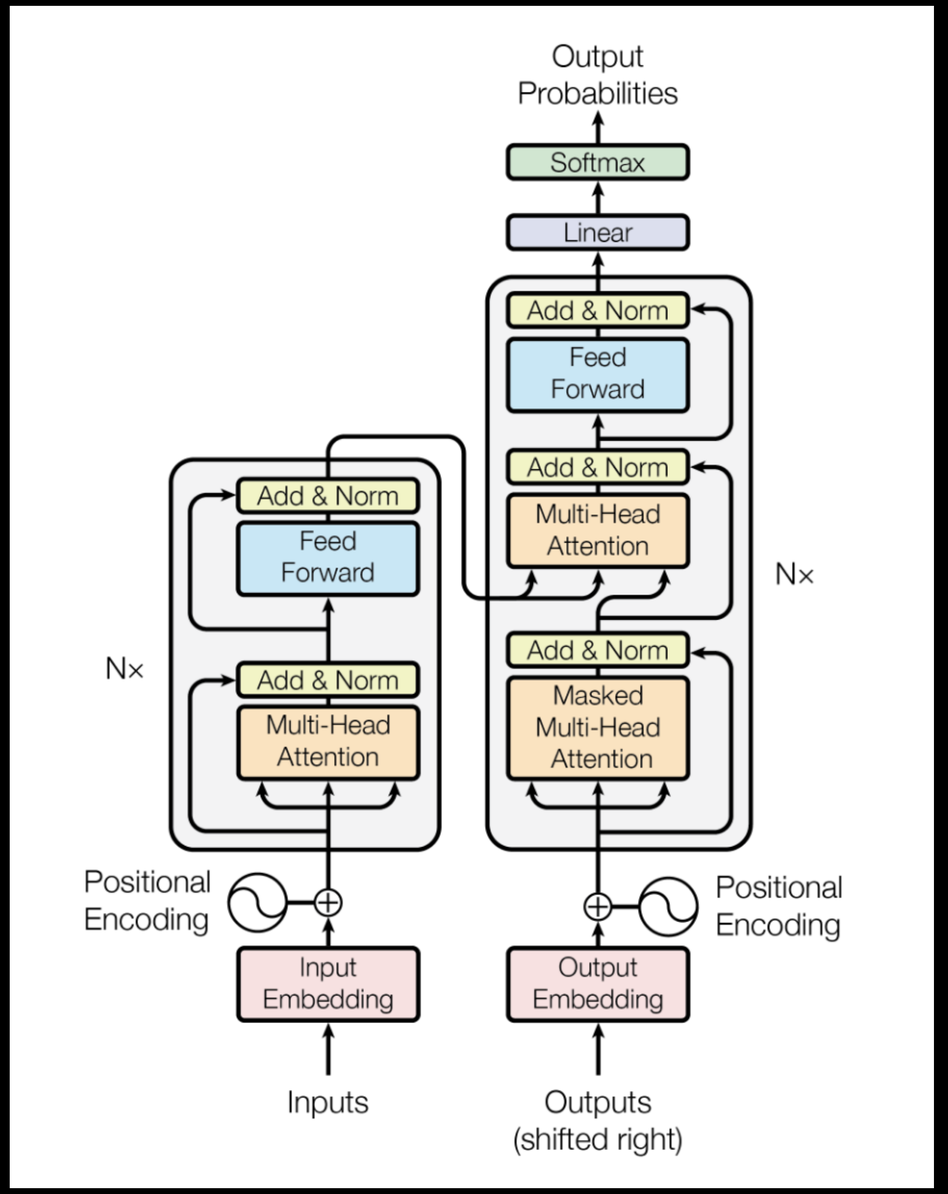
Attention Is All You Need

Vaswani et al., 2017

Ray Groshan

Why is this paper so important?

The Transformer Model



Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., Kaiser, Ł., & Polosukhin, I. (2017). *Attention Is All You Need*. https://arxiv.org/pdf/1706.03762

Summary

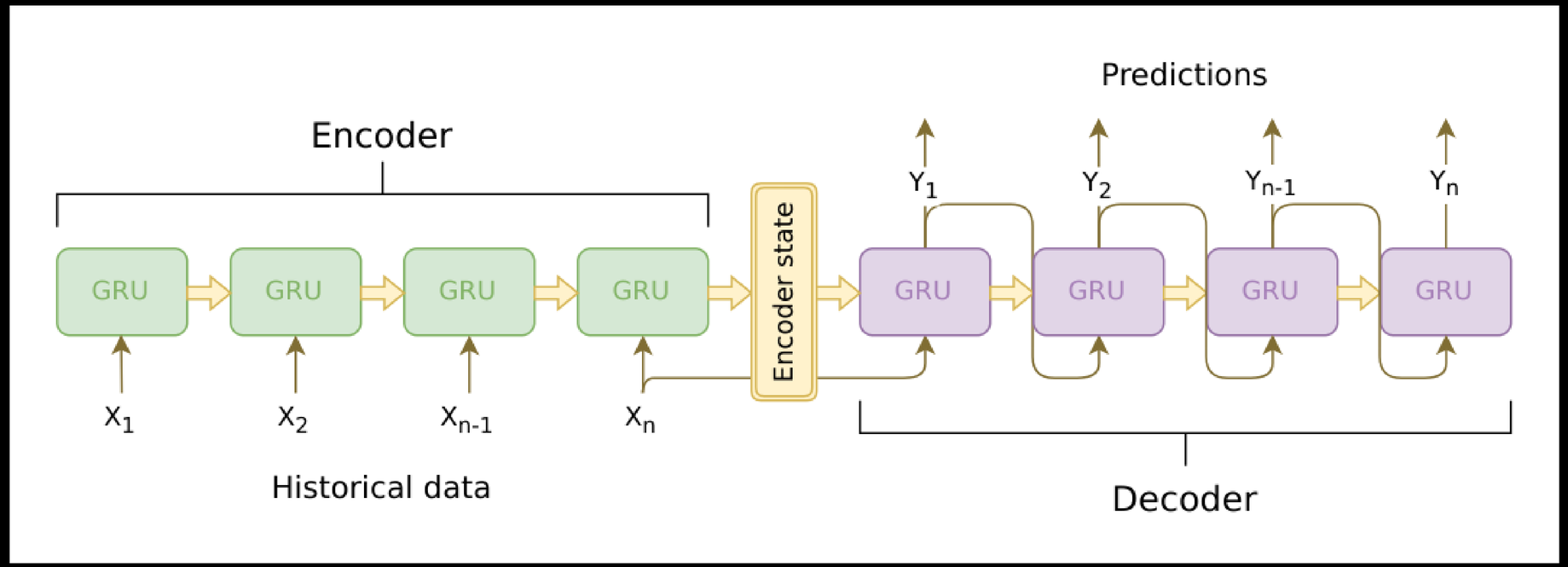
“The first transduction model relying entirely on self-attention to compute representations of input and output without using sequence-aligned RNNs or convolution.”

Vaswani et al., 2017

“The first transduction model relying entirely on self-attention to compute representations of input and output without using sequence-aligned RNNs or convolution.”

Vaswani et al., 2017

Transduction (Seq2Seq) Models

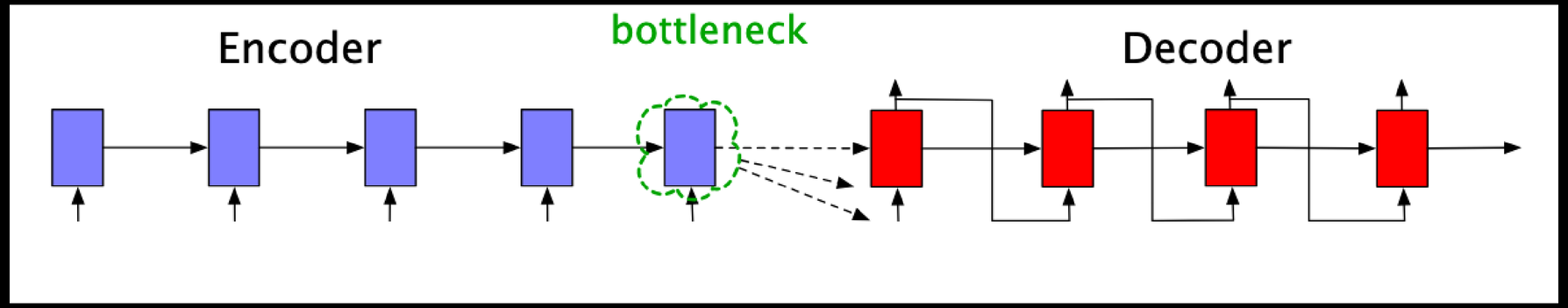


sooftware. (2020). *GitHub - sooftware/seq2seq: PyTorch implementation of the RNN-based sequence-to-sequence architecture.* GitHub. https://github.com/sooftware/seq2seq?tab=readme-ov-file

“The first transduction model relying entirely on self-attention to compute representations of input and output without using sequence-aligned RNNs or convolution.”

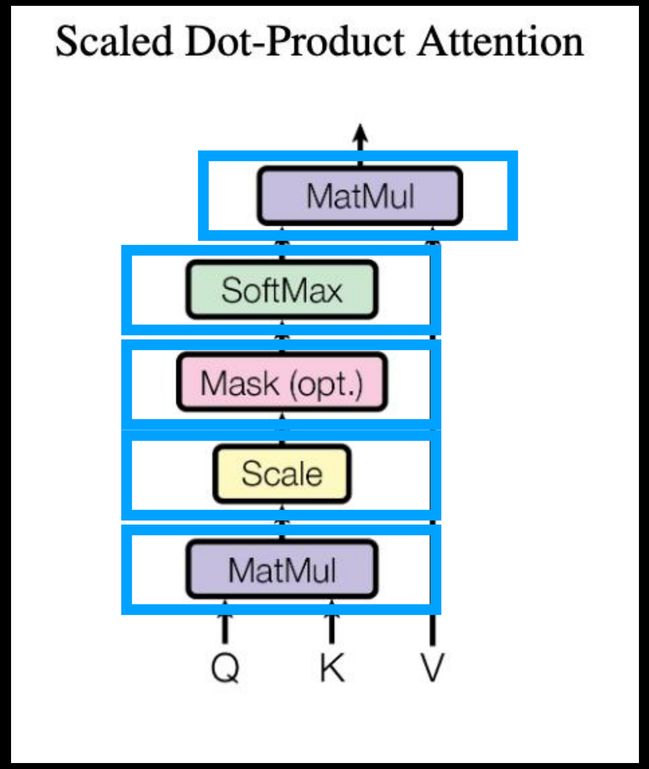
Vaswani et al., 2017

The Bottleneck Problem



Daniel Jurafsky and James H. Martin. 2025. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models, 3rd edition. Online manuscript released August 24, 2025. https://web.stanford.edu/~jurafsky/slp3.

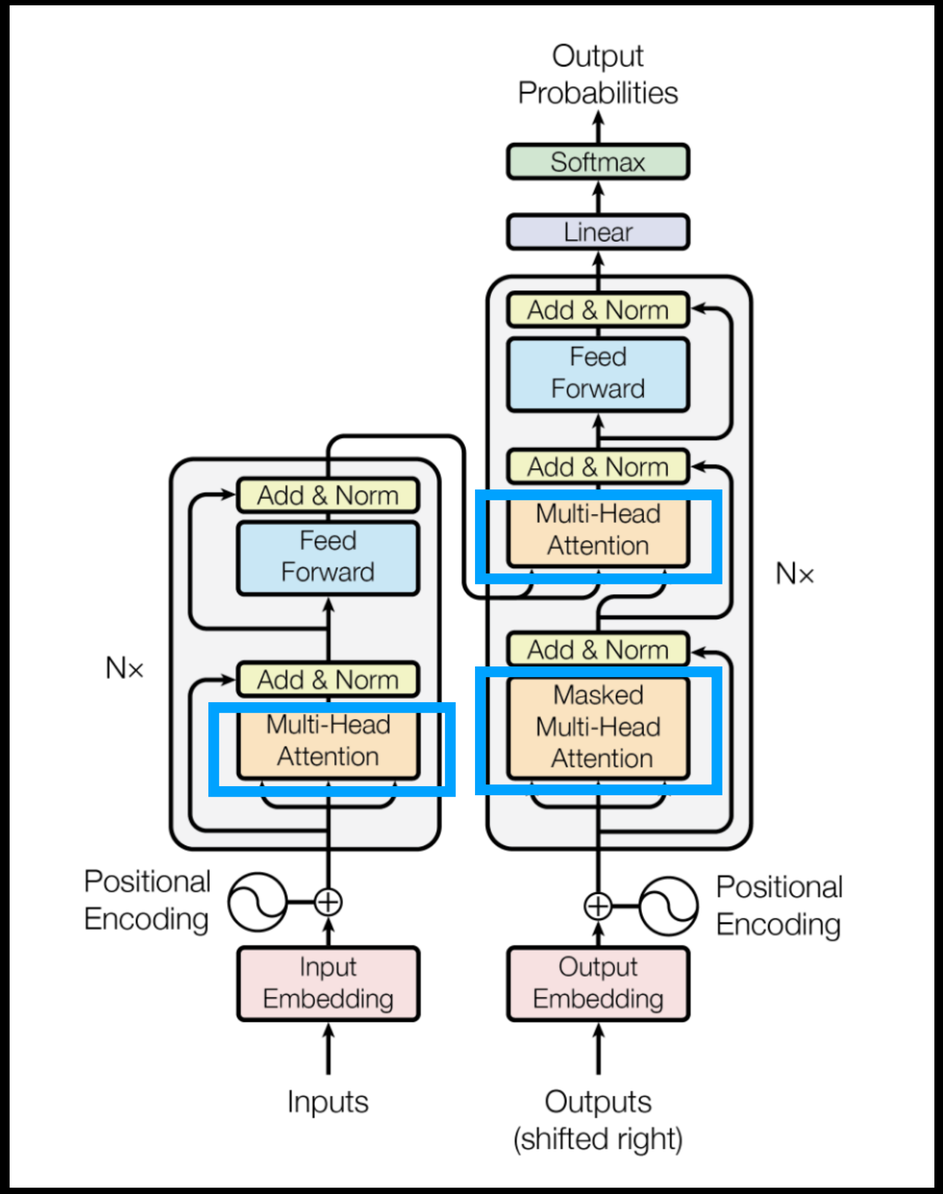
Self-Attention



Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., Kaiser, Ł., & Polosukhin, I. (2017). *Attention Is All You Need*. https://arxiv.org/pdf/1706.03762

“The first transduction model relying entirely on self-attention to compute representations of input and output without using sequence-aligned RNNs or convolution.”

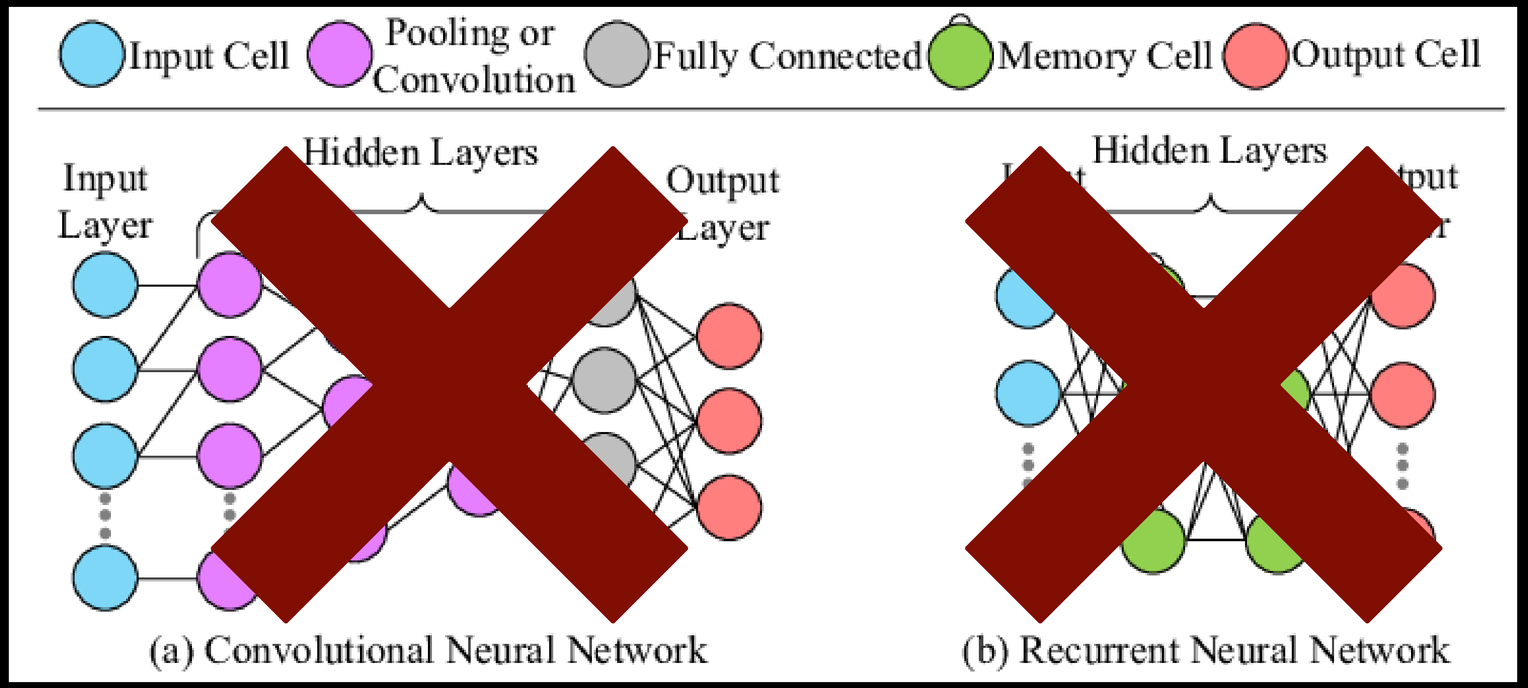
Vaswani et al., 2017



Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., Kaiser, Ł., & Polosukhin, I. (2017). *Attention Is All You Need*. https://arxiv.org/pdf/1706.03762

“The first transduction model relying entirely on self-attention to compute representations of input and output without using sequence-aligned RNNs or convolution.”

Vaswani et al., 2017

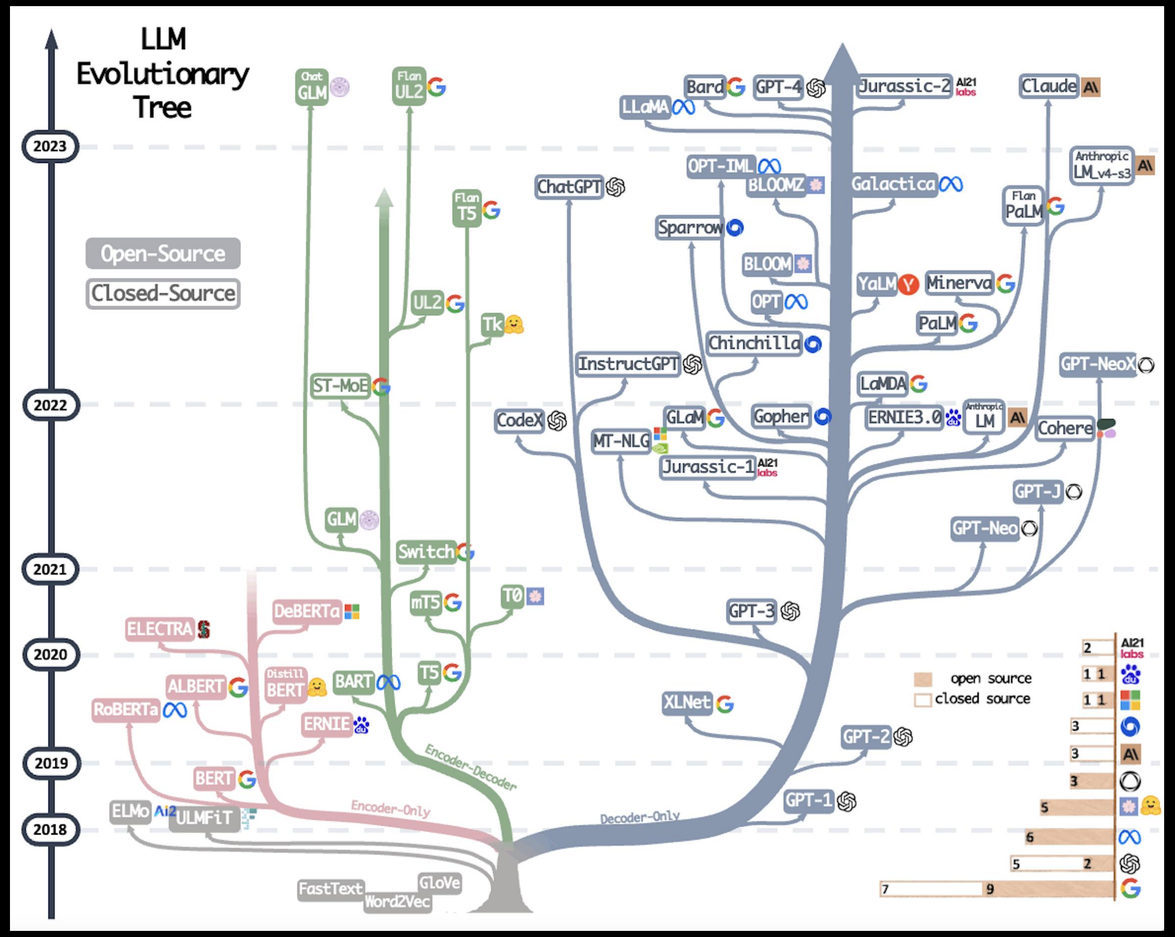


Zhang, Ke & Ying, Hanbo & Dai, Hong-Ning & Li, Lin & Peng, Yuangyuang & Guo, Keyi & Yu, Hongfang. (2021). Compacting Deep Neural Networks for Internet of Things: Methods and Applications. IEEE Internet of Things Journal. PP. 1-1. 10.1109/JIOT.2021.3063497.

Strengths & Weaknesses

Strengths

- Revolutionized the field
- Training can be parallelized
- ↑ efficiency
- ↓ computational complexity & costs



Jensen, P. A. (n.d.). *LLM Evolutionary Tree. LLM Proliferation.* Blog.biocomm.ai. https://blog.biocomm.ai/2023/05/14/open-source-proliferation-llm-evolutionary-tree/

Weaknesses

- No ethics or limitations sections
- The use of the BLEU score as a metric

BLEU Score	Interpretation
< 10	Almost useless
10 - 19	Hard to get the gist
20 - 29	The gist is clear, but has significant grammatical errors
30 - 40	Understandable to good translations
40 - 50	High quality translations
50 - 60	Very high quality, adequate, and fluent translations
> 60	Quality often better than human

Evaluate models. (2024). Google Cloud. https://cloud.google.com/translate/docs/advanced/automl-evaluate

Sources

CMSC 491/691 - Interactive Fiction and Text Generation - UMBC. (2025). Laramartin.net. https://laramartin.net/interactive-fiction-class

Evaluate models. (2024). Google Cloud. https://cloud.google.com/translate/docs/advanced/automl-evaluate

Graves, A. (2012). *Sequence Transduction with Recurrent Neural Networks*. ArXiv.org. *https://arxiv.org/abs/1211.3711*

sooftware. (2020). *GitHub - sooftware/seq2seq: PyTorch implementation of the RNN-based sequence-to-sequence architecture.* GitHub. *https://github.com/sooftware/seq2seq?tab=readme-ov-file*

Jensen, P. A. (n.d.). *LLM Evolutionary Tree. LLM Proliferation.* Blog.biocomm.ai. https://blog.biocomm.ai/2023/05/14/open-source-proliferation-llm-evolutionary-tree/

Daniel Jurafsky and James H. Martin. 2025. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models, 3rd edition. Online manuscript released August 24, 2025. https://web.stanford.edu/~jurafsky/slp3.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., Kaiser, Ł., & Polosukhin, I. (2017). *Attention Is All You Need*. https://arxiv.org/pdf/1706.03762

Zhang, Ke & Ying, Hanbo & Dai, Hong-Ning & Li, Lin & Peng, Yuangyuang & Guo, Keyi & Yu, Hongfang. (2021). Compacting Deep Neural Networks for Internet of Things: Methods and Applications. IEEE Internet of Things Journal. PP. 1-1. 10.1109/JIOT.2021.3063497.