PHD: Pixel-Based Language Modeling of Historical Documents

Nadav Borenstein, Phillip Rust, Desmond Elliott, Isabelle Augenstein

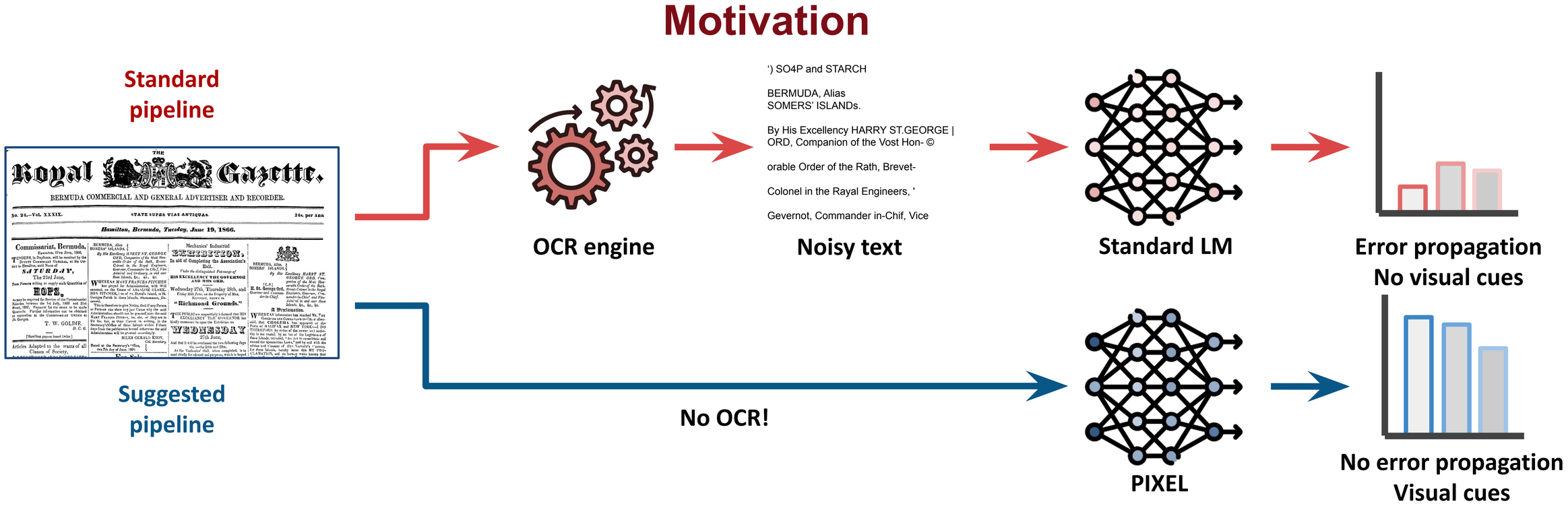
Department of Computer Science
University of Copenhagen
Denmark

nb@di.ku.dk | p.rust@di.ku.dk | de@di.ku.dk | augenstein@di.ku.dk

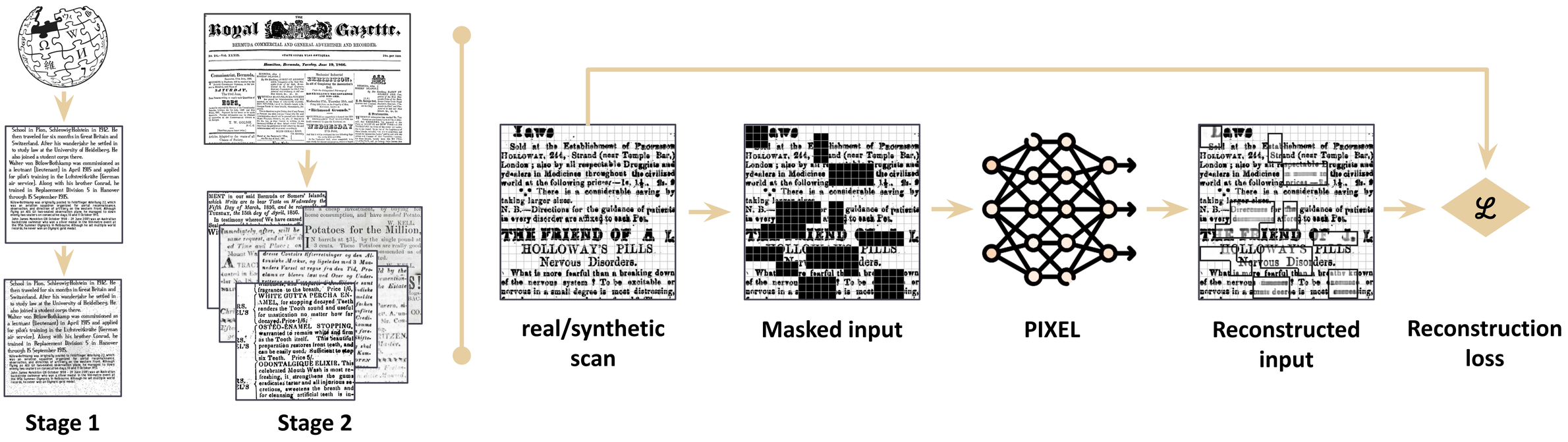
ArXiv Link

UNIVERSITY OF COPENHAGEN

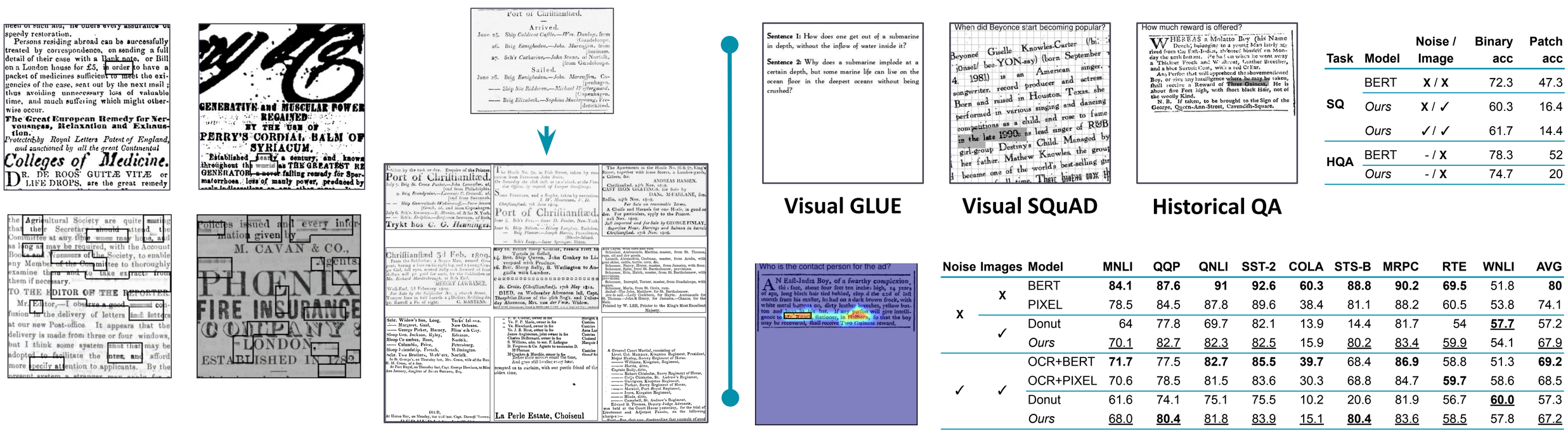
Motivation



Method



Evaluation



Task	Model	Noise / Image	Binary acc	Patch acc
SQ	BERT	X / X	72.3	47.3
	Ours	X / ✓	60.3	16.4
	Ours	✓ / ✓	61.7	14.4
HQA	BERT	- / X	78.3	52
	Ours	- / X	74.7	20

Noise Images		Model	MNLI	QQP	QNLI	SST-2	COLA	STS-B	MRPC	RTE	WNLI	AVG
X	X	BERT	**84.1**	**87.6**	**91**	**92.6**	**60.3**	**88.8**	**90.2**	**69.5**	51.8	**80**
		PIXEL	78.5	84.5	87.8	89.6	38.4	81.1	88.2	60.5	53.8	74.1
	✓	Donut	64	77.8	69.7	82.1	13.9	14.4	81.7	54	**57.7**	57.2
		Ours	70.1	82.7	82.3	82.5	15.9	80.2	83.4	59.9	54.1	67.9
✓	✓	OCR+BERT	**71.7**	77.5	**82.7**	**85.5**	**39.7**	68.4	**86.9**	58.8	51.3	**69.2**
		OCR+PIXEL	70.6	78.5	81.5	83.6	30.3	68.8	84.7	**59.7**	58.6	68.5
		Donut	61.6	74.1	75.1	75.5	10.2	20.6	81.9	56.7	**60.0**	57.3
		Ours	68.0	**80.4**	81.8	83.9	15.1	**80.4**	83.6	58.5	57.8	67.2

Conclusions

- **PHD is an OCR-free language encoder** designed for analysing historical documents at the pixel level.
- It Uses a novel pre-training method – **a combination of synthetic scans and real historical newspapers**
- PHD can reconstruct masked image patches, and has language understanding capabilities.
- **PHD is successfully applied to historical and modern tasks**

Limitations

- English only datasets
- Training set contains a mixture of modern and historical texts. The historical corpus is not diverse enough
- Quantitative evaluation of reconstructed image regions is unclear
- Limited compute and dataset size, the full potential of the approach is not explored