

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?



ACVSS 2025

Home

Apply & Attend

FAQ

Sponsoring

Past editions



AFRICAN COMPUTER VISION SUMMER SCHOOL

13-23 July 2025, AIMS, Kigali, Rwanda

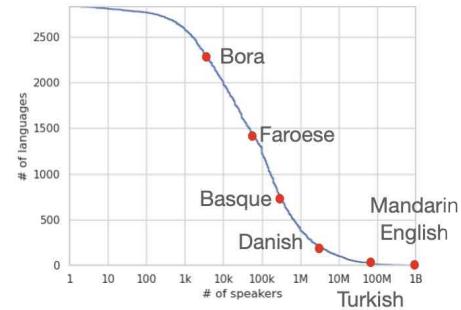
The **African Computer Vision Summer School** (ACVSS) unites outstanding African students and researchers with leading computer vision and AI experts.



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



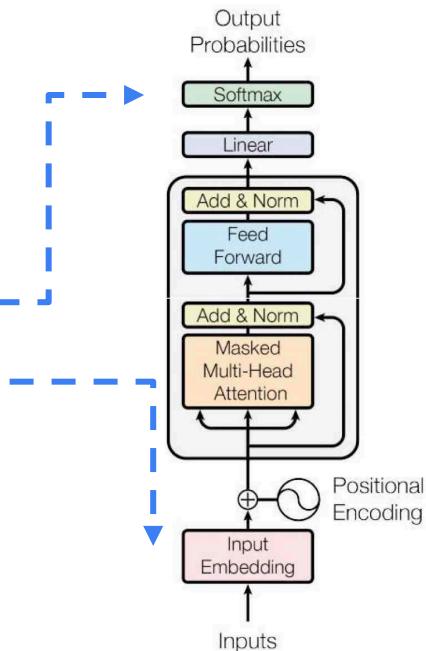
The Vocabulary Bottleneck

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

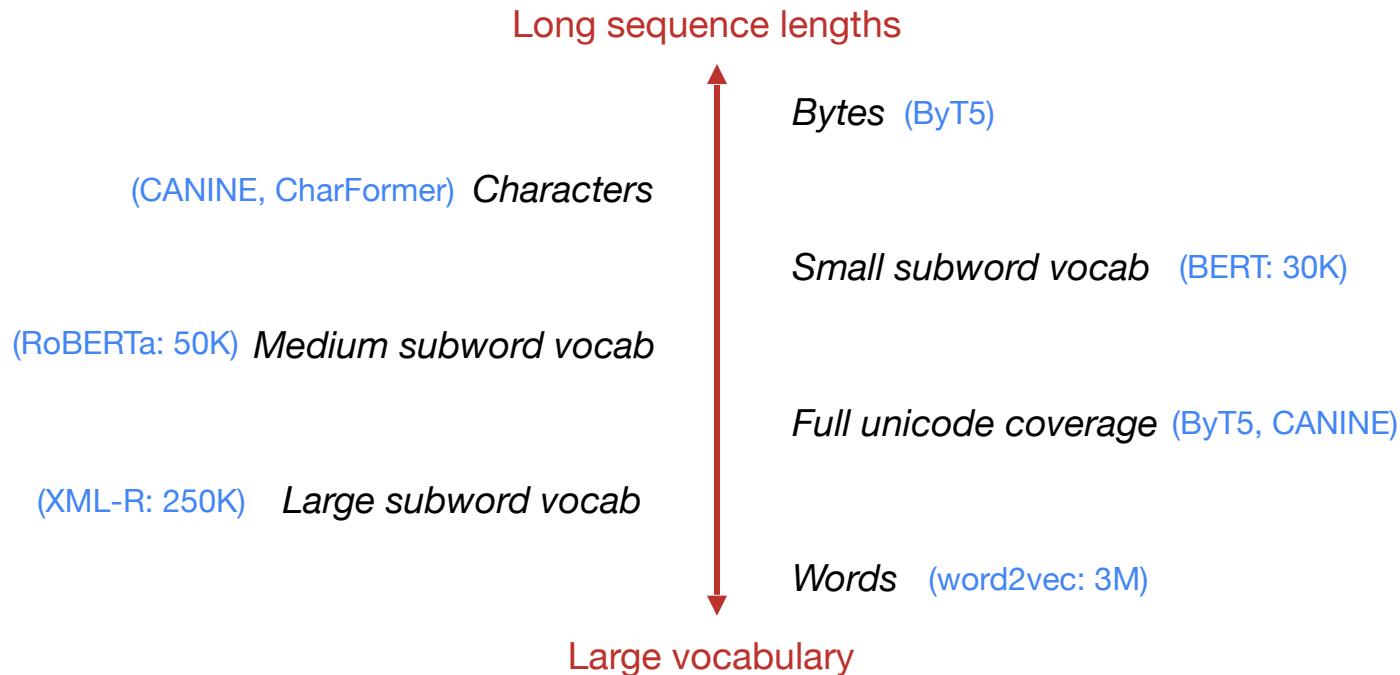
1. *Representational bottleneck* in the Embedding layer
2. *Computational bottleneck* in the token output layer

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

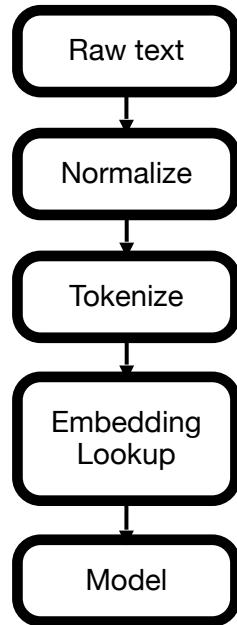
All words in vocabulary



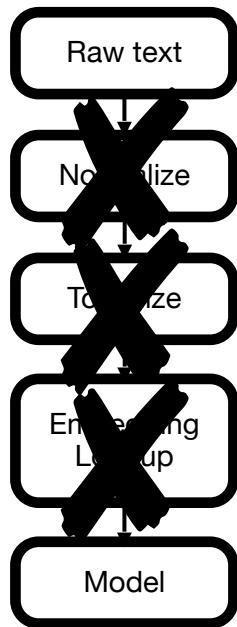
Where's the sweet spot?



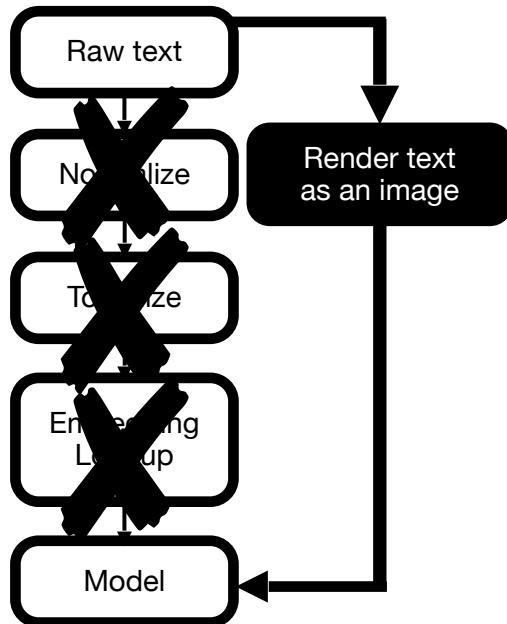
Alternative: treat language as vision



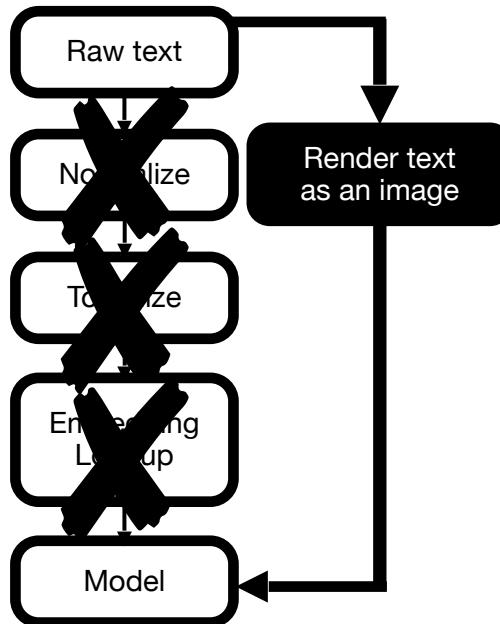
Alternative: treat language as vision



Alternative: treat language as vision

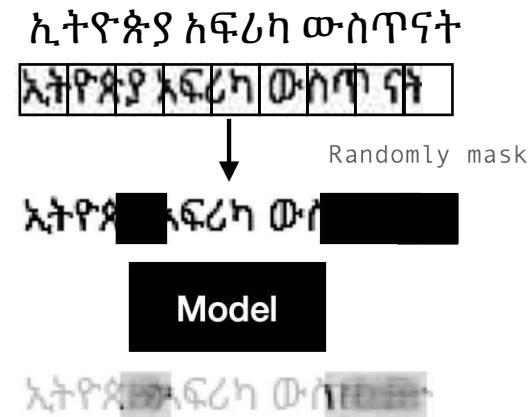
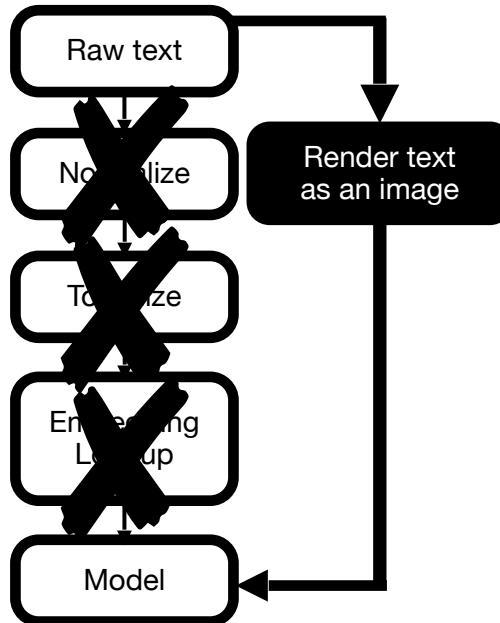


Alternative: treat language as vision

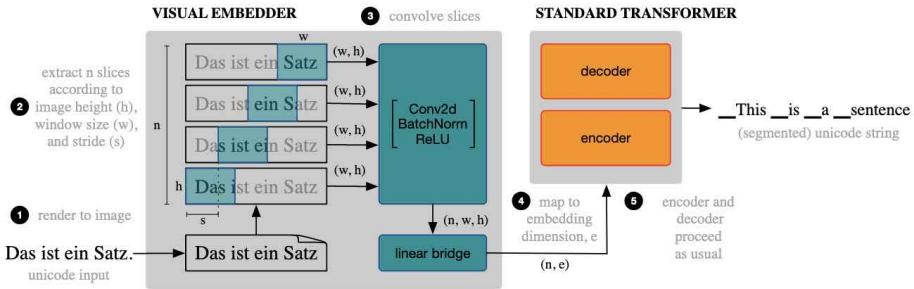


አትኩኩ አፍሪካ ወ-ሰጣጥ
አትኩኩ አፍሪካ ወ-ሰጣጥ

Alternative: treat language as vision



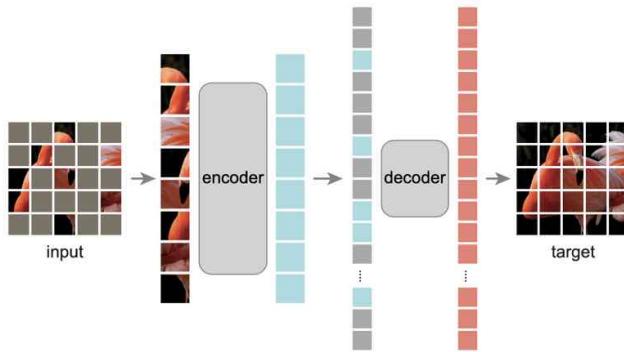
Inspiration



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

Montias ... pumps a lot of energy into his nicely nuanced narrative and surrounds himself with a cast of quirky -- but not stereotyped -- street characters.

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 **COO** enjoys eating warm oatmeal for lunch and dinner.

The Model

16pixel x 16pixel patch

Google Noto Fonts

PyGame / PangoCairo

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



1 Render Text as Image



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

The Model

16pixel x 16pixel patch

Google Noto Fonts

PyGame / PangoCairo

2 Projection + Position Embedding



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

1 Render Text as Image

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

The Model



- 3 CLS Embedding & Span Mask m patches
- 2 Projection + Position Embedding

My cat enjoys eating warm oatmeal for lunch and dinner.

16pixel x 16pixel patch

Google Noto Fonts

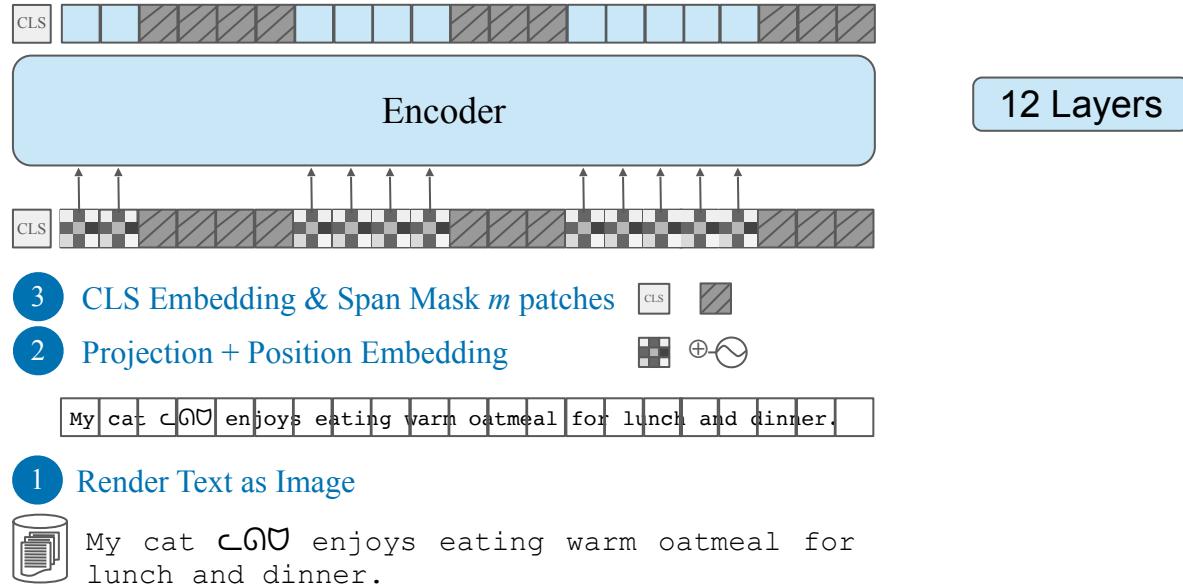
PyGame / PangoCairo



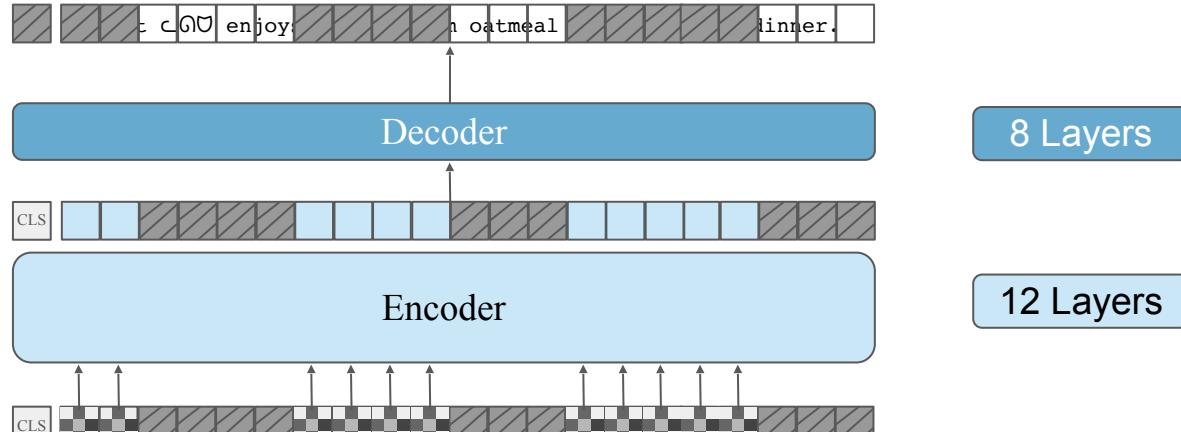
My cat enjoys eating warm oatmeal for lunch and dinner.

- 1 Render Text as Image

The Model



The Model



- 3 CLS Embedding & Span Mask m patches
- 2 Projection + Position Embedding

My cat c GO enjoys eating warm oatmeal for lunch and dinner.

16pixel x 16pixel patch

Google Noto Fonts

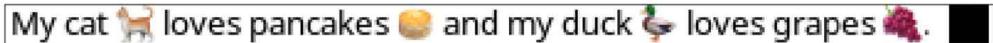
PyGame / PangoCairo

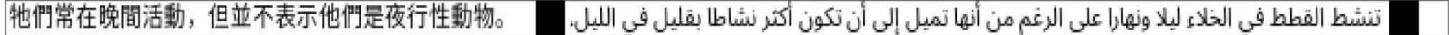


My cat c GO enjoys eating warm oatmeal for lunch and dinner.

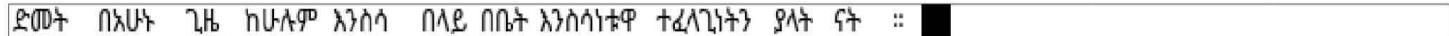
- 1 Render Text as Image

Flexible Text Renderer

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



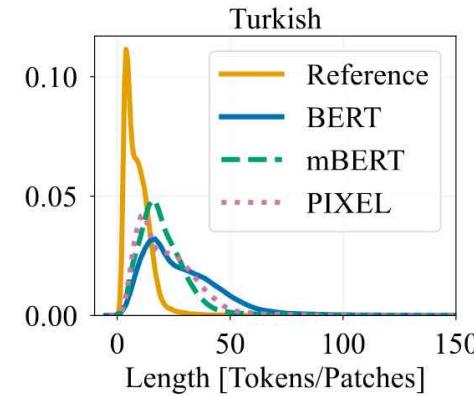
- Word-level rendering



Rendered Text is Compact

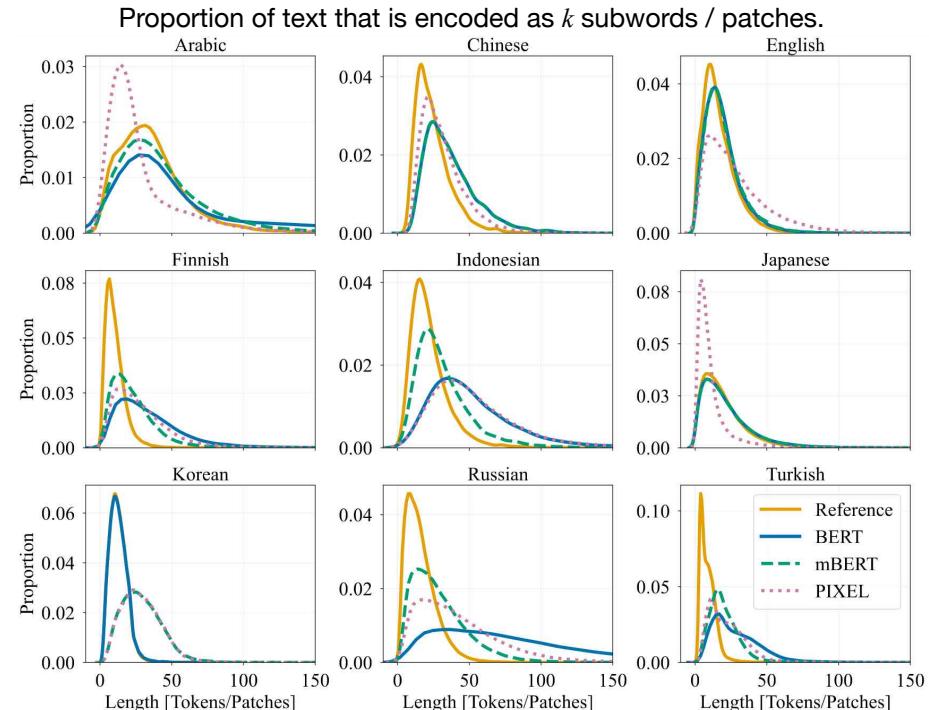
- PIXEL encoding produces sequence lengths that are at least as long as BERT.
 - Universal Dependencies datasets with human reference segmentations
 - No length penalty for languages, unlike some LLMs (Ahia et al. 2023)

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



Rendered Text is Compact

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



“Embedding” Layer

2 Projection + Position Embedding



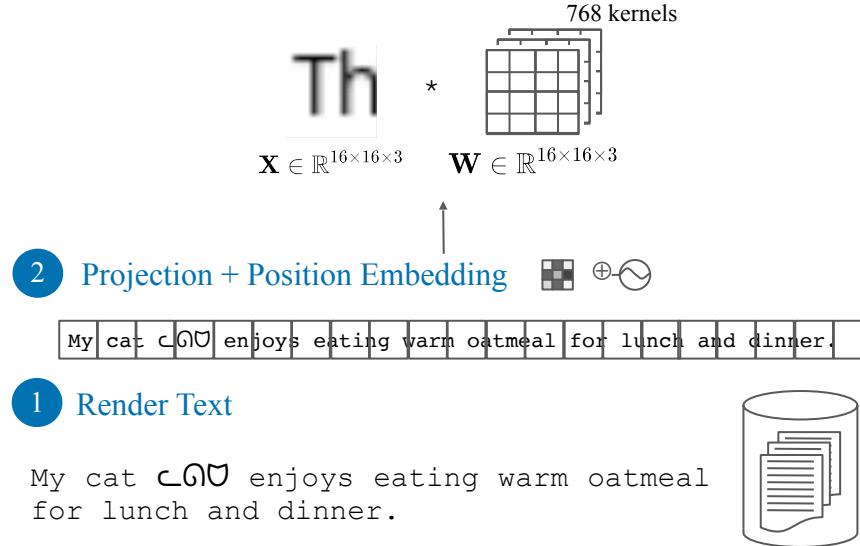
My cat 🐱 enjoys eating warm oatmeal for lunch and dinner.

1 Render Text

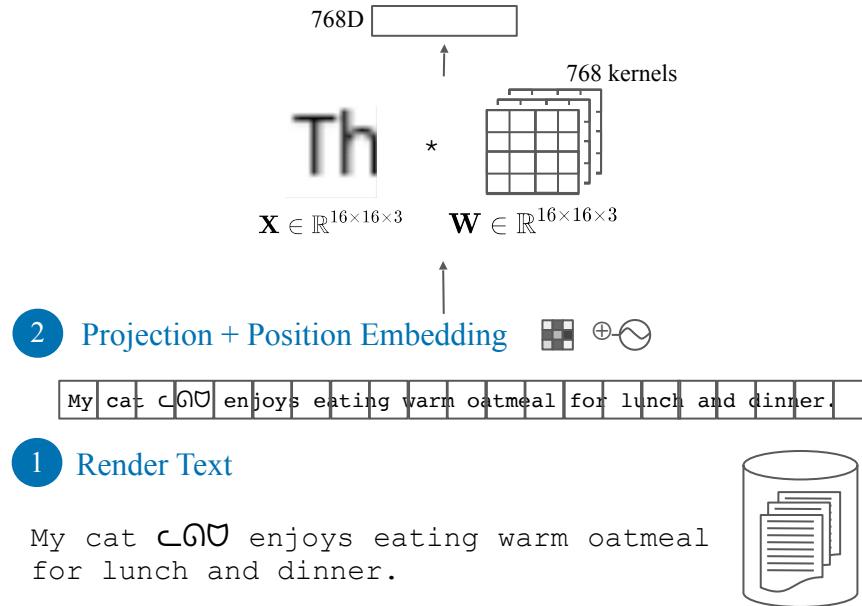
My cat 🐱 enjoys eating warm oatmeal
for lunch and dinner.



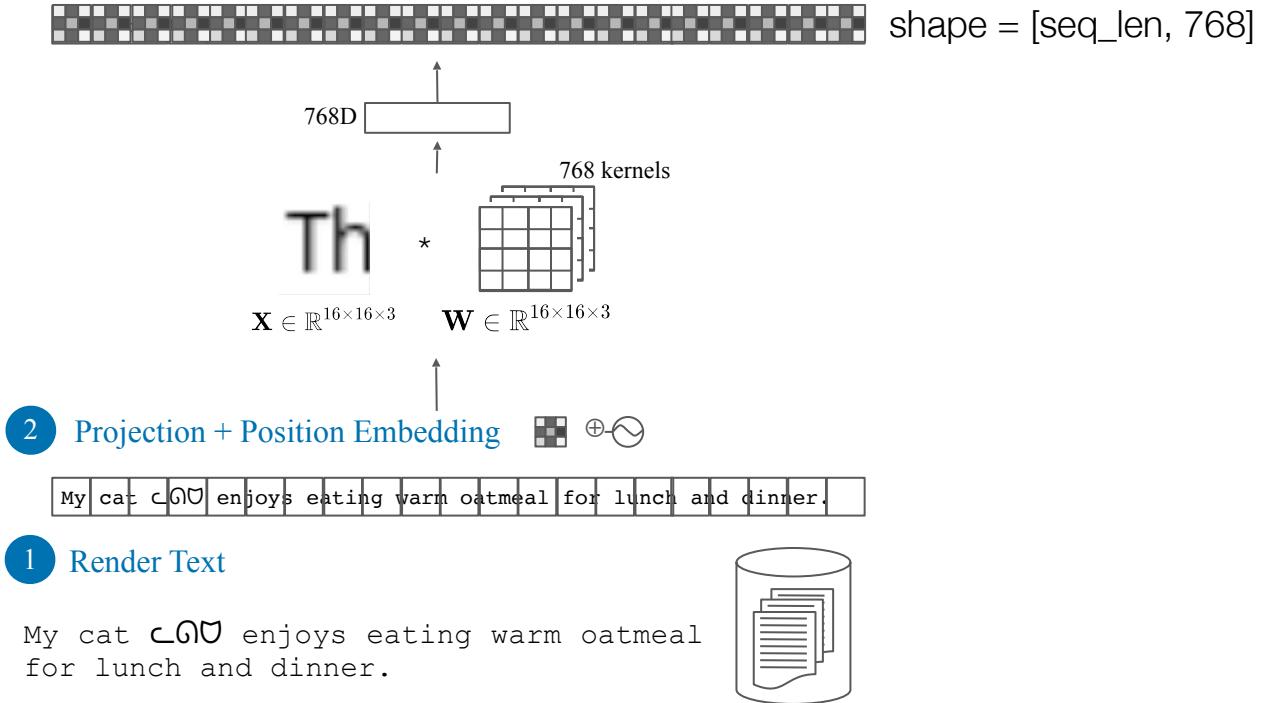
“Embedding” Layer



“Embedding” Layer

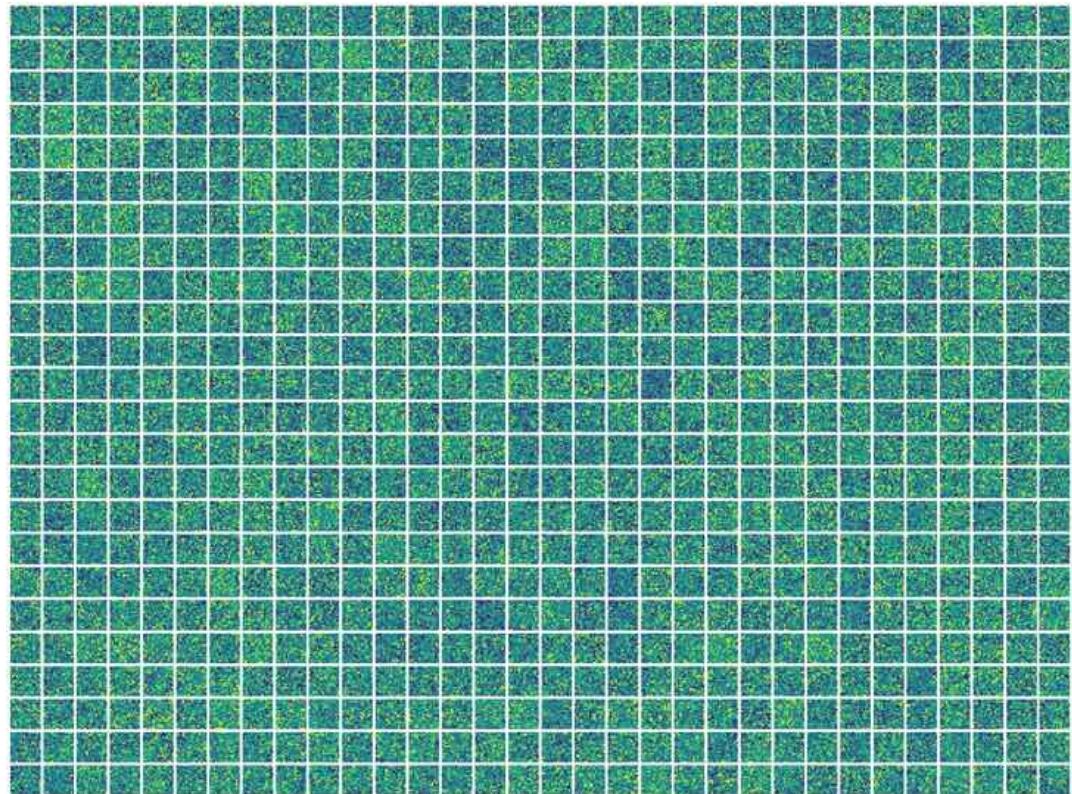


“Embedding” Layer



Visualization of Convolution Kernels

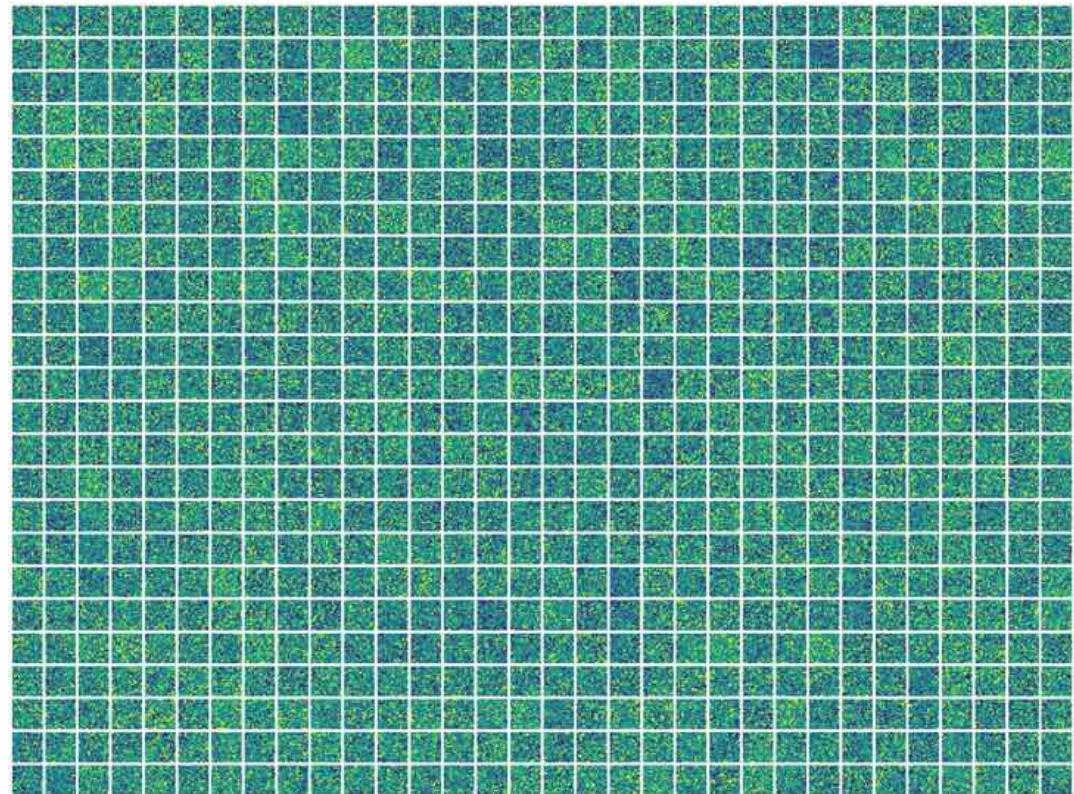
1. Some kernels learn about the presence / absence of any pixels.
2. Many kernels capture horizontal strokes
3. Only a few kernels capture curved shapes (*likely due to letters rendered across patch boundaries*)



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

Visualization of Convolution Kernels

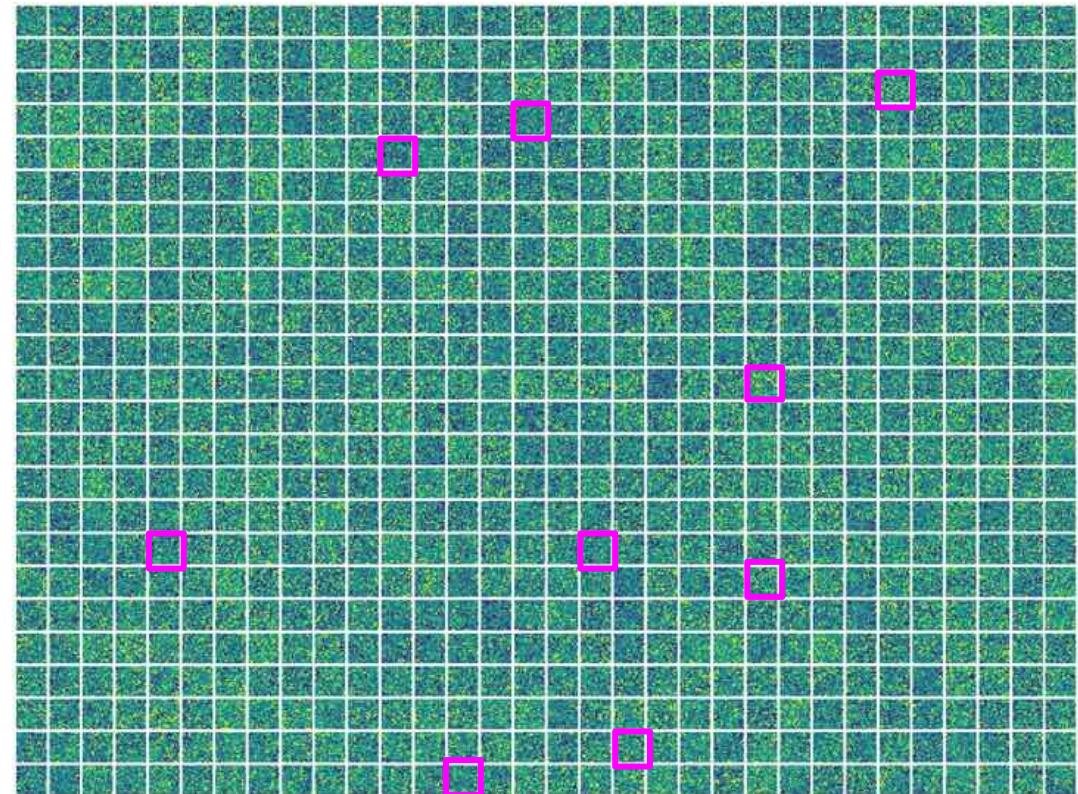
1. Some kernels learn about the presence / absence of any pixels.
2. Many kernels capture horizontal strokes
3. Only a few kernels capture curved shapes (*likely due to letters rendered across patch boundaries*)



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

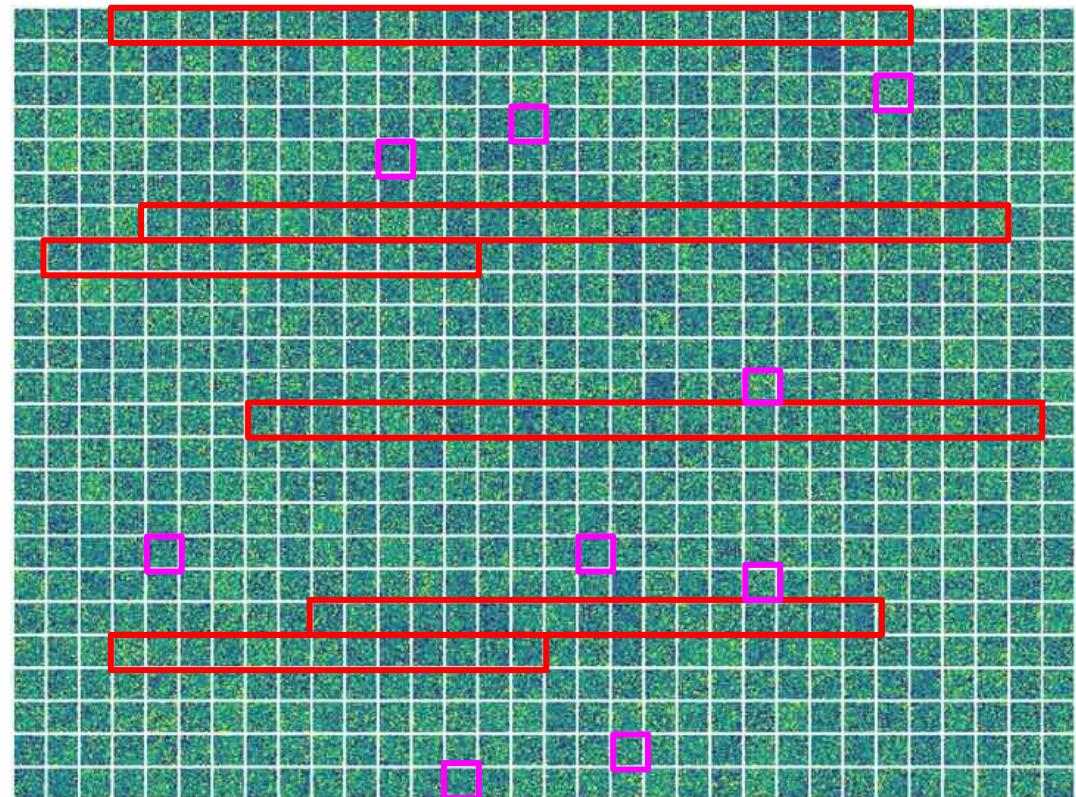
Visualization of Convolution Kernels

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*)



Visualization of Convolution Kernels

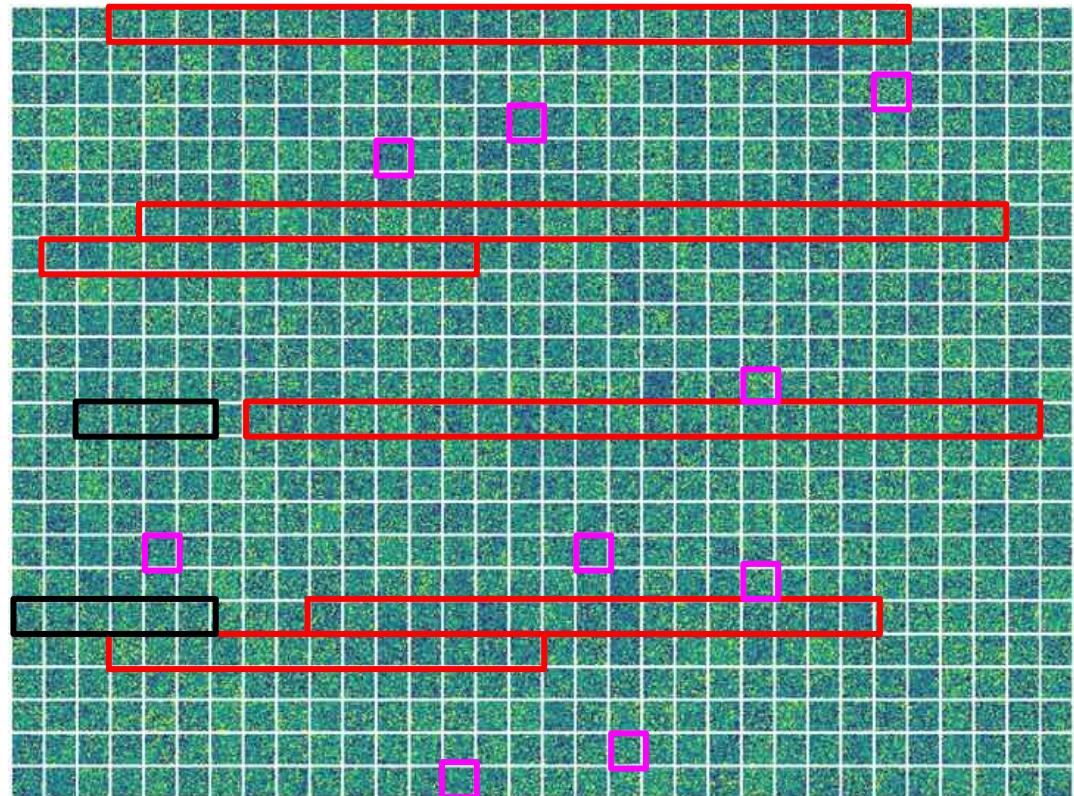
1. Some kernels learn about the presence / absence of any pixels.
2. Many kernels capture horizontal strokes
3. Only a few kernels capture curved shapes (*likely due to letters rendered across patch boundaries*)



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

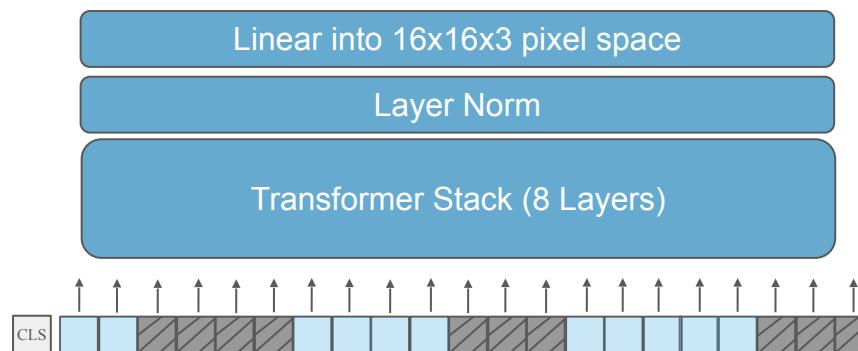
Visualization of Convolution Kernels

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*)

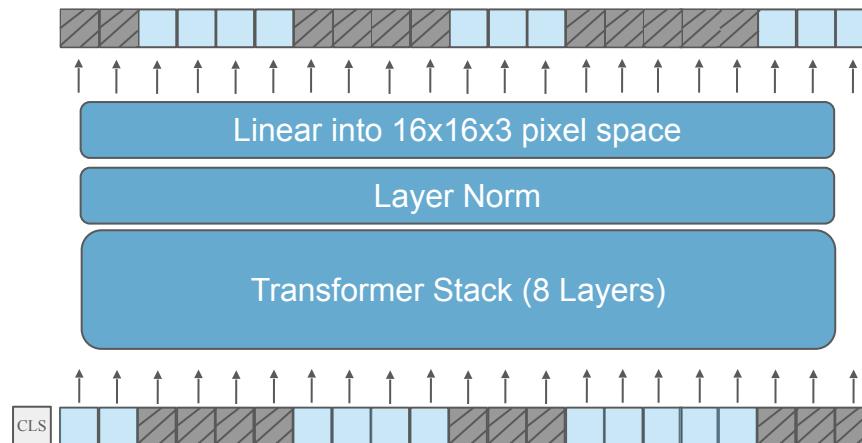


Evolution of Conv2D weights during pretraining step 10K-1M

Objective Function



Objective Function

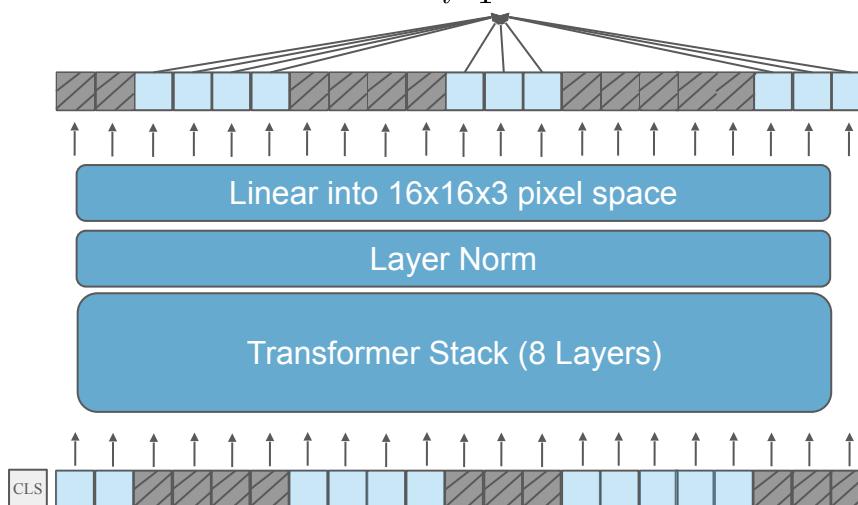


Objective Function

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

Penguins are designed to be streamlined and hydrodynamic, so having long legs would add expanding. Having short legs with webbed feet to act like runners, helps to give them that the do-like figure. If we compare bird anatomy with humans, we would see something peculiar. By taking a look at the side-by-side image in Figure 1, you can see how their leg bones relate to ours. What most people mistake for knees are actually the ankles of birds. This gives the illusion that bird knees bend opposite of ours. The knees are actually tucked up inside the boxes of the bird! So how does this look inside of a penguin? In the images below, you can see boxes surrounding the penguins' knees.

Penguins are designed to be streamlined and hydrodynamic, so having long legs would add expanding. Having short legs with webbed feet to act like runners, helps to give them that the do-like figure. If we compare bird anatomy with humans, we would see something peculiar. By taking a look at the side-by-side image in Figure 1, you can see how their leg bones relate to ours. What most people mistake for knees are actually the ankles of birds. This gives the illusion that bird knees bend opposite of ours. The knees are actually tucked up inside the boxes of the bird! So how does this look inside of a penguin? In the images below, you can see boxes surrounding the penguins' knees.

Penguins are designed to be streamlined and hydrodynamic, so having long legs would add expanding. Having short legs with webbed feet to act like runners, helps to give them that the do-like figure. If we compare bird anatomy with humans, we would see something peculiar. By taking a look at the side-by-side image in Figure 1, you can see how their leg bones relate to ours. What most people mistake for knees are actually the ankles of birds. This gives the illusion that bird knees bend opposite of ours. The knees are actually tucked up inside the boxes of the bird! So how does this look inside of a penguin? In the images below, you can see boxes surrounding the penguins' knees.

100K steps

500K steps

1M steps

A new type of generative model

Penguins are designed to be streamlined and hydrodynamic, so having long legs would add expanding. Having short legs with webbed feet to act like runners, helps to give them that the do-like figure. If we compare bird anatomy with humans, we would see something peculiar. By taking a look at the side-by-side image in Figure 1, you can see how their leg bones are opposite to ours. What most people mistake for knees are actually the ankles of birds. This gives the illusion that bird knees bend opposite of ours. The knees are actually tucked up inside the boxes of the bird! So how does this look inside of a penguin? In the images below, you can see boxes surrounding the penguins' knees.

Penguins are designed to be streamlined and hydrodynamic, so having long legs would add expanding. Having short legs with webbed feet to act like runners, helps to give them that the do-like figure. If we compare bird anatomy with humans, we would see something peculiar. By taking a look at the side-by-side image in Figure 1, you can see how their leg bones are opposite to ours. What most people mistake for knees are actually the ankles of birds. This gives the illusion that bird knees bend opposite of ours. The knees are actually tucked up inside the boxes of the bird! So how does this look inside of a penguin? In the images below, you can see boxes surrounding the penguins' knees.

Penguins are designed to be streamlined and hydrodynamic, so having long legs would add expanding. Having short legs with webbed feet to act like runners, helps to give them that the do-like figure. If we compare bird anatomy with humans, we would see something peculiar. By taking a look at the side-by-side image in Figure 1, you can see how their leg bones are opposite to ours. What most people mistake for knees are actually the ankles of birds. This gives the illusion that bird knees bend opposite of ours. The knees are actually tucked up inside the boxes of the bird! So how does this look inside of a penguin? In the images below, you can see boxes surrounding the penguins' knees.

100K steps

500K steps

1M steps

A new type of generative model

Penguins are designed to be streamlined and hydrodynamic, so having long legs would add expanding. Having short legs with webbed feet to act like runners, helps to give them that the do-like figure. If we compare bird anatomy with humans, we would see something peculiar. By taking a look at the side-by-side image in Figure 1, you can see how their leg bones are close to ours. What most people mistake for knees are actually the ankles of birds. This gives the illusion that bird knees bend opposite of ours. The knees are actually tucked up inside the bones of the bird! So how does this look inside of a penguin? In the images below, you can see boxes surrounding the penguins' knees.

Penguins are designed to be streamlined and hydrodynamic, so having long legs would add expanding. Having short legs with webbed feet to act like runners, helps to give them that the do-like figure. If we compare bird anatomy with humans, we would see something peculiar. By taking a look at the side-by-side image in Figure 1, you can see how their leg bones are close to ours. What most people mistake for knees are actually the ankles of birds. This gives the illusion that bird knees bend opposite of ours. The knees are actually tucked up inside the bones of the bird! So how does this look inside of a penguin? In the images below, you can see boxes surrounding the penguins' knees.

Penguins are designed to be streamlined and hydrodynamic, so having long legs would add expanding. Having short legs with webbed feet to act like runners, helps to give them that the do-like figure. If we compare bird anatomy with humans, we would see something peculiar. By taking a look at the side-by-side image in Figure 1, you can see how their leg bones are close to ours. What most people mistake for knees are actually the ankles of birds. This gives the illusion that bird knees bend opposite of ours. The knees are actually tucked up inside the bones of the bird! So how does this look inside of a penguin? In the images below, you can see boxes surrounding the penguins' knees.

100K steps

500K steps

1M steps

Pretraining

- **English Dataset:** English Wikipedia and Books Corpus
- **Masking:** 25% Span Masking
- **Maximum sequence length:** 529 patches (16×8464 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: Wàn lǐ Chángchéng)

Downstream Tasks

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

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

Downstream Tasks

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

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

Similar pretraining setup

Downstream Tasks

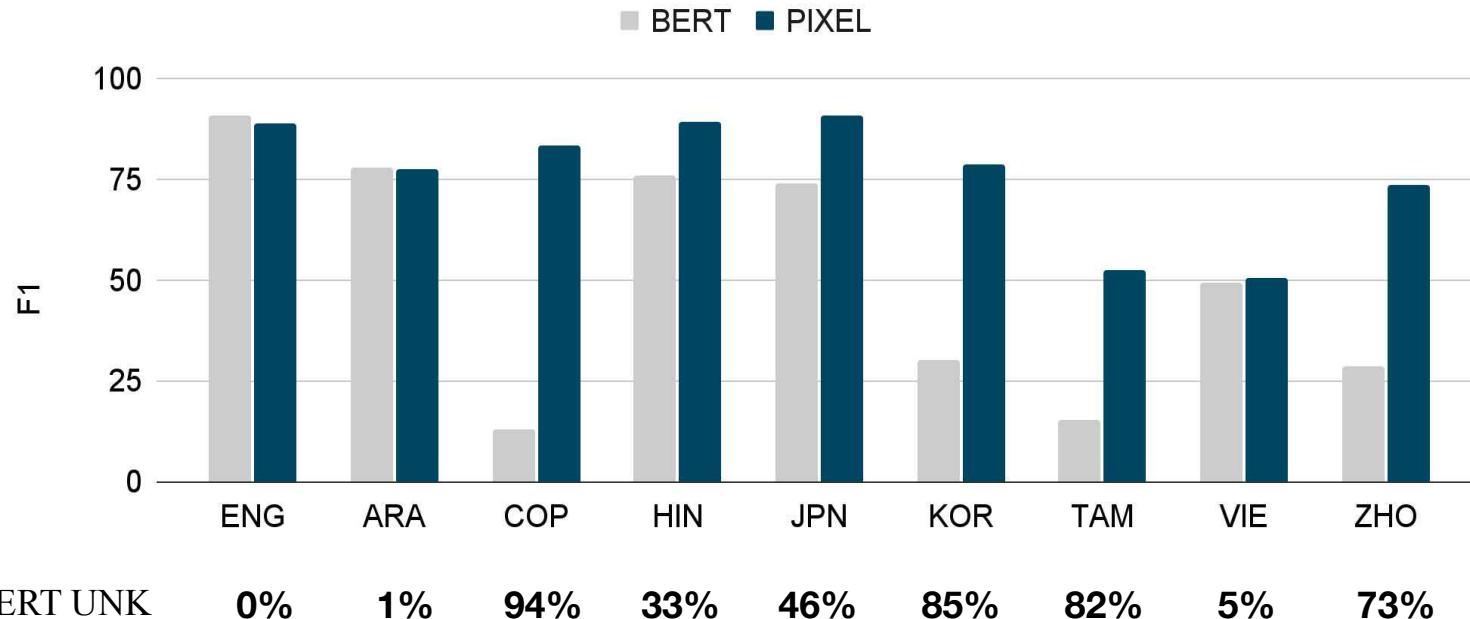
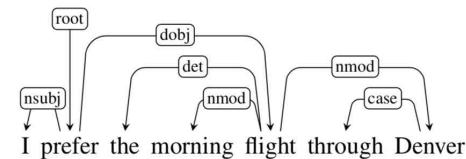
- **Datasets:** Universal Dependencies, MasakhaNER, GLUE, Zeroé
- **Models:**

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

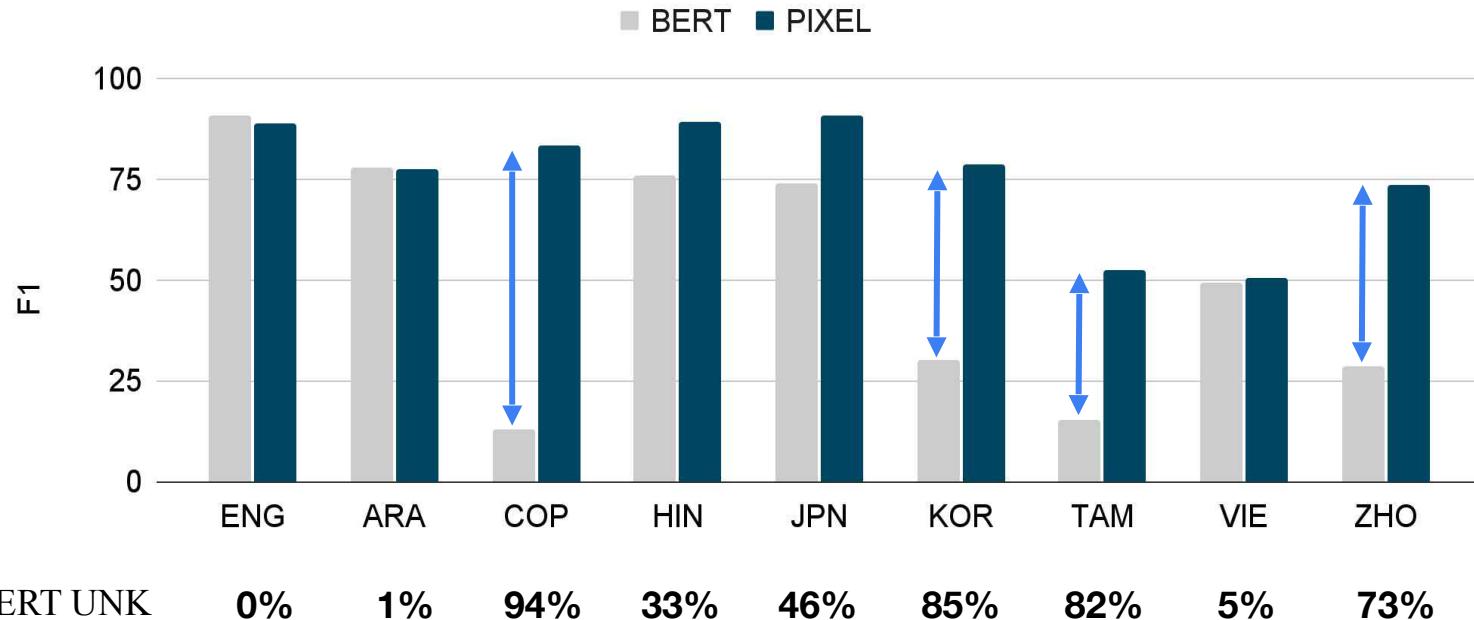
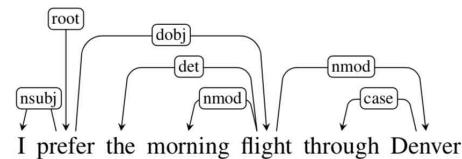
Similar pretraining setup

Tries to solve the same problem using UTF-32

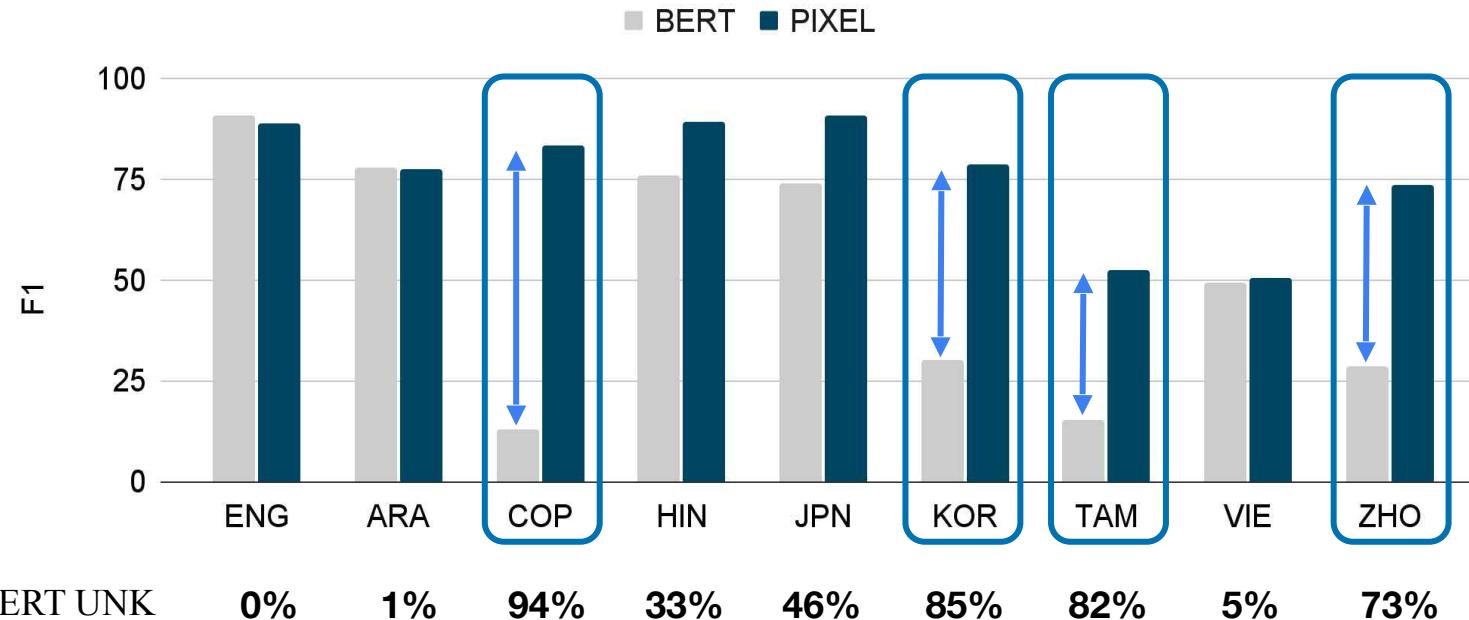
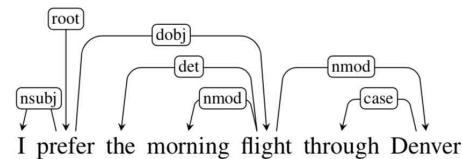
Dependency Parsing Results



Dependency Parsing Results



Dependency Parsing Results

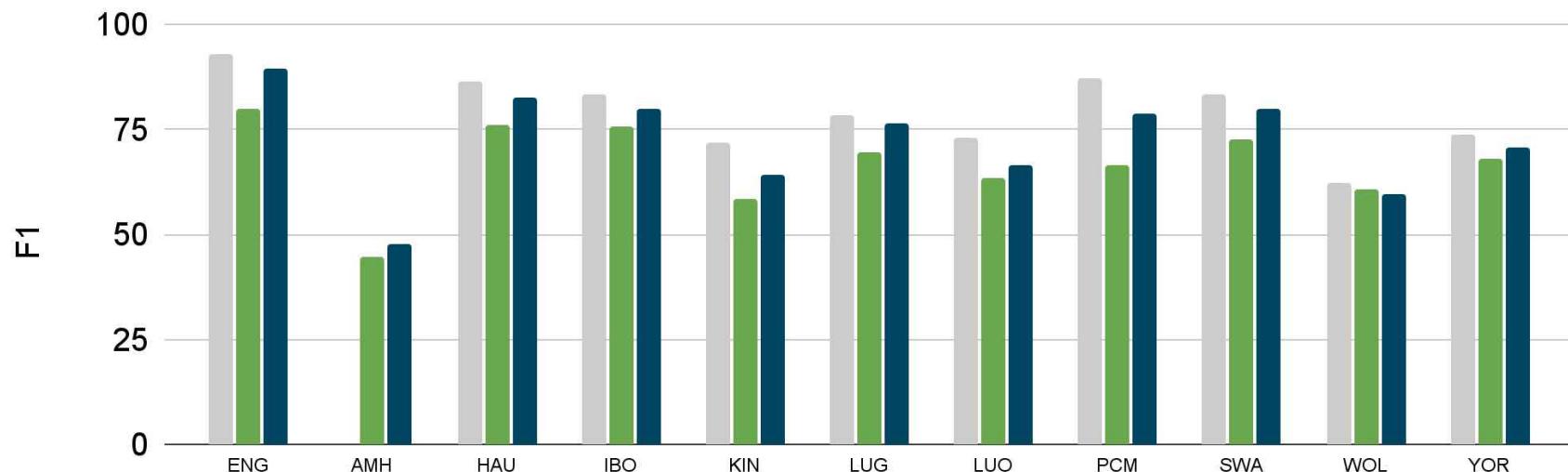


PIXEL vastly outperforms BERT on unseen scripts

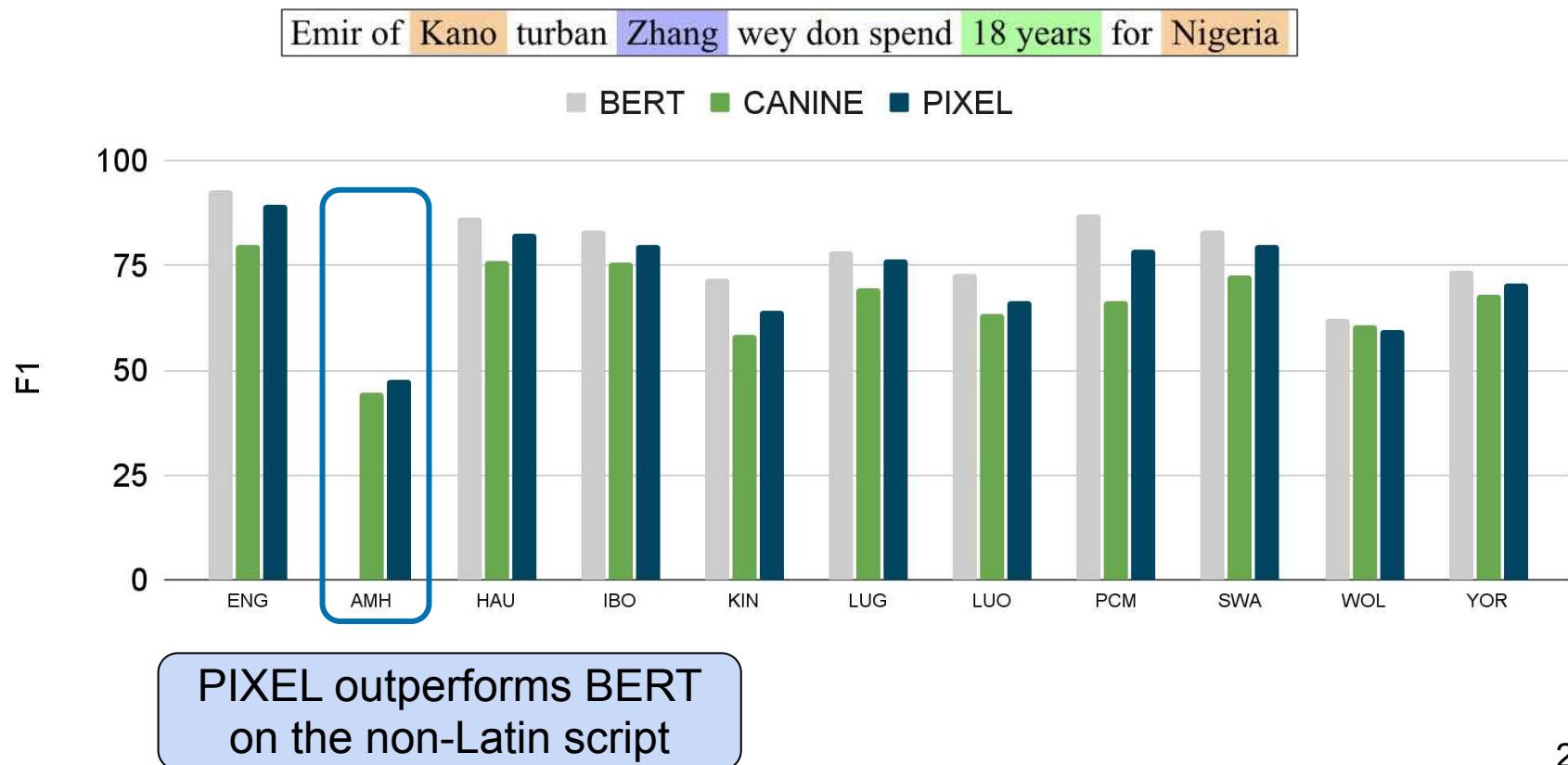
Named Entity Recognition in African Languages

Emir of Kano turban Zhang wey don spend 18 years for Nigeria

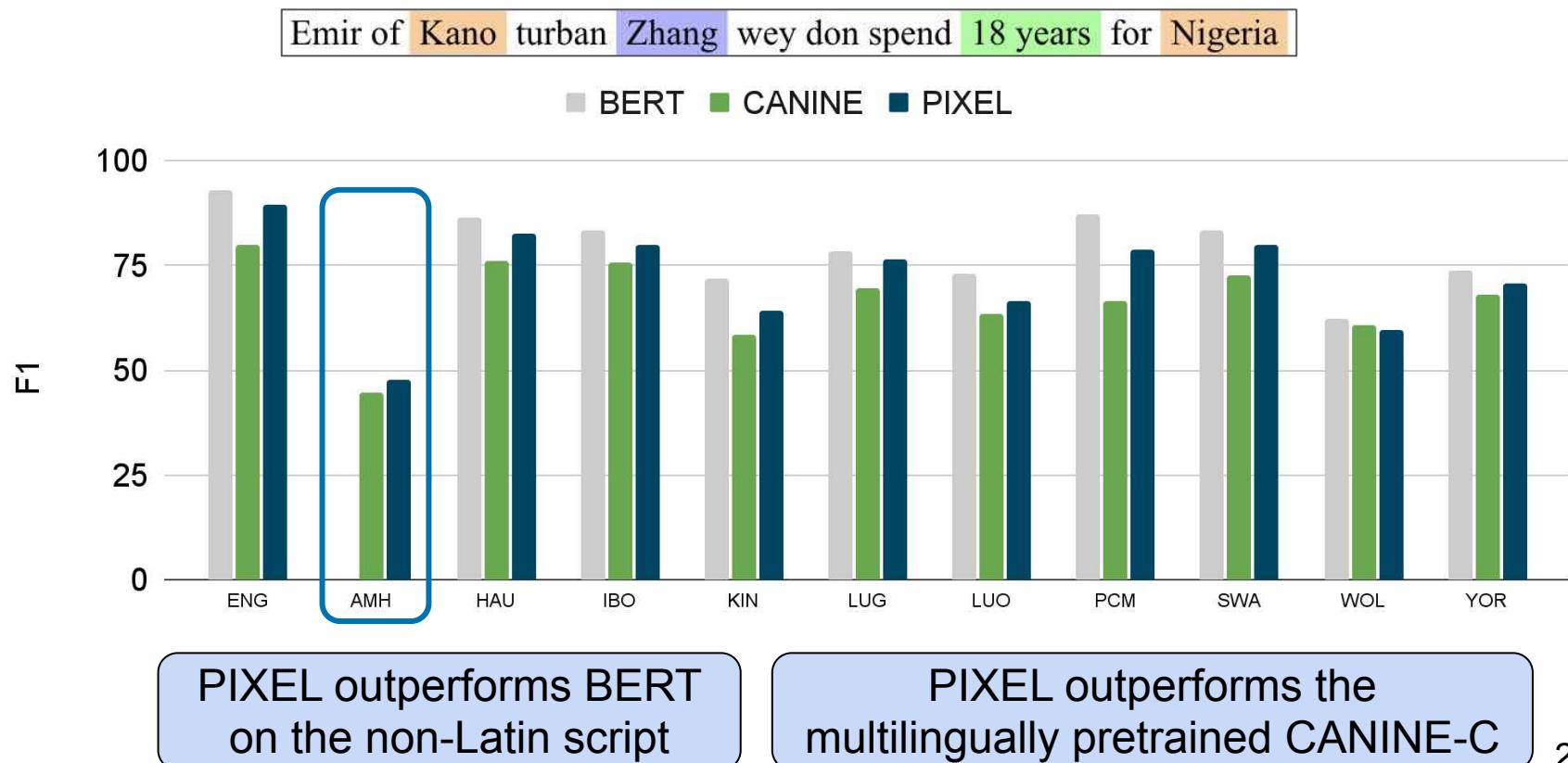
BERT CANINE PIXEL



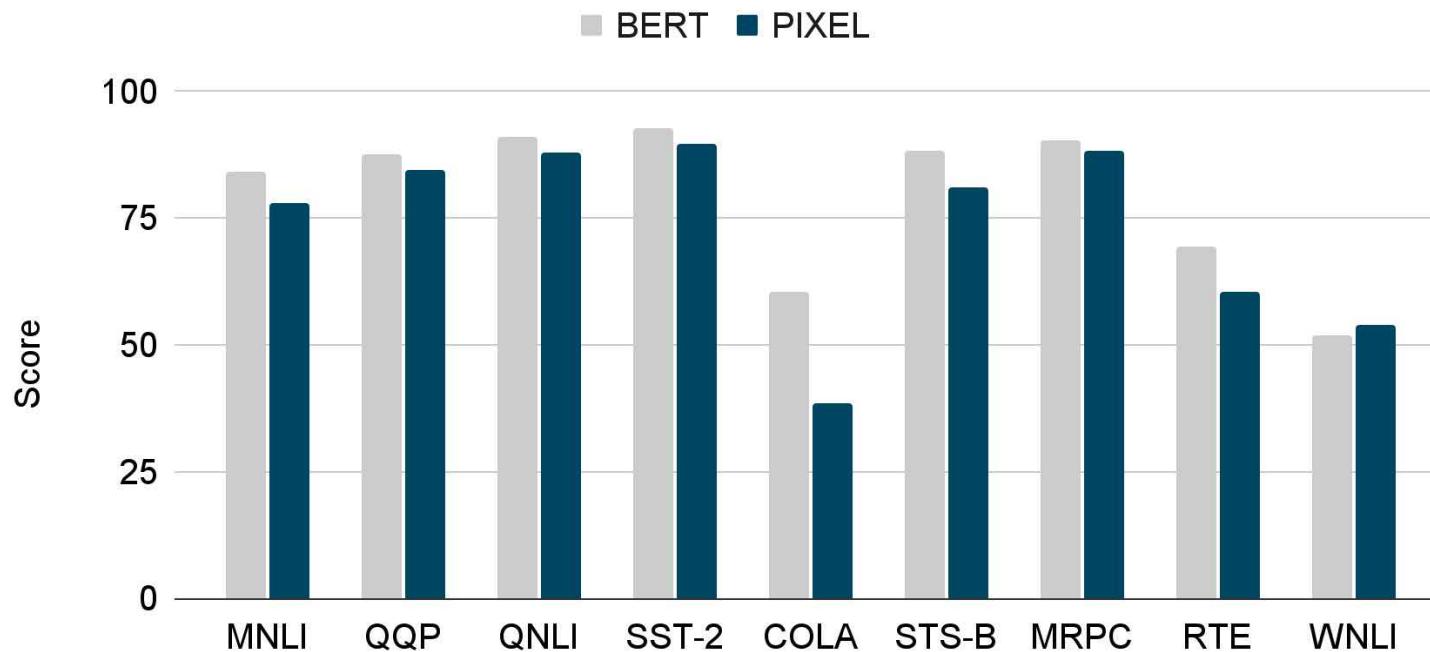
Named Entity Recognition in African Languages



Named Entity Recognition in African Languages



GLUE: Sentence-level Understanding



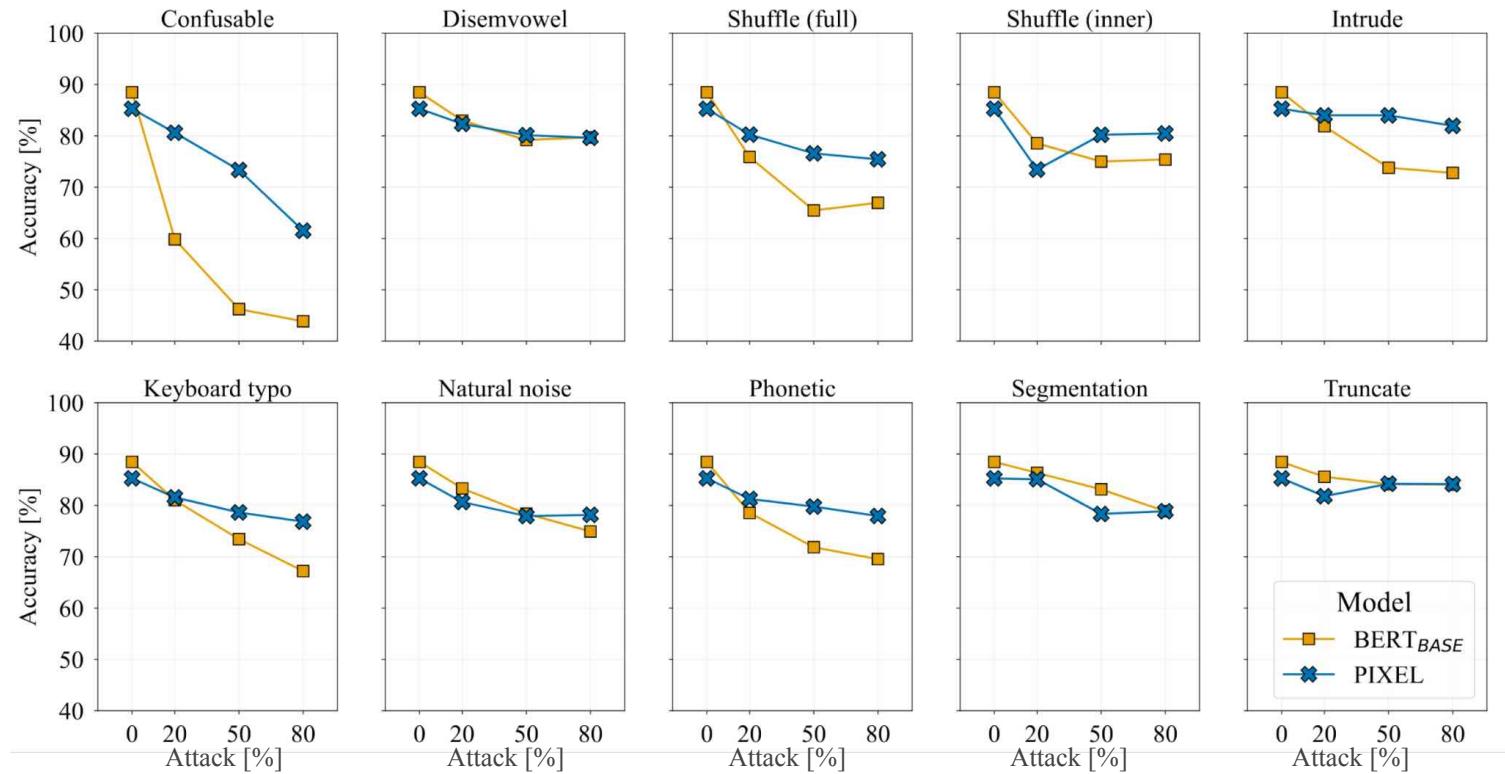
BERT outperforms PIXEL on English sentence-level tasks

Ädu3rsarīał attacks

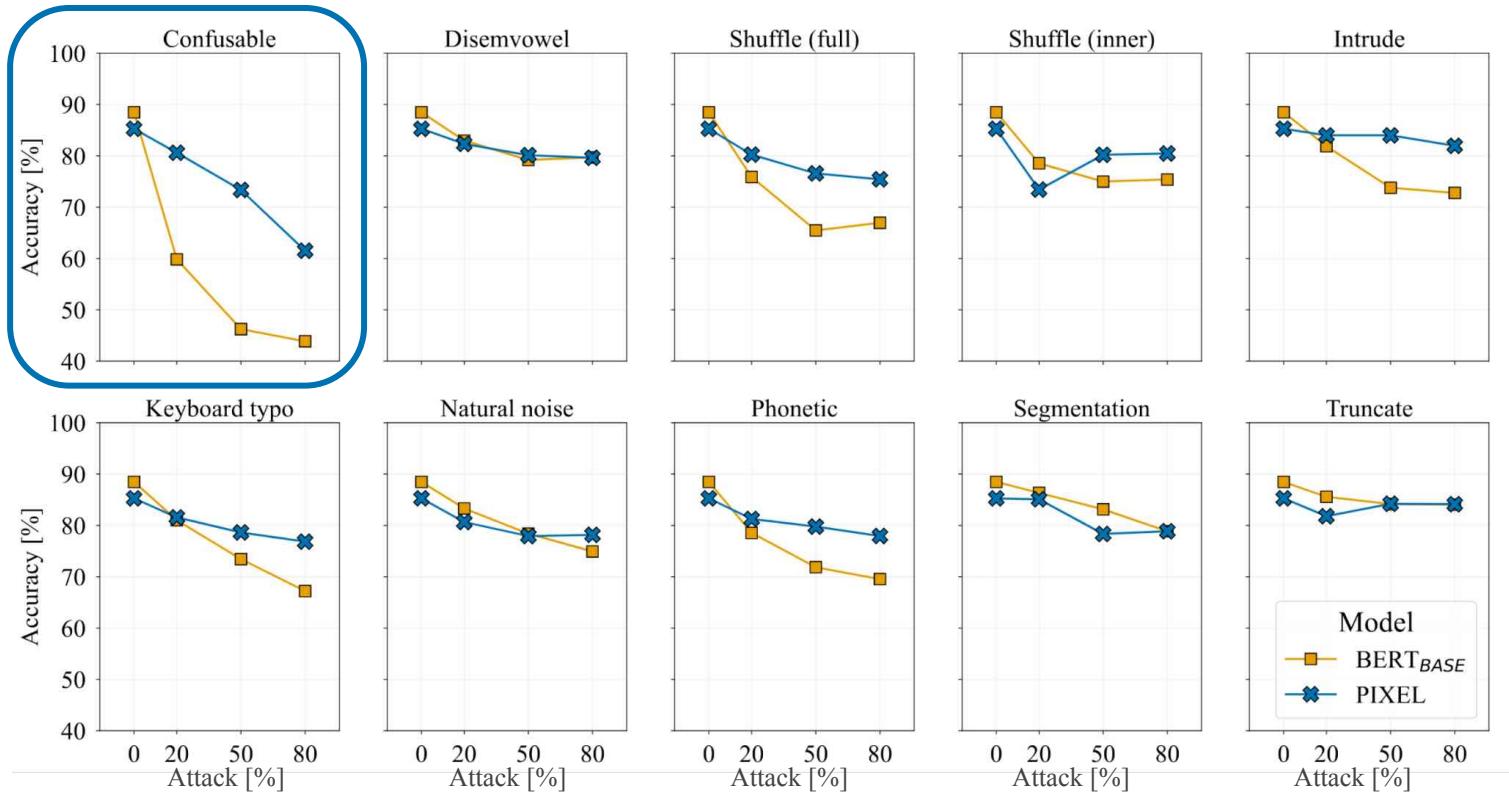
- How well does PIXEL deal with orthographic text attacks?

Attack	Sentence
NONE	Penguins are designed to be streamlined
CONFUSABLE	<i>Penguins are desiged to be streamlined</i>
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 dwsinged 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

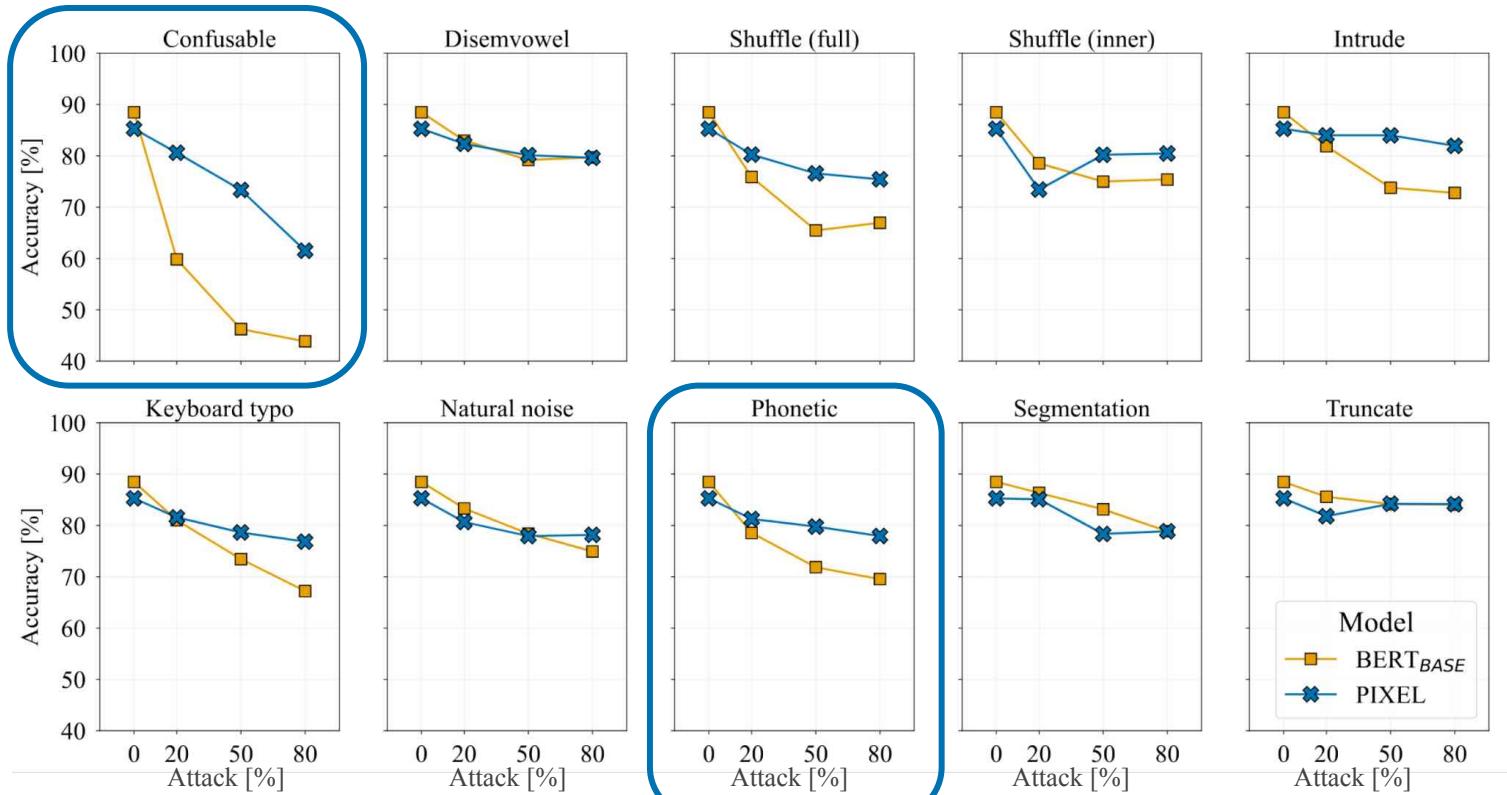
Results on Zeroé (SNLI)



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):

I must be growi ng small again. ■■■

(b) Structured rendering (BIGRAMS):

I must be gr ow in g sm al l ag ai n. ■■■

(c) Structured rendering (MONO):

I mu st b e gr ow in g sm al l ag ai n. ■■■

(d) Structured rendering (WORDS):

I mu st be growi ng sm al l ag ai n. ■■■

Alternative Rendering Strategies

(a) Continuous rendering (CONTINUOUS):

I must be growing small again. ■■■

(b) Structured rendering (BIGRAMS):

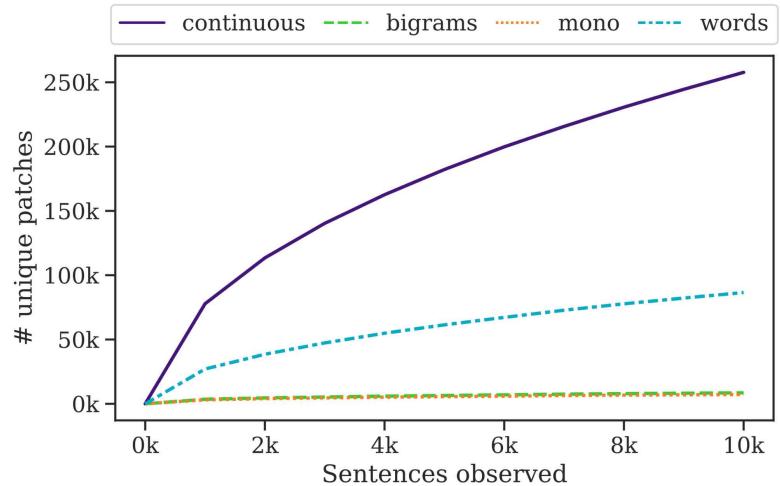
I must be gr owing smal l ag ai n. ■■■

(c) Structured rendering (MONO):

I mu st b e gr owing sm al l ag ai n. ■■■

(d) Structured rendering (WORDS):

I must be growi ng sm all again. ■■■



Alternative Rendering Strategies

(a) Continuous rendering (CONTINUOUS):

I must be growing small again. ■■■

(b) Structured rendering (BIGRAMS):

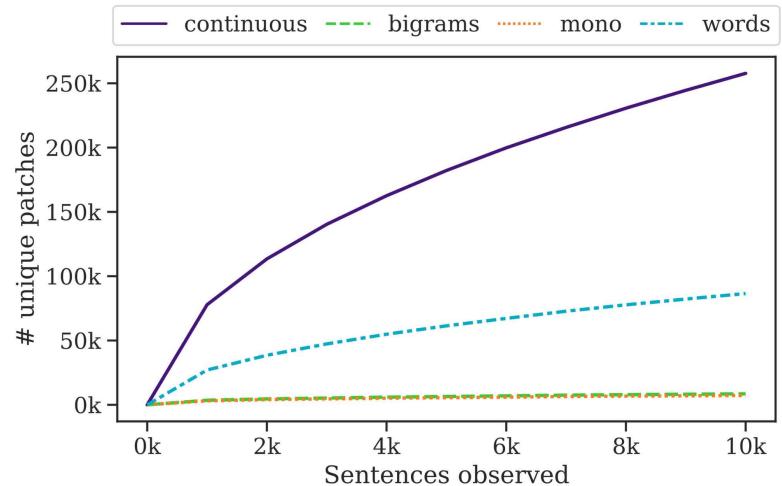
I must be gr owing sm al l ag ai n. ■■■

(c) Structured rendering (MONO):

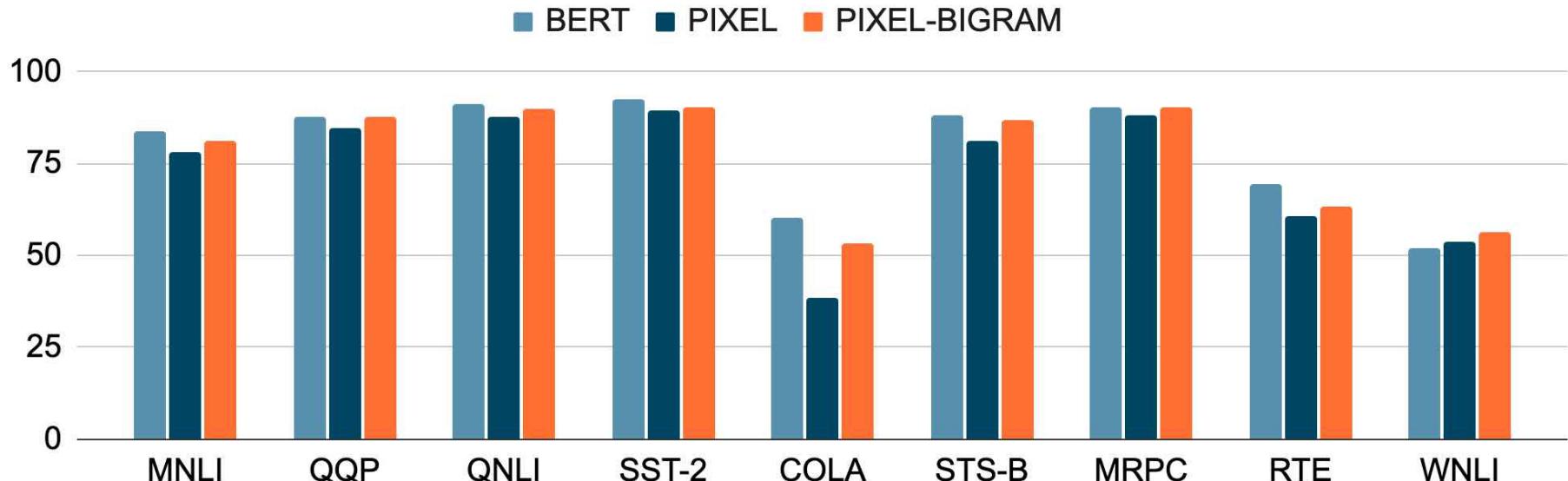
I mu st b e gr owing sm al l ag ai n. ■■■

(d) Structured rendering (WORDS):

I must be grow ing sm all ag ain. ■■■



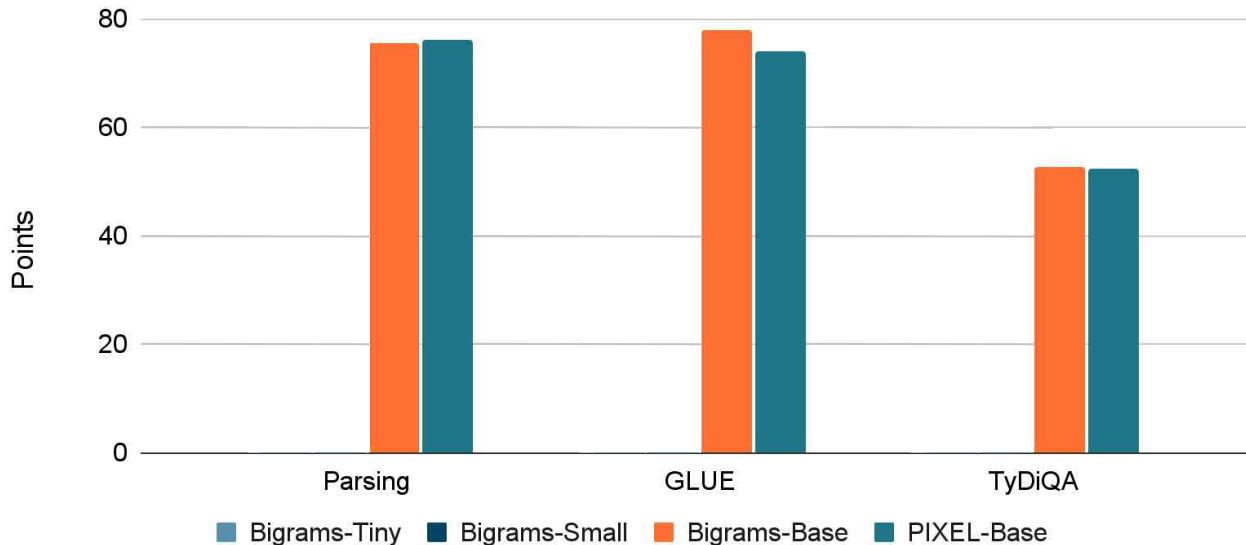
GLUE (revisited)



Bigram text rendering produces better pixel models

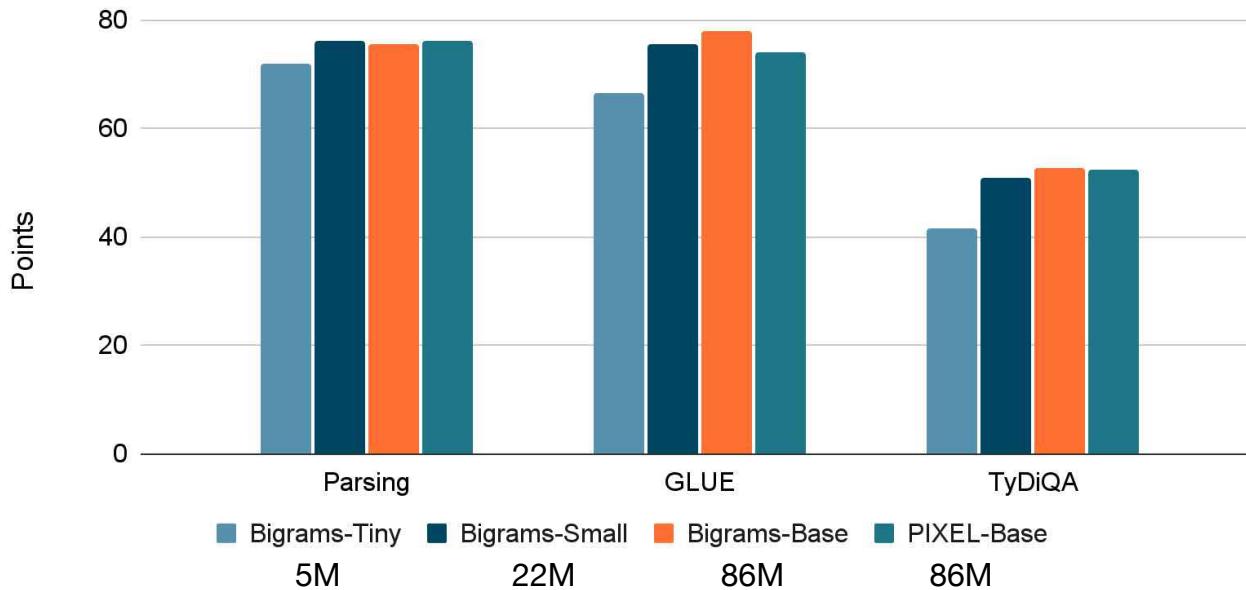
Scaling Down ↓

- 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



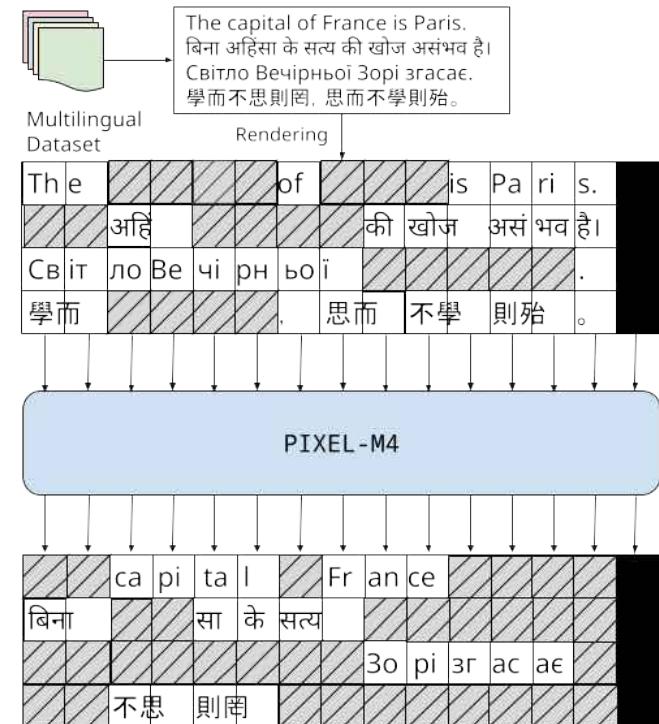
P. Rust



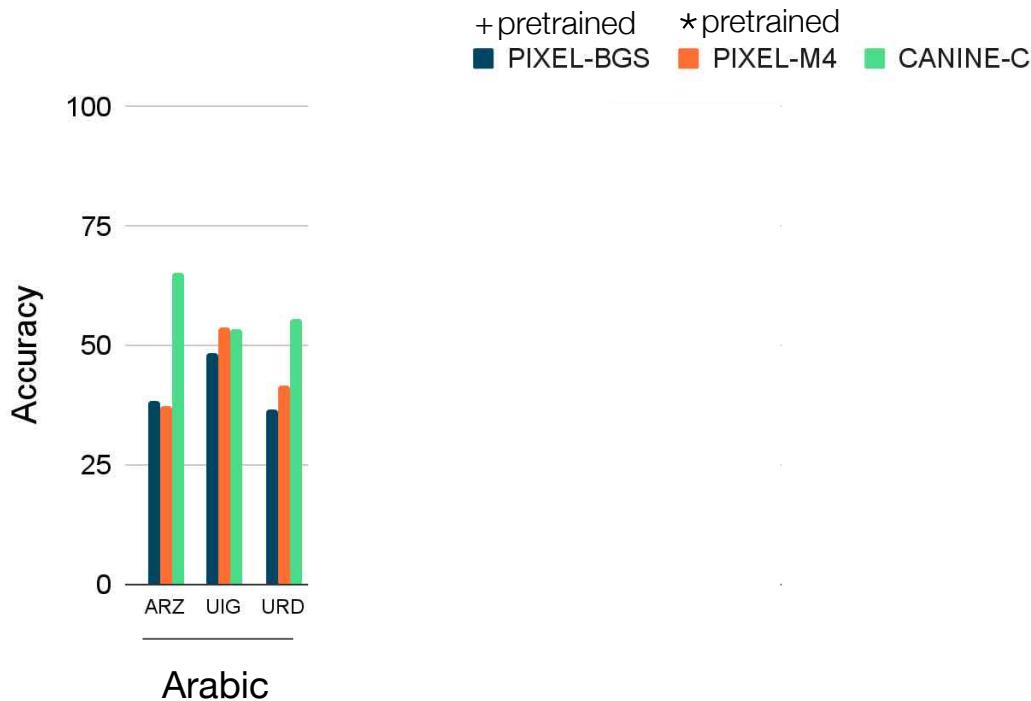
D. Elliott

PIXEL-M4

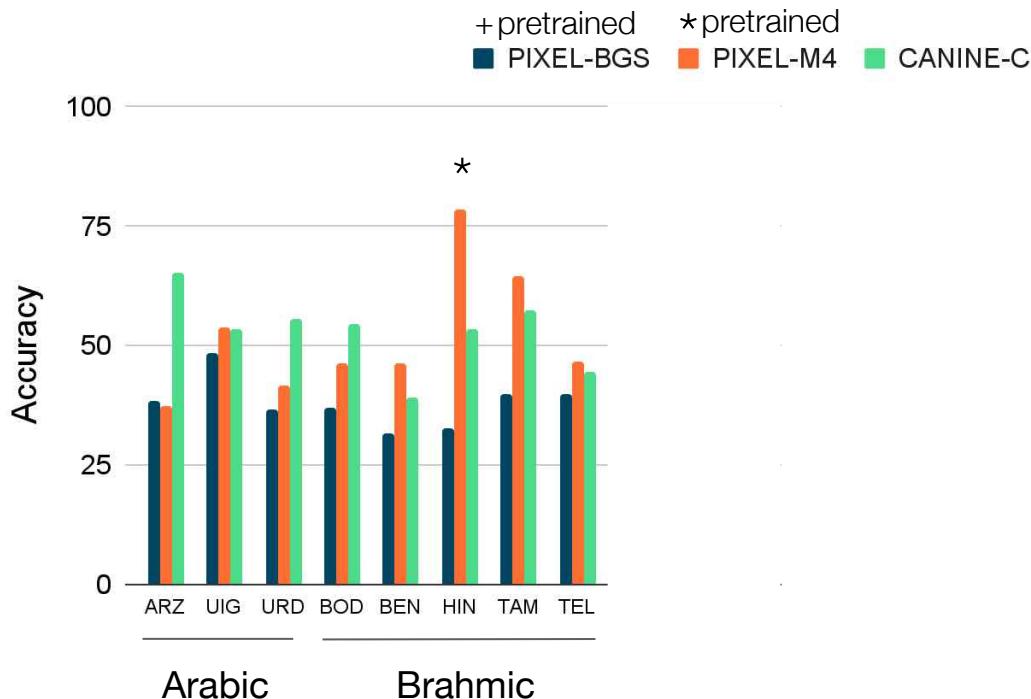
- Same architecture and hyperparameters as PIXEL-BIGRAMS
- But, pretrained on four visually diverse scripts sourced from mC4
 - Latin - English
 - Han - Simplified Chinese
 - Cyrillic - Ukrainian
 - Brahmic - Hindi



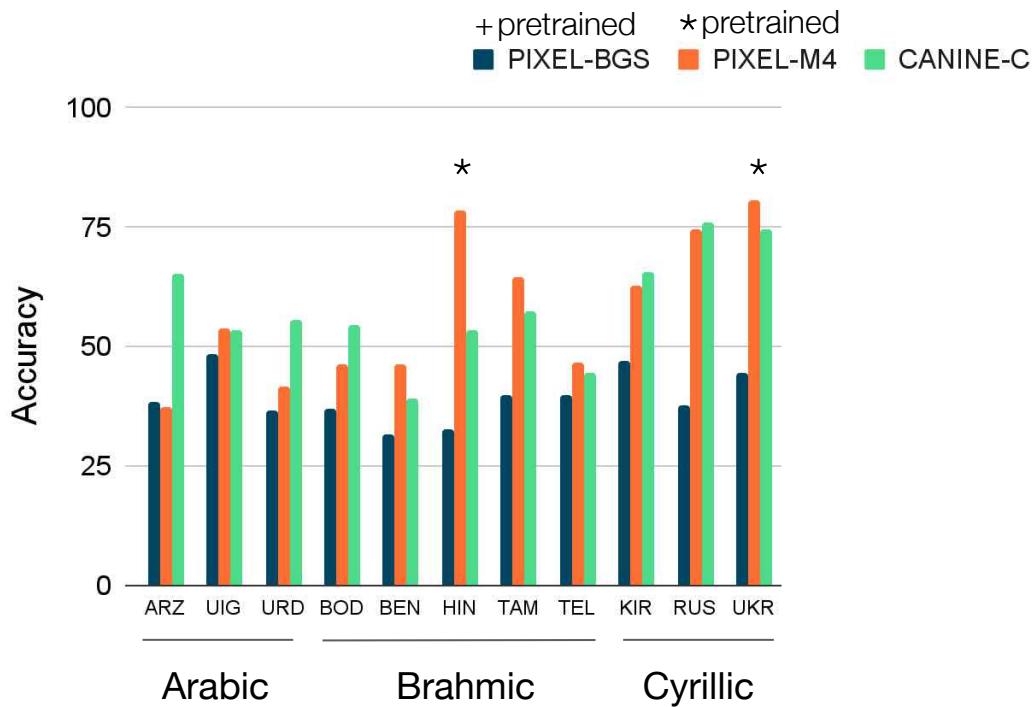
Text Classification on SIB-200



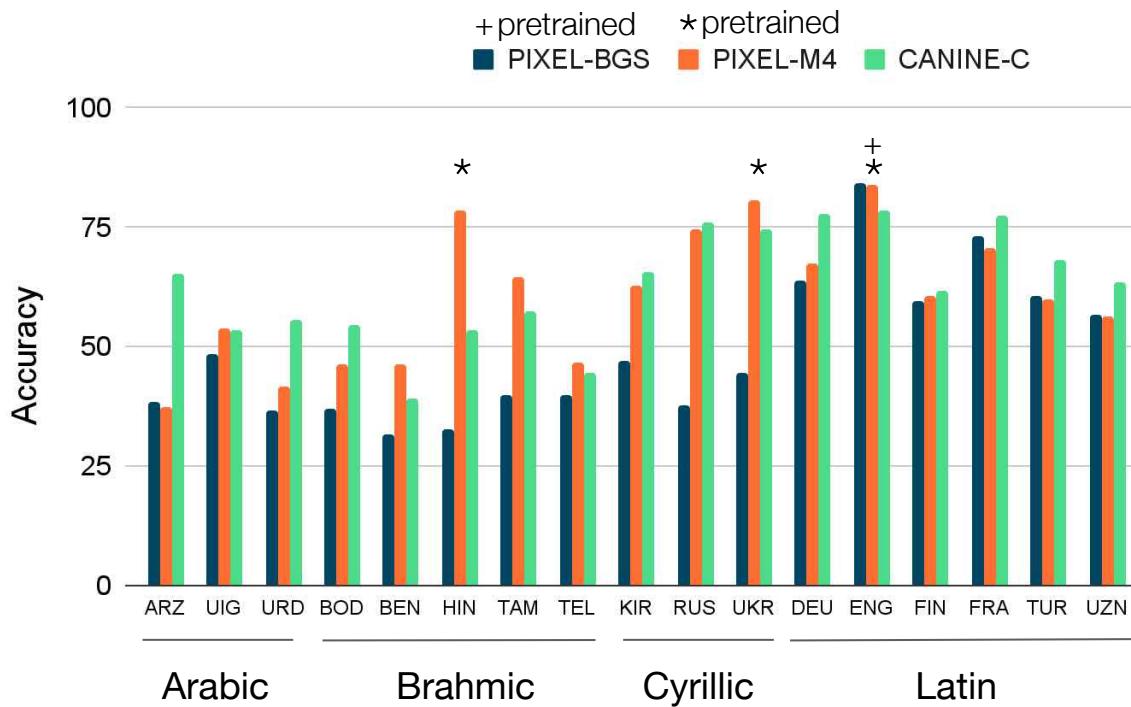
Text Classification on SIB-200



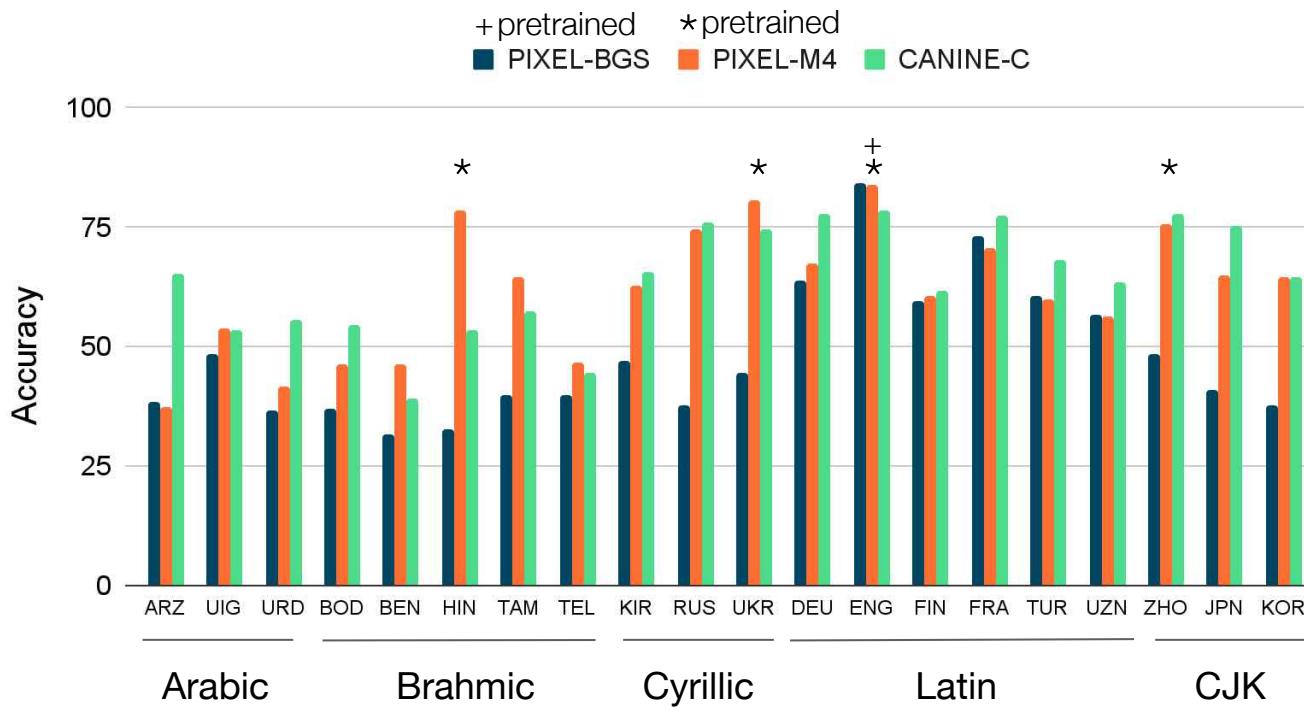
Text Classification on SIB-200



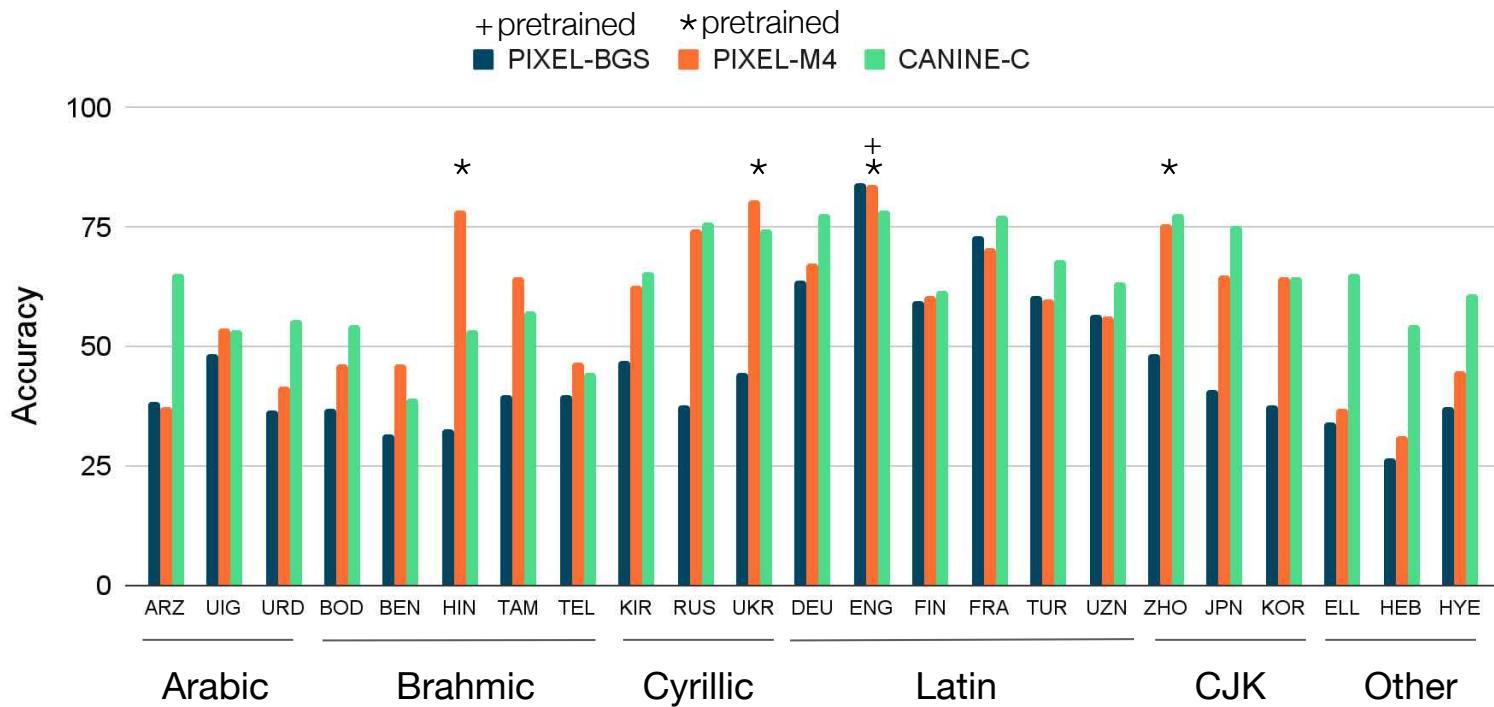
Text Classification on SIB-200



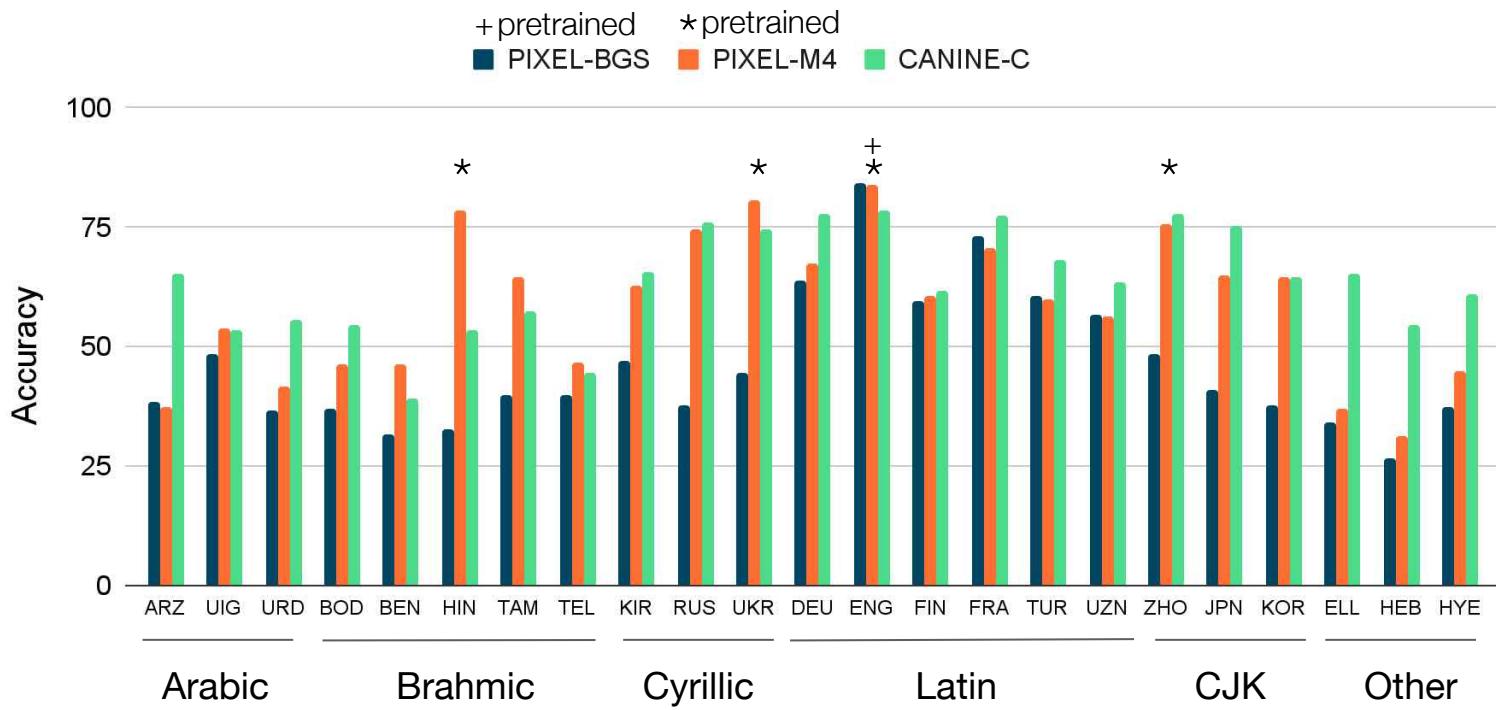
Text Classification on SIB-200



Text Classification on SIB-200

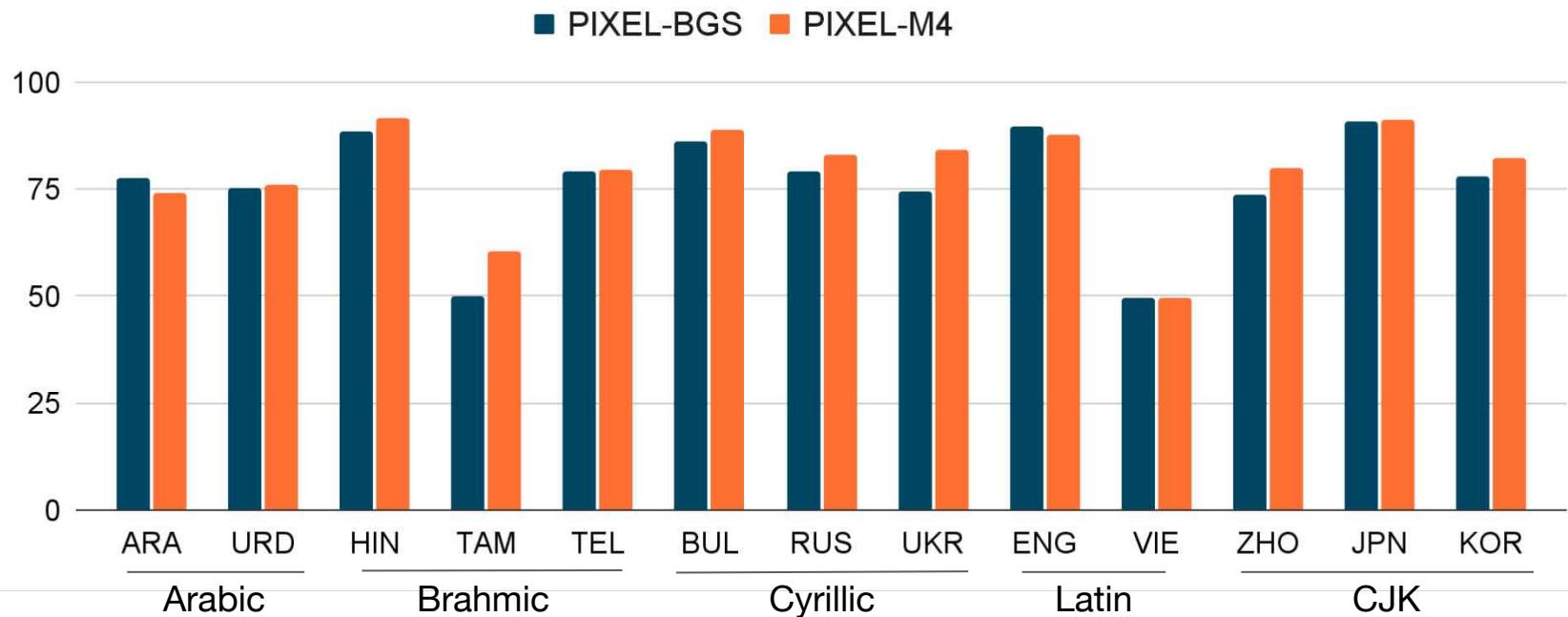


Text Classification on SIB-200



Multilingual pretraining is very helpful for sentence classification

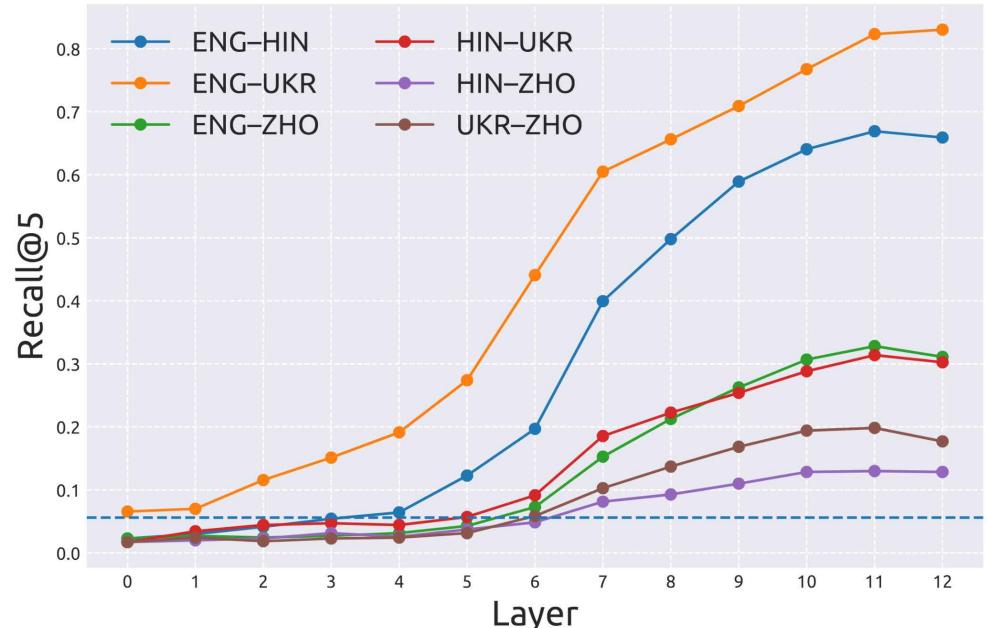
Dependency Parsing



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: Pixel-based Language Modeling of Historical Documents

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



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

sionally blogs such as Arcade, a humanities site published by Stanford University. From 2012 to 2016, he hosted a radio show webcast by Alanna Heiss's Clocktower Productions. In autumn 2020, an article he wrote for The Creative Independent was widely disseminated on the internet. Called 19 things I'd tell people contemplating starting a record label (after running one for 19 years) it was a mix of advice, warnings, and personal history gleaned from almost two decades of operating Brassland. It was followed by an appearance on the Third Story podcast.

Sickmon's war service took him to Tokyo during the occupation of Japan where he served as one of the "Monuments Men" under General Douglas MacArthur's

terminated by the All England Club in 1981 in order for The Championships, Wimbledon to be held. Since then the club has been nomadic, moving to Osterley and Greenford before settling in Acton and playing their matches at Wasps FC's Tuford Avenue Sports Ground. By 2012, the club had downsized to running only one team.

A number of players for the New Zealand national rugby union team have played for London New Zealand including Doug Rollerton, Terry Morrison and Paul Sapeford. In recognition of their history, the club have been granted privileges from both the Rugby Football Union and the New Zealand Rugby. They are the only rugby team aside of New Zealand national representative teams that wears the silver fern as their crest and the RFU exempted them from the overseas player quotas, prior to their abolition. The club have also taken part in a number of New Zealand government

aving been estranged from her father's family for most of her life, Andrea is intrigued. But what exactly is the Bancroft's involvement with "Genesis," a mysterious person working to destabilize the geopolitical balance at the risk of millions of lives? In a series of devastating coincidences, Andrea and Belknap come together and must form an uneasy alliance if they are to uncover the truth behind "Genesis"—before it is too late.

Girls' BMX was part of the cycling at the 2010 Summer Youth Olympics program. The event consisted of a seeding round, then elimination rounds where after three races the top 4

swimmers have so far achieved qualifying standards in the following events (up to a maximum of 2 swimmers in each event at the Olympic Qualifying Time (OQT), and potentially 1 at the Olympic Selection Time (OST)):

Venezuela has entered one athlete into the table tennis competition at the Games. Gremlins Arvelo secured the Olympic spot in the women's singles by virtue of her top six finish at the 2016 Latin American Qualification Tournament in Santiago, Chile.

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

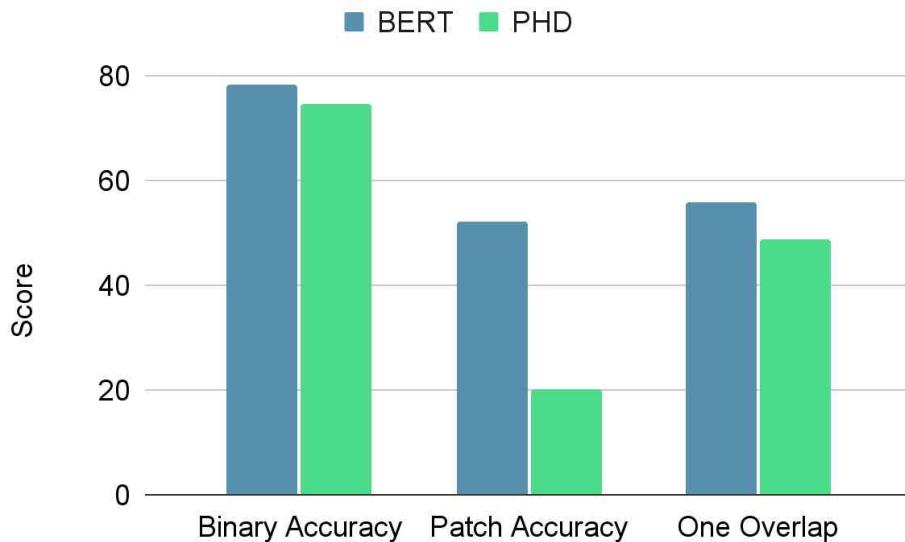
How much reward is offered?

WHEREAS a Molatto Boy (his Name Dench) belonging to a young Man lately arrived from the East-Indies, absented himself on Monday the 27th Instant. He has on when he went away a Thicker Frock and Waistcoat, Leather breeches, and a blue Surtuit Coat, with a red Colian.

Any Person that will apprehend the abovementioned Boy, or give any intelligence where he may be taken, shall receive a Reward of Three Guineas. He is about five Feet high, with short black Hair, not of the woolly Kind.

N. B. If taken, to be brought to the Sign of the George, Queen-Anne-Street, Cavendish-Square.

Results



What other rewards were offered?

RUN AWAY,
From the Ship BRITANNIA, Capt. Scott,
Commander, on Friday the 25th Instant,
TWO Negro Men, the one named
LEWIS, near Six Feet high, and two
Holes in his Ears; the other about Five Feer
Six Inches high, he has two or three Particular
Sears between his Eyebrows, and his Teeth are
filed down like a Saw between every Tooth. If
any Body will bring them to Mess. MUER and
CLANDEK, Merchants, in Nicholas Lane,
shall be **handsomely rewarded**.

Who is the contact person for the ad?

Last week run away from his Master **J. Bromley, Esq;**
of Bookham in Surrey, his Negro Man **Henry,** alias
Harry Johnson, aged about 35 Years, tall of stature, sometimes
wears a Perriwig, speaks English well, in a blue
Livery with pearl Buttons, and has taken with him several
of his Masters Goods. Whoever secures him, and
gives notice to his Master aforesaid, or to **Mr. Richard**
Sheppard in **Lothbury, London,** shall have a Guinea Re-
ward.

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

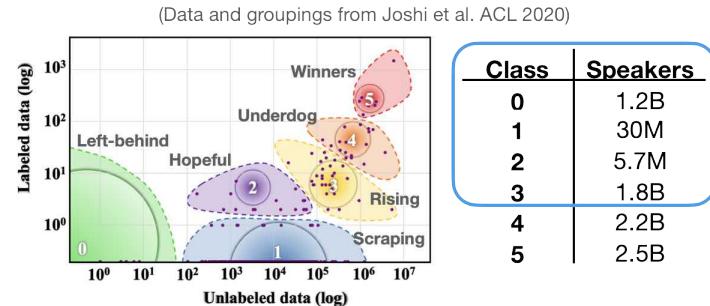
1. General-purpose transferable representations of text can be learned directly from rendered text without any additional linguistic supervision
2. Careful consideration of how to structure the rendering of the text (“tokenization”) is important: bigrams >> unrestricted rendering

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
- 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 word, given the unmasked context.
- Why is it possible to learn a good model by predicting pixels?

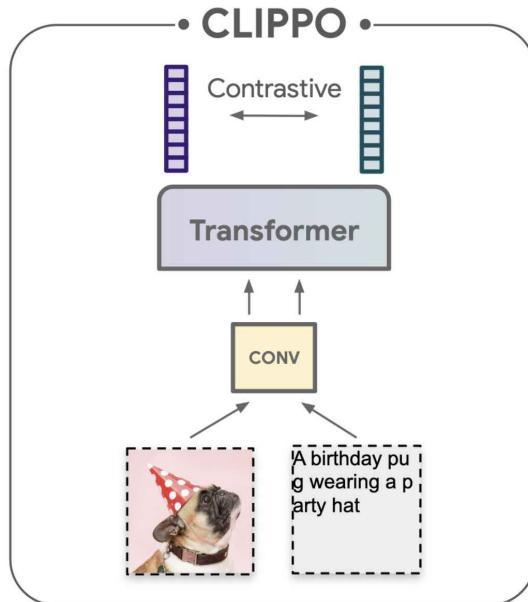
$$-\sum_{m \in M} \log p(x_m | \mathbf{x}_{\setminus m})$$

BERT: Masked Language Modelling

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

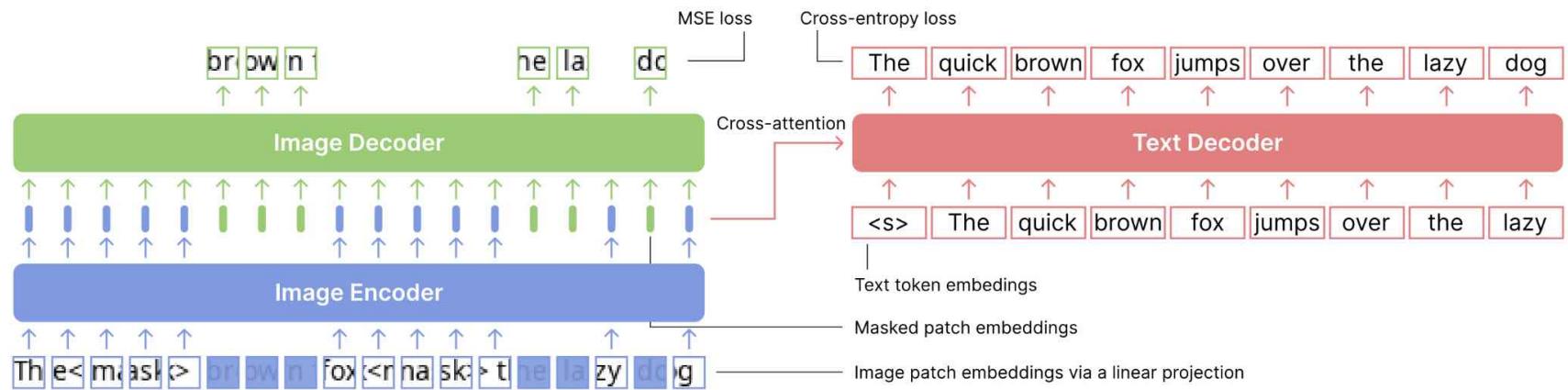
PIXEL: Masked Autoencoding

Joint Multimodal Processing



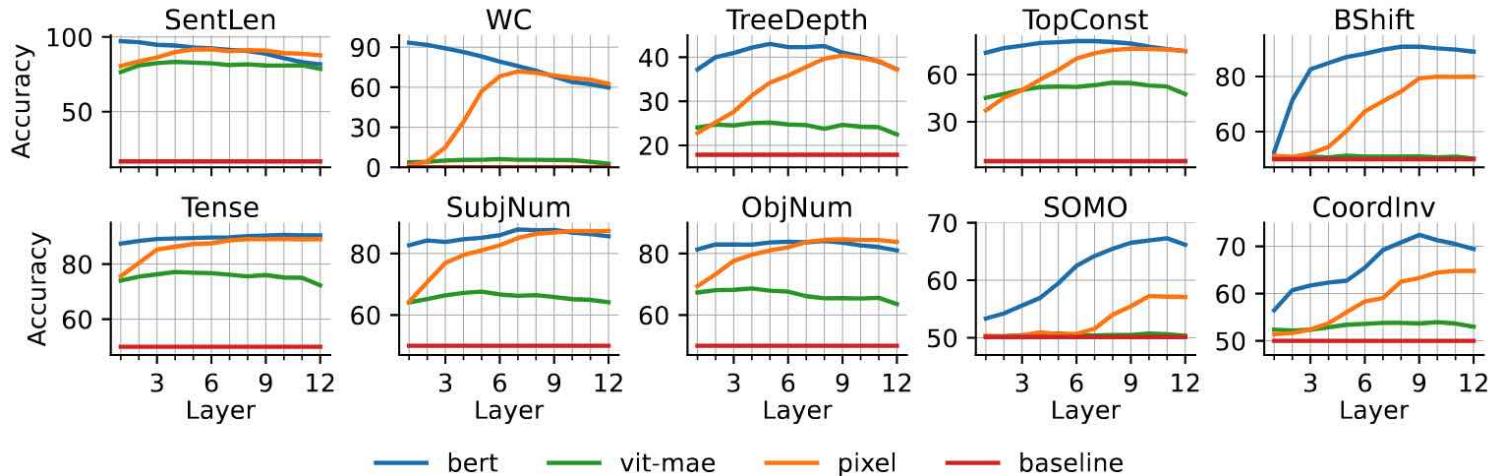
Patch and Text Prediction

- Combine patch and token prediction



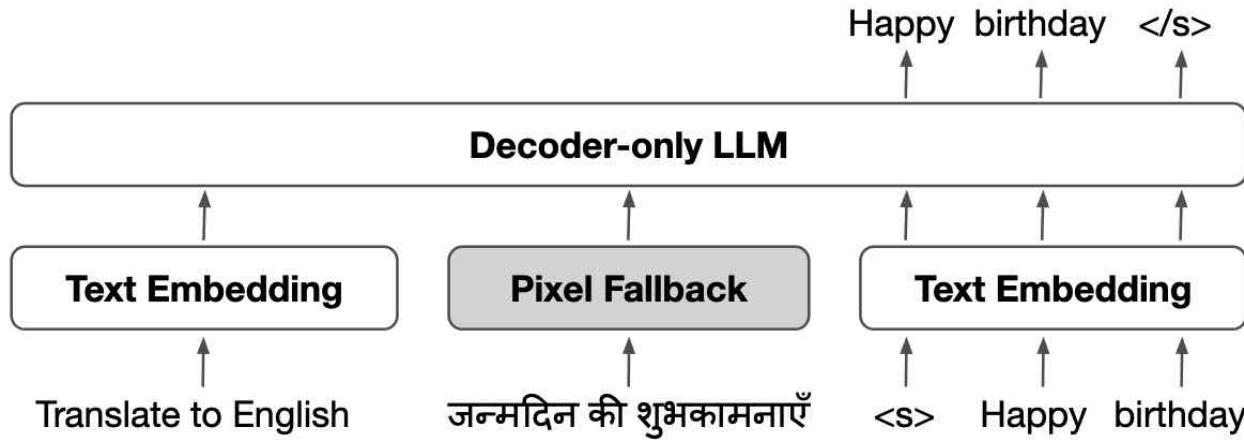
Pixology

- 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