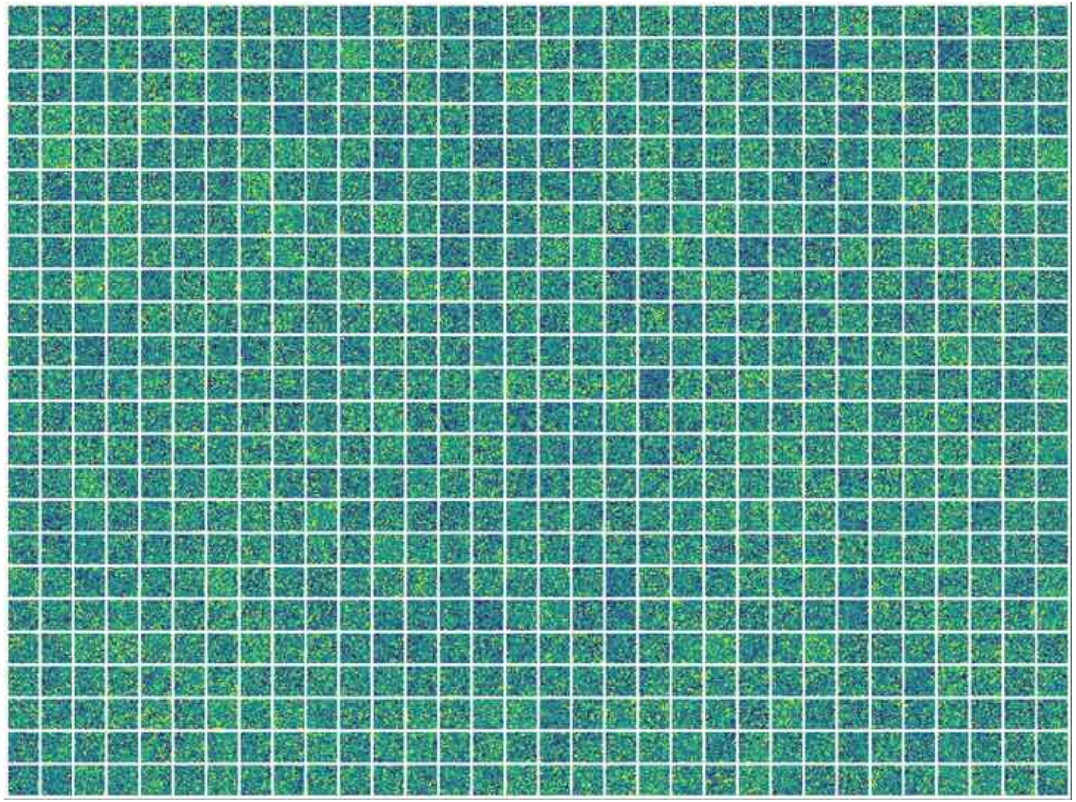


Visualization of Convolution Kernels

1. Some kernels learn about the presence / absence of any pixels.

2. Many kernels capture horizontal strokes

3. Only a few kernels capture curved shapes (*likely due to letters rendered across patch boundaries*)



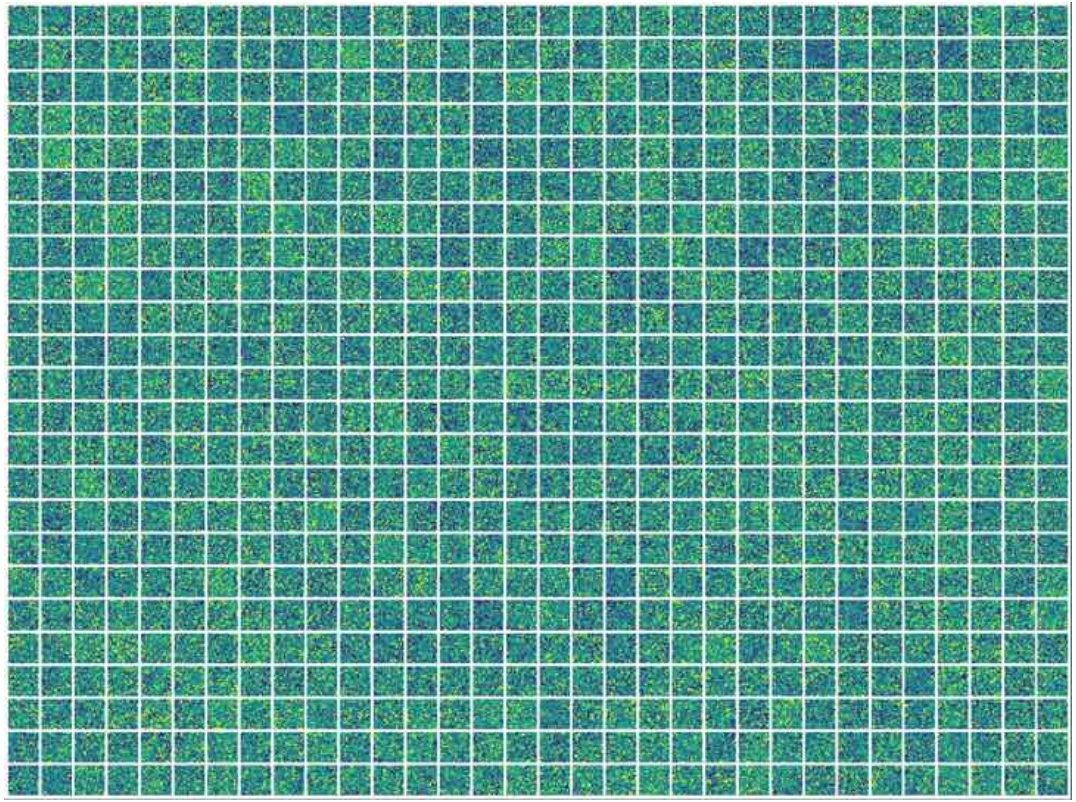
Evolution of Conv2D weights during pretraining step 10K–1M

Visualization of Convolution Kernels

1. Some kernels learn about the presence / absence of any pixels.

2. Many kernels capture horizontal strokes

3. Only a few kernels capture curved shapes (*likely due to letters rendered across patch boundaries*)



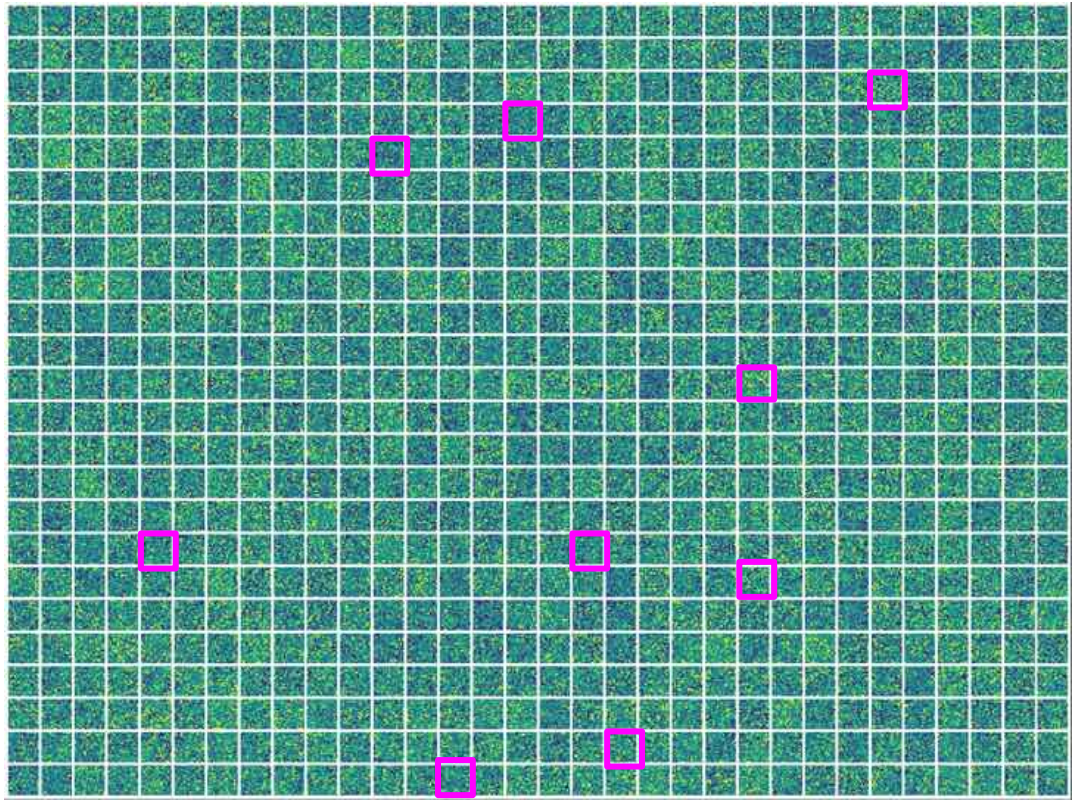
Evolution of Conv2D weights during pretraining step 10K–1M

Visualization of Convolution Kernels

1. Some kernels learn about the presence / absence of any pixels.

2. Many kernels capture horizontal strokes

3. Only a few kernels capture curved shapes (*likely due to letters rendered across patch boundaries*)



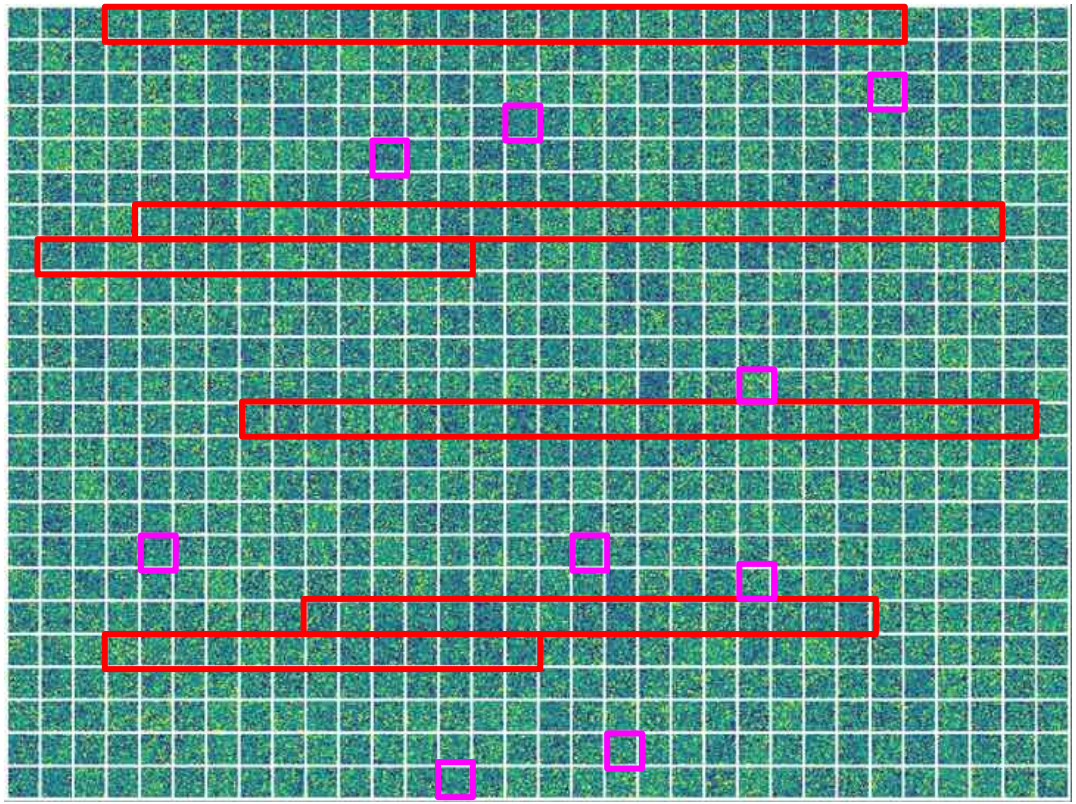
Evolution of Conv2D weights during pretraining step 10K–1M

Visualization of Convolution Kernels

1. Some kernels learn about the presence / absence of any pixels.

2. Many kernels capture horizontal strokes

3. Only a few kernels capture curved shapes (*likely due to letters rendered across patch boundaries*)



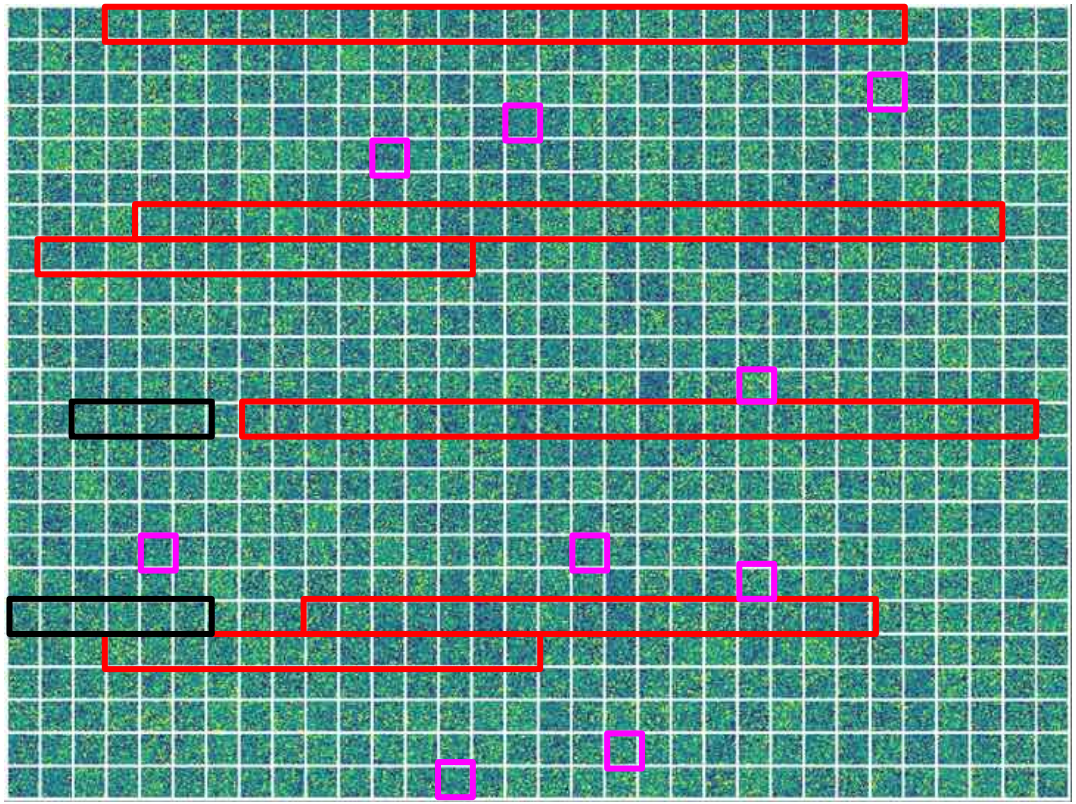
Evolution of Conv2D weights during pretraining step 10K–1M

Visualization of Convolution Kernels

1. Some kernels learn about the presence / absence of any pixels.

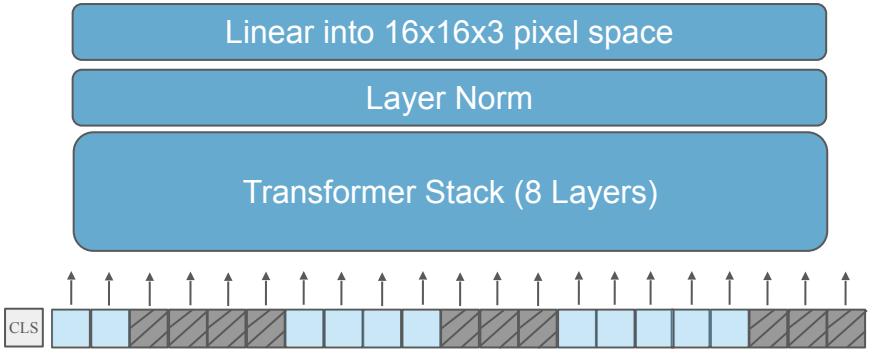
2. Many kernels capture horizontal strokes

3. Only a few kernels capture curved shapes (*likely due to letters rendered across patch boundaries*)

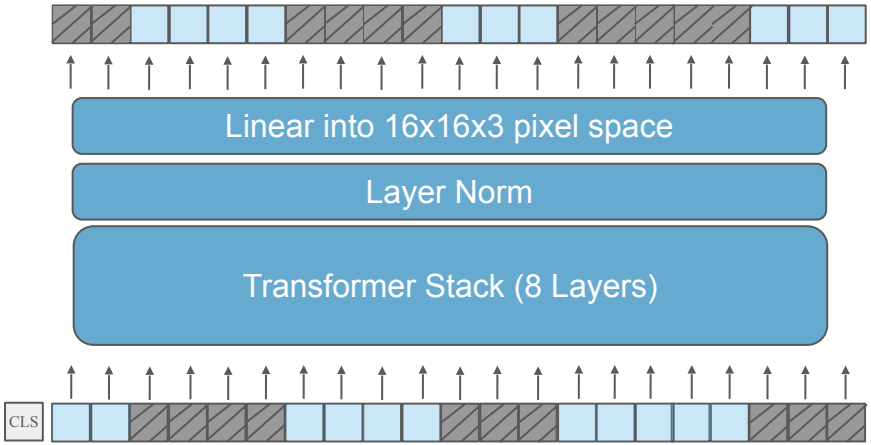


Evolution of Conv2D weights during pretraining step 10K–1M

Objective Function



Objective Function

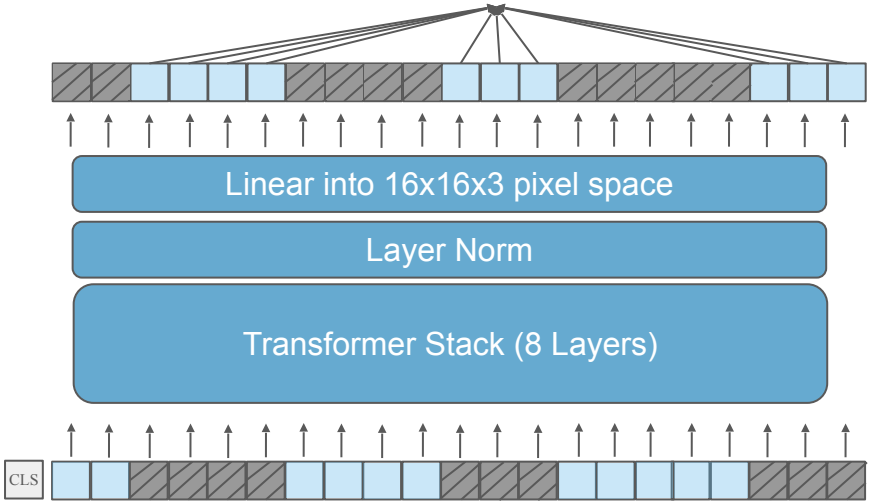


Objective Function

Mean square error loss over M randomly masked patches

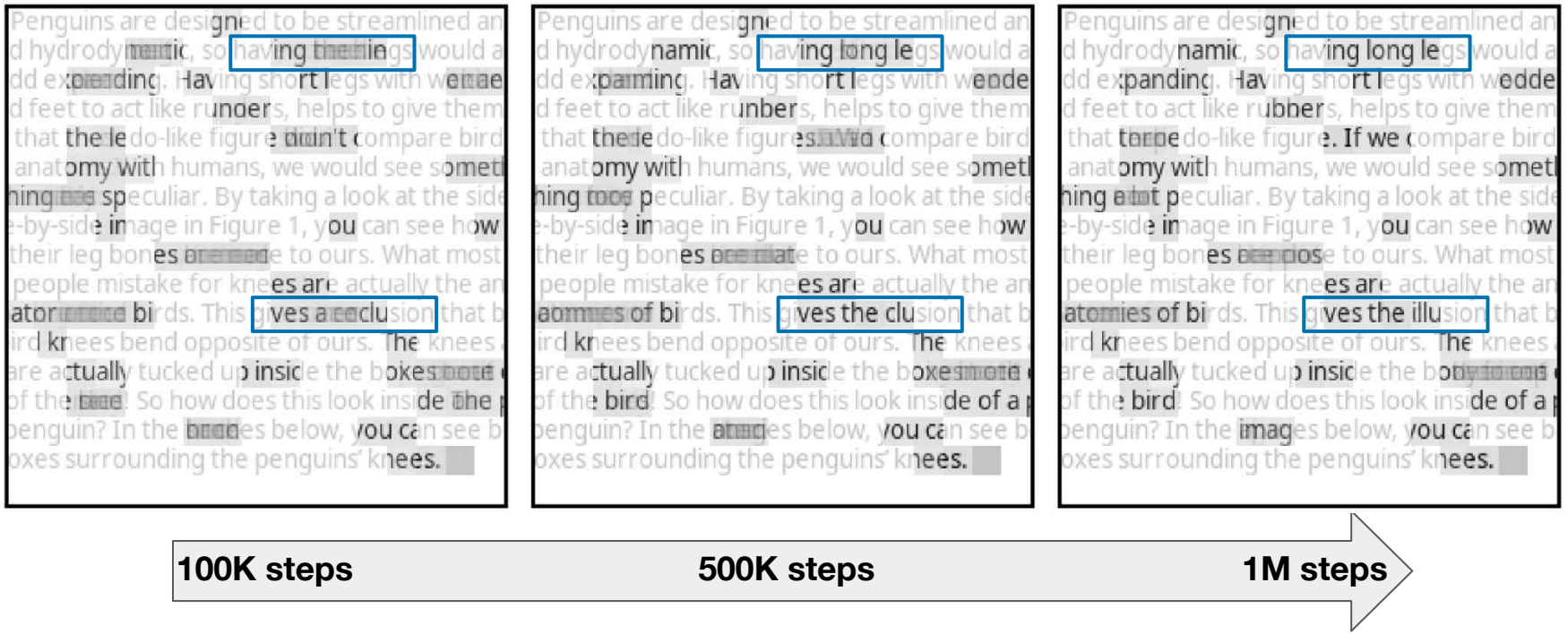
$$\text{MSE} := \frac{1}{M}\sum_{i=1}^{m}(X_i - \hat{X}_i|\mathbf{X}_{\setminus m})^2$$

No Softmax normalization

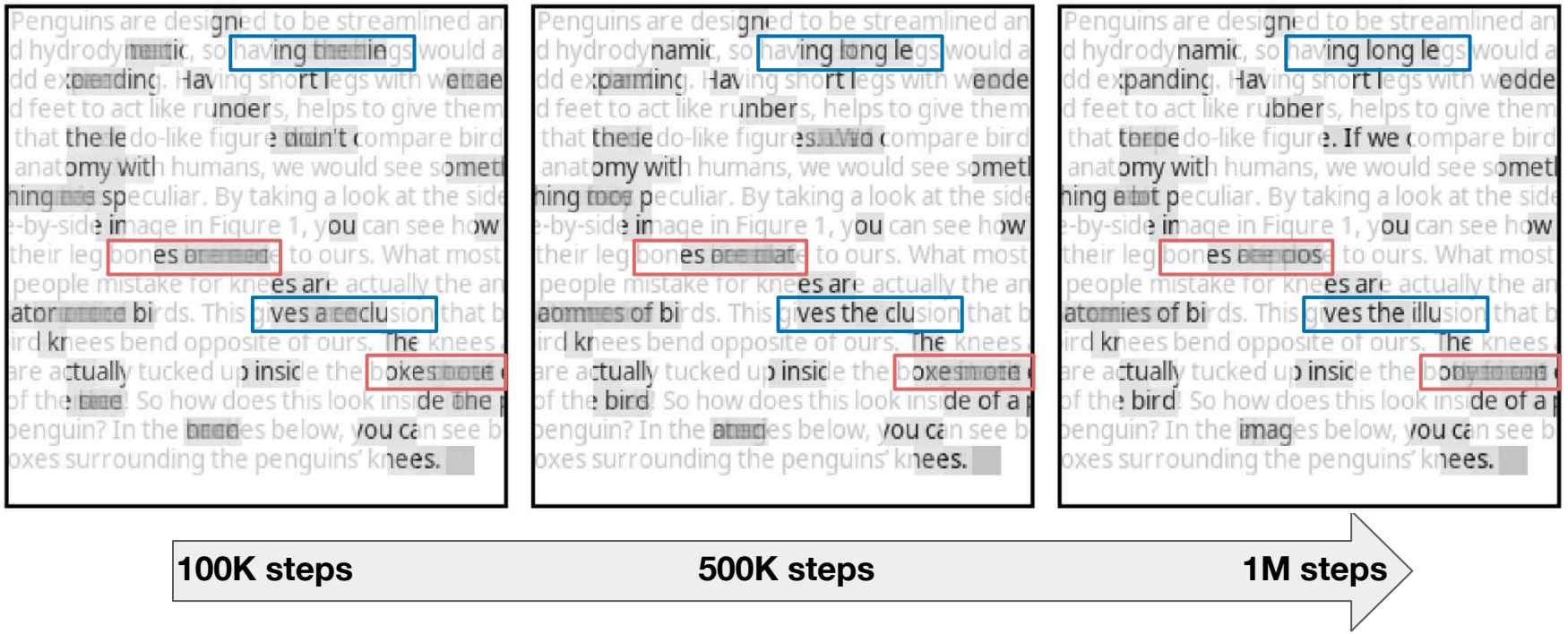


A new type of generative model

A new type of generative model



A new type of generative model



Pretraining

- **English Dataset**: English Wikipedia and Books Corpus
- **Masking:** 25% Span Masking
- **Maximum sequence length**: 529 patches (`16x8464` pixels)
- **Compute**: 8 x 40GB A100 GPUs for 8 days
- **Parameters**: 86M encoder + 26M decoder

There is only 0.05% non-English text in our pretraining data (estimated by Blevins and Zettlemoyer 2022)

The **Great Wall of China** (traditional Chinese: 萬里長城; simplified Chinese: 万里长城; pinyin: *Wànlǐ Chángchéng*)

Downstream Tasks

- **Datasets**: Universal Dependencies, MasakhaNER, GLUE, Zeroé
- **Models:**

	Parameters	**Pretraining Data**
$\text{PIXEL}_{\text{BASE}}$	86M	English Wikipedia + Bookcorpus
$\text{BERT}_{\text{BASE}}$	110M	—
CANINE-C	127M	104-languages from Wikipedia

Downstream Tasks

- **Datasets**: Universal Dependencies, MasakhaNER, GLUE, Zeroé
- **Models:**

	Parameters	**Pretraining Data**
$\text{PIXEL}_{\text{BASE}}$	86M	English Wikipedia + Bookcorpus
$\text{BERT}_{\text{BASE}}$	110M	—
CANINE-C	127M	104-languages from Wikipedia

Similar pretraining setup

Downstream Tasks

- **Datasets**: Universal Dependencies, MasakhaNER, GLUE, Zeroé

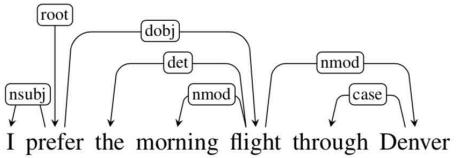
- **Models:**

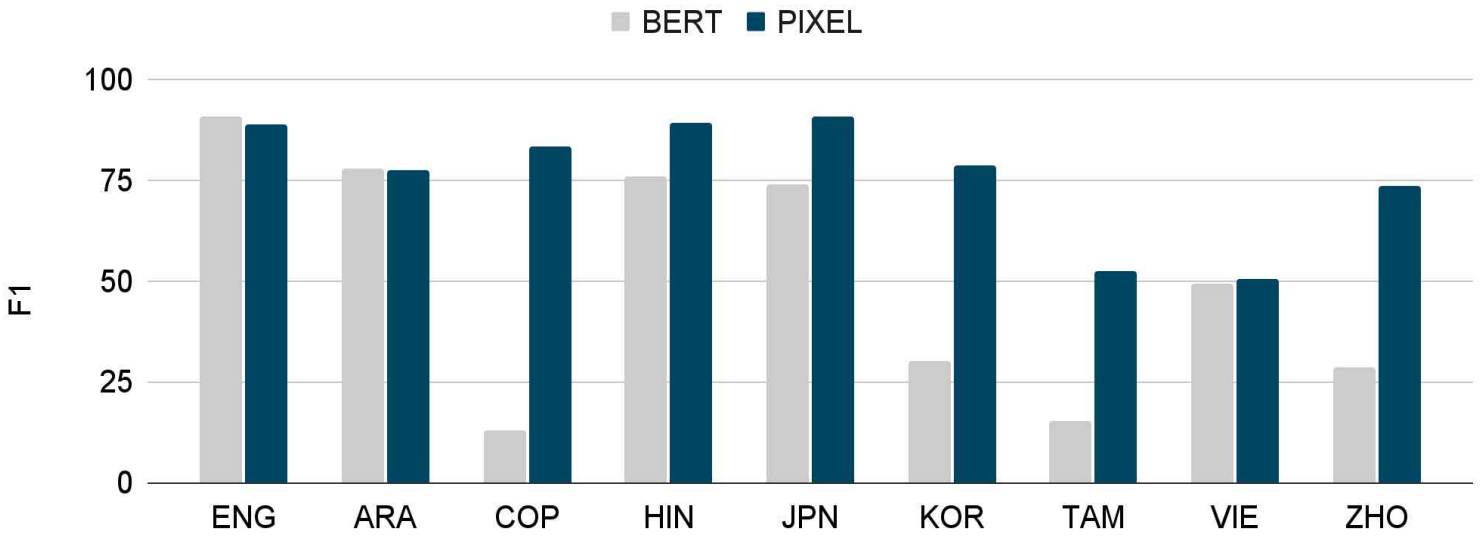
	Parameters	**Pretraining Data**
$\text{PIXEL}_{\text{BASE}}$	86M	English Wikipedia + Bookcorpus
$\text{BERT}_{\text{BASE}}$	110M	—
CANINE-C	127M	104-languages from Wikipedia

Similar pretraining setup

Tries to solve the same problem using UTF-32

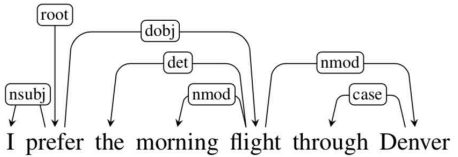
Dependency Parsing Results

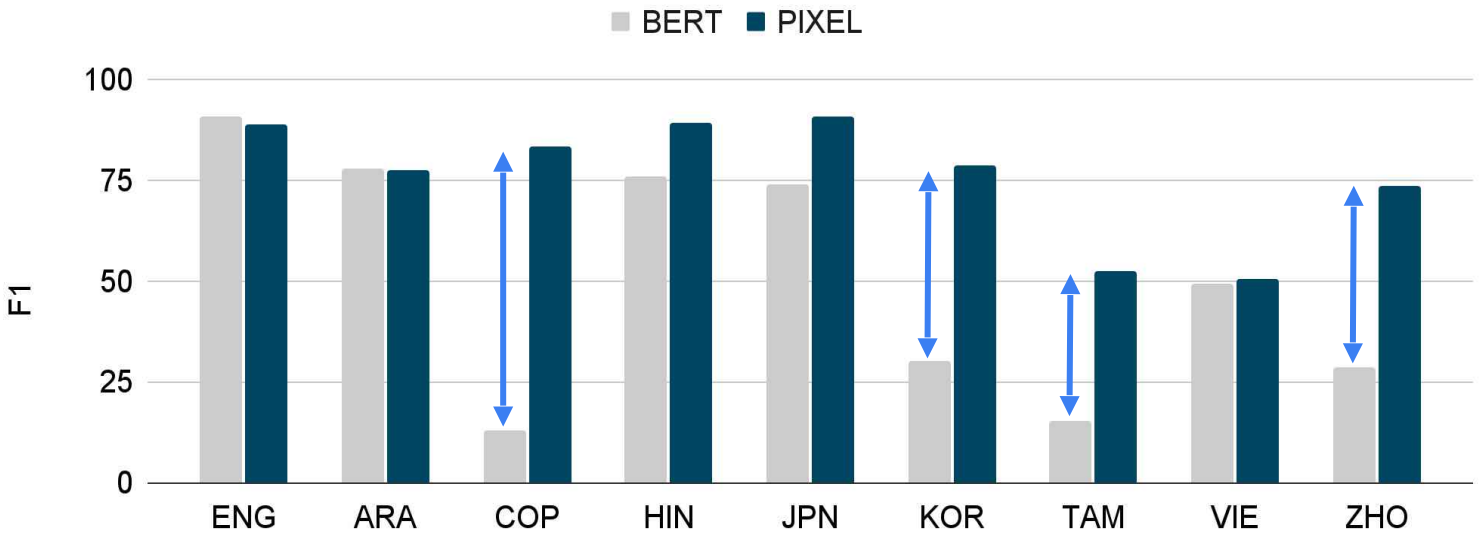




BERT UNK	0%	1%	94%	33%	46%	85%	82%	5%	73%

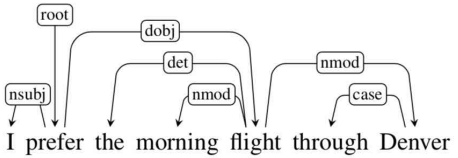
Dependency Parsing Results

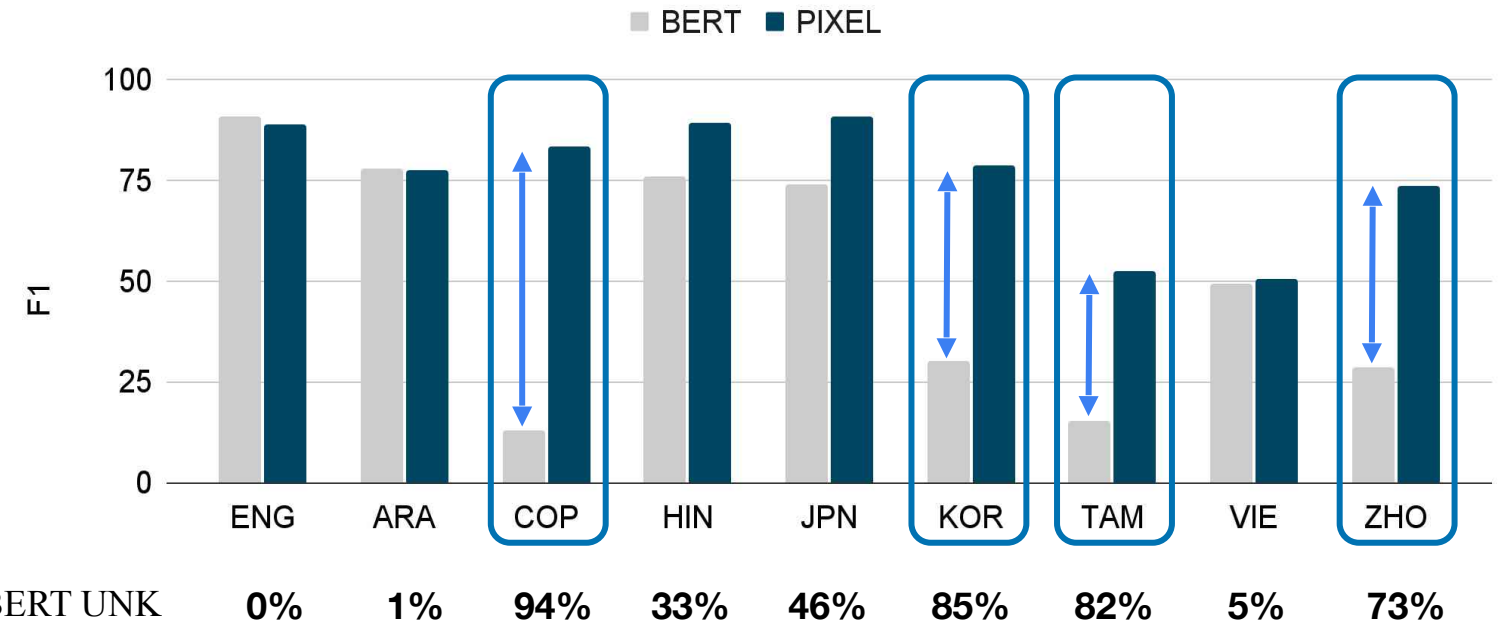




BERT UNK	0%	1%	94%	33%	46%	85%	82%	5%	73%

Dependency Parsing Results

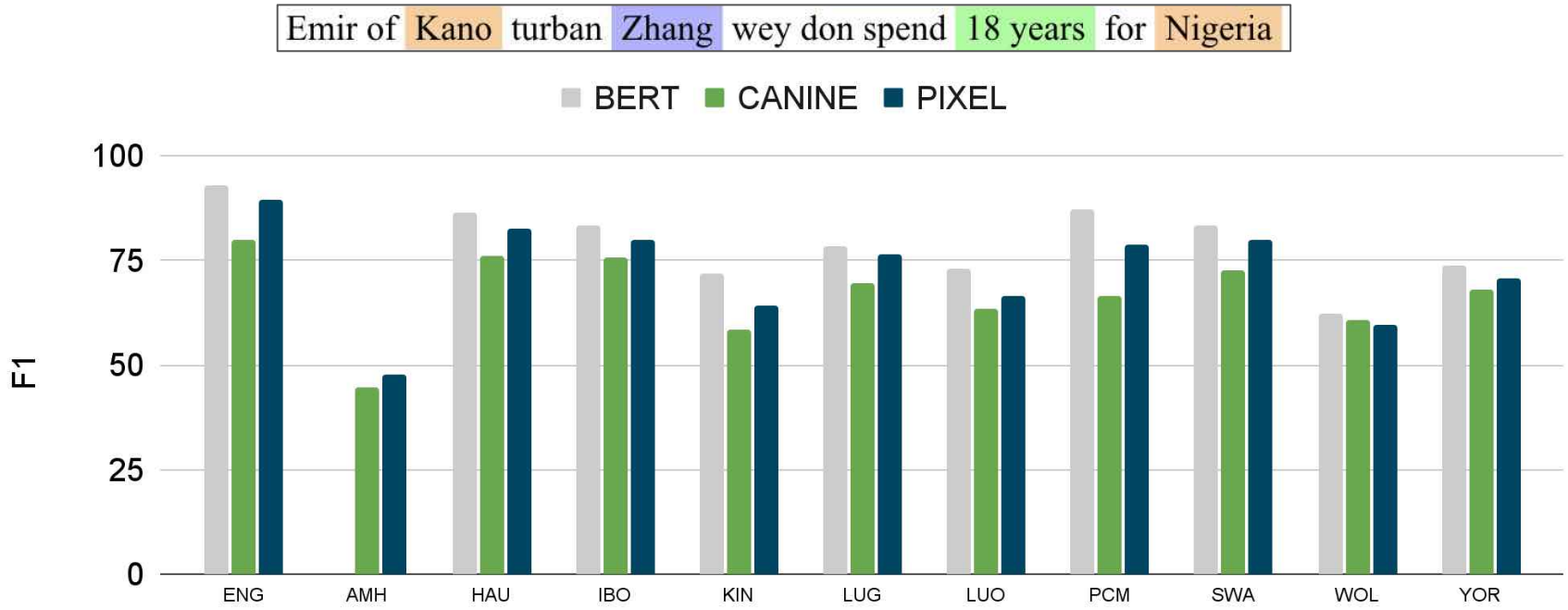




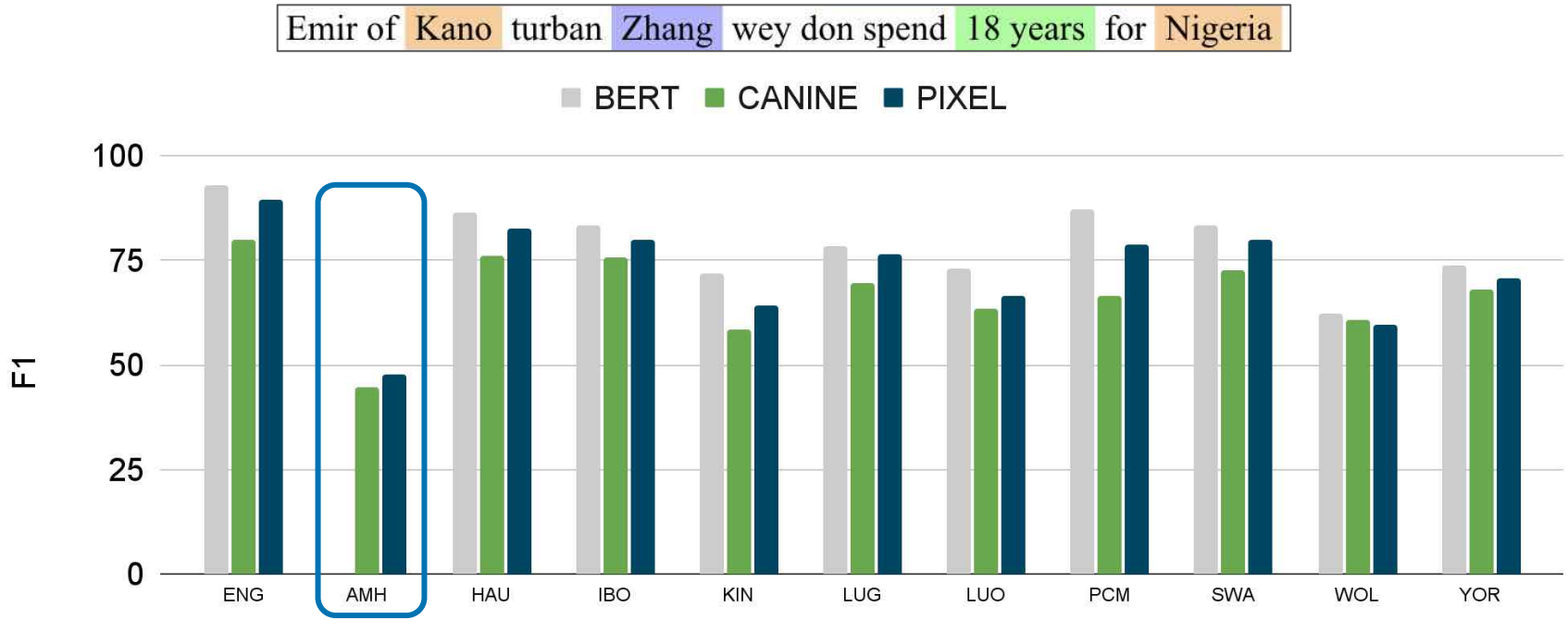
BERT UNK	0%	1%	94%	33%	46%	85%	82%	5%	73%

PIXEL vastly outperforms BERT on unseen scripts

Named Entity Recognition in African Languages

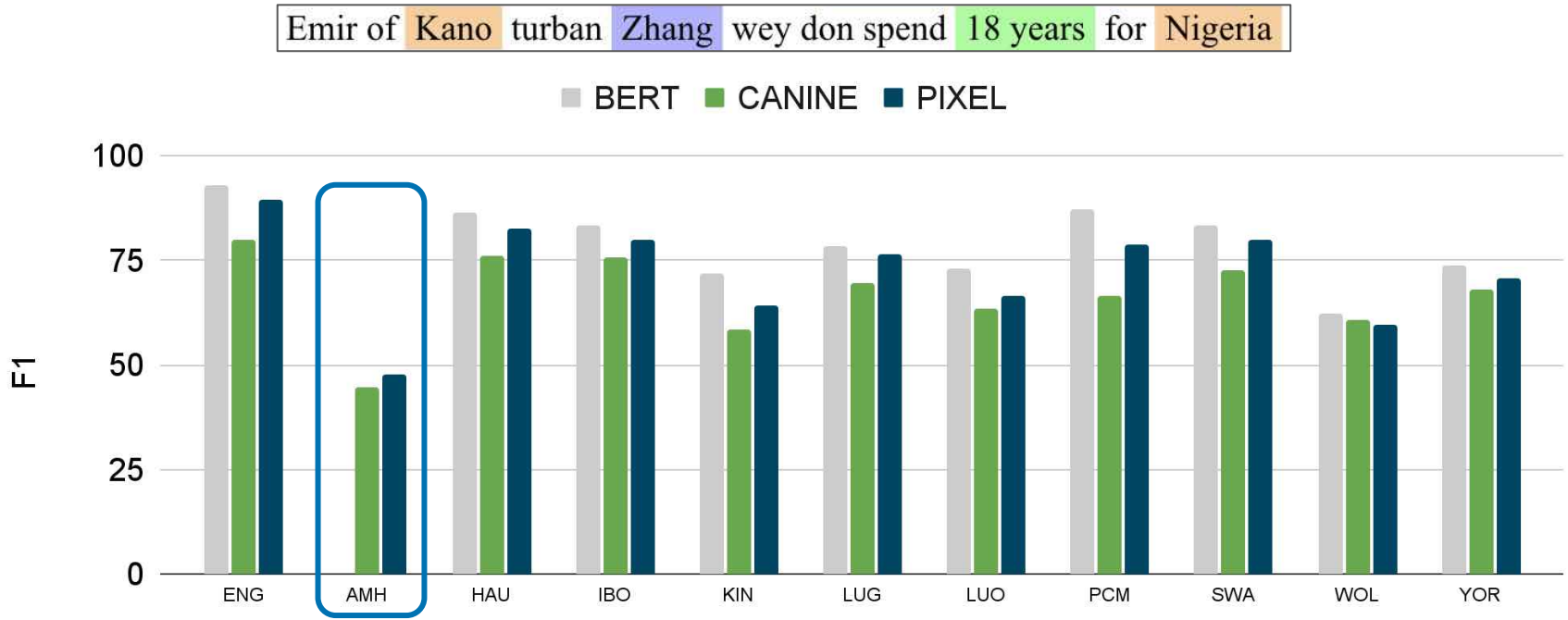


Named Entity Recognition in African Languages



PIXEL outperforms BERT on the non-Latin script

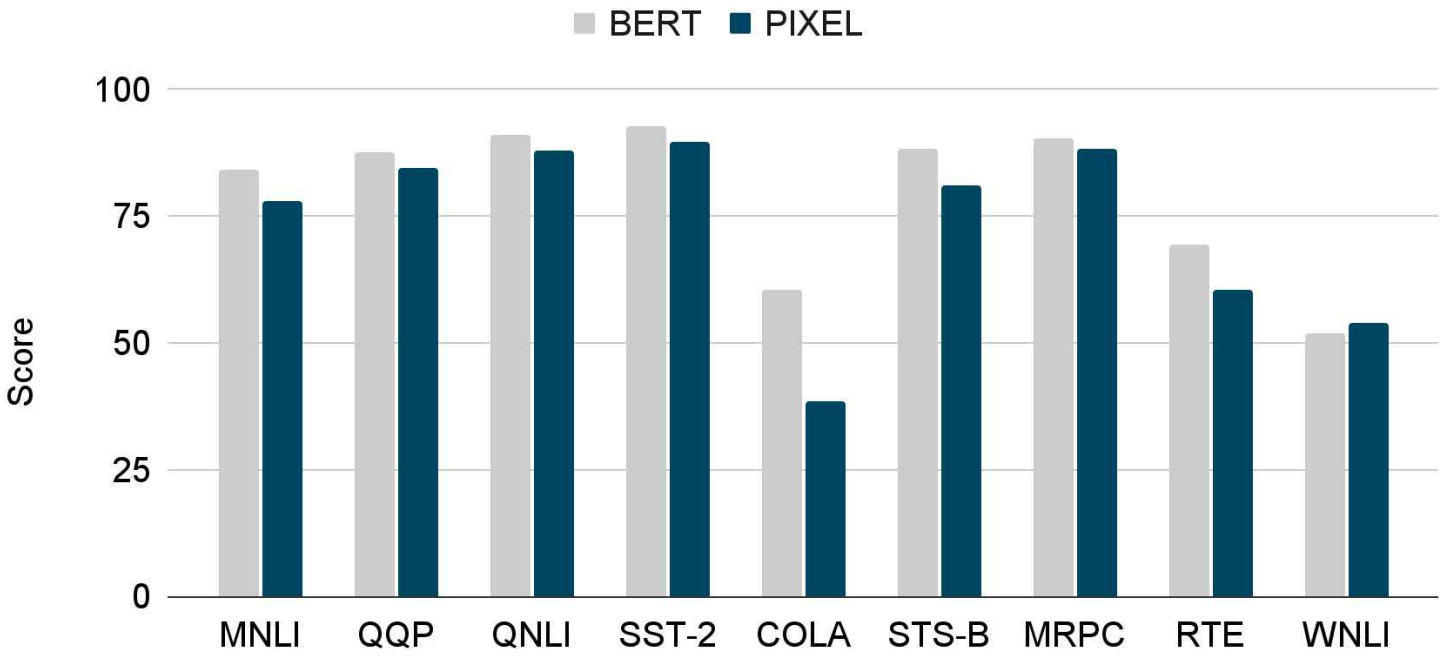
Named Entity Recognition in African Languages



PIXEL outperforms BERT on the non-Latin script

PIXEL outperforms the multilingually pretrained CANINE-C

GLUE: Sentence-level Understanding



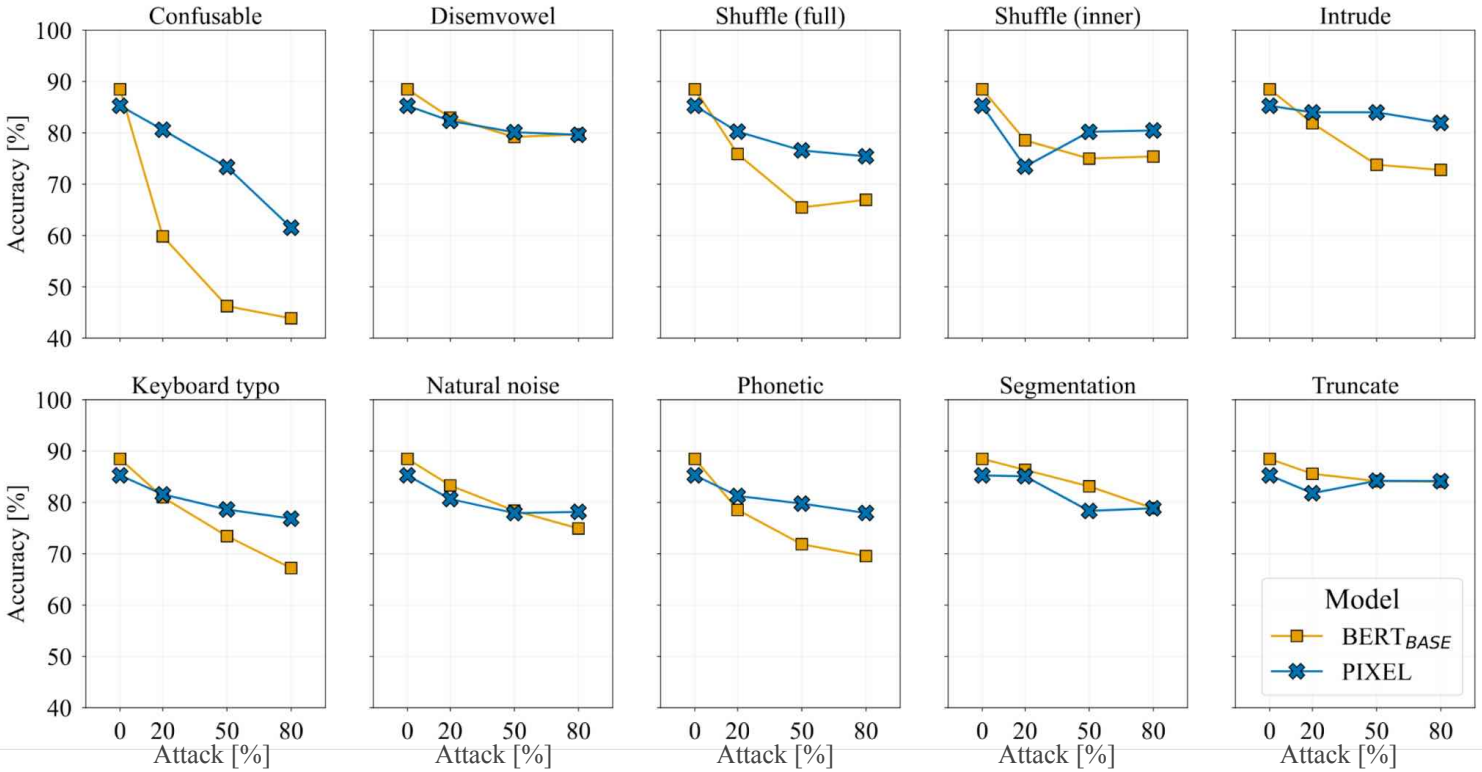
BERT outperforms PIXEL on English sentence-level tasks

Ädʊɜʳrsarīał at̲tacḳṣ

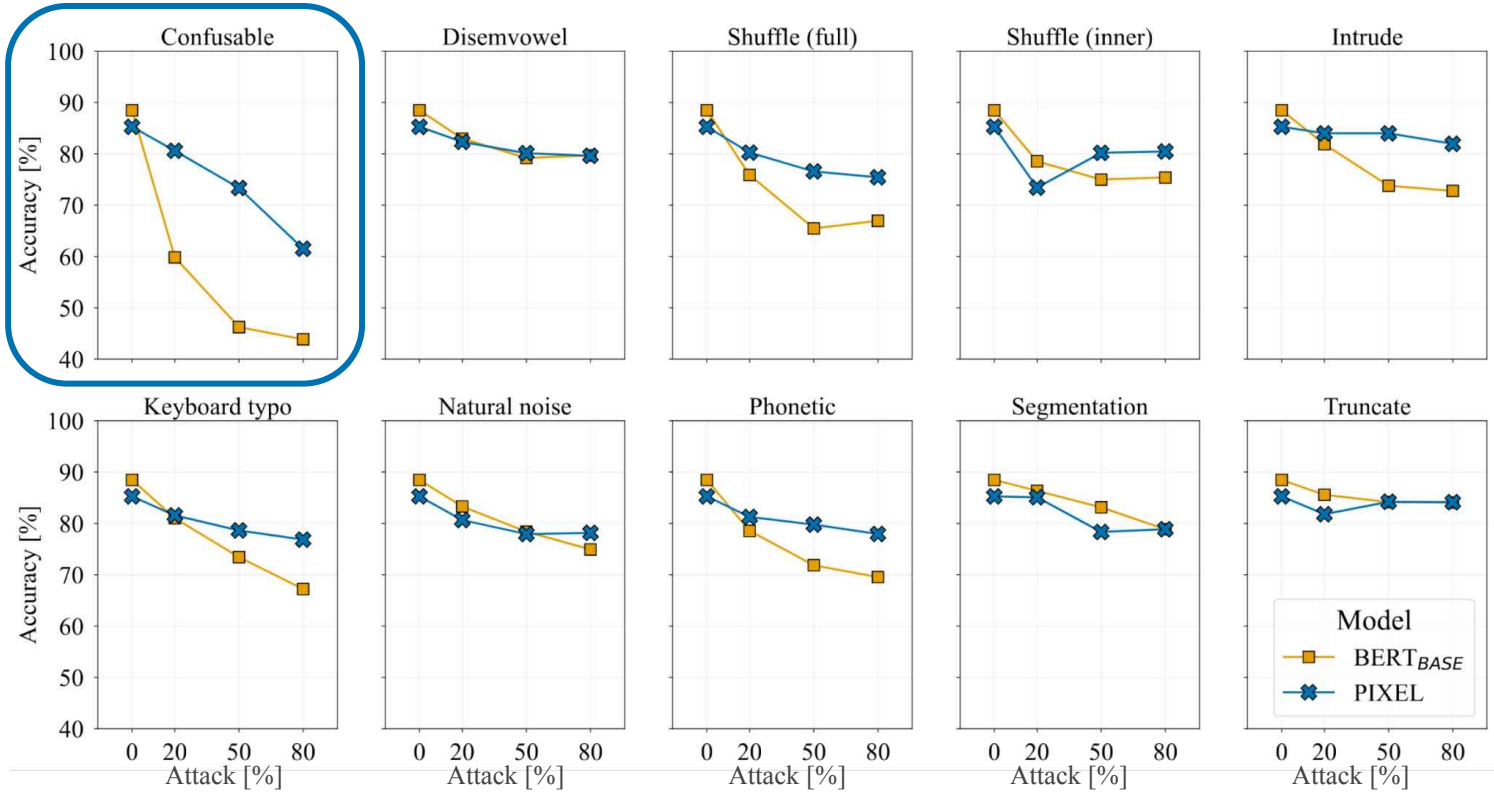
- How well does PIXEL deal with orthographic text attacks?

Attack	Sentence
NONE	Penguins are designed to be streamlined
CONFUSABLE	Pɛɱgɯns are 𝒹esigɳɛ𝒹 to be ʂtreaмlɪɳed
SHUFFLE (INNER)	Pegnuins are dnesiged to be sieatrnmled
SHUFFLE (FULL)	ngePnius rae dsgednei to be etimaslernd
DISEMVOWEL	Pngns r dsgnd to be strmlnd
INTRUDE	Pe‘nguins a{re d)esigned t;o b*e stre<amlined
KEYBOARD TYPO	Penguinz xre dwsigned ro ne streamllned
NATURAL NOISE	Penguijs ard design4d ti bd streamlinfd
TRUNCATE	Penguin are designe to be streamline
SEGMENTATION	Penguinsaredesignedtobestreamlined
PHONETIC	Pengwains’s ar dhiseind te be storimlignd

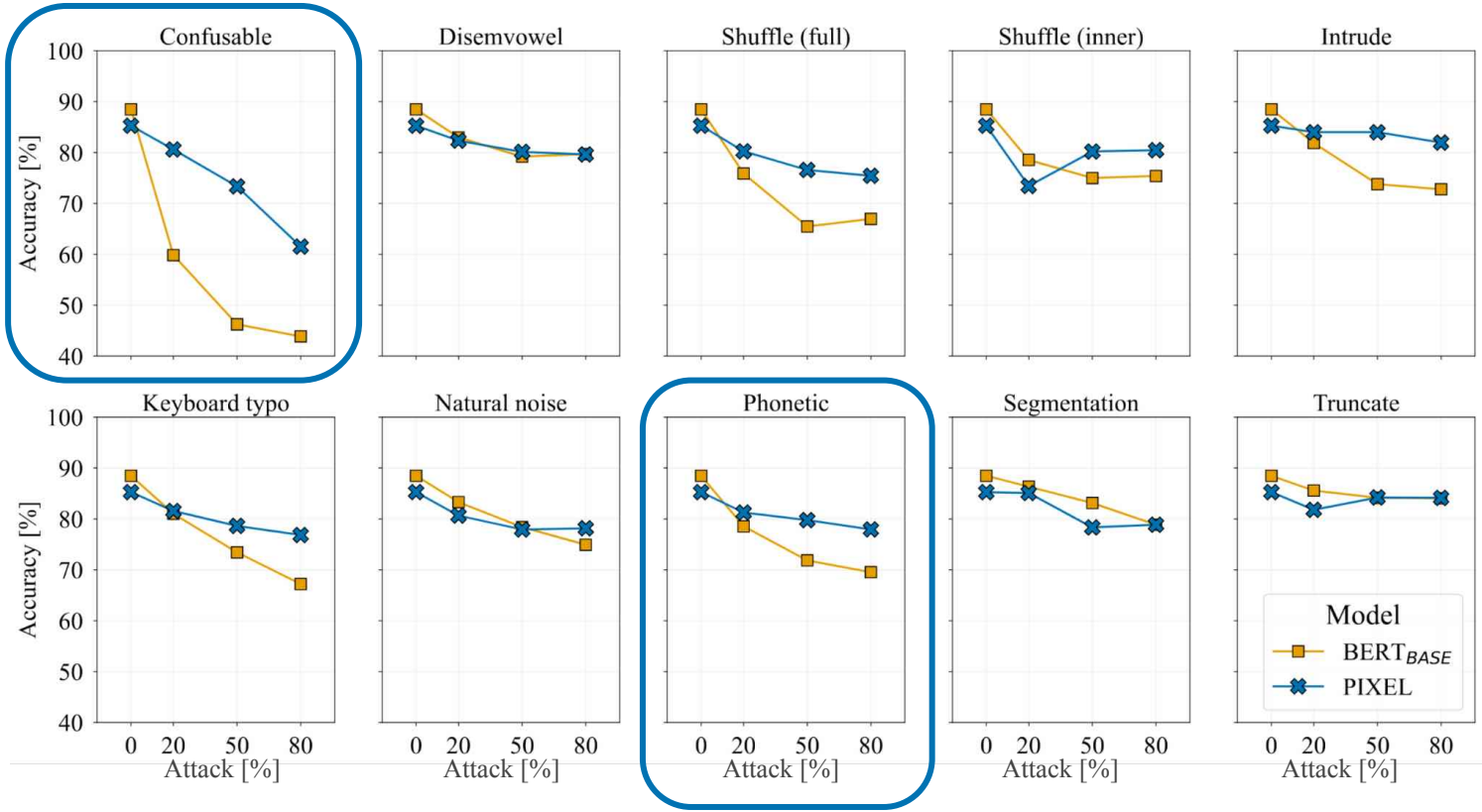
Results on Zeroé (SNLI)



Results on Zeroé (SNLI)



Results on Zeroé (SNLI)



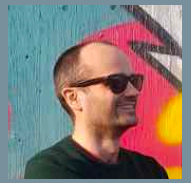
Text Rendering Strategies for Pixel Language Models

EMNLP 2023


J. F. Lotz


E. Salesky


P. Rust


D. Elliott

Text Rendering Matters

- Our original text renderer produces many nearly-identical patches
  - This is representation- and compute-wasteful

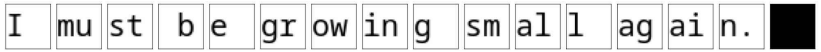
Can we do better?

Alternative Rendering Strategies

(a) Continuous rendering (CONTINUOUS):

(b) Structured rendering (BIGRAMS):

(c) Structured rendering (MONO):



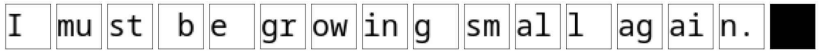
(d) Structured rendering (WORDS):

Alternative Rendering Strategies

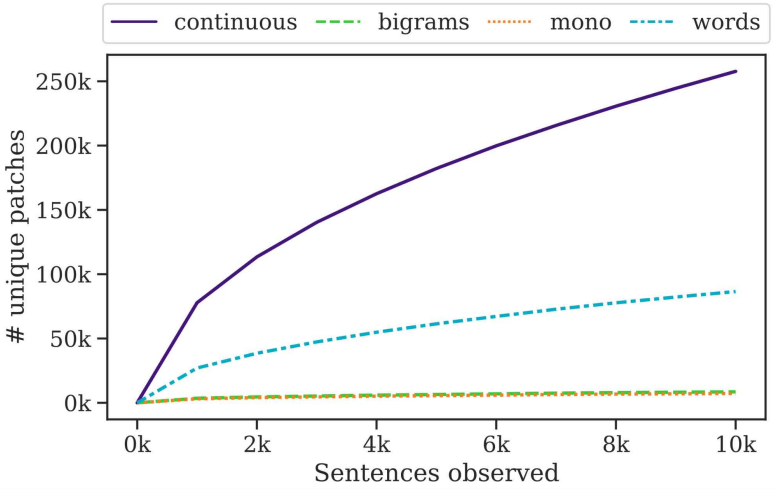
(a) Continuous rendering (CONTINUOUS):

(b) Structured rendering (BIGRAMS):

(c) Structured rendering (MONO):



(d) Structured rendering (WORDS):

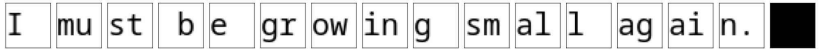


Alternative Rendering Strategies

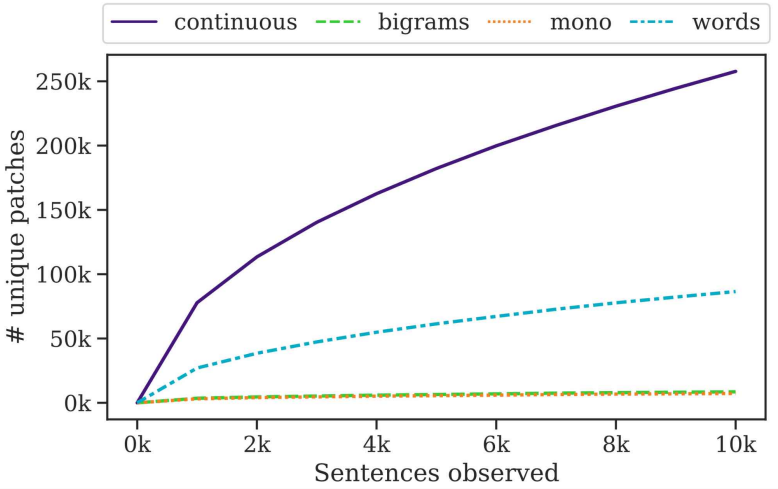
(a) Continuous rendering (CONTINUOUS):

(b) Structured rendering (BIGRAMS):

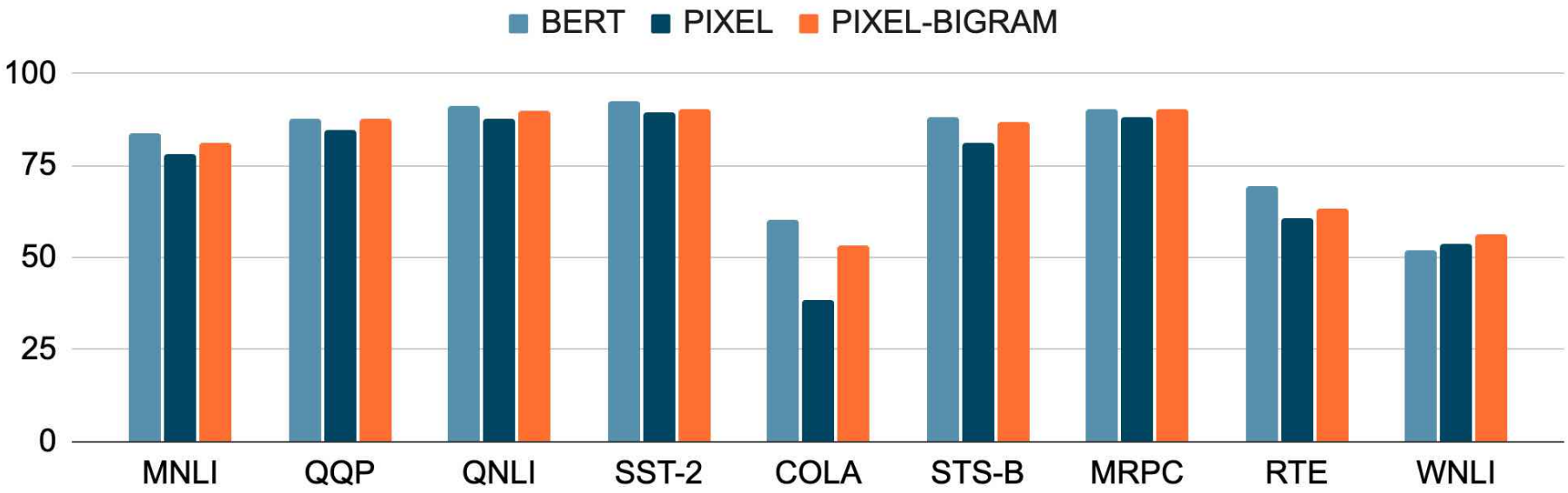
(c) Structured rendering (MONO):



(d) Structured rendering (WORDS):



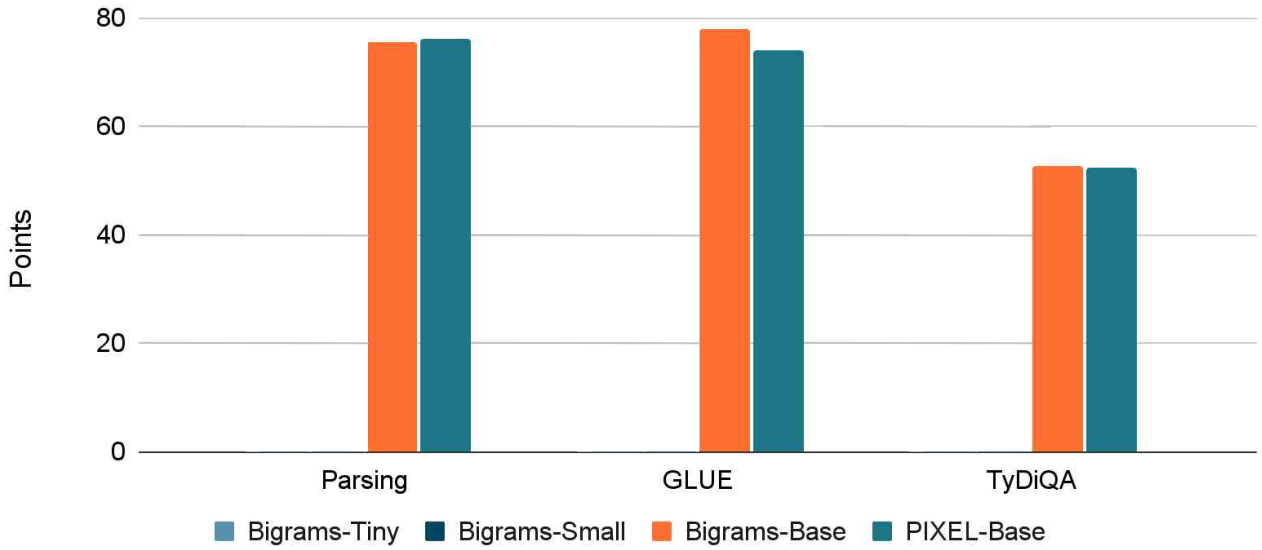
GLUE (revisited)



Bigram text rendering produces better pixel models

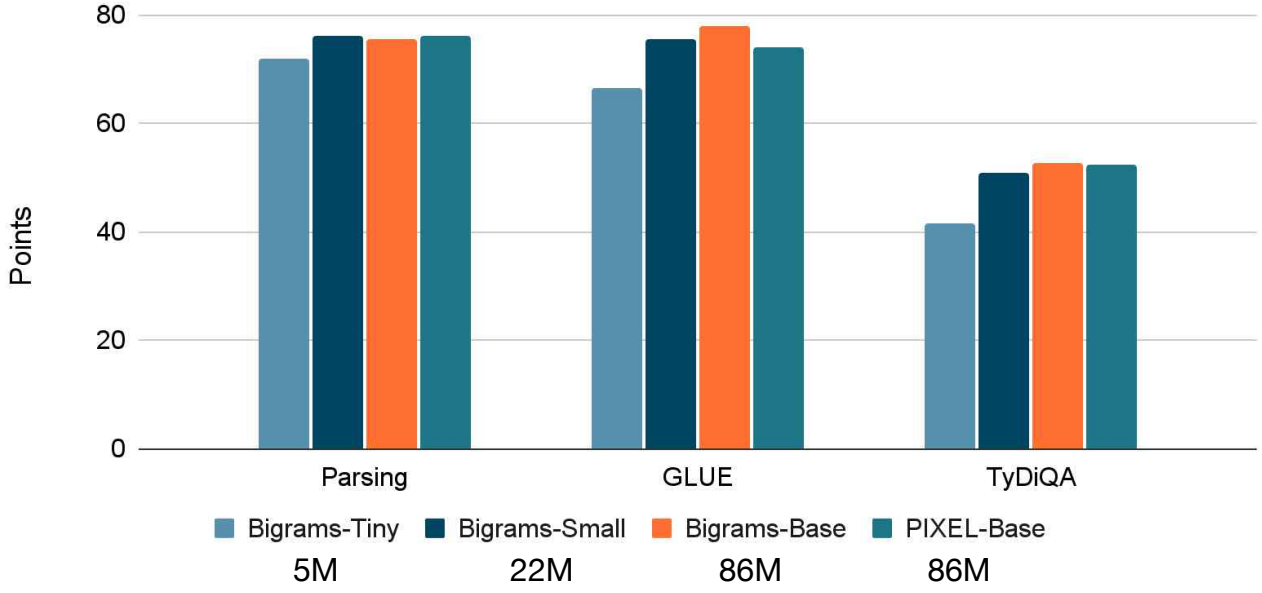
Scaling Down ↓

- Better text rendering can create effective models at smaller scales



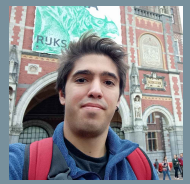
Scaling Down ↓

- Better text rendering can create effective models at smaller scales



Multilingual Pretraining for Pixel Language Models

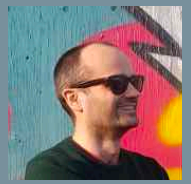
arXiv:2505.21265


I. Kesen


J. F. Lotz

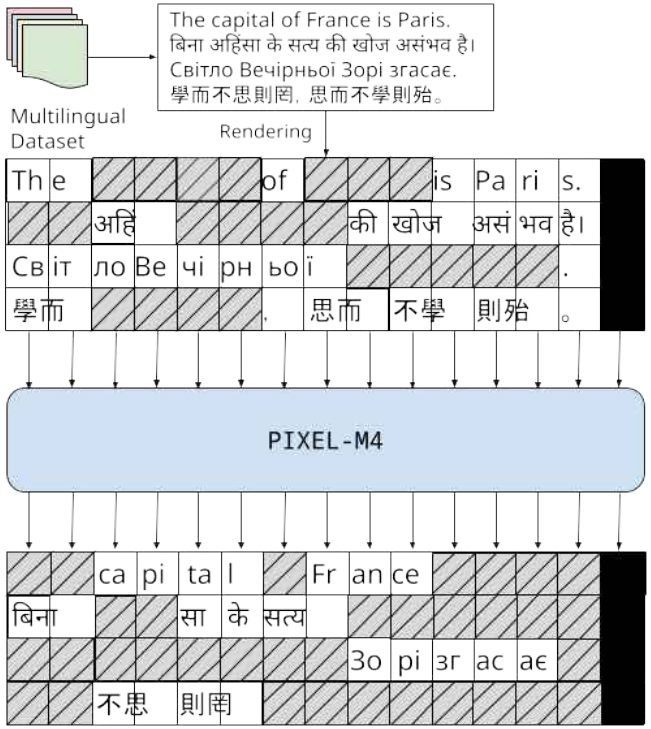

I. Ziegler


P. Rust

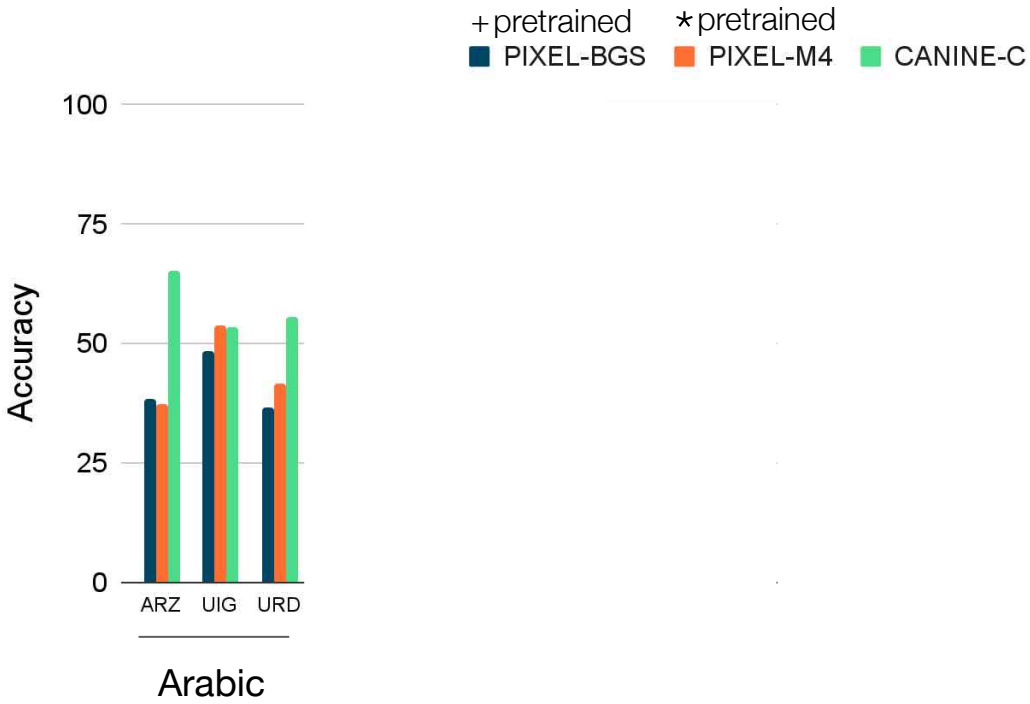

D. Elliott

PIXEL-M4

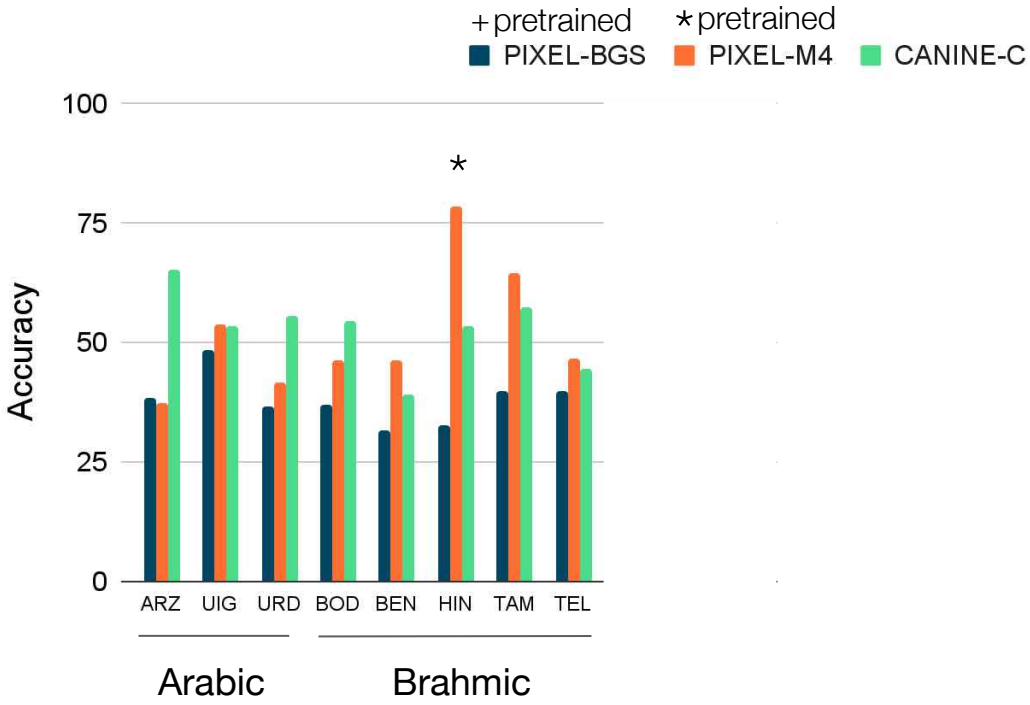
- Same architecture and hyperparameters as PIXEL-BIGRAMS
- But, pretrained on four visually diverse scripts sourced from mC4
  - Latin - English
  - Han - Simplified Chinese
  - Cyrillic - Ukrainian
  - Brahmic - Hindi



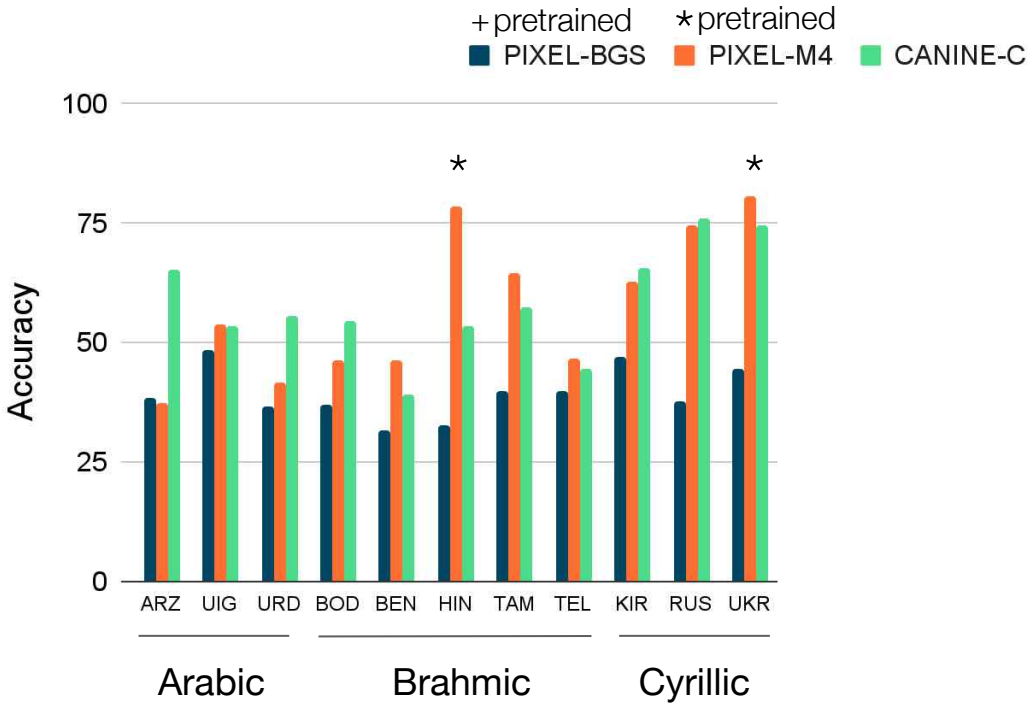
Text Classification on SIB-200



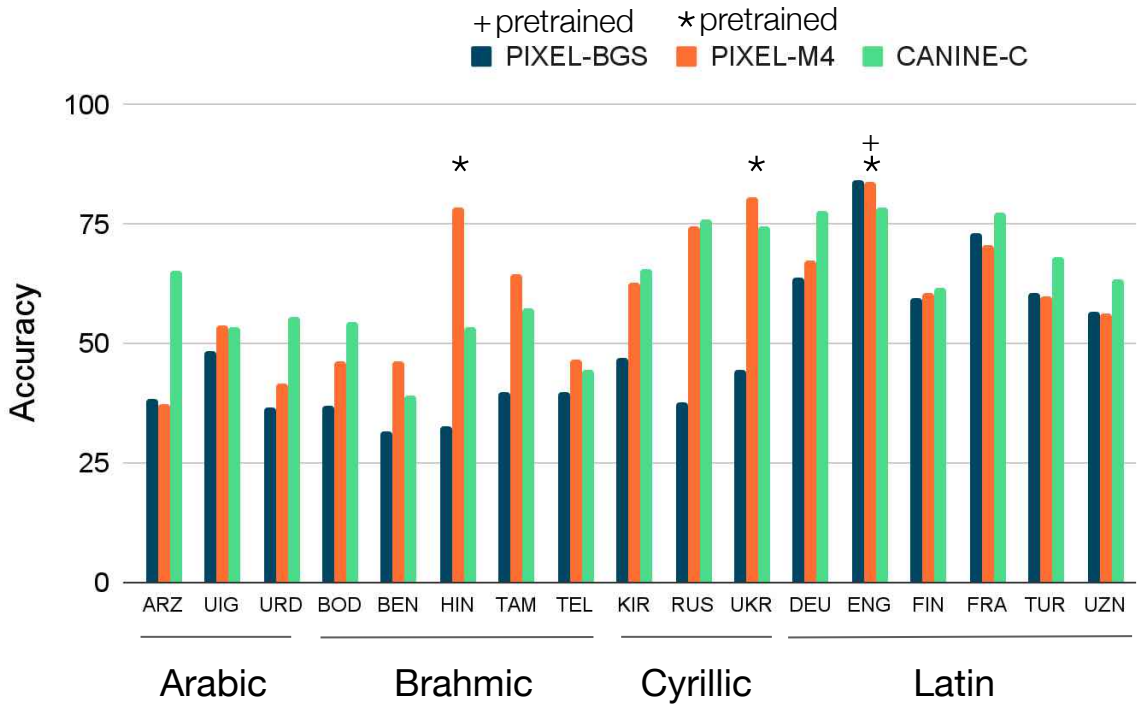
Text Classification on SIB-200



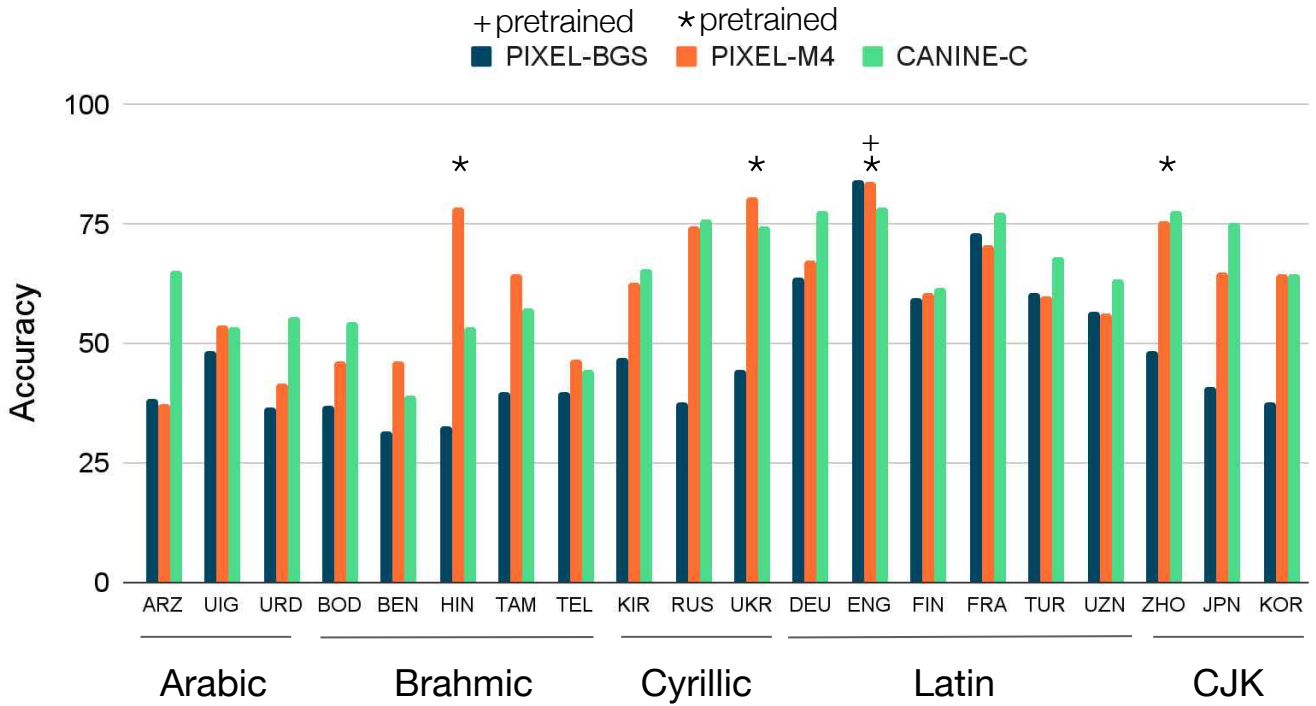
Text Classification on SIB-200



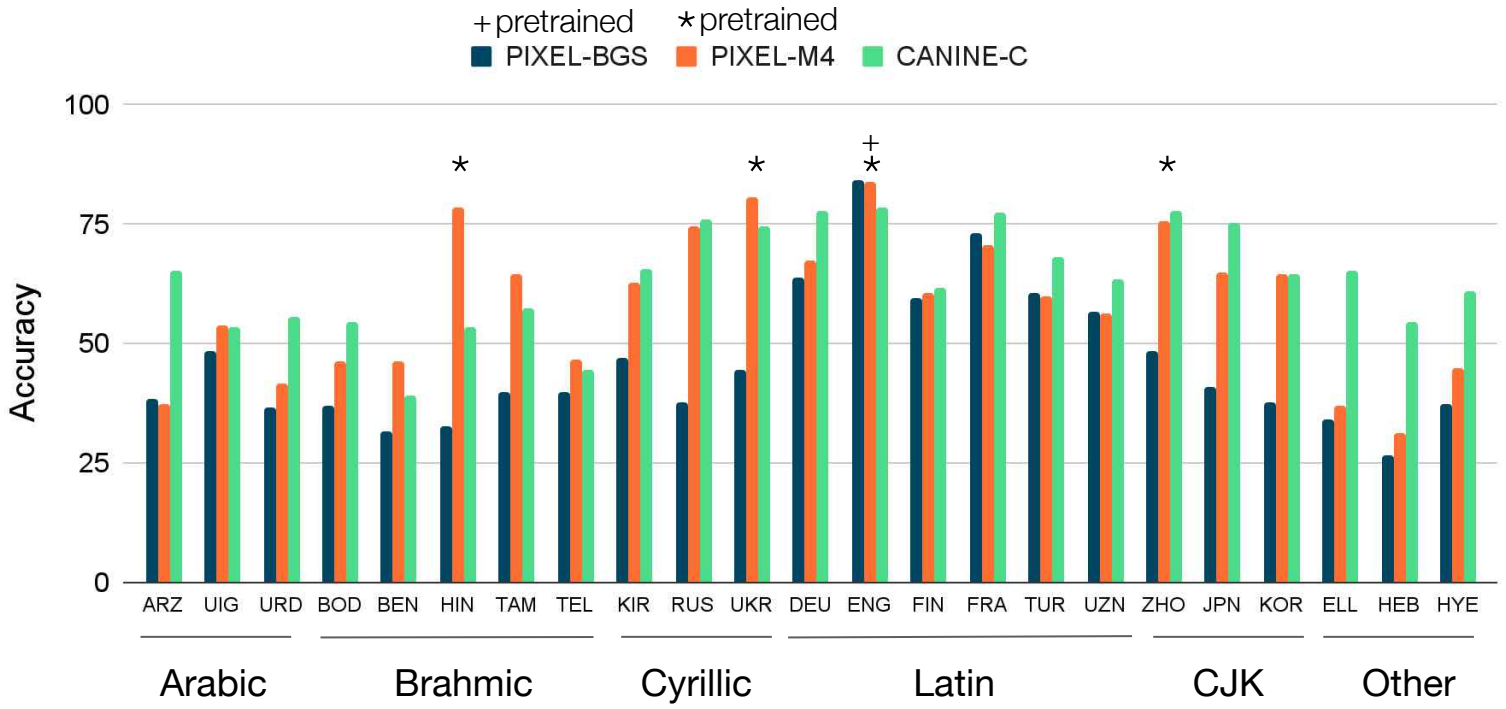
Text Classification on SIB-200



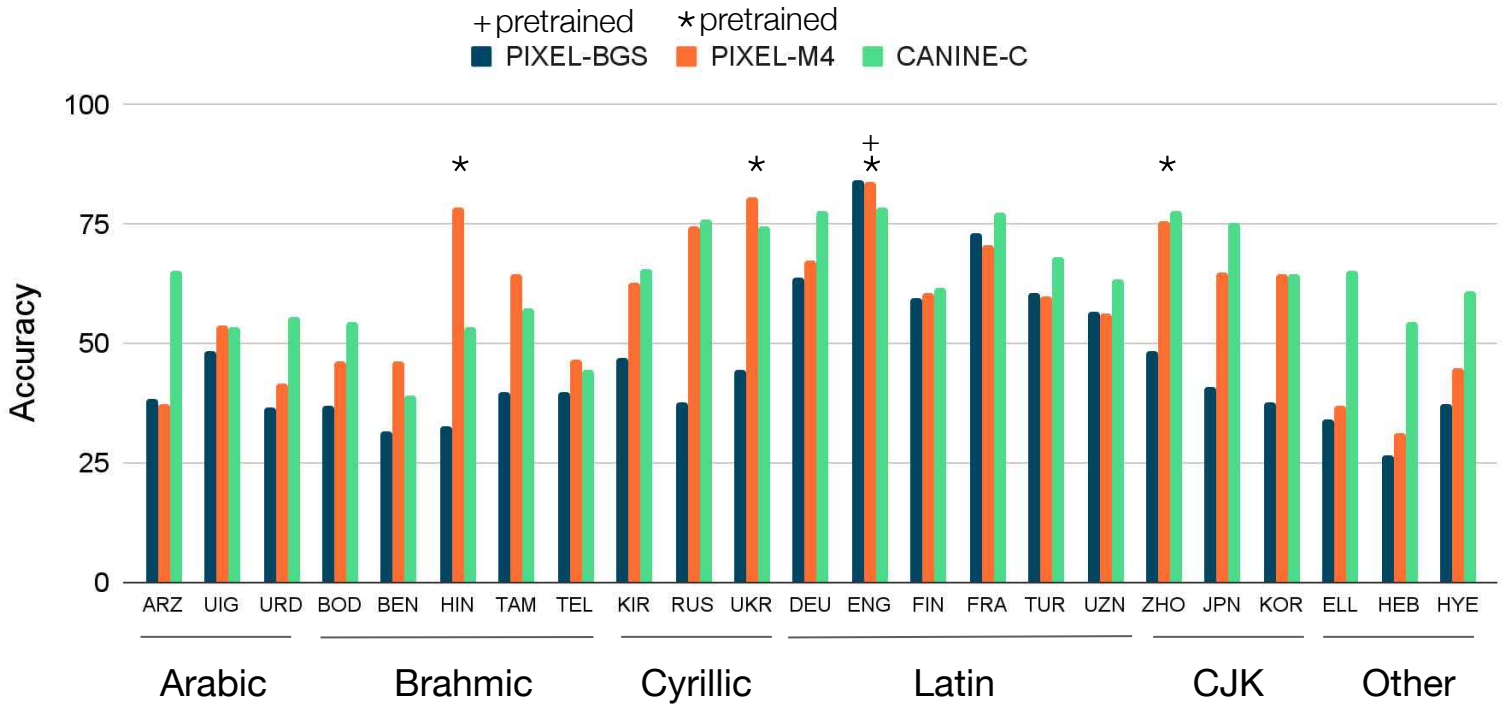
Text Classification on SIB-200



Text Classification on SIB-200

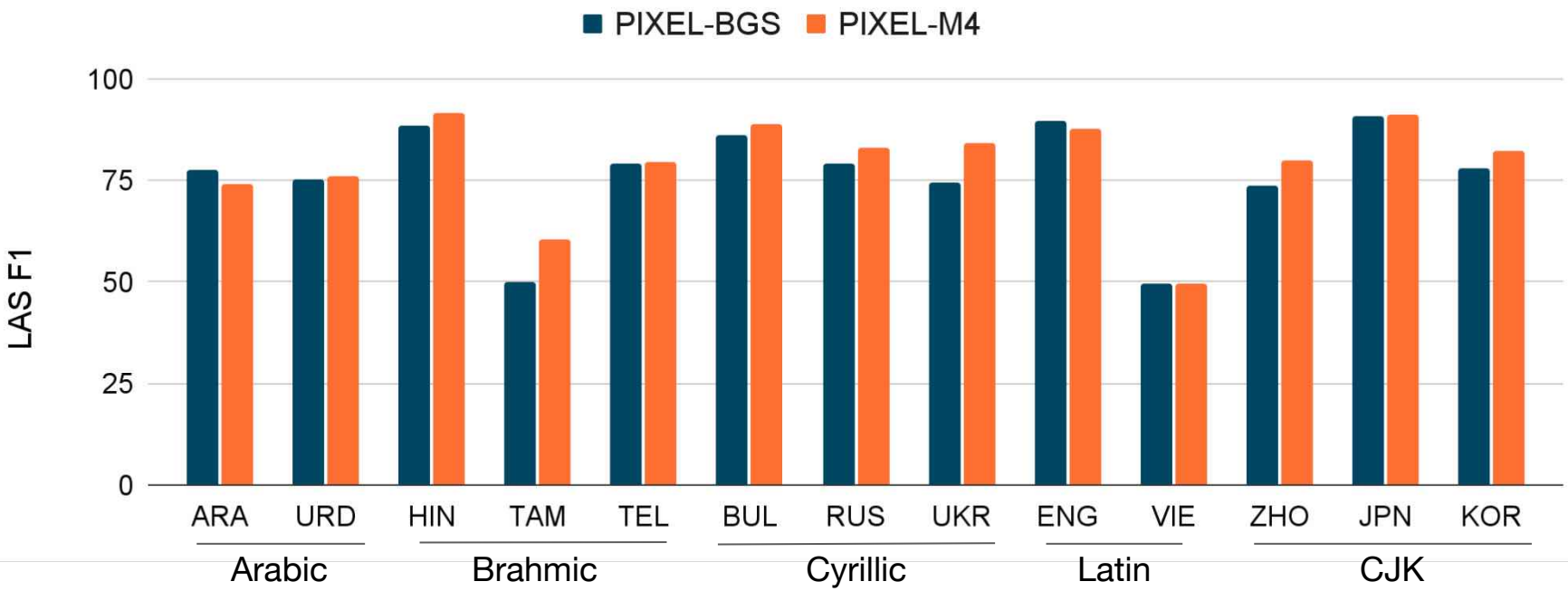


Text Classification on SIB-200



Multilingual pretraining is very helpful for sentence classification

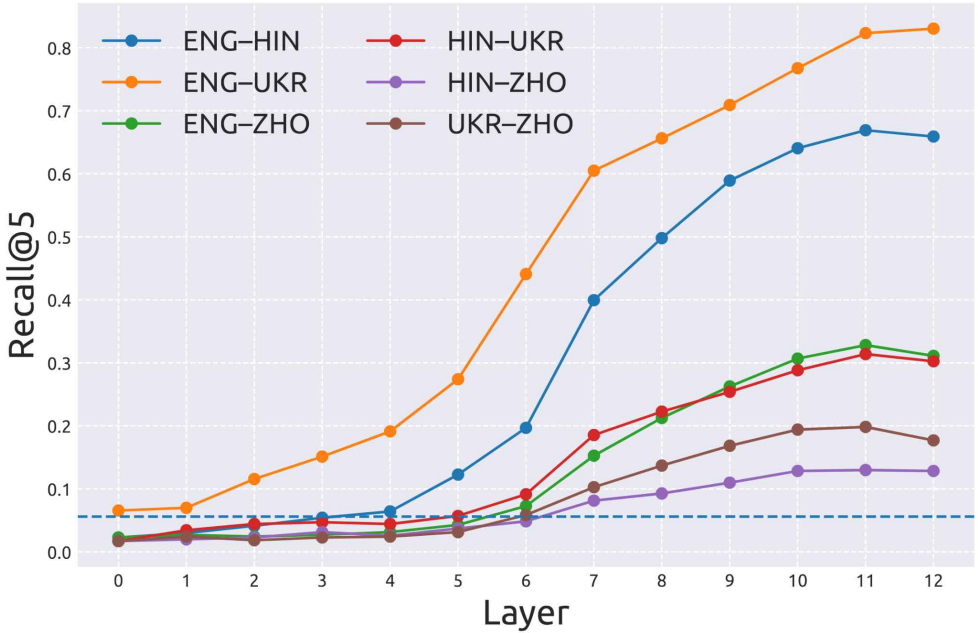
Dependency Parsing



Multilingual pretraining helps for non-Latin script languages

Zero-shot Sentence Retrieval Analysis

- Encode the sentences from the SIB-200 dataset in the four pretraining languages and measure the cosine similarity between the encoded data
- Multilingual pretraining leads to better “semantic” representation of text, but English seems to be *a pivot*

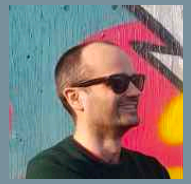


PHD: **P**ixel-based Language Modeling of **H**istorical **D**ocuments

EMNLP 2023


N. Borenstein


P. Rust


D. Elliott


I. Augenstein

Warning: This part of the talk contains dataset samples that are racist in nature.

Historical Document Processing

- Worldwide efforts to digitize historic documents (Groesen 2015)
- Typical pipeline for enabling access is:
  a. Scan documents into high-quality digital formats
  b. Perform OCR on those documents (one-off process)
  c. Search through documents using OCR annotations

What if we could do this without OCR?

Caribbean Newspapers, 1718–1876

- Collaboration with researchers that are interested in tracking newspapers notices about escaped slaves
  - What was the given name?
  - What reward was offered?
  - Who was the contact person?
- Dataset of 1.65M scanned pages


THE Royal Gazette.

BERMUDA COMMERCIAL AND GENERAL ADVERTISER AND RECORDER.

No. 24.—Vol. XXXIX. STATE SUPER VIAS ANTIQUAS. 24s. per Ann

Hamilton, Bermuda, Tuesday, June 19, 1866.

Commissariat, Bermuda,

HOPS.

T. W. GOLDIE, D. C. G.

St. Georges, General Store.

THE UNDERSIGNED ARE RECEIVING

OXBORROW & HUGHES.

For Sale,

For Sale, BY "ELIZA BARSS," AND IN STORE.

B. W. WALKER.

O. C. DUNSCOMBE

Ex barque Eliza Barss, FROM NEW YORK, 10 Barrels TAR.

THE SUBSCRIBER HAS RECEIVED, SUMMER GOODS FROM LONDON.

FOSTER L. BONNELL

TEAS and COFFEE.

GOSLING BROTHERS,

Mechanics' Industrial EXHIBITION,

WEDNESDAY

C. W. GAUNTLETT,

ICE.

THE SUBSCRIBERS

GOSLING BROS.

VICTORIA HOTEL

SODA WATER.

Bottled Soda Water,

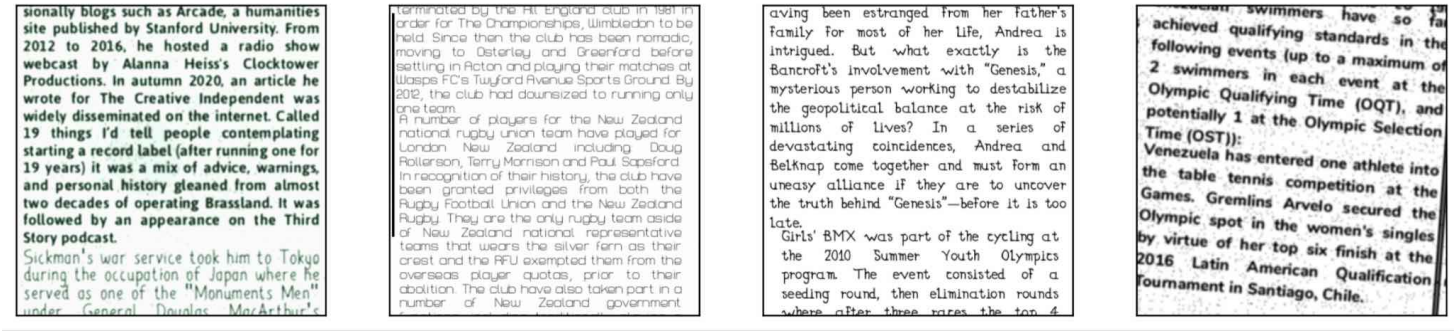
A Proclamation.

MILES GERALD KEON.

A Proclamation.

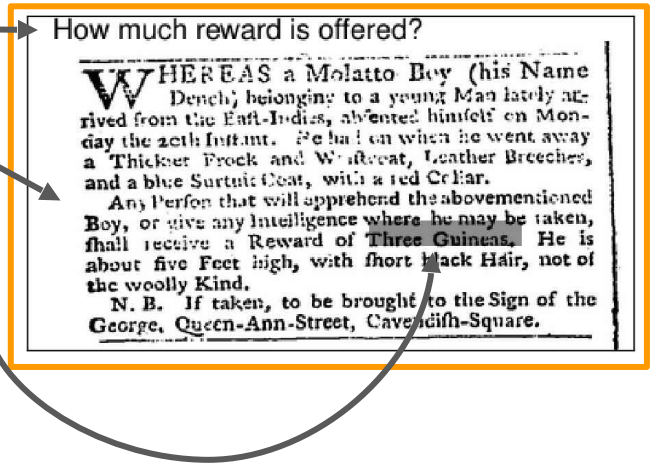
PHD: PIXEL for Historical Documents

- Historical document-aware Pretraining
  - Mixture of scanned newspapers and synthetic newspaper-like text generated from Wikipedia and Bookcorpus datasets
  - All input data is scaled to 368x368 and split into 16x16 patches

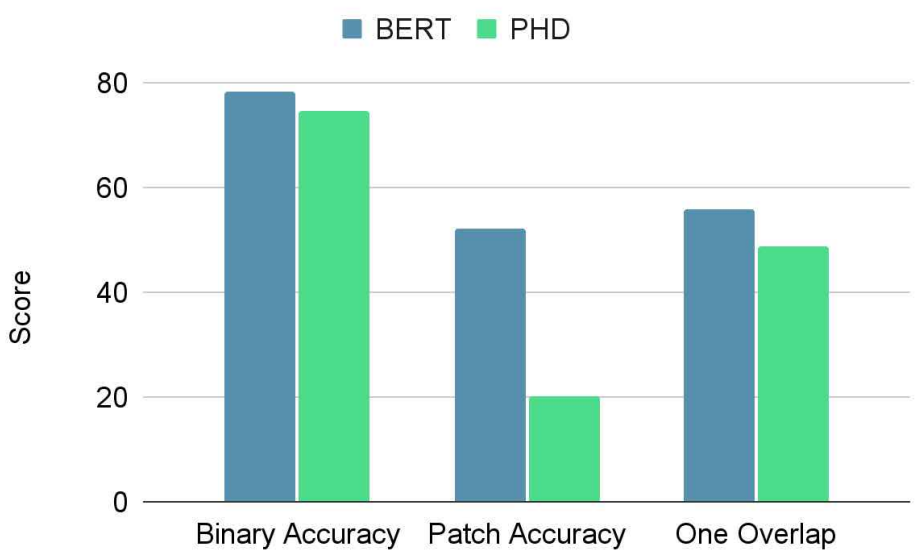


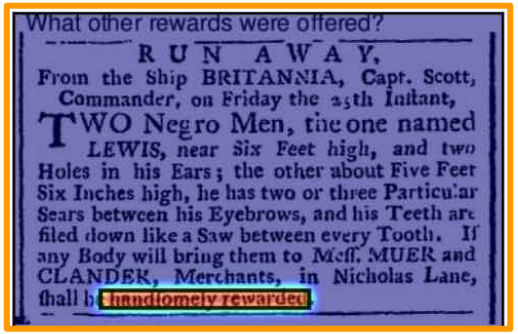
Visual Question Answering in Newspapers

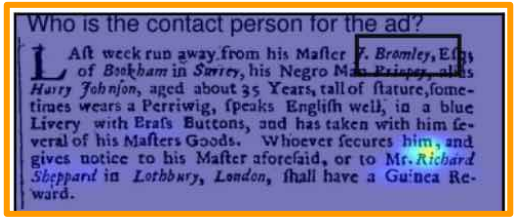
- Frame this as a Visual Question Answering Task
  - Render the question
  - Render the clipping on a canvas
  - Annotate context of answer
- Train the model to predict the label of the answer



Results







Surprisingly good performance compared to a model trained on manually transcribed text

Wrap-up

Overall Insights

1. General-purpose transferable representations of text can be learned directly from rendered text without any additional linguistic supervision

Overall Insights

1. General-purpose transferable representations of text can be learned directly from rendered text without any additional linguistic supervision

2. Careful consideration of how to structure the rendering of the text (“tokenization”) is important: bigrams >> unrestricted rendering

Overall Insights

1. General-purpose transferable representations of text can be learned directly from rendered text without any additional linguistic supervision

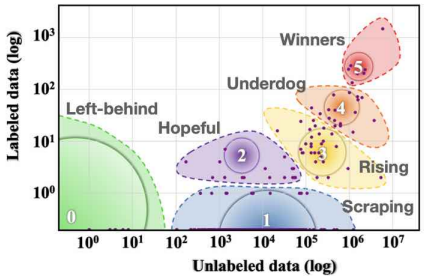
2. Careful consideration of how to structure the rendering of the text ("tokenization") is important: bigrams >> unrestricted rendering

3. Multilingual pretraining without any other architectural changes can improve performance on both seen and unseen scripts

Looking Ahead

- We need a clearer vision of *who* is the target audience of tokenization-free language models

(Data and groupings from Joshi et al. ACL 2020)



Class	Speakers
0	1.2B
1	30M
2	5.7M
3	1.8B
4	2.2B
5	2.5B

- We need a better definition of *unseen language*
  - writing script / language family / orthographic similarity
- Scale up the multilingual pretraining of pixel models
- Better understanding of the type of language learned with pixels
- Develop new methods for building pixel-token hybrids

The Bigger Question

- Masked Language Modelling is **classic distributional semantics** because it models the identity of a masked work, given the unmasked context.
- Why is it possible to learn a good model by predicting pixels?

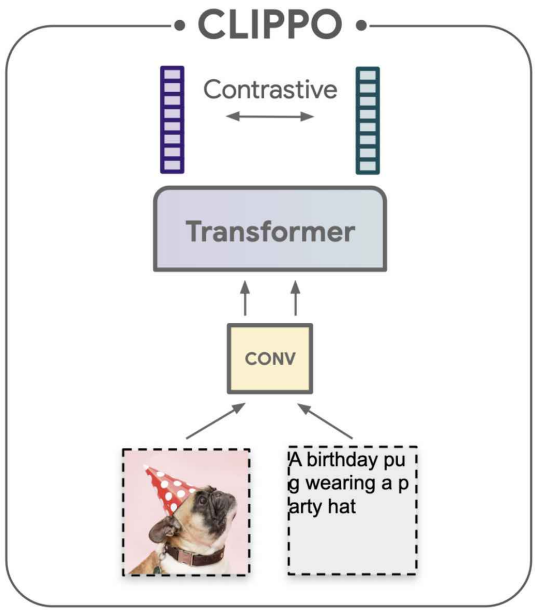
$$-\sum_{m \in M} \log p(x_m | \mathbf{x}_{\setminus m})$$

BERT: Masked Language Modelling

$$\frac{1}{M} \sum_{i=1}^{m} (X_i - \hat{X}_i | \mathbf{X}_{\setminus m})^2$$

PIXEL: Masked Autoencoding

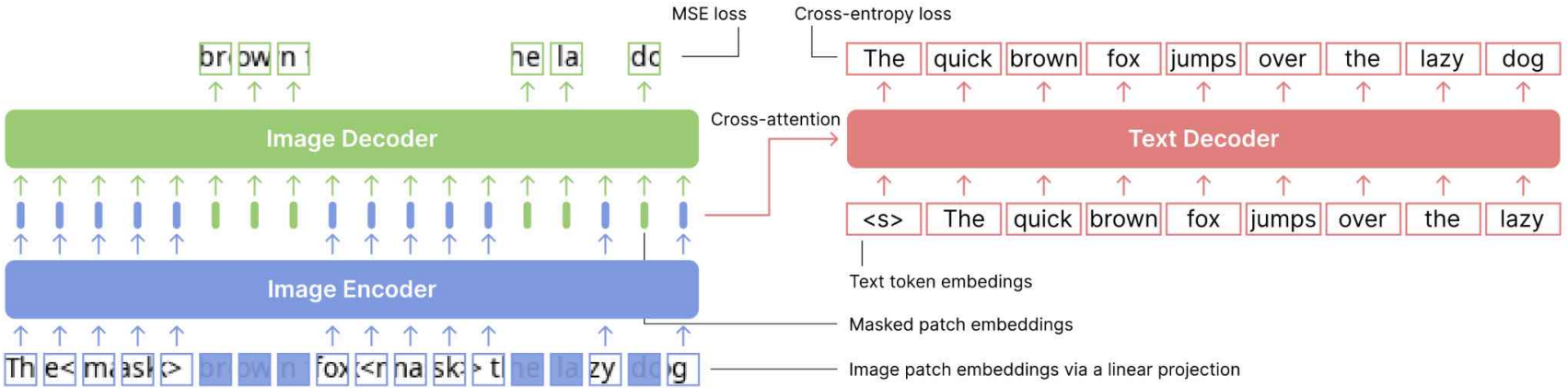
Joint Multimodal Processing



Tschannen et al. CVPR 2023. CLIPPO: Image-and-Language Understanding from Pixels Only.

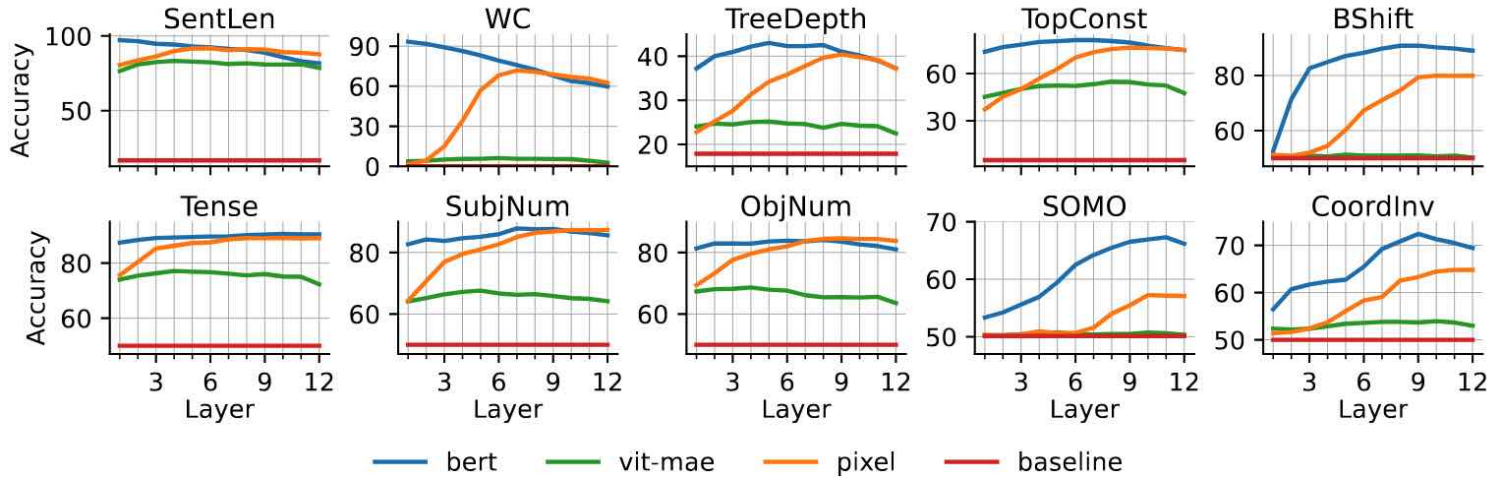
Patch and Text Prediction

- Combine patch and token prediction



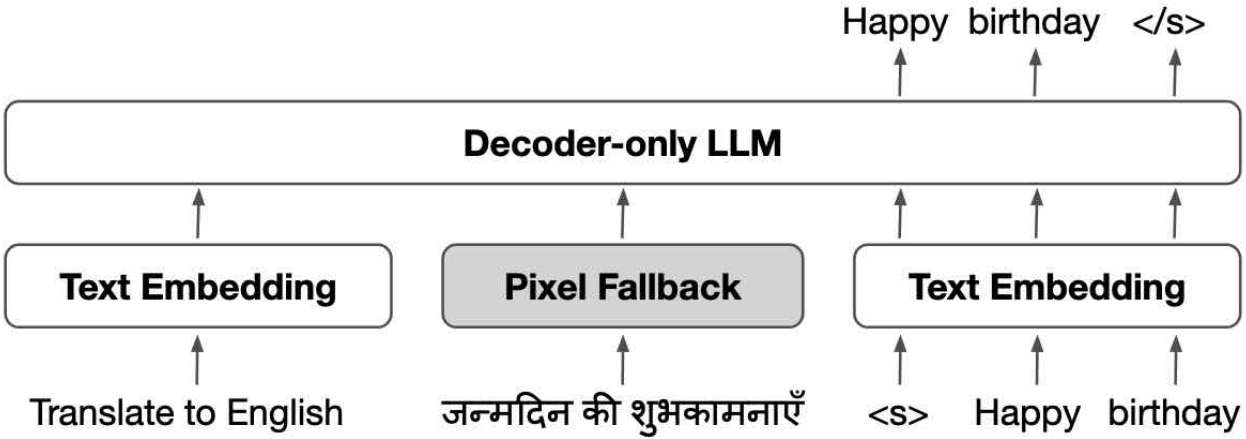
Pixology

- What linguistic knowledge is learned by pixel language models?



Combining Pixels and Tokens

- Handle sub-optimally covered inputs using pixel representations



Conclusions

- PIXEL is a different type of language model that tackles the open vocabulary problem using rendered text
    1. This enables high-quality transfer to different scripts
        - New, unseen languages
        - Different fonts in existing languages
    2. Compact models with as few as 5M parameters
    3. Multilingual pretraining improves performance
    4. Natural interface to scanned documents

References

- P. Rust, J. F. Lotz, E. Bugliarello, E. Salesky, M. de Lhoneux, and D. Elliott. Language Modelling with Pixels. ICLR 2023.
- N. Borenstein, P. Rust, D. Elliott, and I. Augenstein. PHD: Pixel-Based Language Modeling of Historical Documents. EMNLP 2023
- J. F. Lotz, E. Salesky, P. Rust, and D. Elliott. Text Rendering Strategies for Pixel Language Models. EMNLP 2023.
- I. Kesen, J. F. Lotz, I. Ziegler, P. Rust, D. Elliott. Multilingual Pretraining for Pixel Language Models. arXiv: 2505.21265