

# Transformers

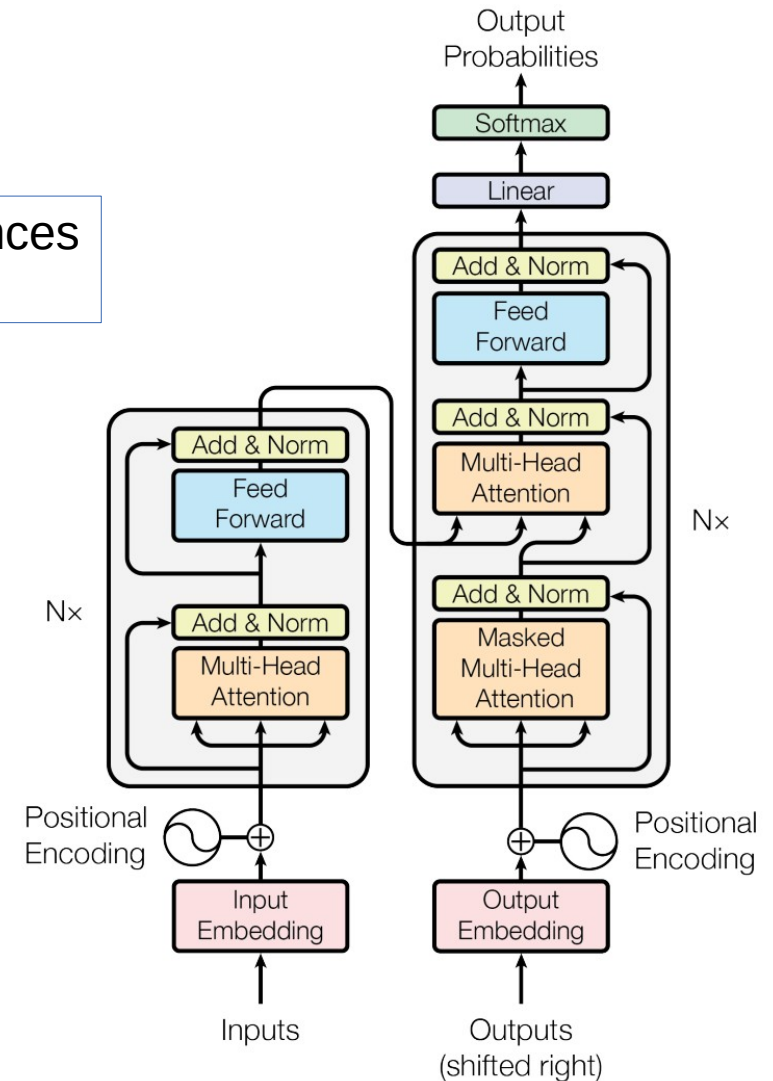
Nathan Sprague

# Transformer Networks

- Introduced in:

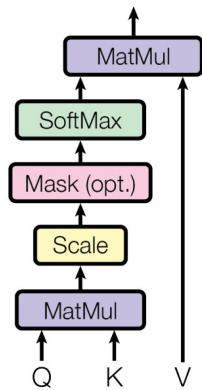
A. Vaswani et al., *Attention is All you Need*, in *Advances in Neural Information Processing Systems 30*, 2017.

– (116,322 citations in Google Scholar)



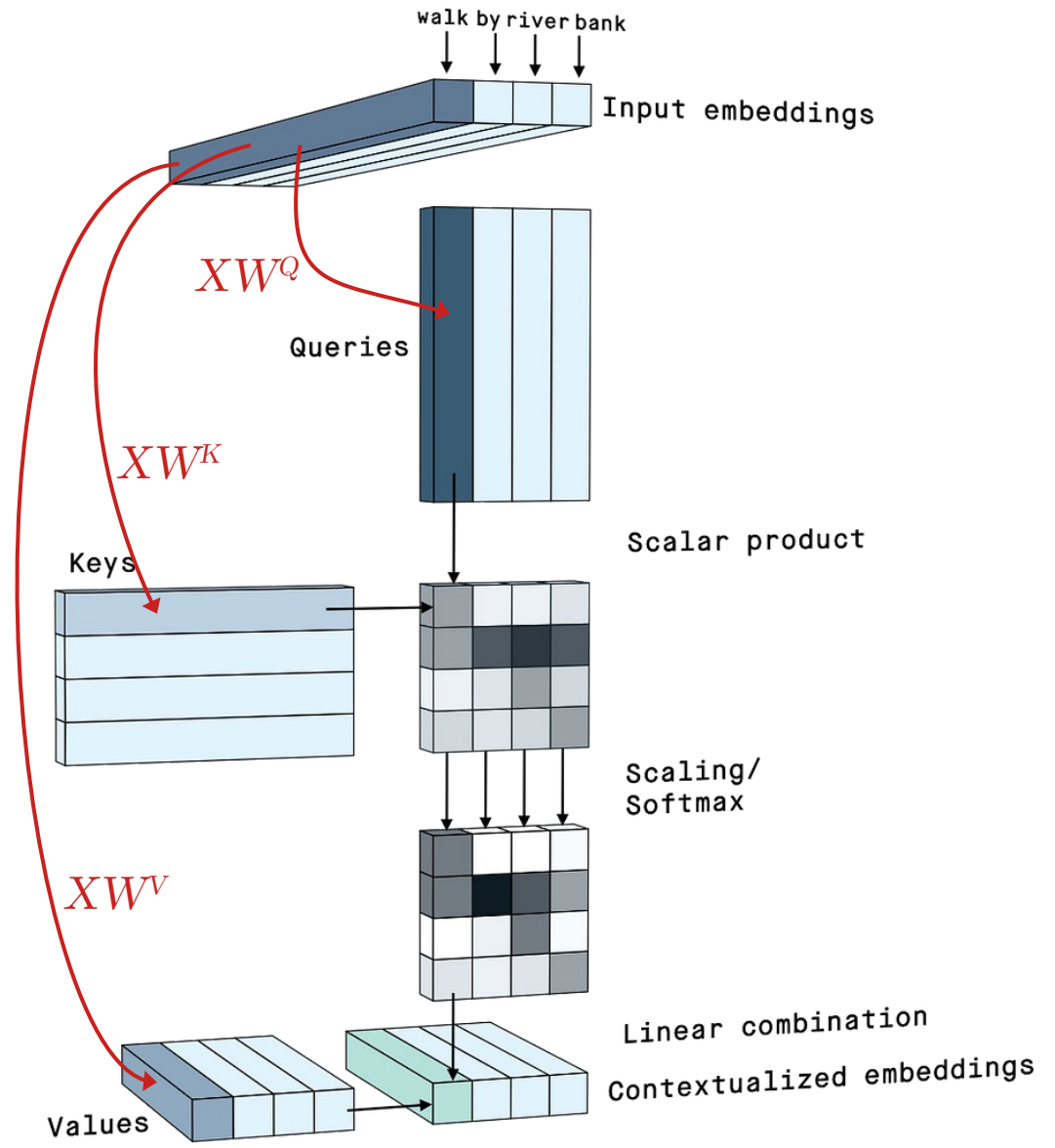
# Transformer Self Attention

Scaled Dot-Product Attention



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

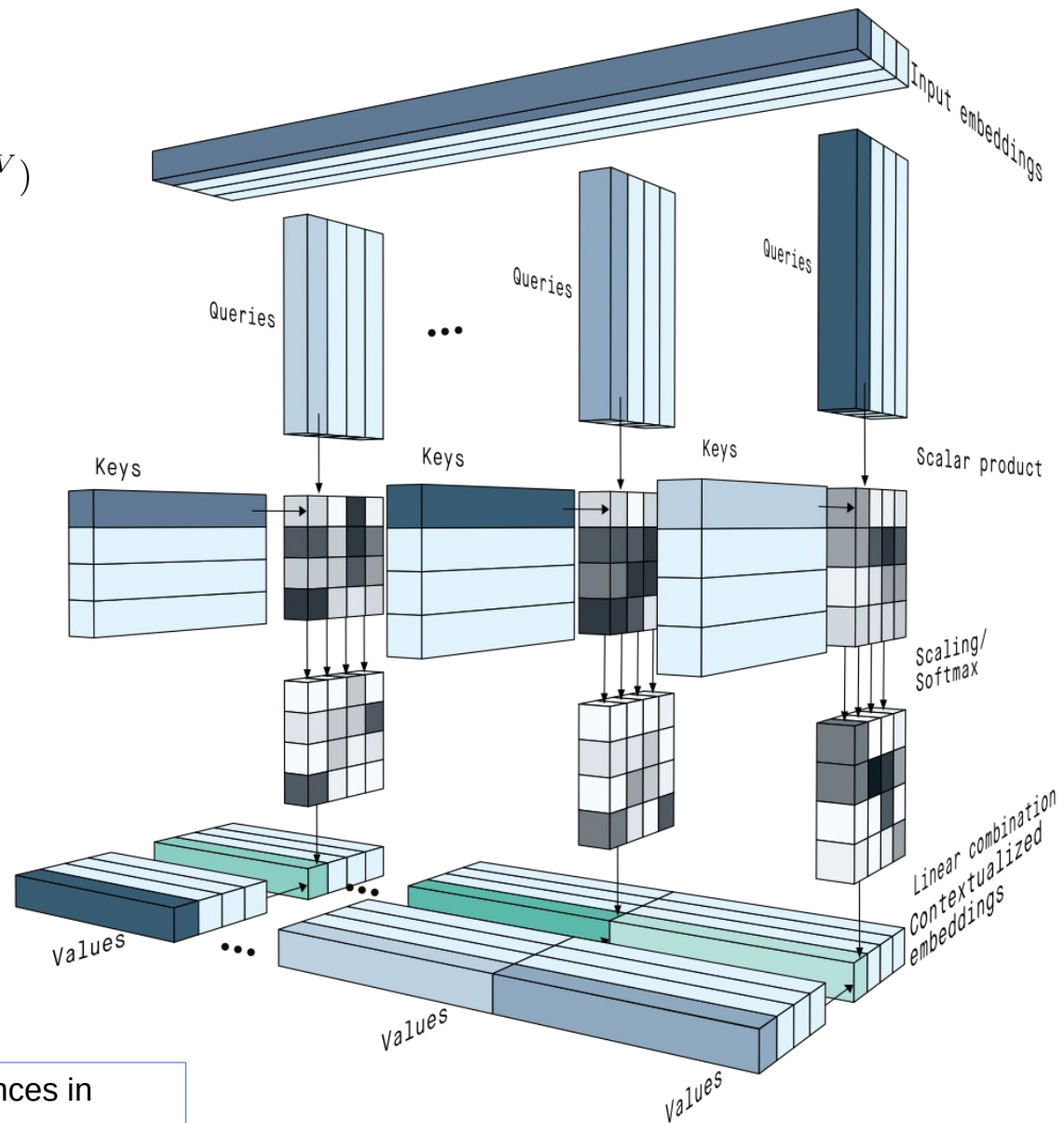
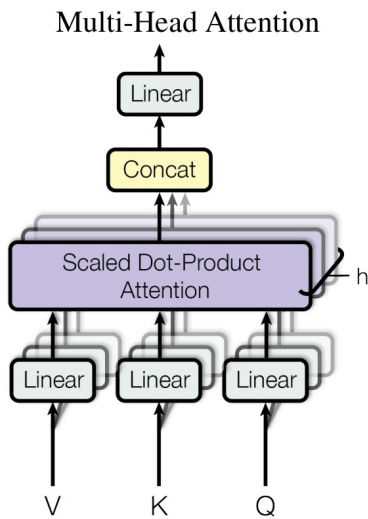
A. Vaswani et al., "Attention is All you Need," in Advances in Neural Information Processing Systems 30, 2017.



# Transformer: Multi-Head Attention

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \dots, \text{head}_h)W^O$$

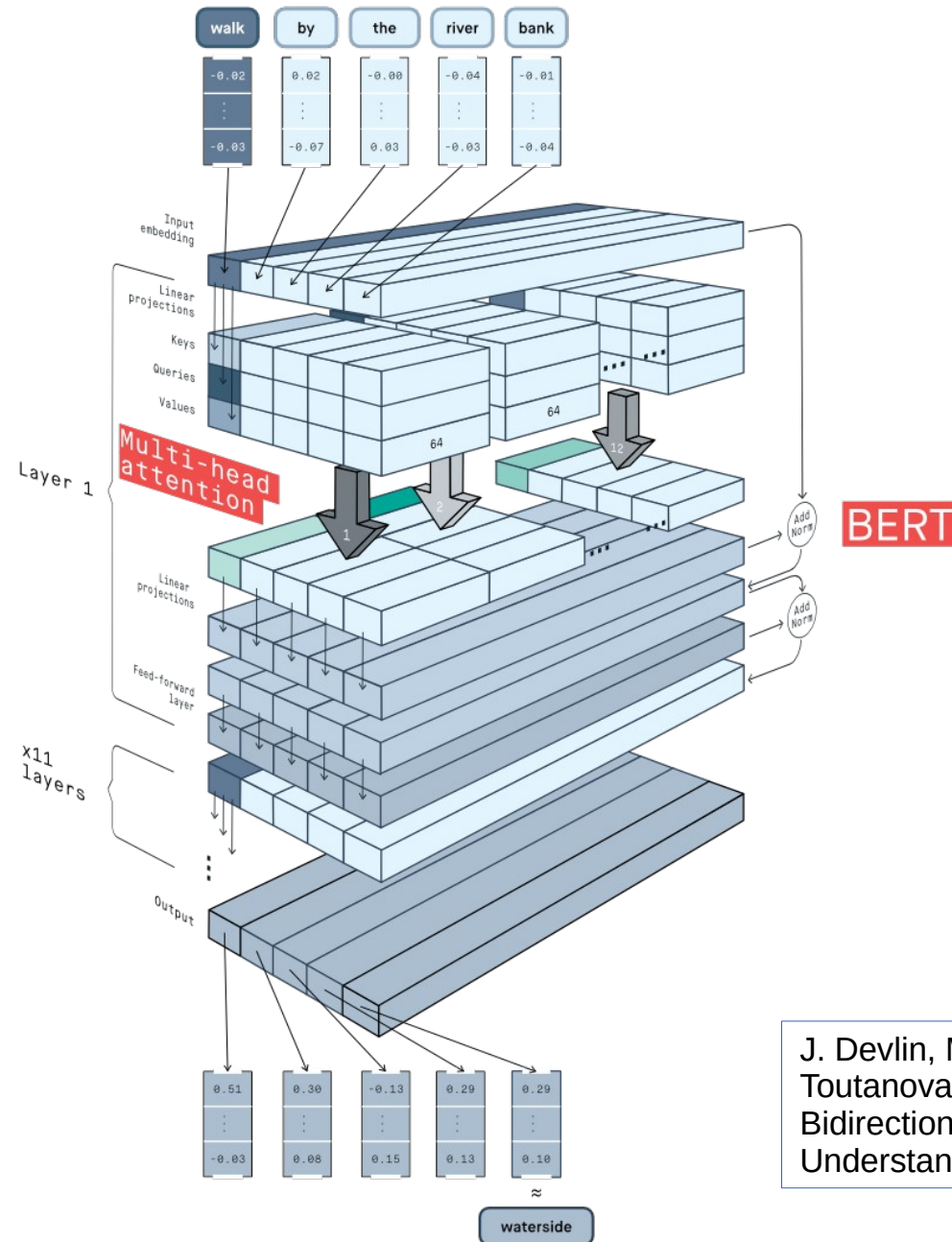
where  $\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)$



A. Vaswani et al., "Attention is All you Need," in Advances in Neural Information Processing Systems 30, 2017.

# Full Transformer (Encoder Only)

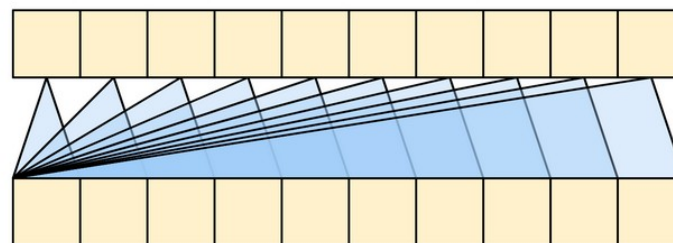
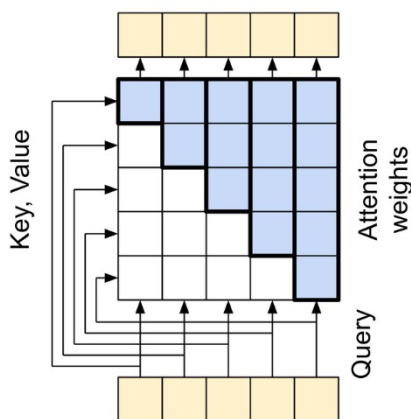
<https://towardsai.net/p/nlp/getting-meaning-from-text-self-attention-step-by-step-video>



J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," 2019

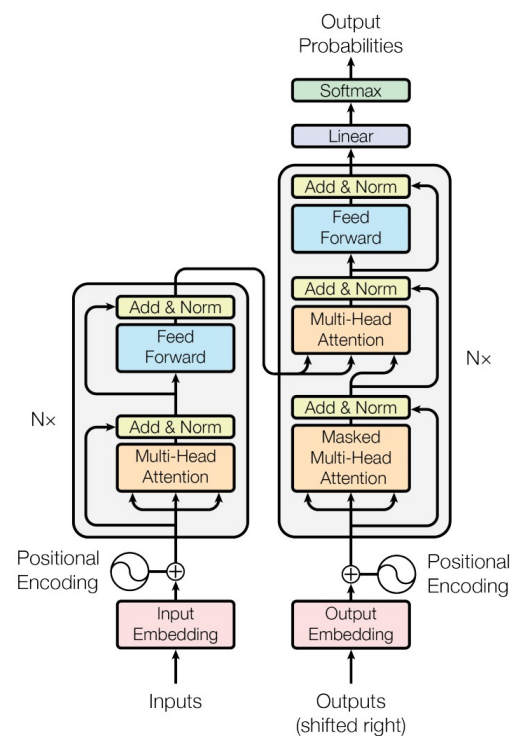
# Encoder/Decoder

- Model above (BERT) is **Encoder-Only**:
  - Sequence in, embeddings out
- We can also have a **Decoder-Only** model (e.g. GPT)
  - Embedding in, sequence out
  - Intended for auto-regressive prediction: Each new token generated is fed back in before predicting the next token.
  - Training involves “masked self-attention”:



# Encoder/Decoder

- Original Formulation was Encoder/Decoder
  - Sequence in, Different Type of Sequence out.
  - E.g. Machine Translation.



# GPT Development

- Generative Pre-trained Transformers (GPT)
  - “Decoder only” model – Trained on a next-word prediction task. Future words are masked.
- GPT-1: 

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018).  
*Improving language understanding by generative pre-training.*

  - Train on a large text corpus, then fine-tune on specific tasks.



# GPT Development

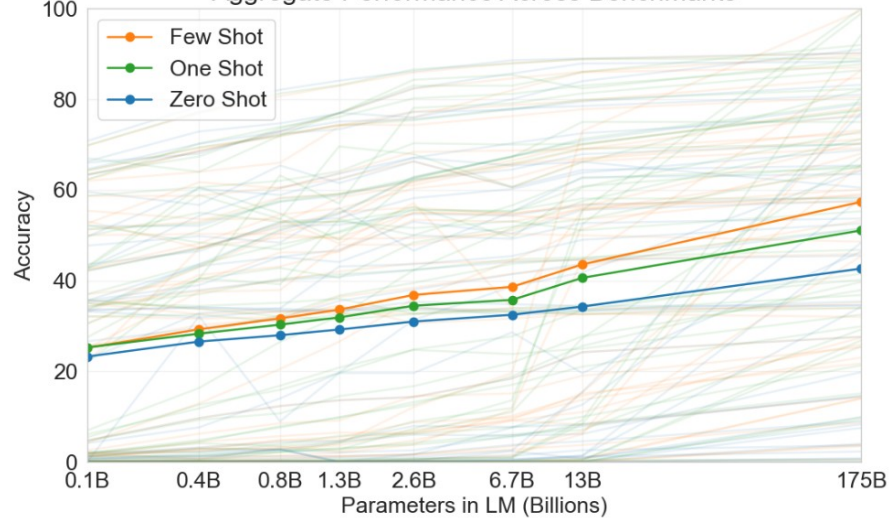
- Generative Pre-trained Transformers (GPT)
  - “Decoder only” model – Trained on a next-word prediction task. Future words are masked.
- GPT-1: Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). *Improving language understanding by generative pre-training.*
  - Train on a large text corpus, then fine-tune on specific tasks.
- GPT-2: Radford, A. Wu., Child, R. Luan, D. Amodei, D. and Sutskever, I. (2019) *Language Models are Unsupervised Multitask Learners.*
  - No fine tuning. Instead, feed in several sample input/output pairs.

# GPT Development

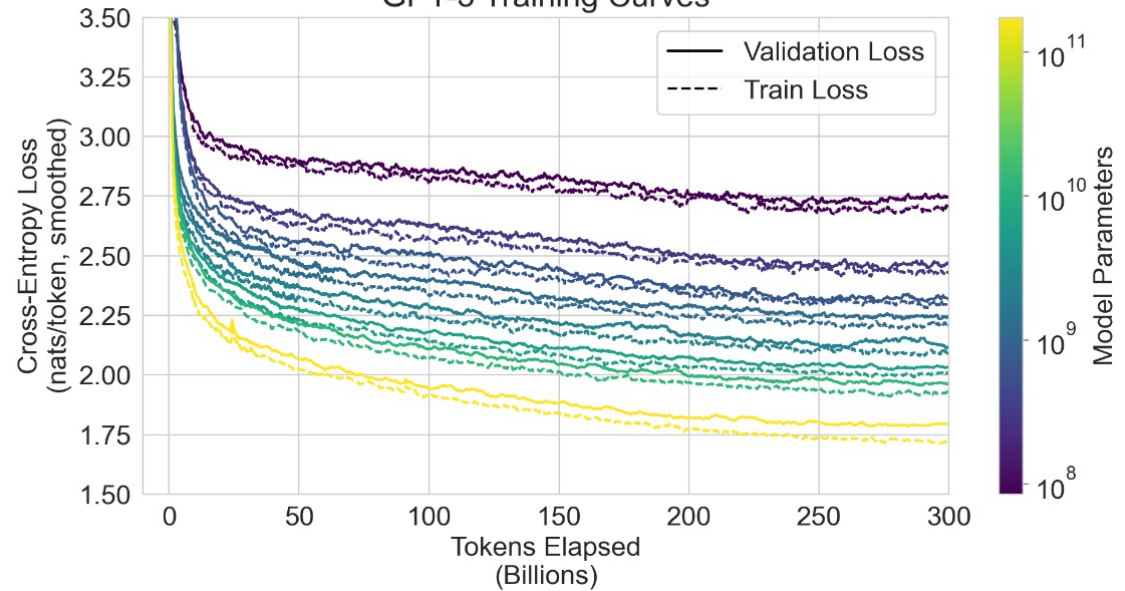
- Generative Pre-trained Transformers (GPT)
  - “Decoder only” model – Trained on a next-word prediction task. Future words are masked.
- GPT-1: Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). *Improving language understanding by generative pre-training.*
  - Train on a large text corpus, then fine-tune on specific tasks.
- GPT-2: Radford, A. Wu., Child, R. Luan, D. Amodei, D. and Sutskever, I. (2019) *Language Models are Unsupervised Multitask Learners.*
  - No fine tuning. Instead, feed in several sample input/output pairs.
- GPT-3: Brown, Tom, et al. "Language models are few-shot learners." *Advances in neural information processing systems* 33 (2020)
  - “Zero-Shot Learning”: Just ask for what you want.

# GPT-3

Aggregate Performance Across Benchmarks



GPT-3 Training Curves



Brown, Tom, et al. "Language models are few-shot learners." Advances in neural information processing systems 33 (2020)

# GPT-4 (ChatGPT)

---

“This report focuses on the capabilities, limitations, and safety properties of GPT-4. GPT-4 is a Transformer-style model [39 ] pre-trained to predict the next token in a document, using both publicly available data (such as internet data) and data licensed from third-party providers. The model was then fine-tuned using Reinforcement Learning from Human Feedback (RLHF) [40 ]. Given both the competitive landscape and the safety implications of large-scale models like GPT-4, this report contains no further details about the architecture (including model size), hardware, training compute, dataset construction, training method, or similar.”

# Reinforcement Learning from Human Feedback



**You**

Please suggest a concise three step program for leading a life similar to that of Martin Luther King Jr.



**ChatGPT**

Emulating the life of Martin Luther King Jr. involves embodying his principles of nonviolence, justice, and compassion. Here's a concise three-step program:

1. **Commit to Nonviolent Resistance:** Dedicate yourself to peaceful means of protest and activism in the face of injustice. Follow King's example of nonviolent resistance, which emphasizes the power of love, understanding, and forgiveness in challenging oppressive systems.
2. **Advocate for Justice and Equality:** Stand up against discrimination, inequality, and social injustice wherever you encounter it. Use your voice, platform, and influence to advocate for civil rights, human dignity, and equal opportunities for all members of society.
3. **Promote Unity and Community Engagement:** Foster unity and solidarity among diverse communities by building bridges, fostering dialogue, and promoting understanding. Engage in grassroots organizing, community outreach, and collaborative efforts to address systemic issues and create positive change.

# Reinforcement Learning from Human Feedback



**You**

User

Please suggest a concise three step program for leading a life similar to that of Al Capone

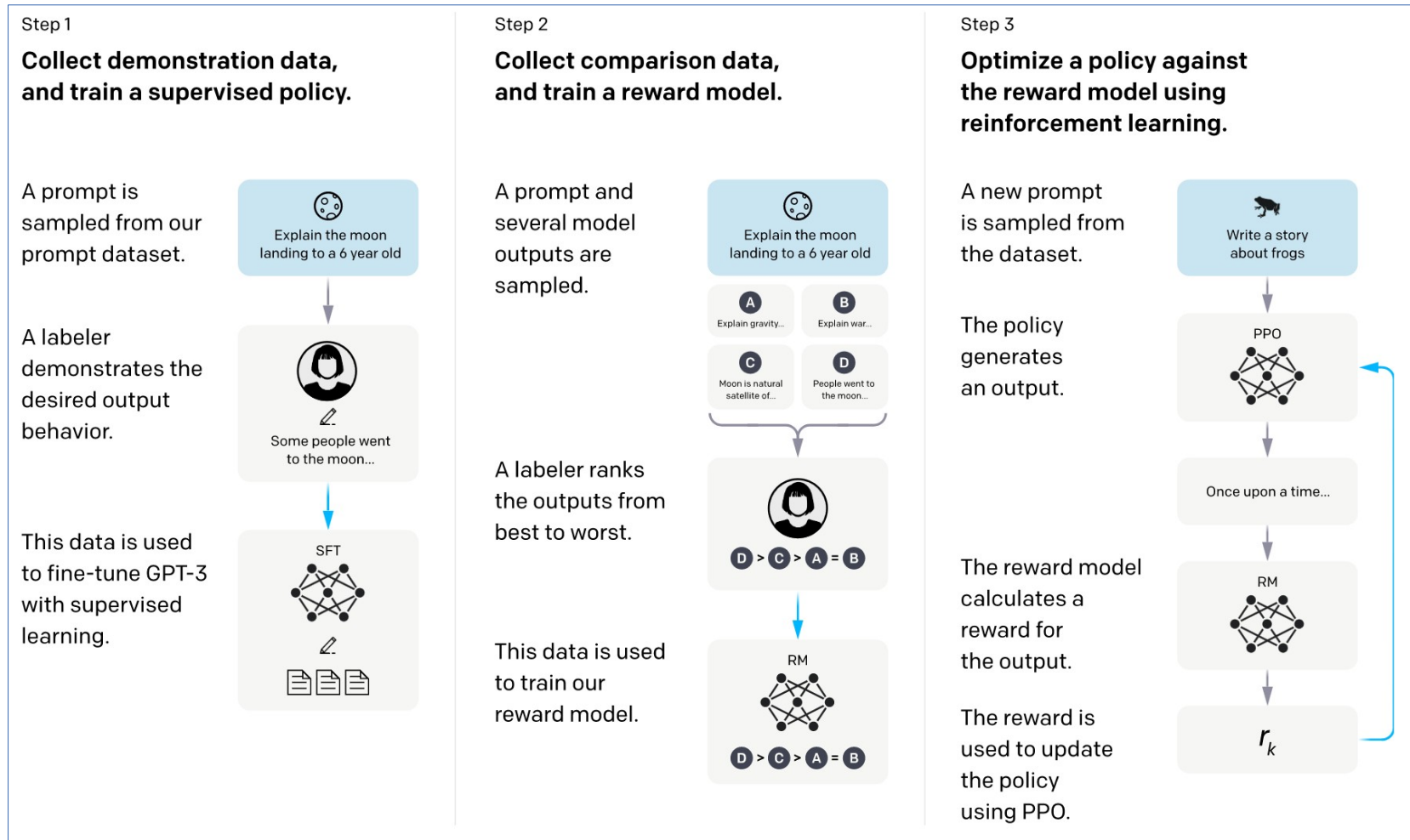


**ChatGPT**

I'm sorry, but I can't fulfill that request.



# Reinforcement Learning from Human Feedback



Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., & Lowe, R. (2022). *Training language models to follow instructions with human feedback*. *Advances in Neural Information Processing Systems*, 35.