Language Modelling with Pixels

Explorations of pixel-based encoding of language

Jonas F. Lotz

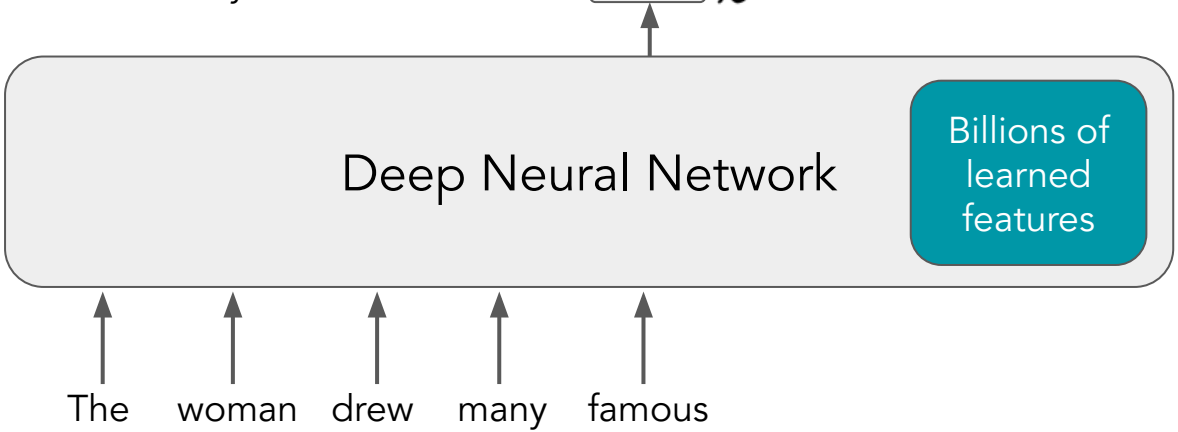
Department of Computer Science, University of Copenhagen

Large Language Models?

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

All words in vocabulary

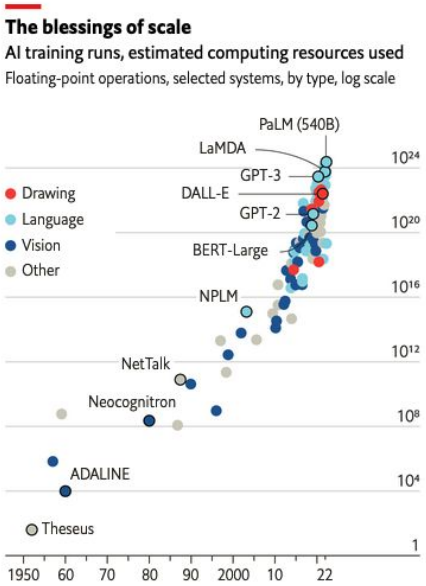
j-dimensional vector z

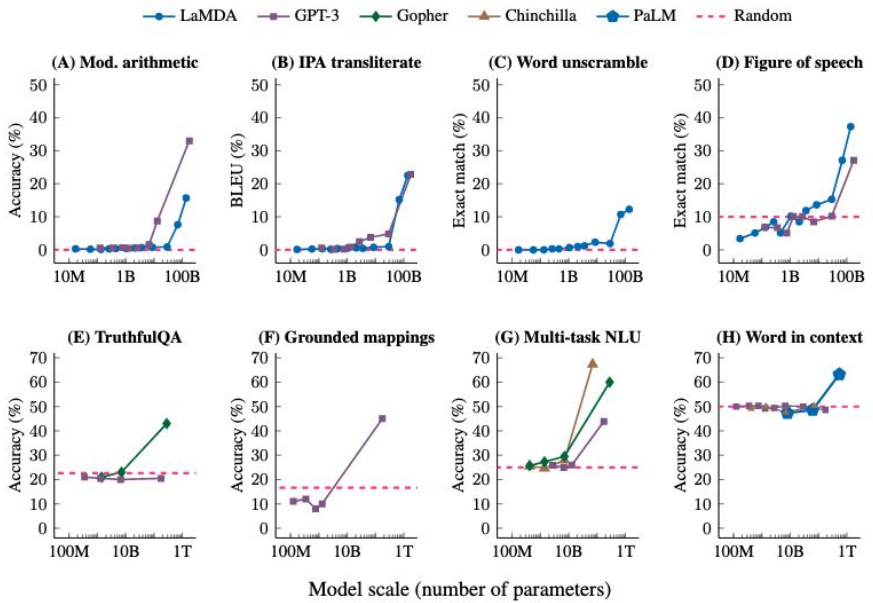


Large Language Models

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

In the Era of Scale

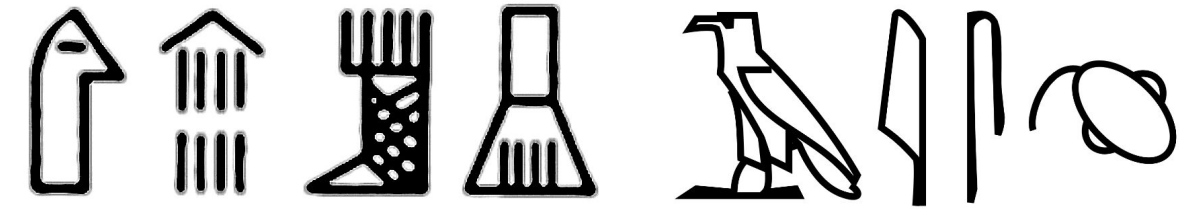




Source: www.economist.com/interactive/briefing/2022/06/11/huge-foundation-models-are-turbo-charging-ai-progress

Emergent Abilities of Large Language Models. *(Wei+ TMLR)*

What's left?

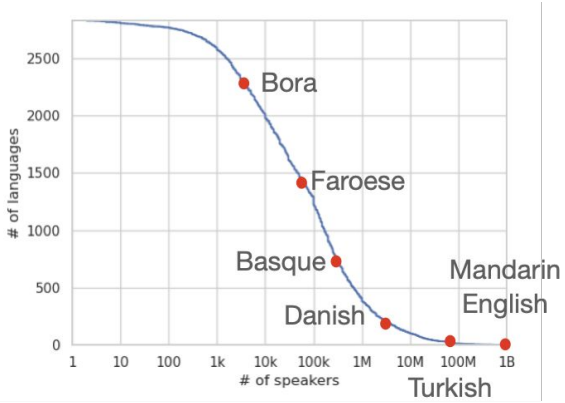


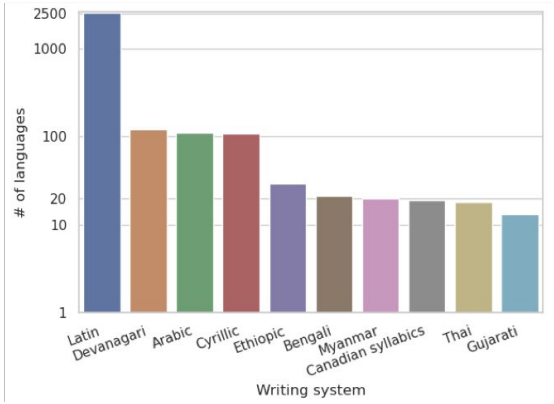
天地玄黃

ABCD अइउण्।

NLP for All Written Languages

- There are 7,000 spoken languages, of which 3,000 are written
  - There is at least 400 languages with >1M speakers
- But NLP only covers 100 languages (van Esch+ LREC22)
  - Lack of technological inclusion for billions of people

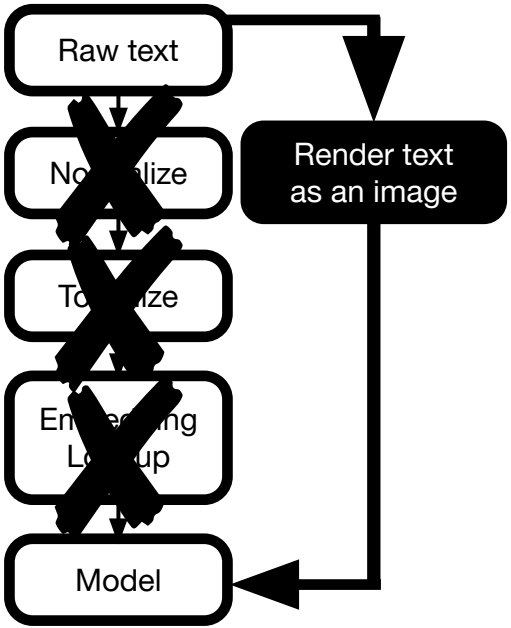




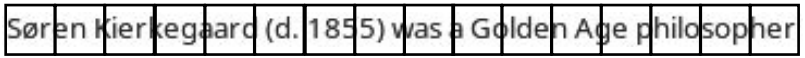
Slide credit: Sebastian Ruder

Today: Pixel-based Language Modelling

- Key insight: treat language processing as visual processing

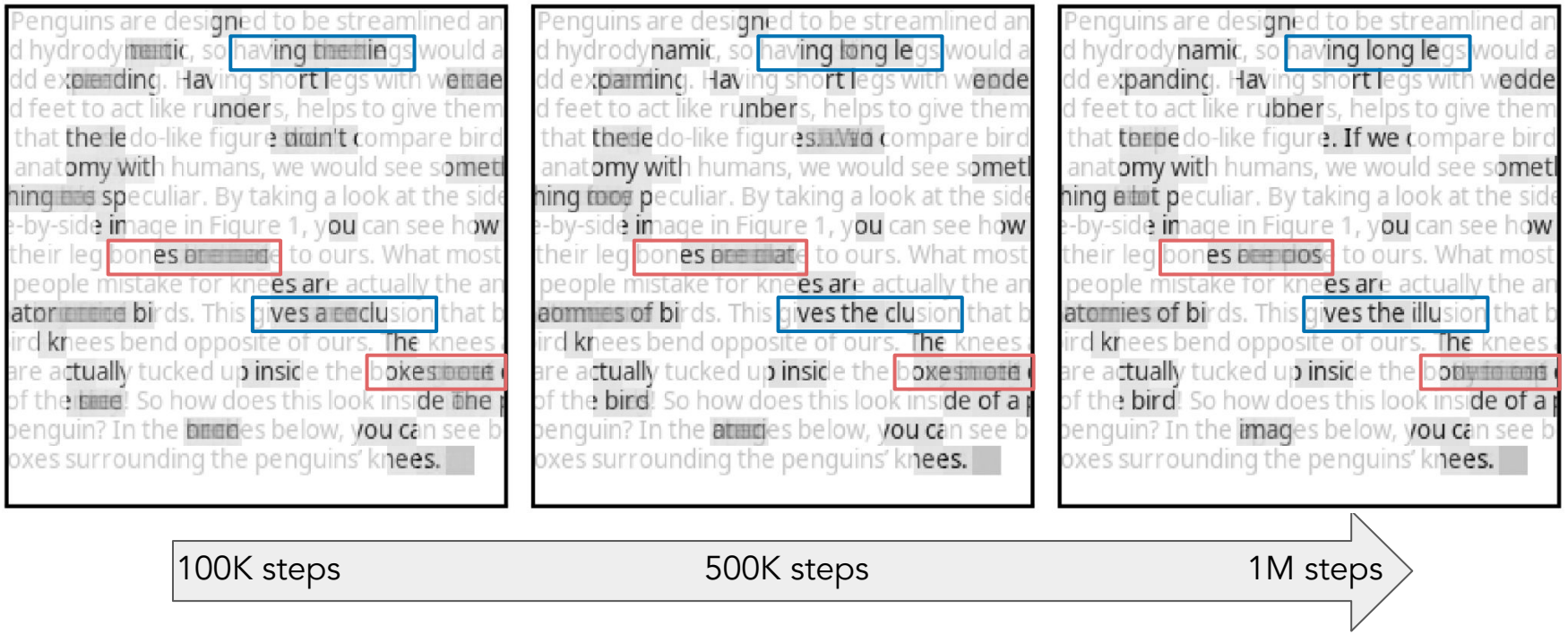


Søren Kierkegaard (d. 1855) was a Golden Age philosopher



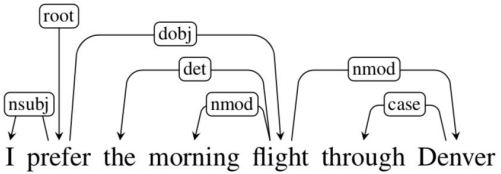
ኢትዮጵያ አፍሪካ ውስጥ ናት

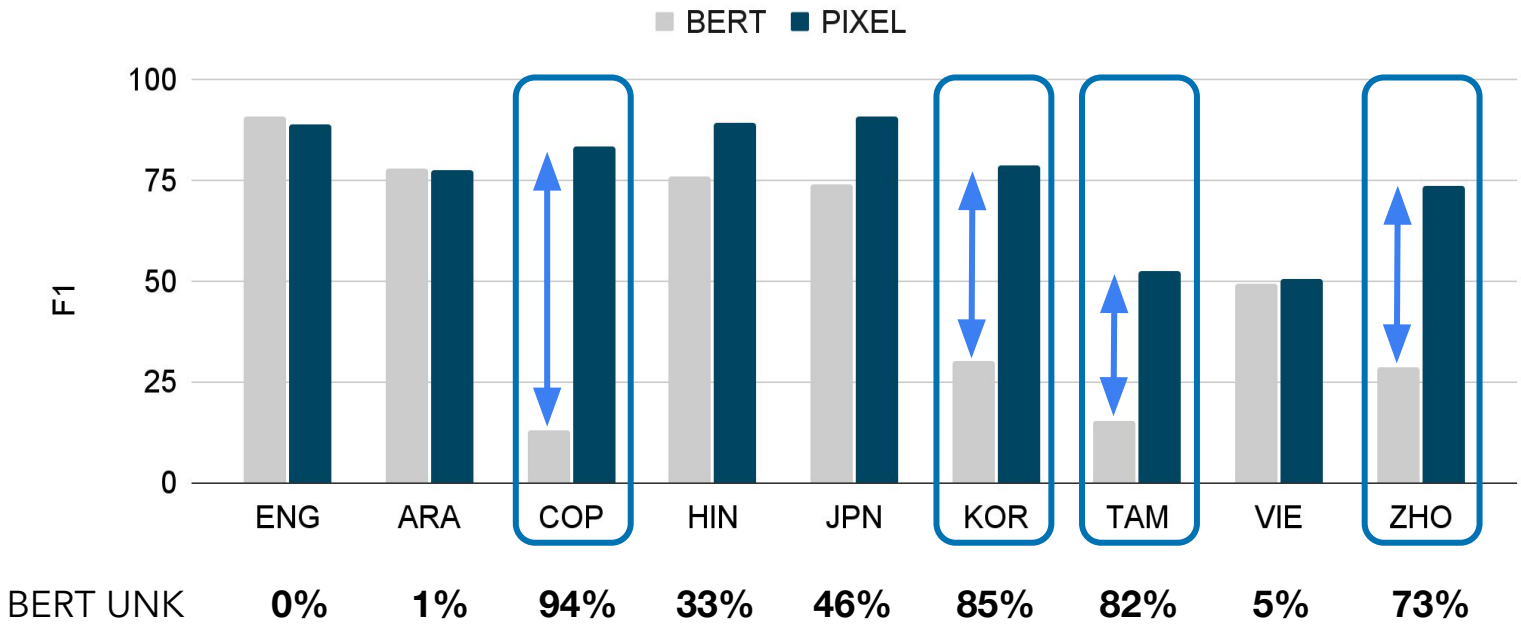
A new type of generative model



Demo: https://huggingface.co/Team-PIXEL/pixel-base

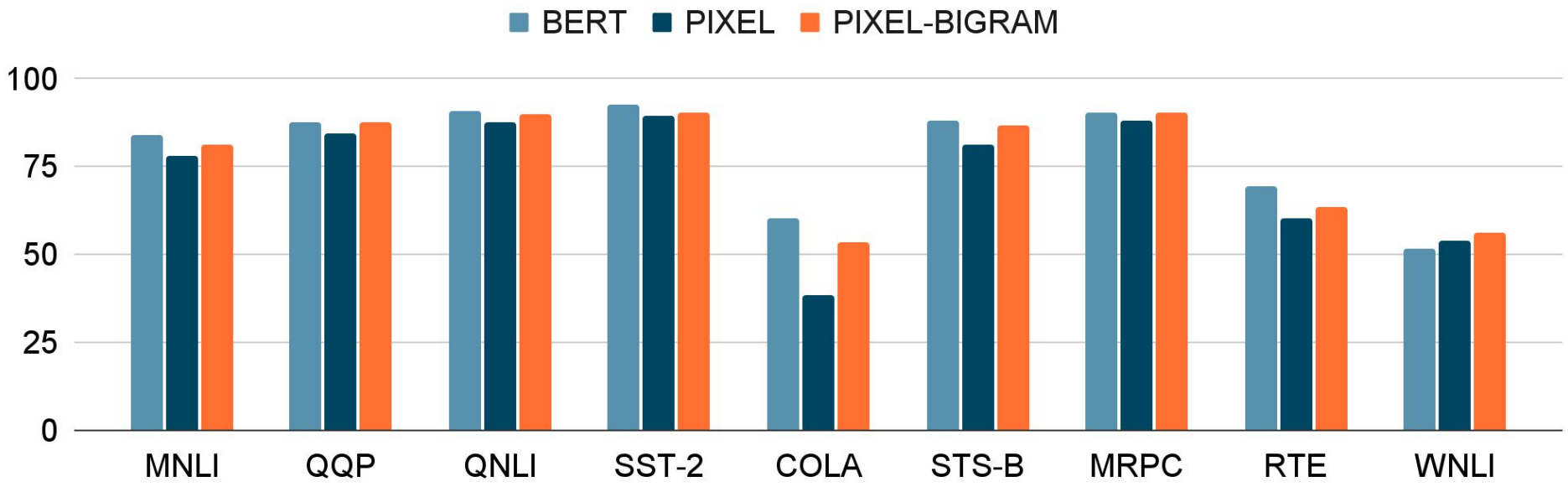
Results: Dependency Parsing





PIXEL vastly outperforms BERT on unseen scripts

Exploring Text Rendering Strategies



Bigram text rendering produces much better models

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)

Pretrained model: https://huggingface.co/Team-PIXEL/pixel-base

Ongoing Work

- Improve sentence level reasoning tasks
  - Contrastive objective or multi-scale modelling
- Scale to multilingual pretraining
  - Better representations from reconstructing multiple scripts?
- Understanding cross-script transfer
  - What causes the PIXEL model transfer well to other scripts?