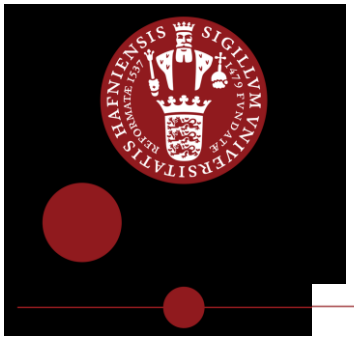
Language Modelling from Pixels



Desmond Elliott

Language and Multimodal Processing Group
Department of Computer Science
University of Copenhagen

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

What is Language Modelling?

- A language model is a statistical model of text and can be used to estimate the probability of a string $\mathbf{x}$ consisting of T tokens.

Tractable with assumptions, e.g. bi-gram model

Claude Shannon. A Mathematical Theory of Communication. 1948.

Modern Language Models

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

All words in vocabulary

j-dimensional vector

Deep Neural Network

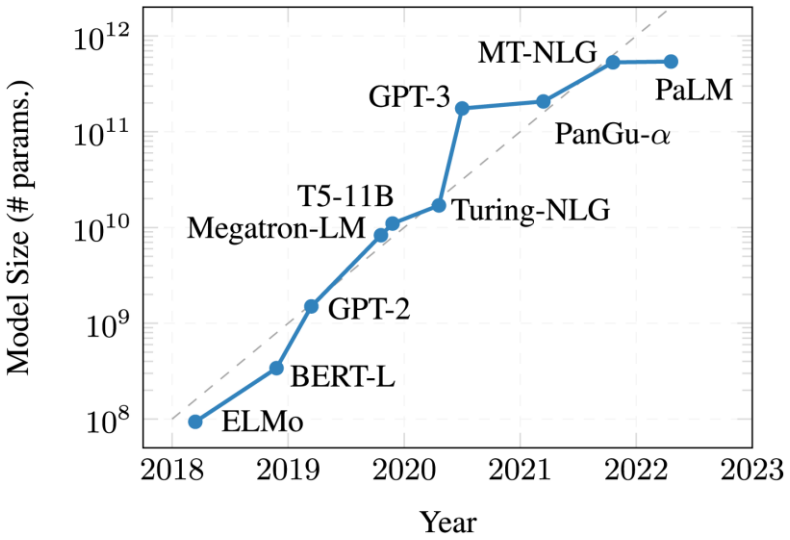
Billions of trained parameters

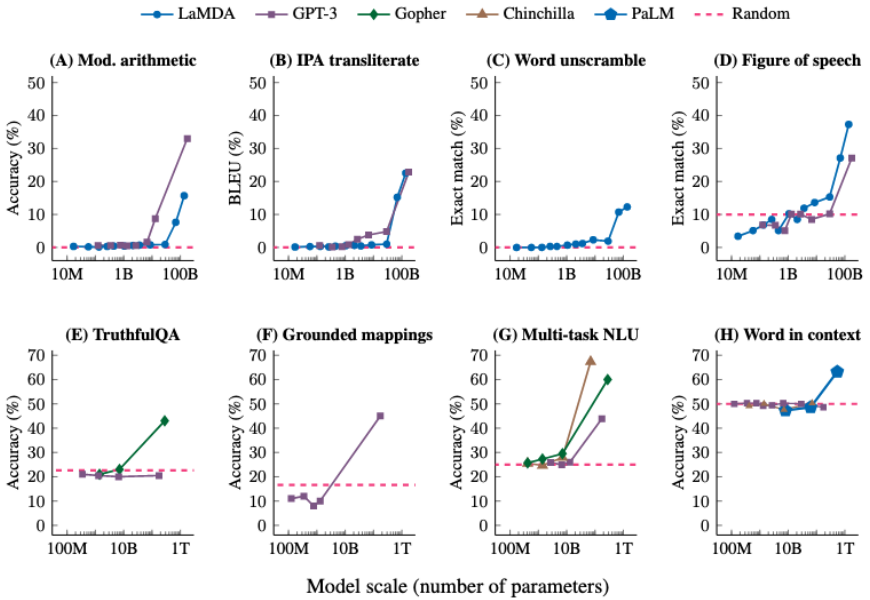
The woman drew many famous

Okay but why is it useful?

- Spelling and grammar checking
- Machine translation
- Web search
- Text prediction
- Topic modelling
- Chatbots

Current Status: Scale

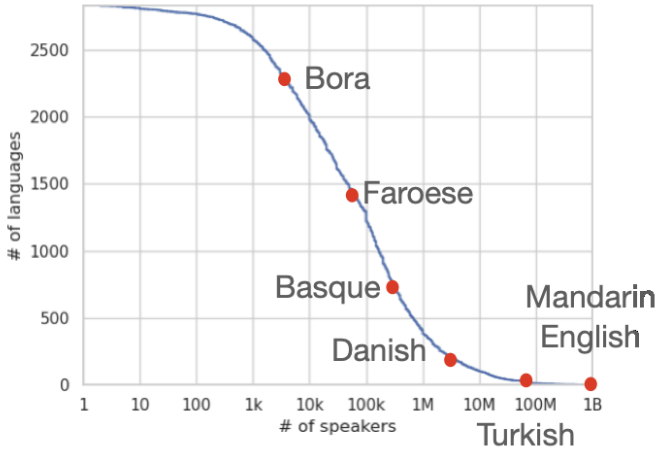




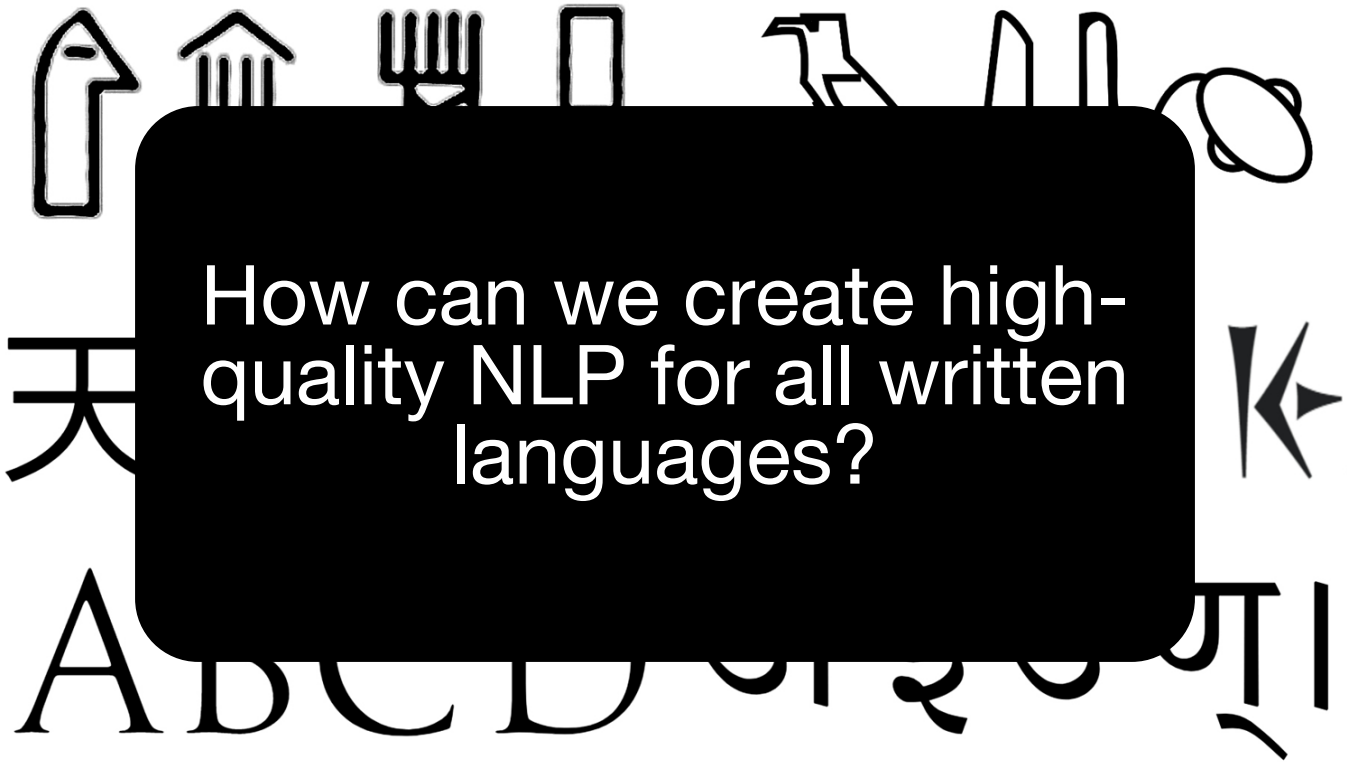
Treviso et al. 2023. Efficient Methods for Natural Language Processing: A Survey. TACL
Wei et al. 2022. Emergent Abilities of Large Language Models. TMLR

NLP for **All** Written Languages

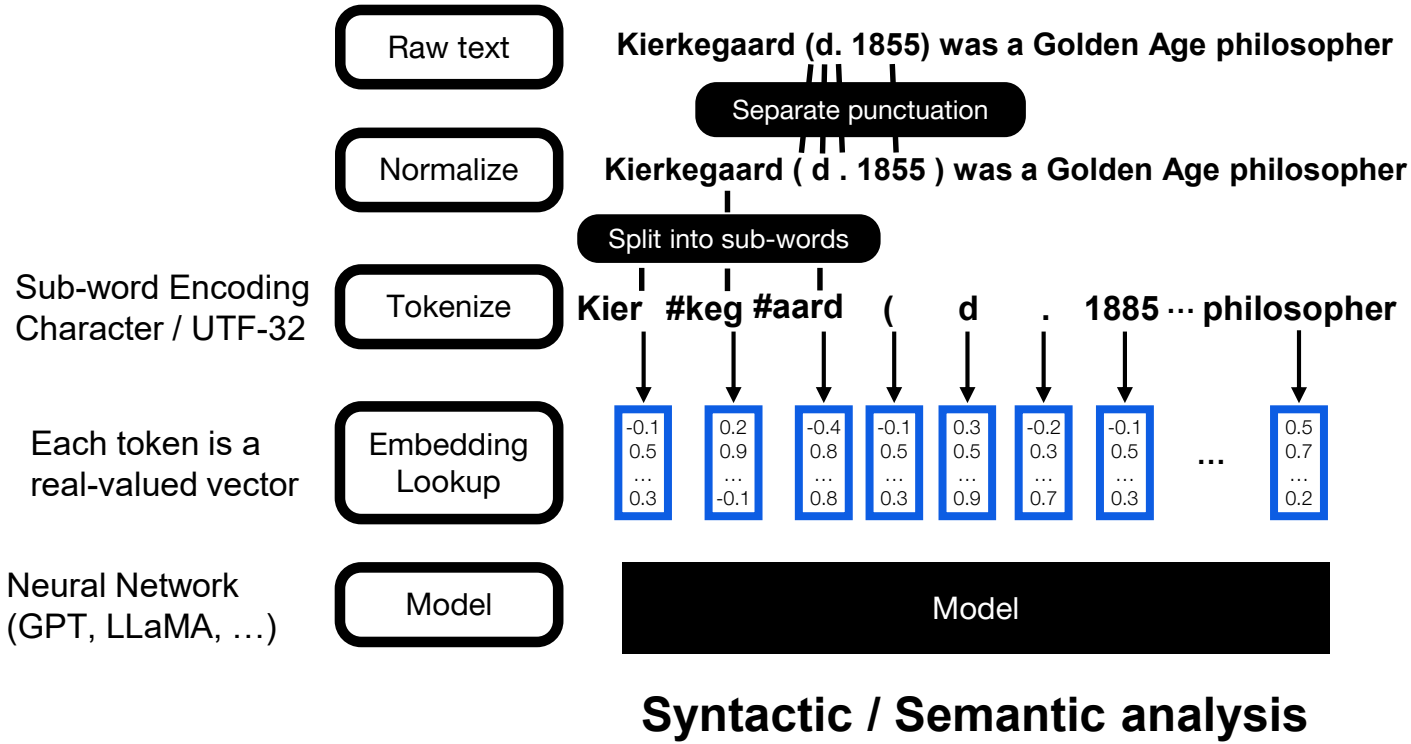
- There are 3,000 written languages
  - 400 with >1M speakers
- NLP usually covers 100 languages
  - Technological exclusion for billions



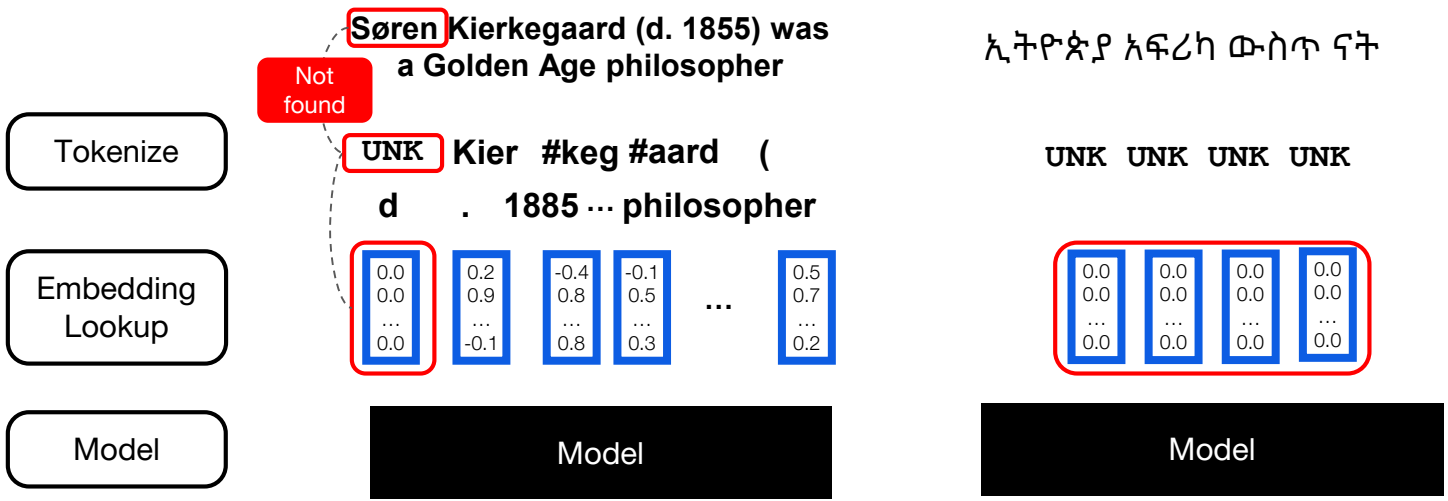
van Esch et al. Writing System and Speaker Metadata for 2,800+ Language Varieties. LREC 2022.
Joshi et al. The State and Fate of Linguistic Diversity and Inclusion in the NLP World. ACL 2020.


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

NLP is a pipeline …



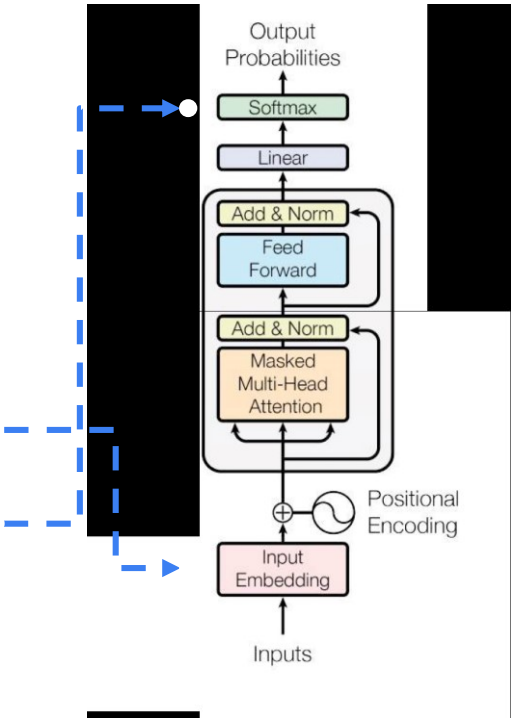
… that is easily broken



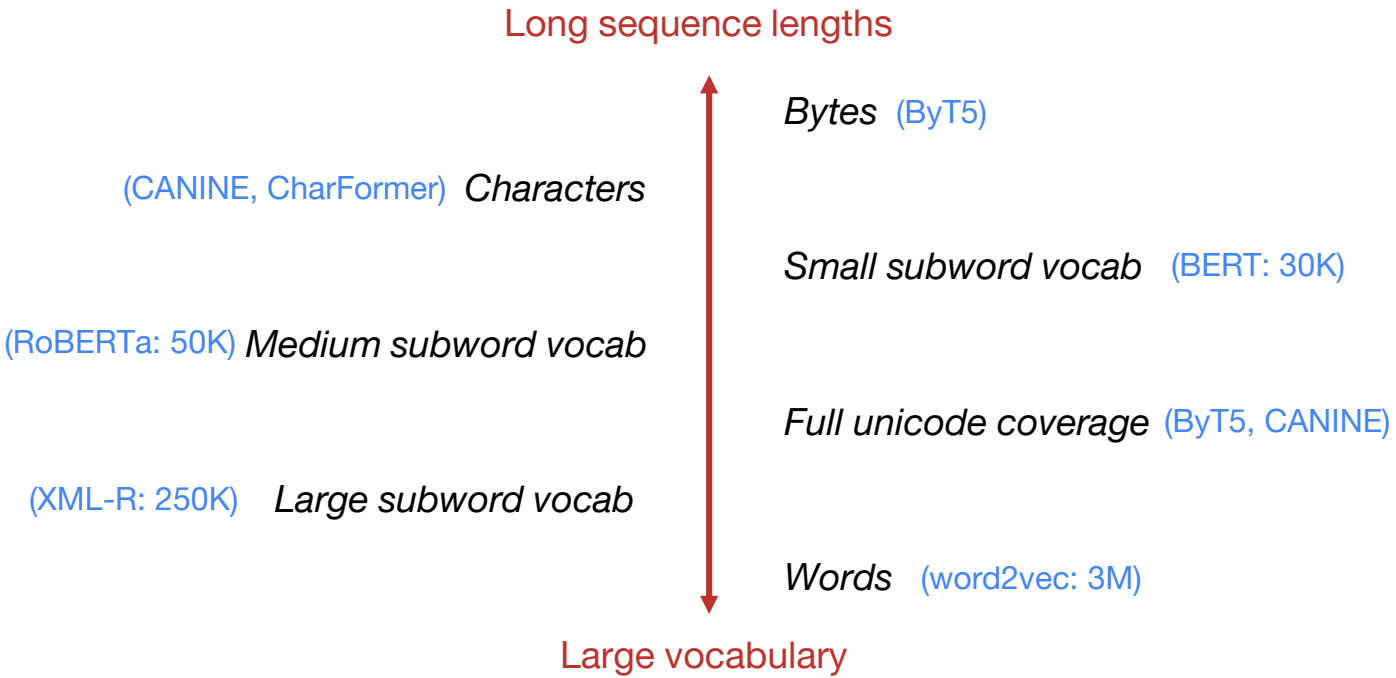
This issue disproportionately affects low-resource languages

The Vocabulary Bottleneck

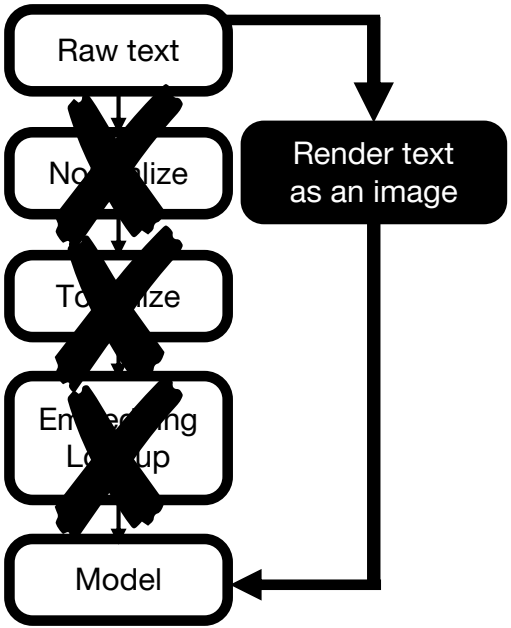
- NLP is an **open vocabulary problem** and the ability of a model is determined by its vocabulary:
  1. tokens, characters, sub-words, etc.
- This creates a bottleneck in two places:
  1. *Representational bottleneck* in the Embedding layer
  2. *Computational bottleneck* in the Output layer

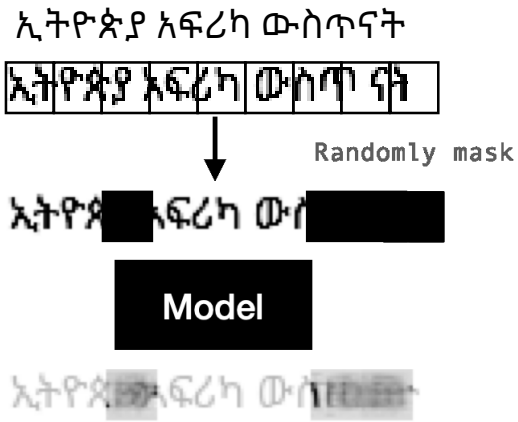


Where's the sweet spot?



Treat language as vision

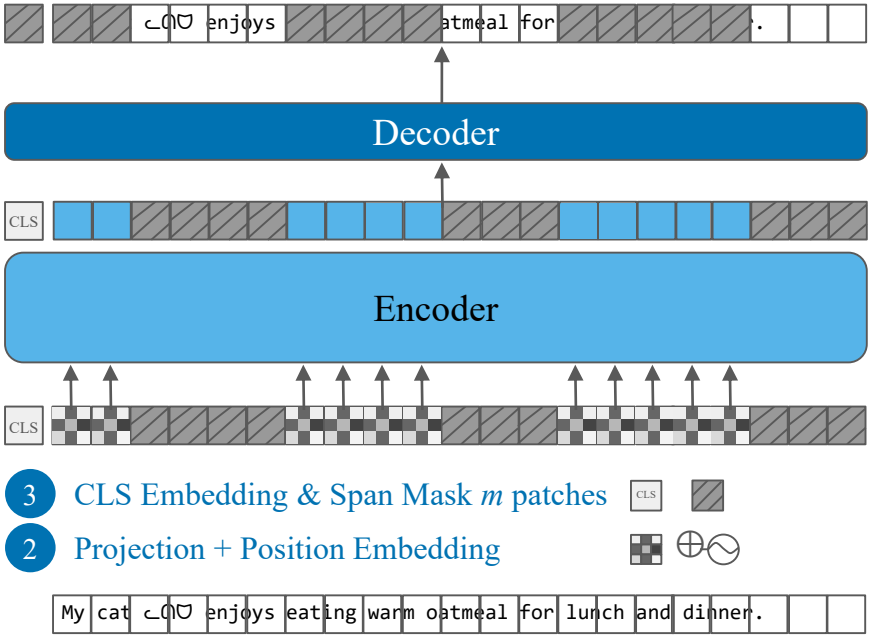




Overview

1. Pixel Language Model
2. Text Rendering Matters
3. Historical Document Processing

The Model



[redacted]

8 Layers

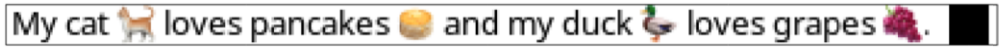
12 Layers

16pixel x 16pixel patch

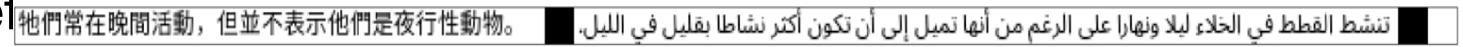
Google Noto Fonts

PyGame / PangoCairo

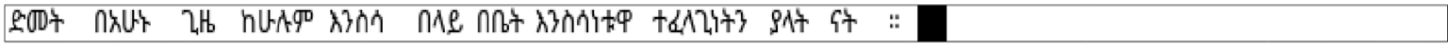
Flexible Text Renderer



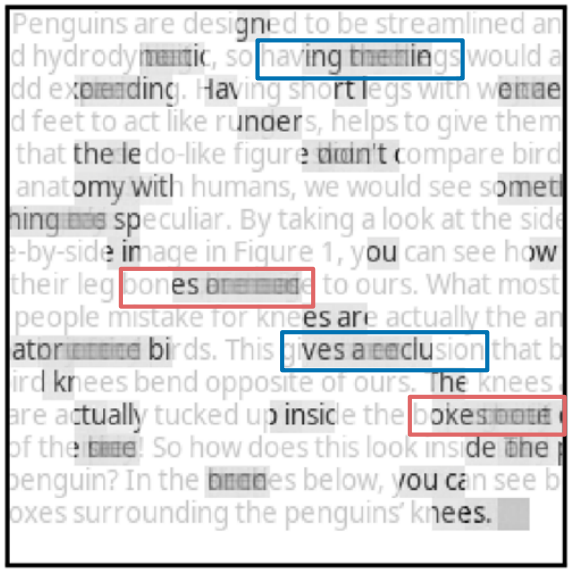
- Emoji
- Lef



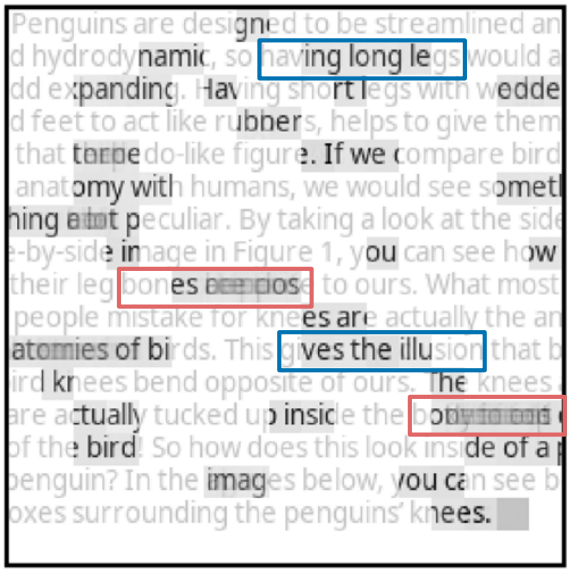
- Wo



A new type of generative model



Try for yourself



100K steps **500K steps** **1M steps**

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

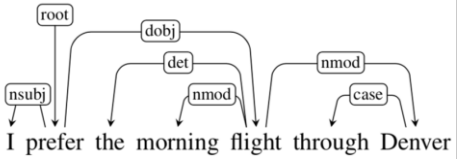
- **Datasets**: Universal Dependencies, MasakhaNER, GLUE, Zeroé
- **Models:**

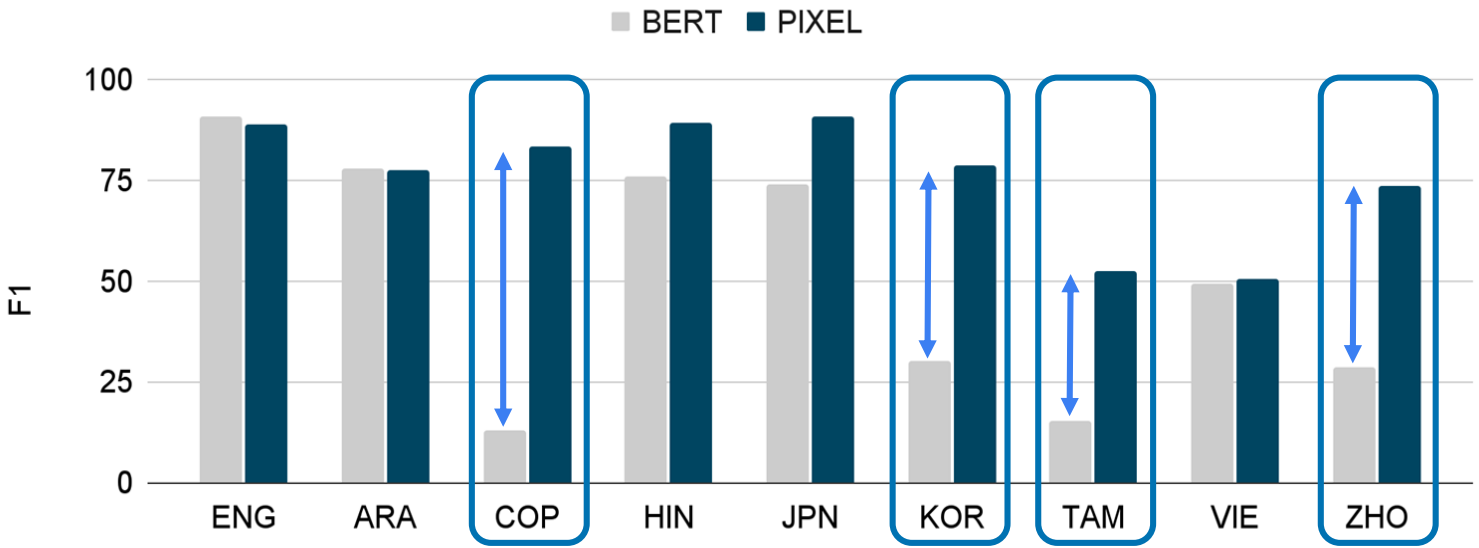
	Parameters	**Pretraining Data**
$\text{PIXEL}_{\text{BASE}}$	86M	English Wikipedia + Bookcorpus
$\text{BERT}_{\text{BASE}}$	110M	—
CANINE-C	127M	104-languages from Wikipedia

Similar pretraining setup

Tries to solve the same problem using UTF-32

Dependency Parsing Results

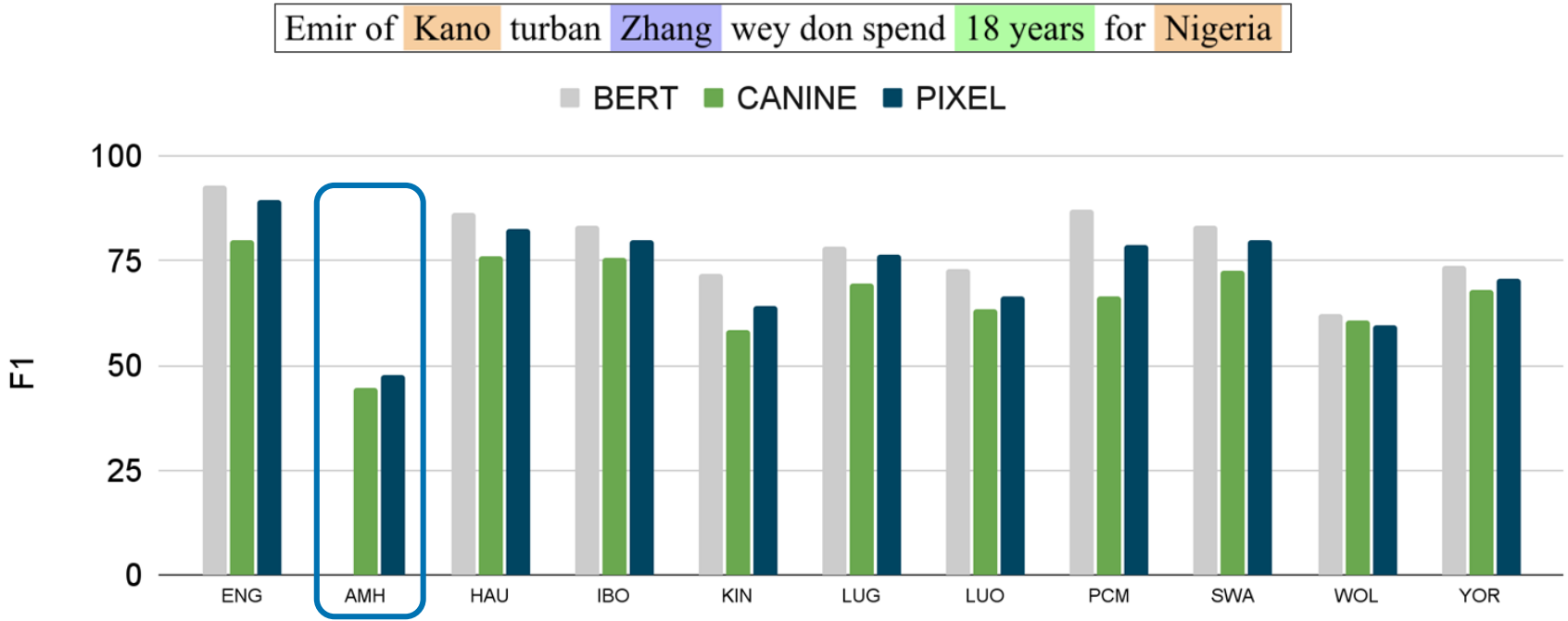




BERT UNK	0%	1%	94%	33%	46%	85%	82%	5%	73%

PIXEL vastly outperforms BERT on unseen scripts

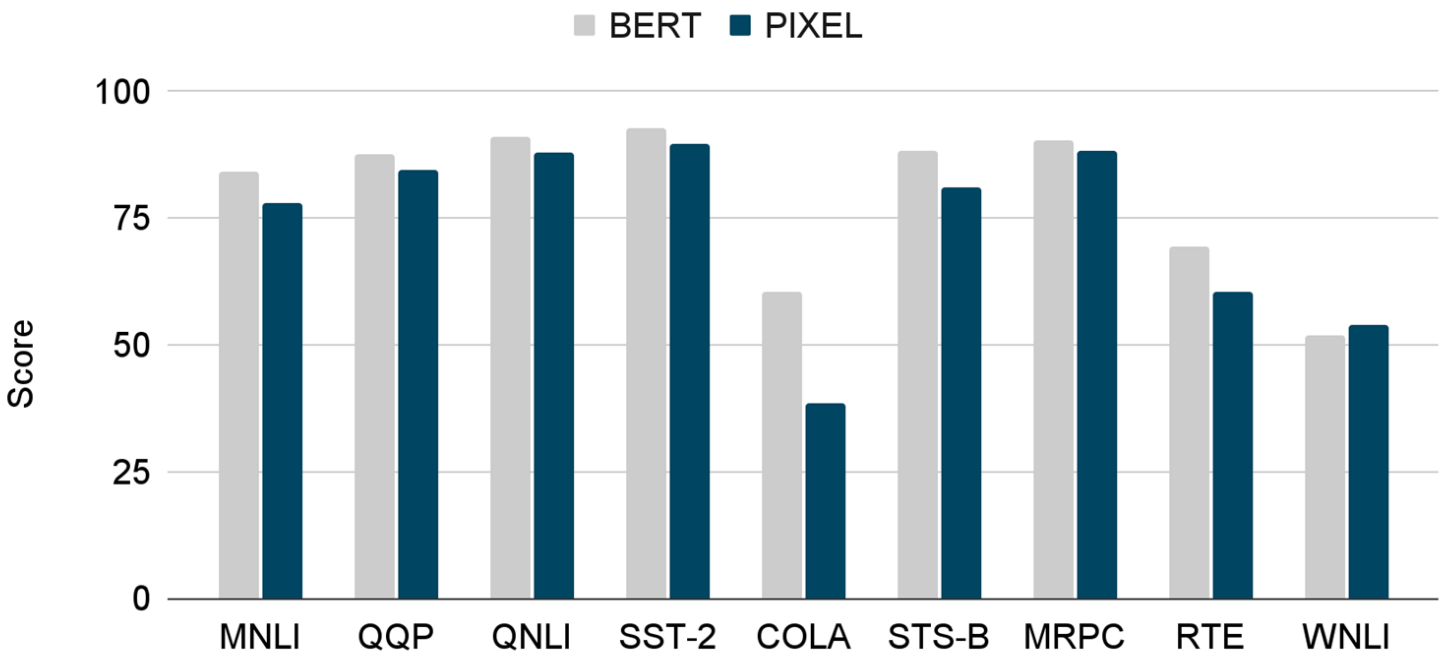
Named Entity Recognition in African Languages



PIXEL outperforms BERT on the non-Latin script

PIXEL outperforms the multilingually pretrained CANINE-C

GLUE: Sentence-level Understanding



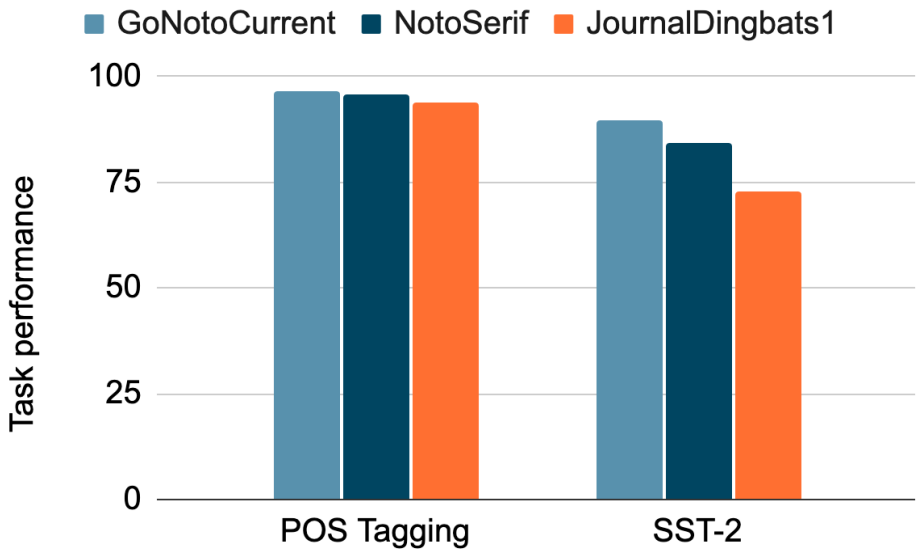
BERT outperforms PIXEL on English sentence-level tasks

Adapting to new fonts

GoNotoCurrent My cat loves oatmeal and pancakes. ■

NotoSerif-Regular My cat loves oatmeal and pancakes. ■

JournalDingbats1 ■



2. Text Rendering Matters

Text Rendering Matters

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

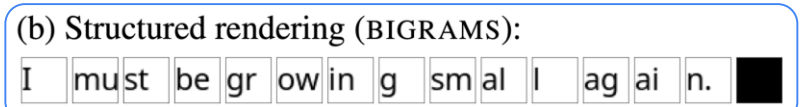
Can we do better?

Lotz et al. Text Rendering Strategies for Pixel Language Models. EMNLP 2023.

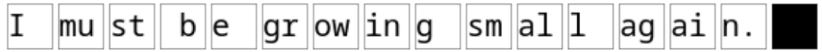
Alternative Text Renderers

(a) Continuous rendering (CONTINUOUS):

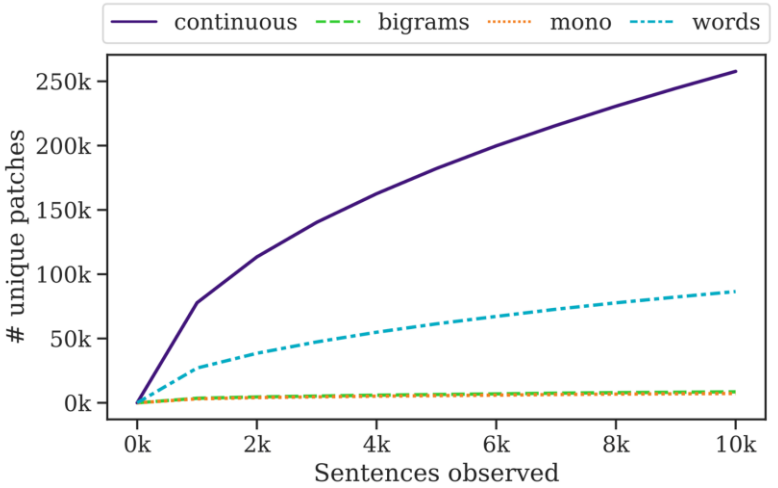
(b) Structured rendering (BIGRAMS):



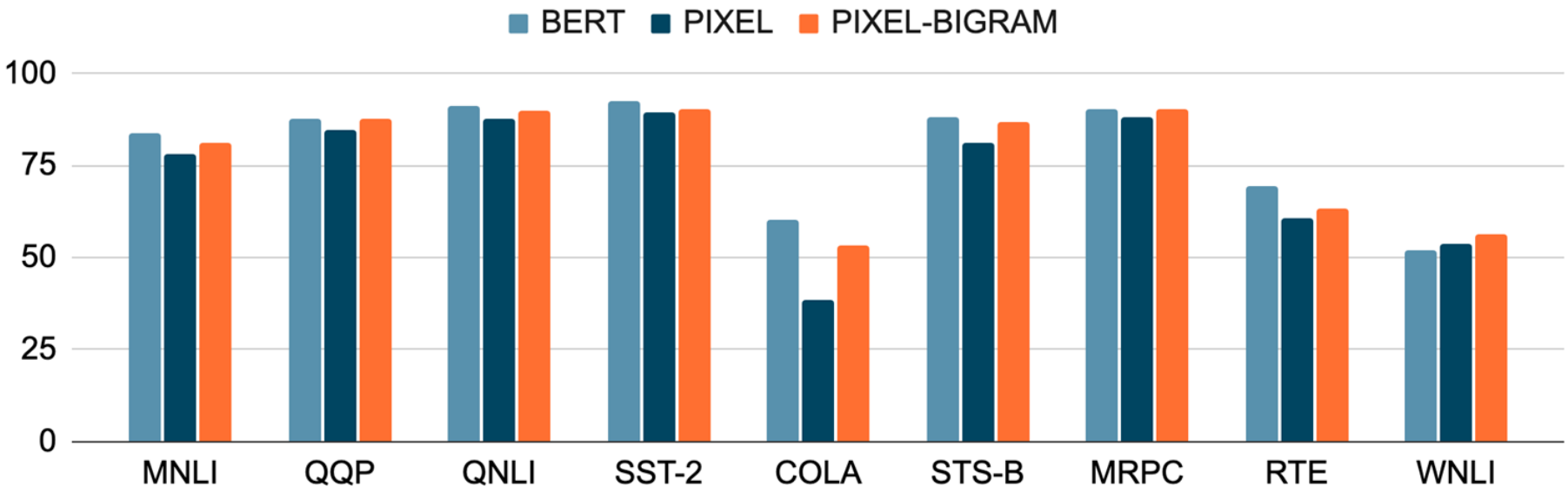
(c) Structured rendering (MONO):



(d) Structured rendering (WORDS):



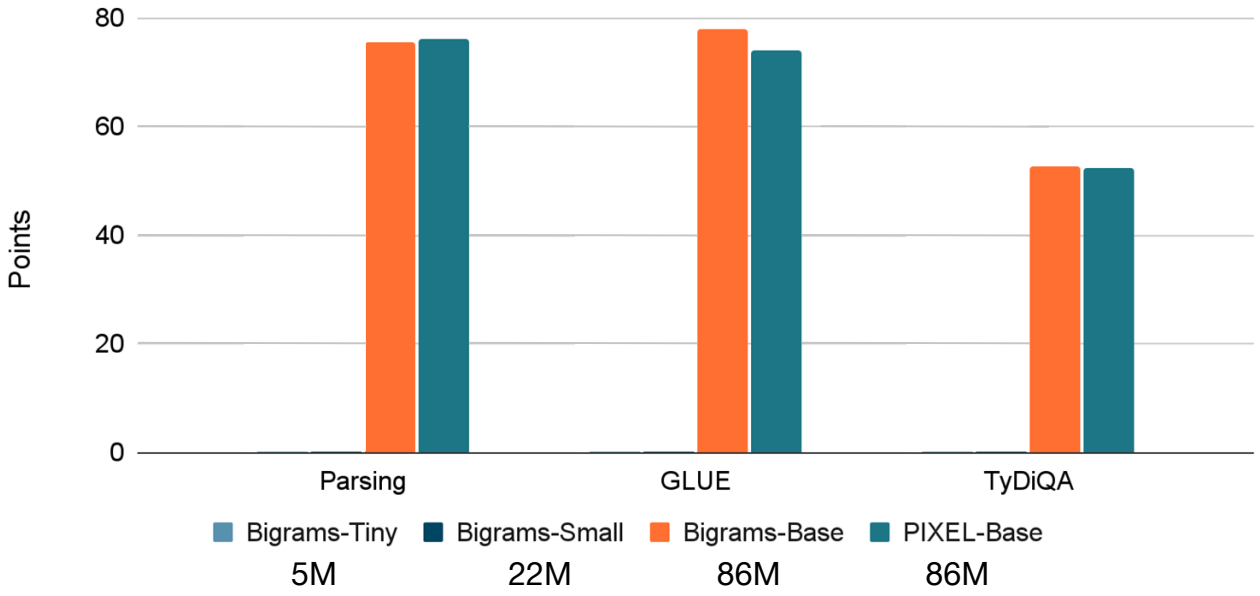
GLUE (revisited)



Bigram text rendering produces better models

Scaling Down

- Better text rendering can create effective models at smaller scales



3. Historical Document Processing

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?

Borenstein et al. PHD: Pixel-based Language modelling of Historical Documents. EMNLP 2023.

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.

St. Georges, General Store.

THE UNDERSIGNED

OXBORROW & HUGHES.

For Sale,

For Sale,

B. W. WALKER.

O. C. DUNSCOMBE

Ex barque Eliza Barss,

THE SUBSCRIBER HAS RECEIVED,

FOSTER L. BONNELL

TEAS and COFFEE.

GOSLING BROTHERS,

Mechanics' Industrial
EXHIBITION,

"Richmond Grounds."

WEDNESDAY

C. W. GAUNTLETT,

ICE.

GOSLING BROS.

VICTORIA HOTEL,

SODA WATER.

A Proclamation.

MILES GERALD KEON.

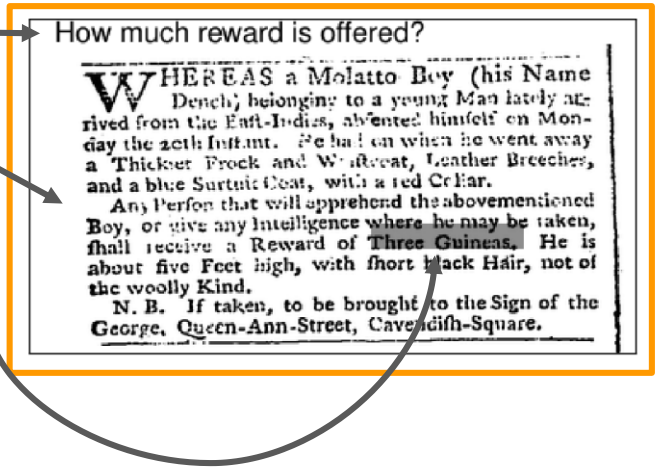
A Proclamation.

PHD: PIXEL for Historical Documents

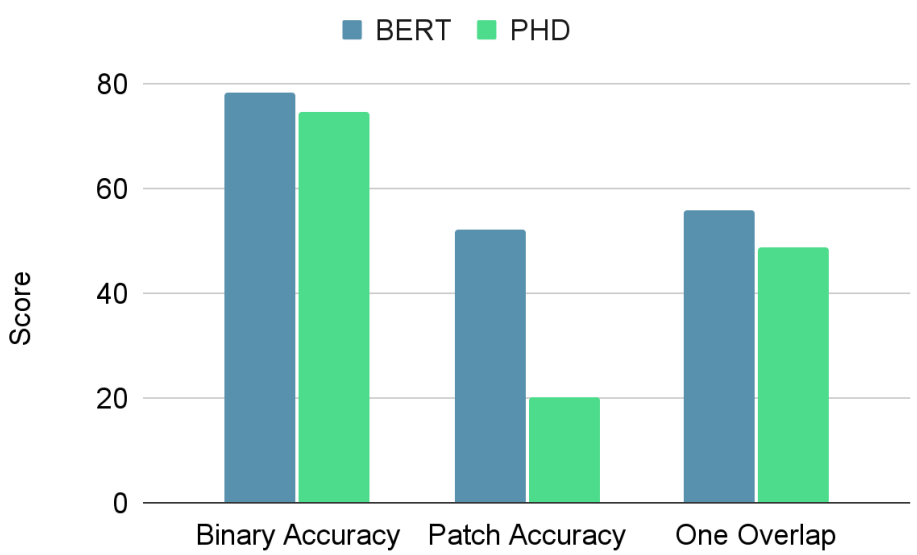
- Historical document-aware Pretraining
  - Mixture of scanned newspapers and synthetic newspaper-like text generated from Wikipedia and Bookcorpus datasets
  - All input data is scaled to 368x368 and split into 16x16 patches

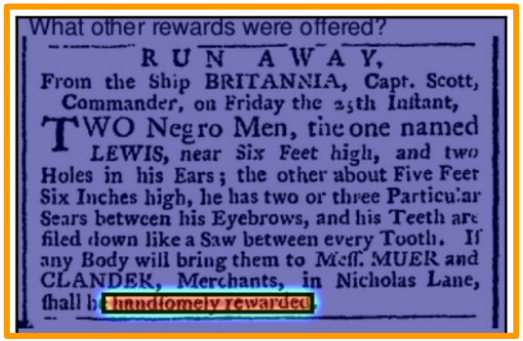
Visual Question Answering in Newspapers

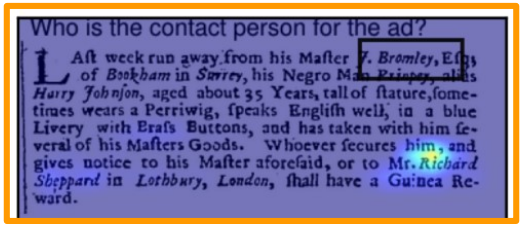
- Frame this as a Visual Question Answering Task
  - Render the question
  - Render the clipping on a canvas
  - Annotate context of answer
- Train the model to predict the label of the answer



Results







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

Conclusions

- PIXEL is a new type of language model that tackles the open vocabulary problem using visually 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. Natural interface to scanned documents

Open Questions

Thanks


N. Borenstein


P. Rust


I. Augenstein


E. Bugliarello


E. Salesky


M. de Lhoneux

Models

VILLUM FONDEN

Code