BIG BIRD AND XLNET

Jonah Winninghoff

THESIS

Two text summarizations are compared using specific metrics and a timer.

Transferred Learnings: Big Bird and XLNet Transformers

Metrics: Recall-Oriented Understudy for Gisting Evaluation (ROUGE)

Timer with CPU 1.6 GHZ

THESIS

Two text summarizations are compared u[...]

Transferred Learnings: Big Bird

Metrics: Recall-Oriented Under[...]

Timer with CPU 1.6 GHZ

BUT SO WHAT?

The Transformers has self-attention mechanism which is expensive to compute longer sequences.

The Google Research team attempts to solve the issue using block sparsity.

Their mathematical assessment shows that the approach reduces quadratic dependency to linear dependency in time or memory term, which is skeptical.
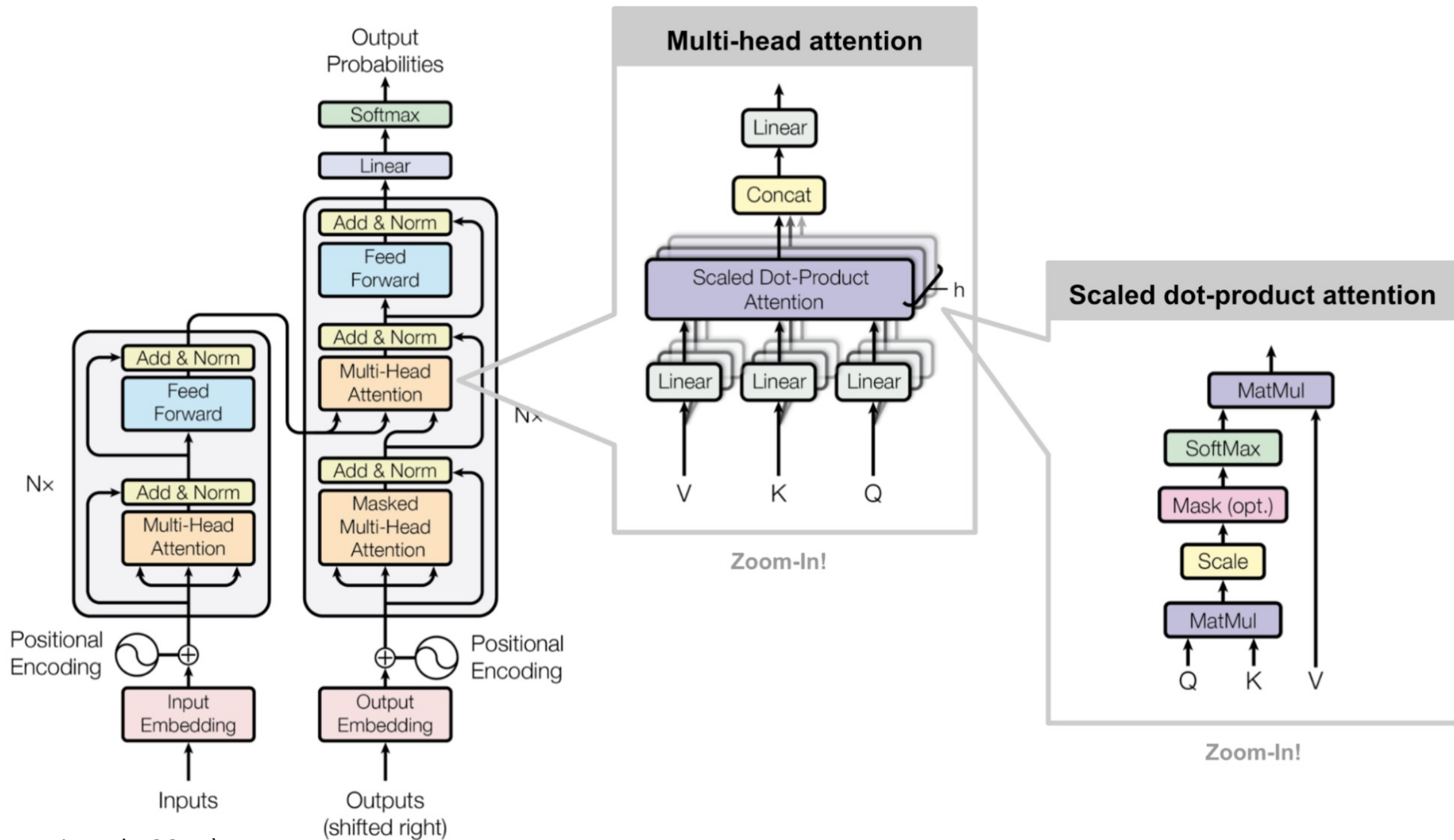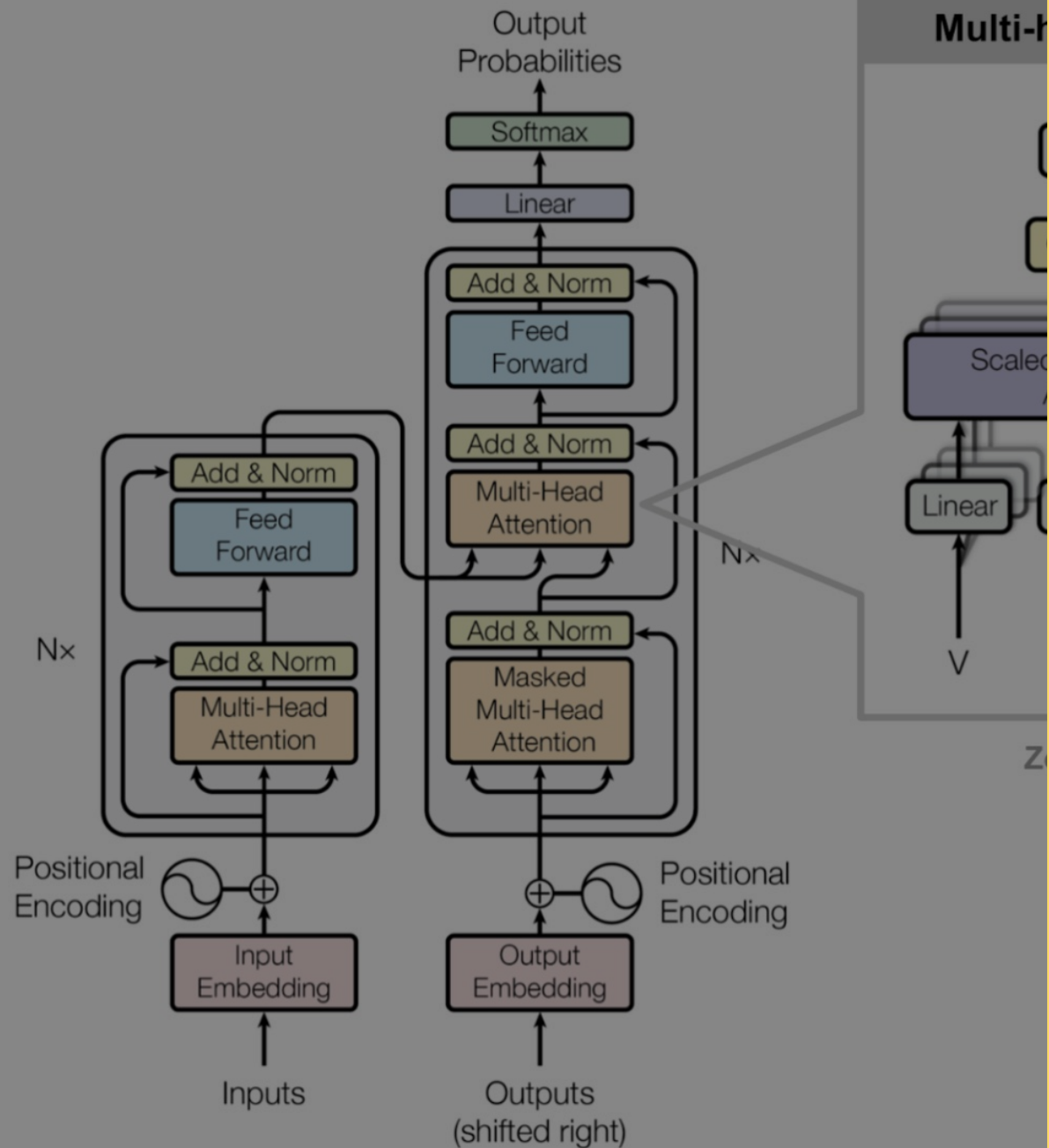
OUTLINE

- Describe Transformers and Big Bird architectures

- Explain method, dataset, and research questions

- Share actionable insights and future research ideas

ısing specific metrics and a timer.

and XLNet Transformers

study for Gisting Evaluation (ROUGE)

Multi-head attention

Scaled dot-product attention

Output
Probabilities

Softmax

Linear

Add & Norm

Feed
Forward

Add & Norm

Multi-Head
Attention

Nx

Add & Norm

Masked
Multi-Head
Attention

Positional
Encoding

Output
Embedding

Outputs
(shifted right)

Add & Norm

Feed
Forward

Nx

Add & Norm

Multi-Head
Attention

Positional
Encoding

Input
Embedding

Inputs

Linear

Concat

Scaled Dot-Product
Attention

h

Linear   Linear   Linear

V   K   Q

MatMul

SoftMax

Mask (opt.)

Scale

MatMul

Q   K   V

Zoom-In!

Zoom-In!

(Vaswani et al., 2017)

TRANSFORMERS ARCHITECTURE

The representation of encoder is the word embedding of X (x_1, ..., x_n), such as article text.

The representation of decoder is the word embedding of Z (z_1, ..., z_n), such as ground-truth summary.

Multi-head attention contains $head_i$ including attention.

FORMULAS:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

$$head_i = Attention\left(QW_i^Q, KW_i^K, VW_i^V\right)$$
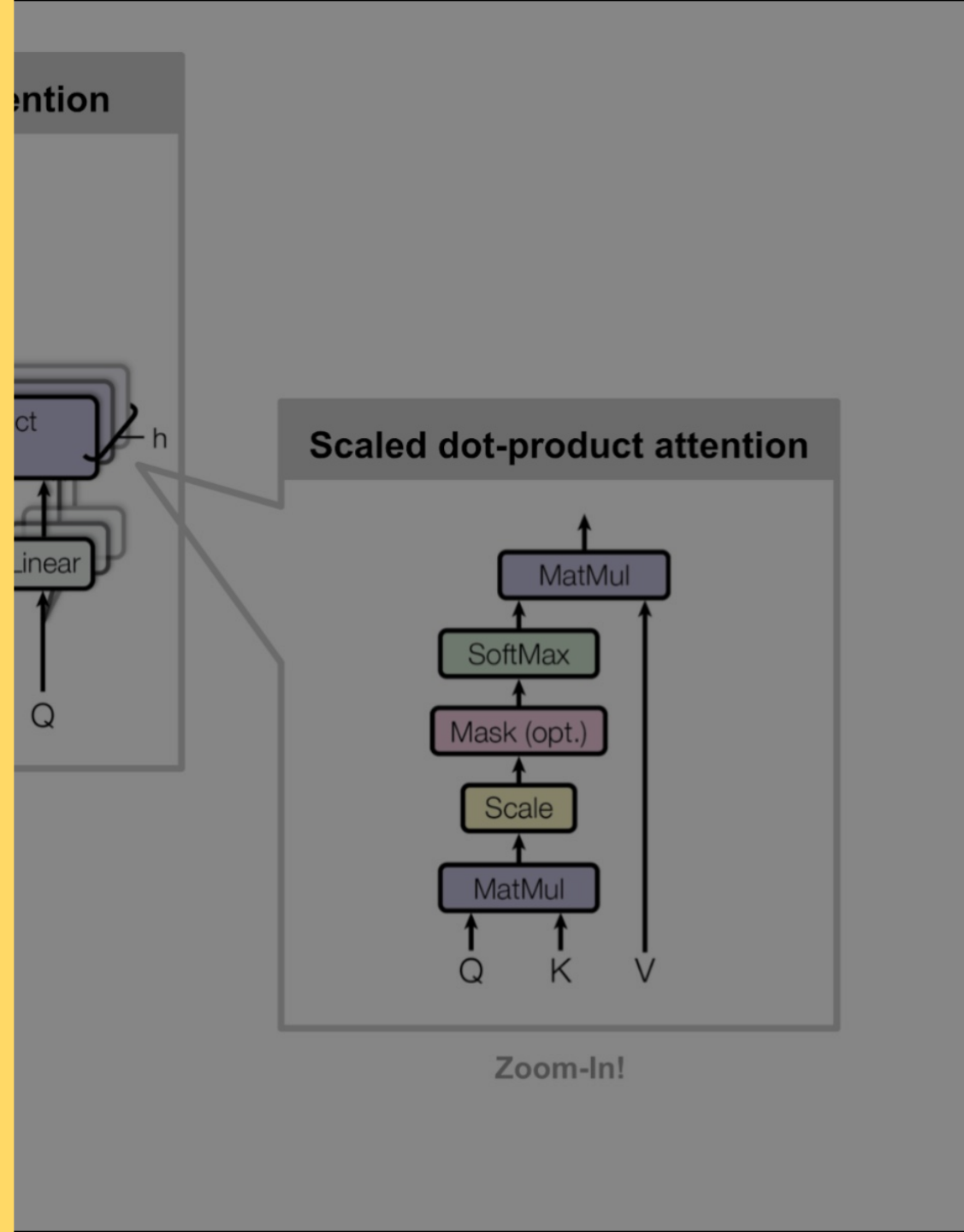
$$MultiHead(Q, K, V) = Concat(head_i, ..., head_h)$$

Q = a matrix of queries
K = a matrix of keys
V = a matrix of values (or weights)
W = a matrix of weights
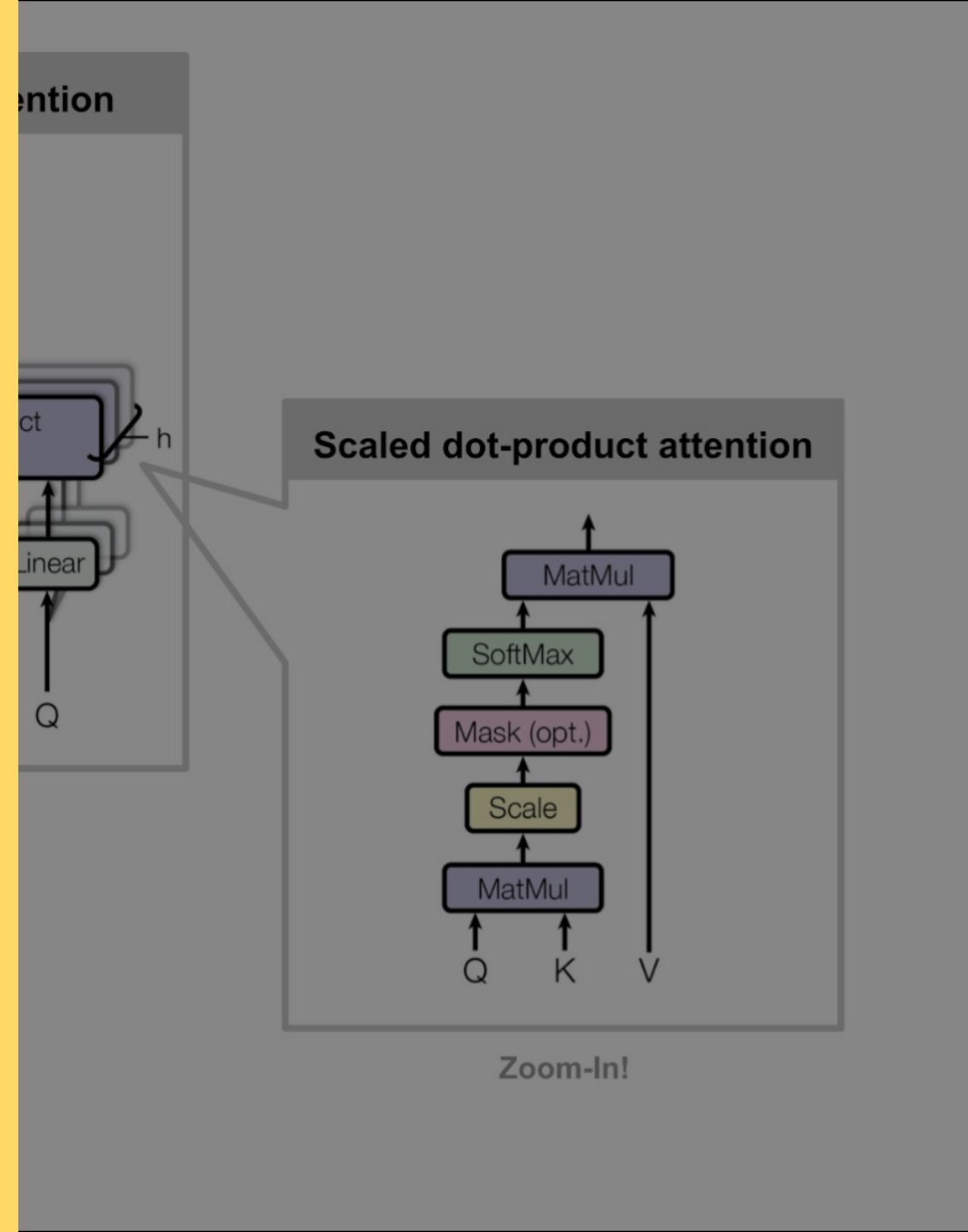d_k = dimensionality of key
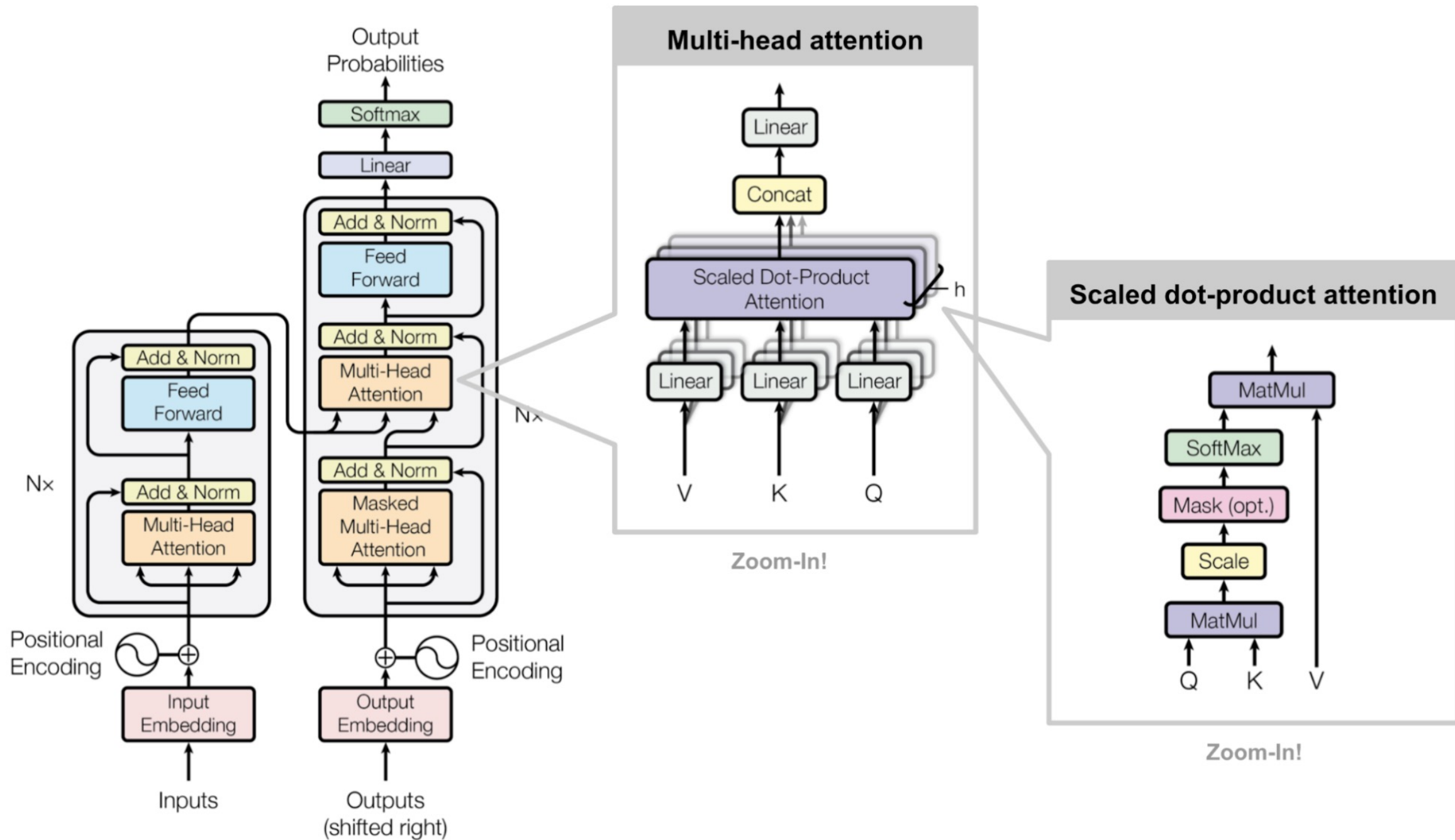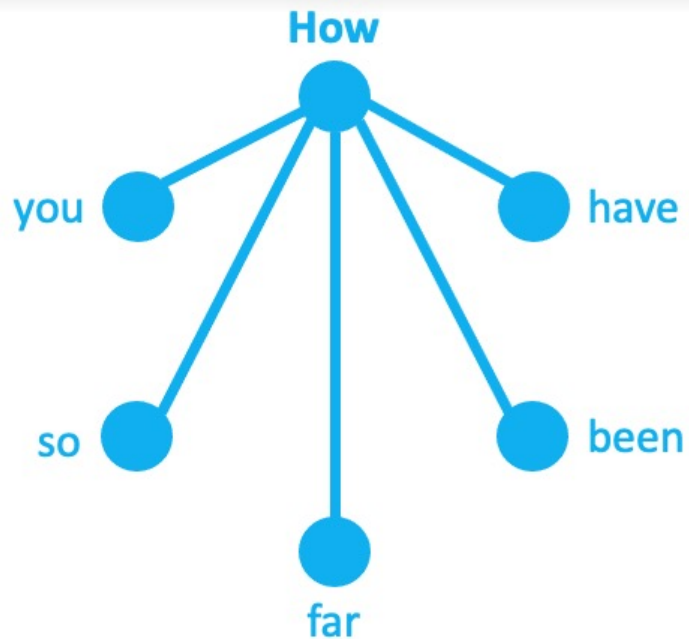

Scaled dot-product attention

Zoom-In!

XLNET

For this research, the XLNet is used. This model is different by the maximum log likelihood of the sequence *wrt* permutation being used.
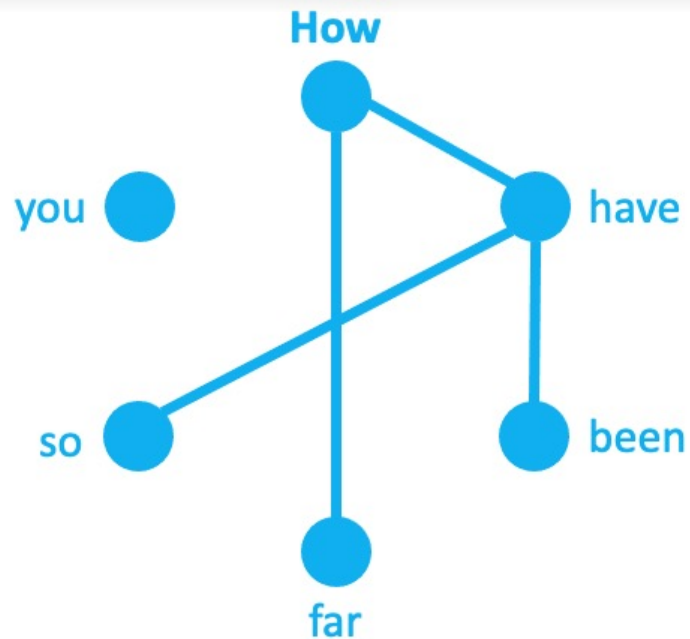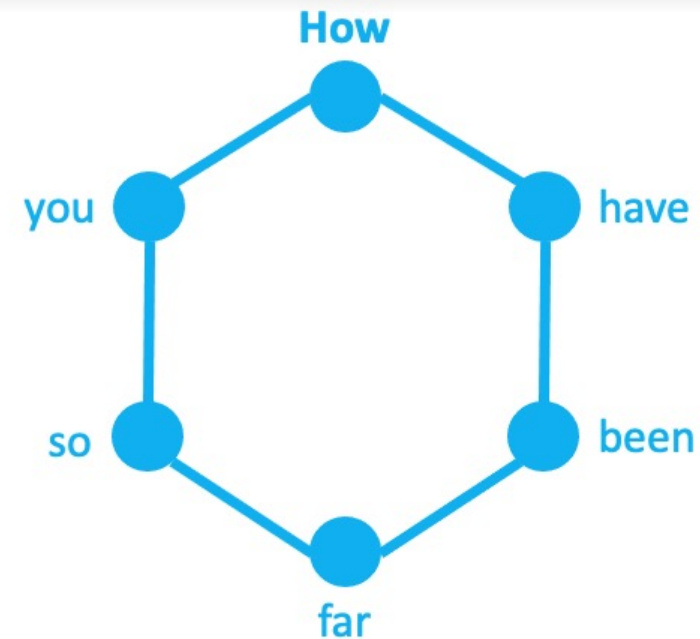

Scaled dot-product attention

Zoom-In!

Output Probabilities

Softmax

Linear

Add & Norm

Feed Forward

Add & Norm

Multi-Head Attention

Add & Norm

Masked Multi-Head Attention

Nx

Add & Norm

Feed Forward

Add & Norm

Multi-Head Attention

Nx

Positional Encoding

Positional Encoding

Input Embedding

Output Embedding

Inputs

Outputs (shifted right)

Multi-head attention

Linear

Concat

Scaled Dot-Product Attention

h

Linear   Linear   Linear

V   K   Q

Zoom-In!

Scaled dot-product attention

MatMul

SoftMax

Mask (opt.)

Scale

MatMul

Q   K   V

Zoom-In!

BIG BIRD ARCHITECTURE Block Sparse Attention

Global Connection

Random Connection

Sliding Connection

Block Sparsity

How

you   have

so   been

far

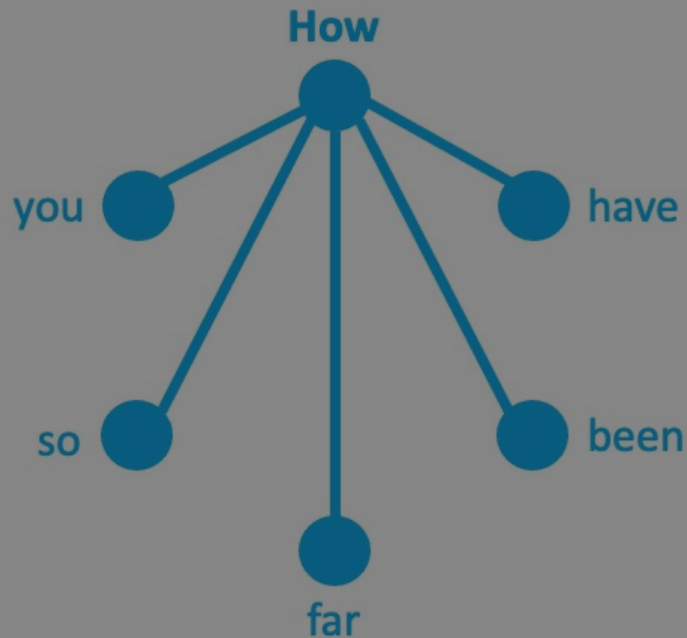Global Connection   Random Connection

BLOCK SPARSITY

The concept of sliding and global connections is not novel but what is new is random connection.
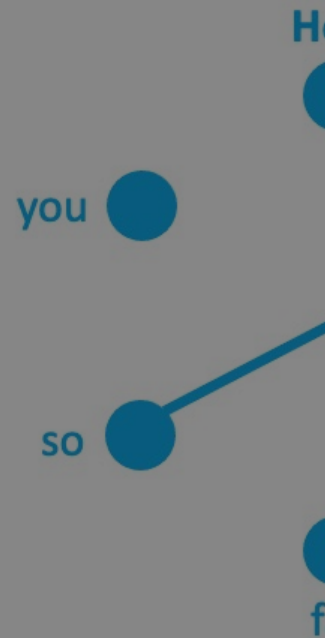
Perhaps, the Google Research team develops the random connection based on CLT and LLN.

For example, predicted summaries become more consistent when the sequence length is longer.

That comes with the price of no theoretical guarantees.

METHOD

Partial NLP data science pipeline and Randomized Controlled Trials (RCTs)

Variables: article, actual abstract, predicted abstract, word counts for each, and type of model.

Target variables: time per predicted abstract (in seconds), ROUGE-1 F1 score, ROUGE-2 F1 score, and ROUGE-L F1 score

METHOD

- Partial NLP data science pipe…

- Variables: article, actual abs… each, and type of model.

- Target variables: time per pr… score, ROUGE-2 F1 score, and…

ROUGE-N

ROUGE-N F1-score is the measure of model accuracy based on the number of matching n-grams between predicted and ground-truth summaries.

For example, ROUGE-1 means the number of matching unigram.

ROUGE-1 means the number of matching the longest common subsequence (LCS).

DATASET

The arXiv journals prepared by TensorFlow is used, which contains *article_id*, *article_text*, and *actual abstract_text*.

Three subsets: testing (6,658 entities), training (119,924 entities), and validation (6,633 entities) sets.

70.8% of tokens in article texts matches NLTK dictionaries while 62.05% in abstract text matches these dictionaries

...ine and Randomized Controlled Trials (RCTs)

...stract, predicted abstract, word counts for

...edicted abstract (in seconds), ROUGE-1 F1
...d ROUGE-L F1 score

DATASET

For this research, validation set is in use. Technically, this set is unseen.

Why?

- Big Bird model is pretrained with Wikipedia dataset.
- XLNet model is pretrained with several datasets other than arXiv.

Random sampling for this set to predict is 110 in total for each model. This data collection takes two days to compute.

ine and Randomized Controlled Trials (RCTs)

stract, predicted abstract, word counts for

redicted abstract (in seconds), ROUGE-1 F1 ROUGE-L F1 score
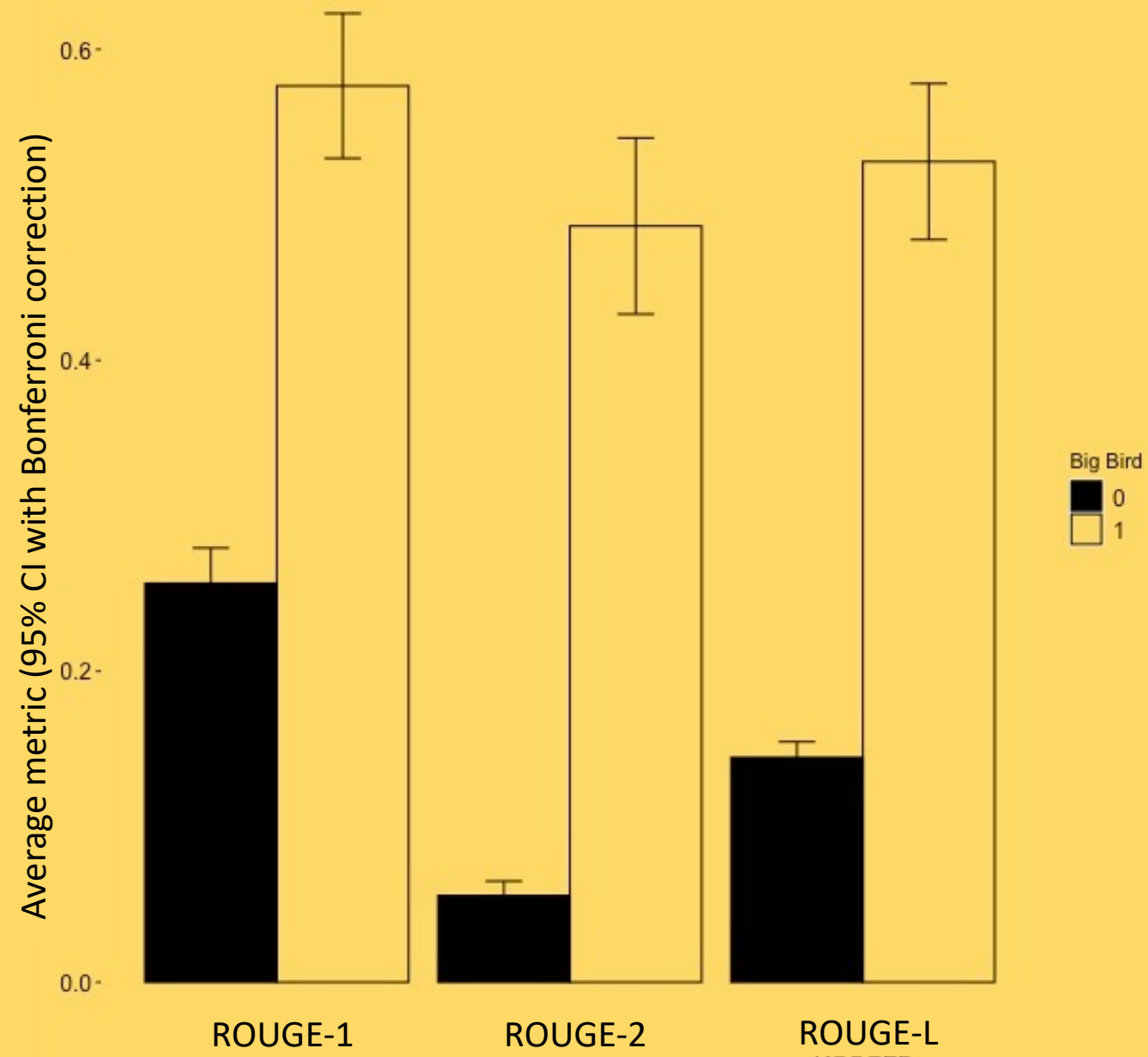
RESEARCH QUESTIONS

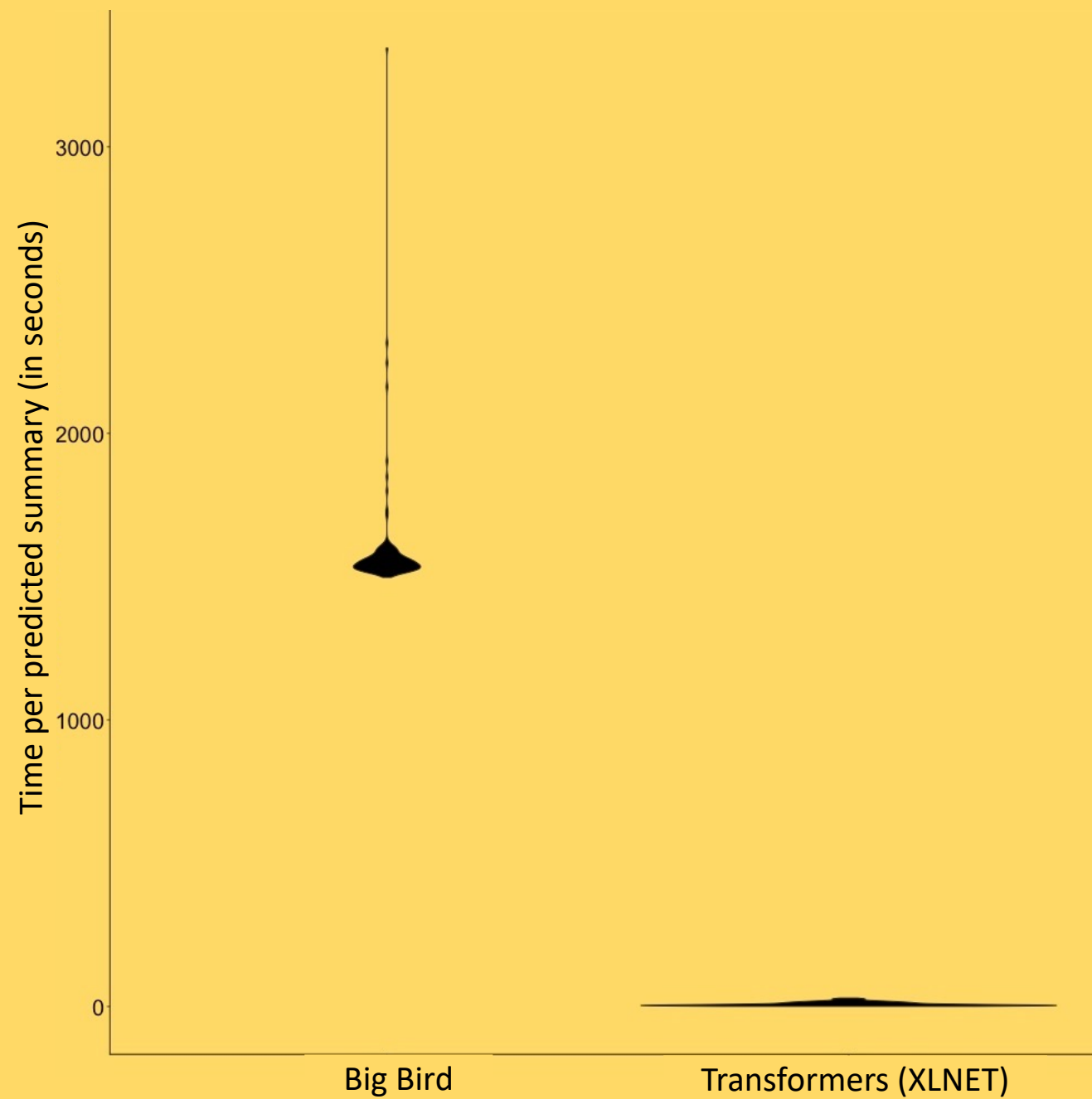**?** Does the Big Bird model outperform XLNet model?

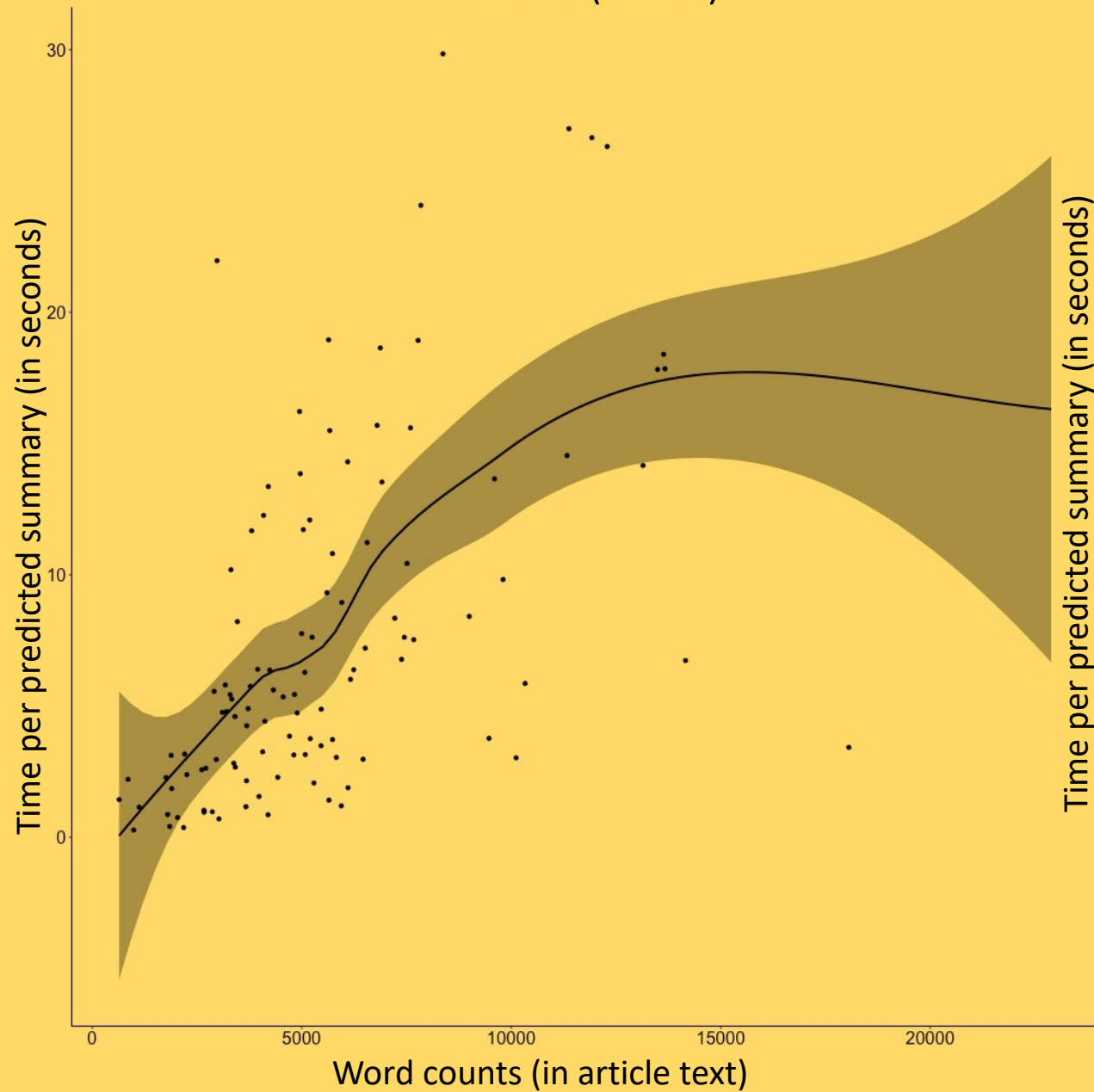**?** Being compared with XLNet model, how fast can Big Bird produce each predicted summary?

? Does the Big Bird successfully reduce this quadratic dependency to linear dependency in sequence term?
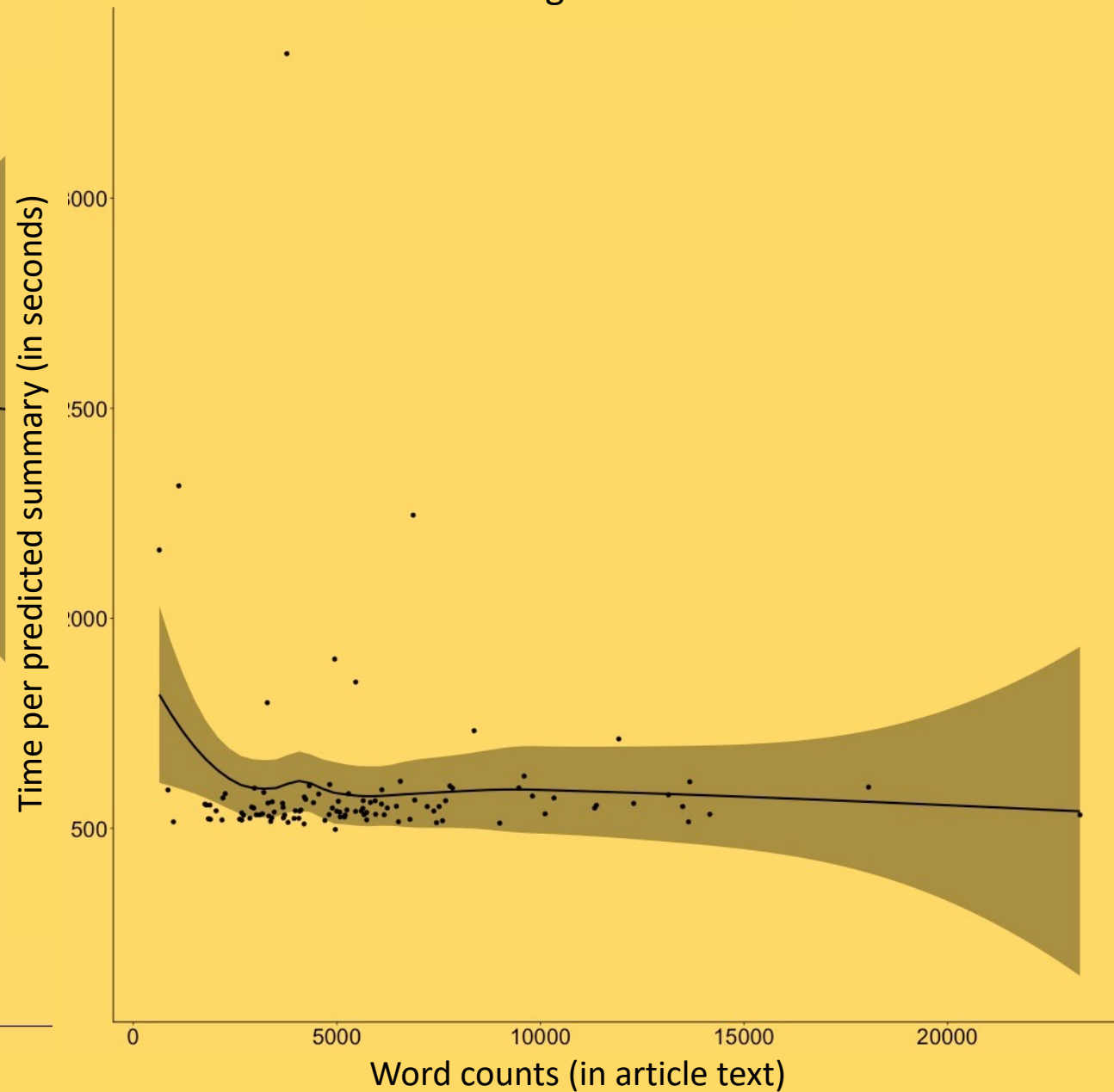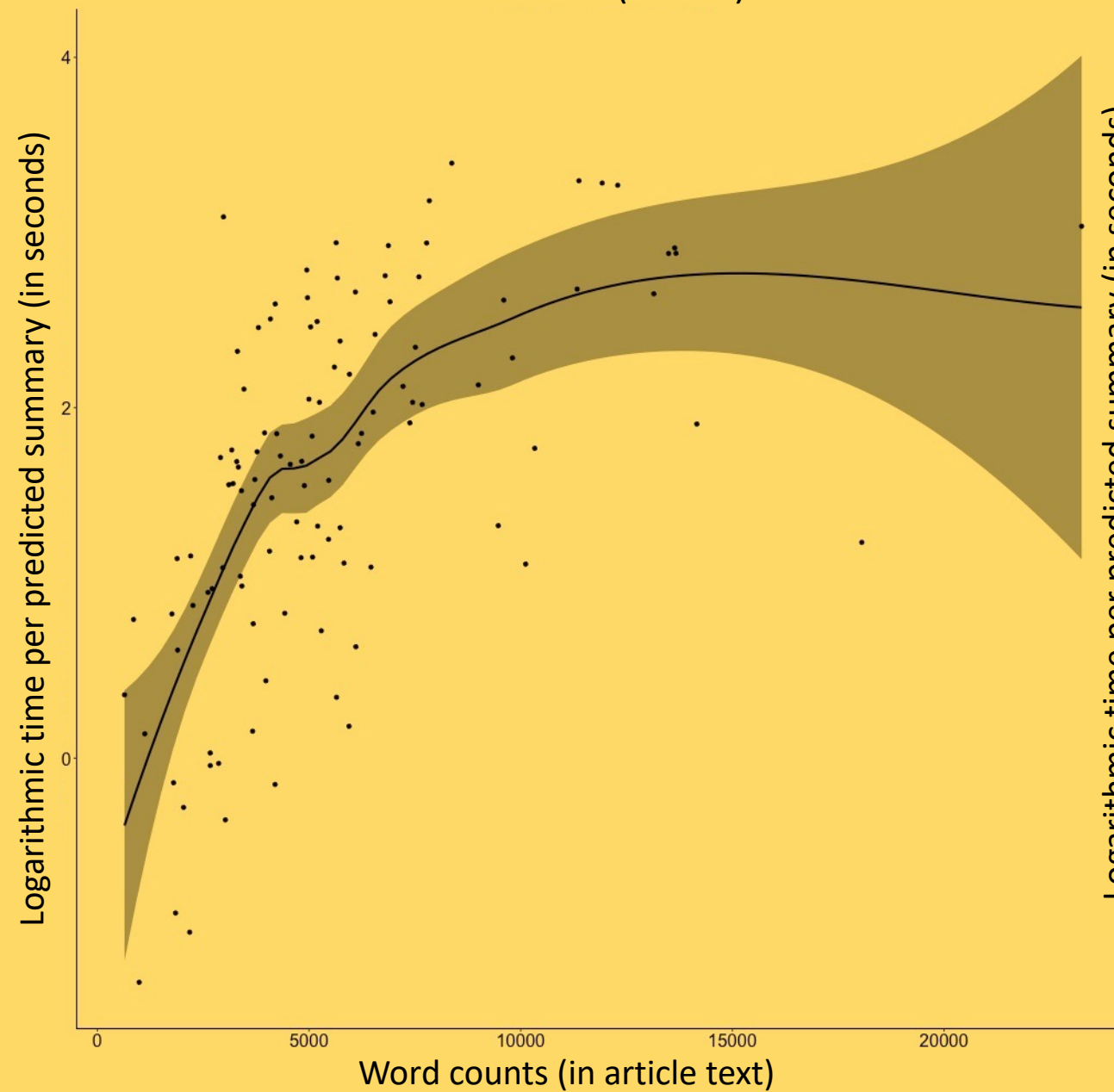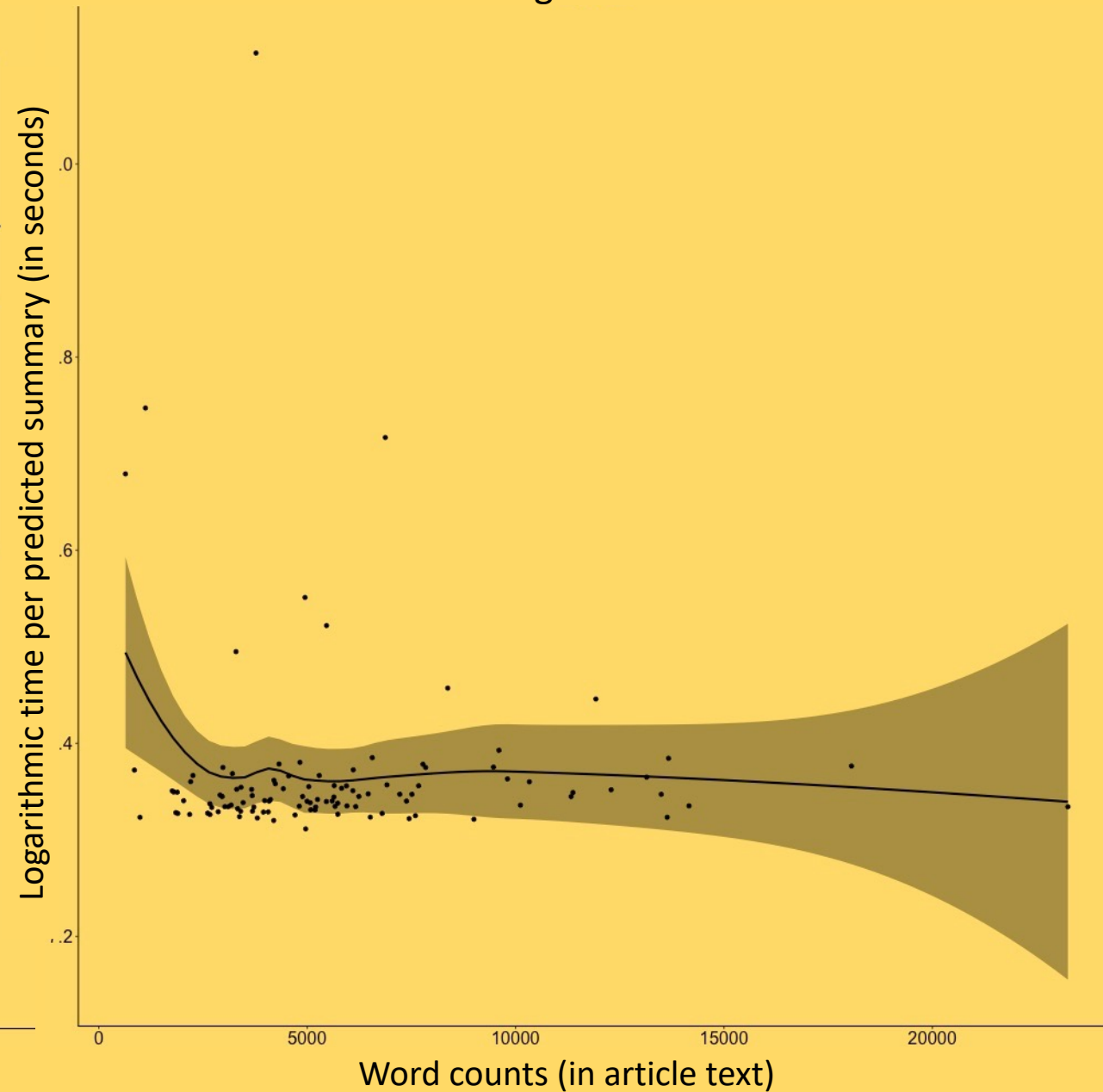
FUTURE RESEARCH AND CONCLUSION

Big Bird model does better with predicting summary and successfully linearize self-attention. However, the speed of this model is 193.04 wpm by median.

The Big Bird algorithm is highly recommended for producing summaries as long as if the cloud environment is used.

To address scalability and redundancy problems, Attention Free Transformer and Bayesian connection need to be tested with block sparsity.

REFERENCES

Brochu, E., Cora, VM., and Freitas, N. "A Tutorial on Bayesian Optimization of Expensive Cost Functions, With Application to Active User Modeling and Hierarchical Reinforcement Learning." arXiv, Dec. 2020. https://www.math.umd.edu/~slud/RITF17/Tutorial_on_Bayesian_Optimization.pdf

Child, R., Gray, S., Radford, A., and Sutskever, I. "Generating Long Sequences with Sparse Transformers." arXiv, 2019. https://arxiv.org/pdf/1904.10509.pdf

Rayner, K., Slattery, TJ., and Bélanger, NN. "Eye movements, the perceptual span, and reading speed." Psychon Bull Rev., Dec. 2010. doi: 10.3758/PBR.17.6.834

Gupta, V. "Understanding BigBird's Block Sparse Attention." Huggingface, Mar. 2021. https://huggingface.co/blog/big-bird

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, Ł., Gomez, AN., Kaiser, L., and Polosukhin, I. "Attention is All You Need." Advances in Neural Information Processing Systems 30. NIPS, 2017. https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., and Le., QV. "XLNet: Generalized Autoregressive Pretraining for Language Understanding." arXiv, Jan. 2020.https://arxiv.org/pdf/1906.08237v2.pdf

Zaheer, M., Guruganesh, G., Dubey, A., Ainslie, J., Alberti, C., Ontanon, S., Philip, P., Ravula, A., Wang, Q., Yang, L., and Amr Ahmed, A. "Big Bird: Transformers for Longer Sequences." Advances in Neural Information Processing Systems 33, NeurIPS, 2020. https://papers.nips.cc/paper/2020/file/c8512d142a2d849725f31a9a7a361ab9-Paper.pdf

Zhai, S., Talbott, W., Srivastava, N., Huang, C., Goh, H., Zhang, R., and Susskind, J. "An Attention Free Transformer." arXiv, Sep. 2021. https://arxiv.org/pdf/2105.14103.pdf