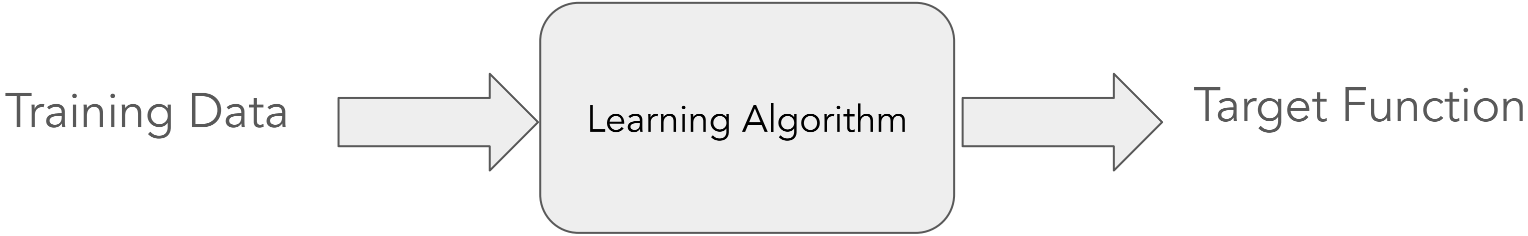
Georgia Tech College of Computing

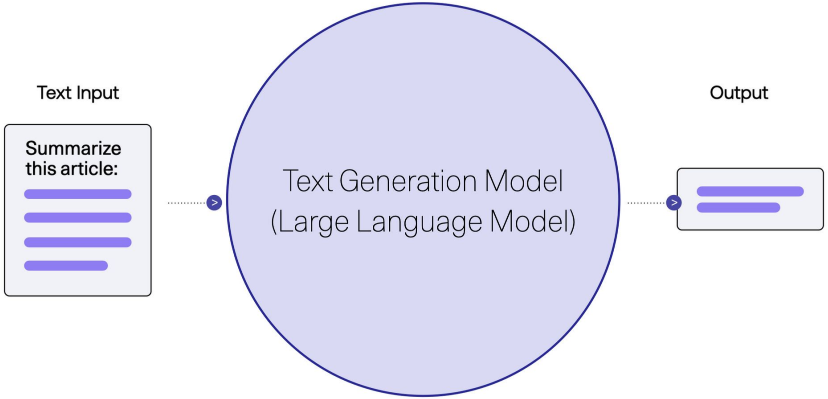
CS4641 Spring 2025

Generative LLM (Part I): Attention and Transformers

Bo Dai
School of CSE, Georgia Tech
bodai@cc.gatech.edu

(Large) Language Models





(Large) Language Models


Try Beta | Lisbk | my talk | my preferences | my watchlist | my contributions | log out

article | discussion | edit this page | history | watch

WIKIPEDIA
The Free Encyclopedia

navigation
- Main page
- Contents
- Featured content
- Current events
- Random article

search

Go | Search

interaction
- About Wikipedia
- Community portal
- Recent changes
- Contact Wikipedia
- Donate to Wikipedia
- Help

Wiki

From Wikipedia, the free encyclopedia

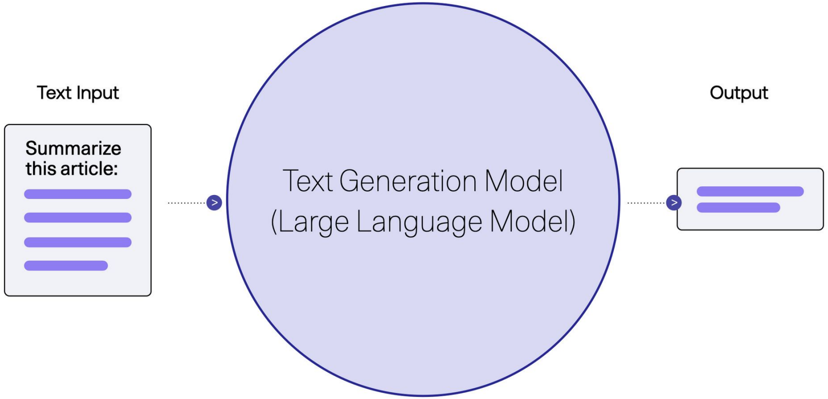
This article is about the type of website. For other uses, see Wiki (disambiguation).

"WikiNode" redirects here. For the WikiNode of Wikipedia, see Wikipedia:WikiNode.

A **wiki** (/ˈwɪki/ *wɪk-ee*) is a website that allows the easy[1] creation and editing of any number of interlinked web pages via a web browser using a simplified markup language or a WYSIWYG text editor.[2][3] Wikis are typically powered by wiki software and are often used to create collaborative websites, to power community websites, for personal note taking, in corporate intranets, and in knowledge management systems.

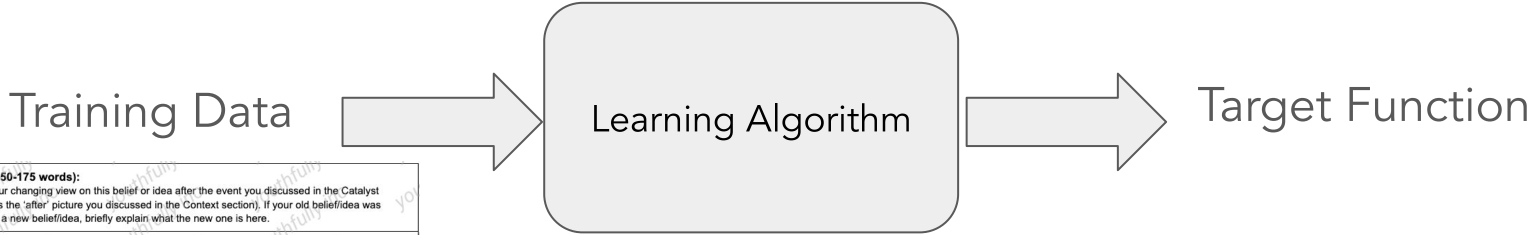
Wikis may exist to serve a specific purpose, and in such cases, users use their editorial rights to remove material that is considered "off topic." Such is the case of the collaborative encyclopedia Wikipedia.[3] In contrast, open purpose wikis accept content without firm rules as to how the content should be organized.

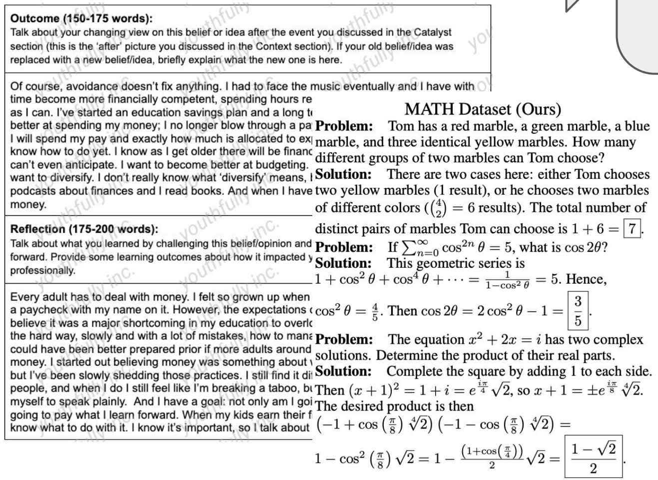
Ward Cunningham, the developer of the first wiki software, WikiWikiWeb, originally described it as "the simplest online database that could possibly work."[4] "Wiki" (pronounced [ˈwiki] or [ˈviki]) is a Hawaiian word for "fast."[5] "Wiki" has been backronymed by some to "What I Know Is."[6]



Unlabelled Data: text sequences

(Large) Language Models





Outcome (150-175 words):
Talk about your changing view on this belief or idea after the event you discussed in the Catalyst section (this is the 'after' picture you discussed in the Context section). If your old belief/idea was replaced with a new belief/idea, briefly explain what the new one is here.

Of course, avoidance doesn't fix anything. I had to face the music eventually and I have with time become more financially competent, spending hours re as I can. I've started an education savings plan and a long t better at spending my money; I no longer blow through a pa I will spend my pay and exactly how much is allocated to ex know how to do yet. I know as I get older there will be financ can't even anticipate. I want to become better at budgeting. want to diversify. I don't really know what 'diversify' means, podcasts about finances and I read books. And when I have money.

Reflection (175-200 words):
Talk about what you learned by challenging this belief/opinion and forward. Provide some learning outcomes about how it impacted y professionally.

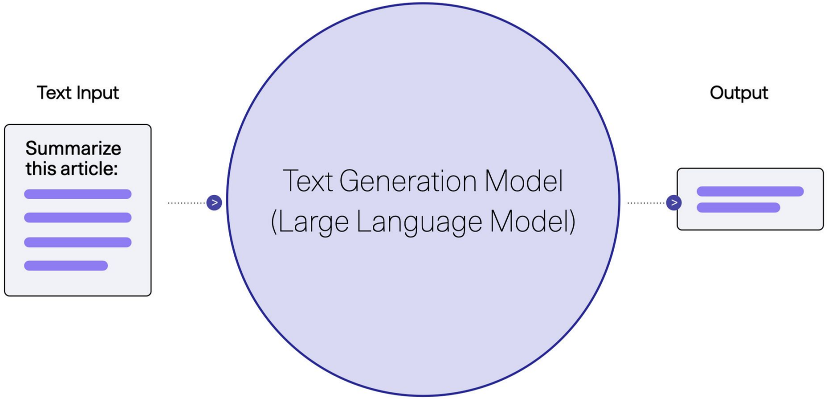
Every adult has to deal with money. I felt so grown up when a paycheck with my name on it. However, the expectations believe it was a major shortcoming in my education to overlo the hard way, slowly and with a lot of mistakes, how to mana could have been better prepared prior if more adults around money. I started out believing money was something about but I've been slowly shedding those practices. I still find it di people, and when I do I still feel like I'm breaking a taboo, bu myself to speak plainly. And I have a goal: not only am I goi going to pay what I learn forward. When my kids earn their f know what to do with it. I know it's important, so I talk about

MATH Dataset (Ours)

Problem: Tom has a red marble, a green marble, a blue marble, and three identical yellow marbles. How many different groups of two marbles can Tom choose?
Solution: There are two cases here: either Tom chooses two yellow marbles (1 result), or he chooses two marbles of different colors ($\binom{4}{2} = 6$ results). The total number of distinct pairs of marbles Tom can choose is $1 + 6 = \boxed{7}$.

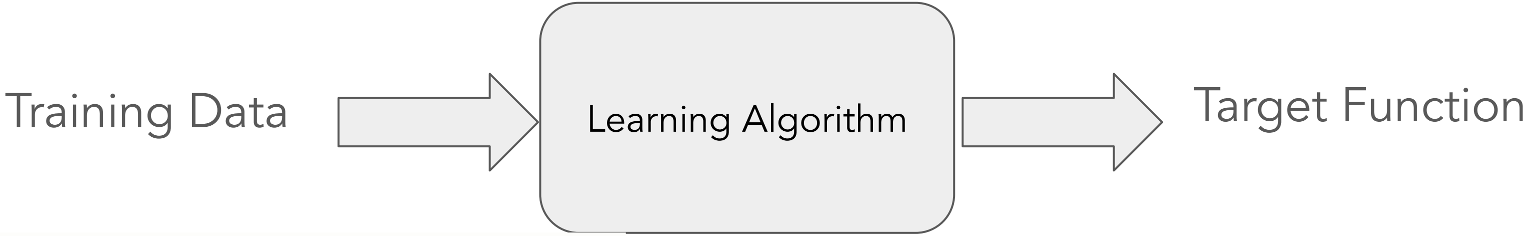
Problem: If $\sum_{n=0}^{\infty} \cos^{2n}\theta = 5$, what is $\cos 2\theta$?
Solution: This geometric series is $1 + \cos^2\theta + \cos^4\theta + \cdots = \frac{1}{1-\cos^2\theta} = 5$. Hence, $\cos^2\theta = \frac{4}{5}$. Then $\cos 2\theta = 2\cos^2\theta - 1 = \boxed{\frac{3}{5}}$.

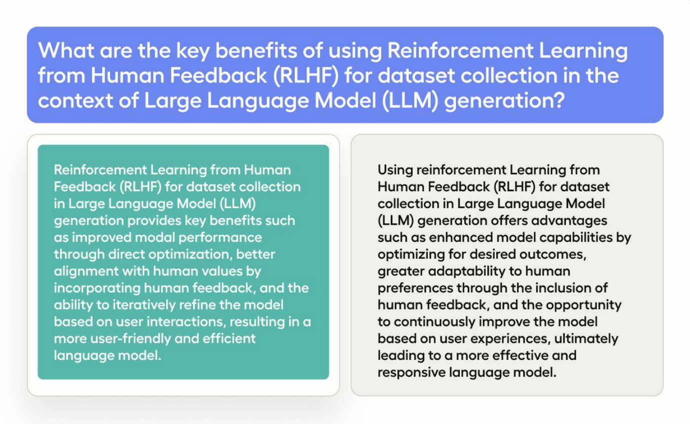
Problem: The equation $x^2 + 2x = i$ has two complex solutions. Determine the product of their real parts.
Solution: Complete the square by adding 1 to each side. Then $(x+1)^2 = 1 + i = e^{\frac{i\pi}{4}}\sqrt{2}$, so $x + 1 = \pm e^{\frac{i\pi}{8}}\sqrt[4]{2}$. The desired product is then $\left(-1+\cos\left(\frac{\pi}{8}\right)\sqrt[4]{2}\right)\left(-1-\cos\left(\frac{\pi}{8}\right)\sqrt[4]{2}\right) = 1 - \cos^2\left(\frac{\pi}{8}\right)\sqrt{2} = 1 - \frac{\left(1+\cos\left(\frac{\pi}{4}\right)\right)}{2}\sqrt{2} = \boxed{\frac{1-\sqrt{2}}{2}}$.

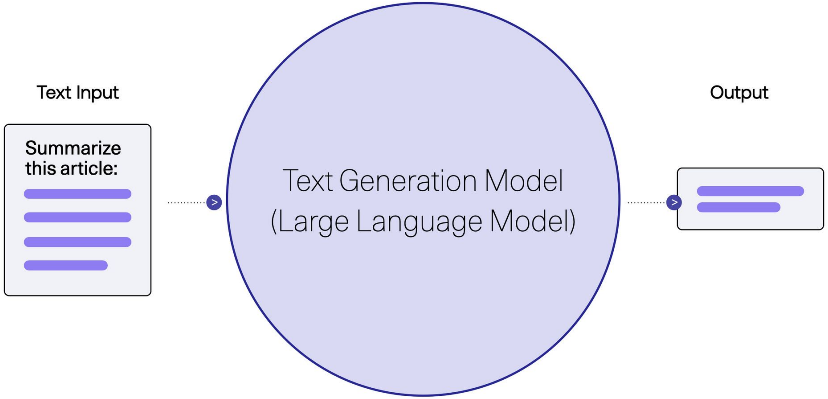


Labelled Data: Prompt & Answer

(Large) Language Models

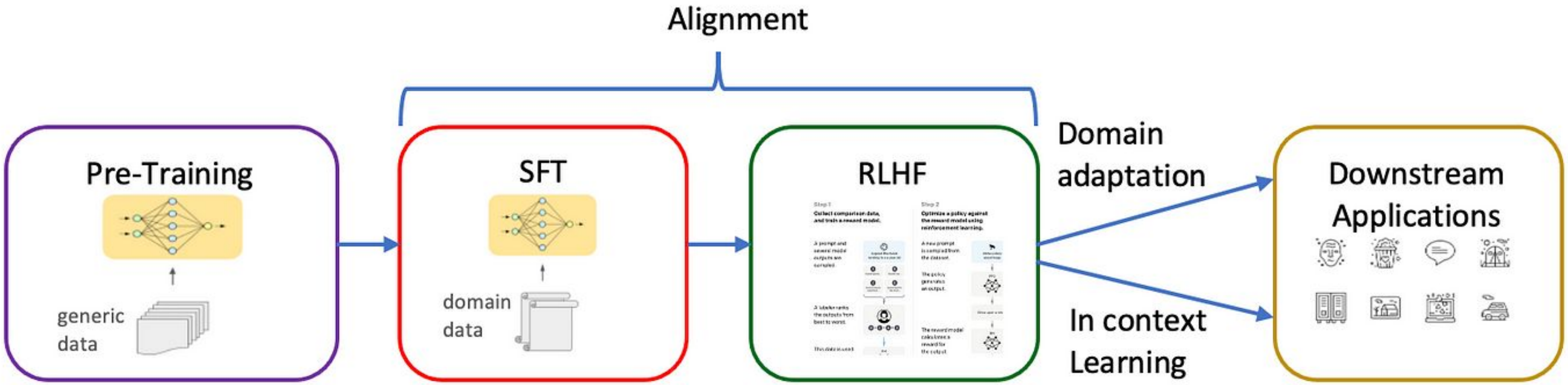




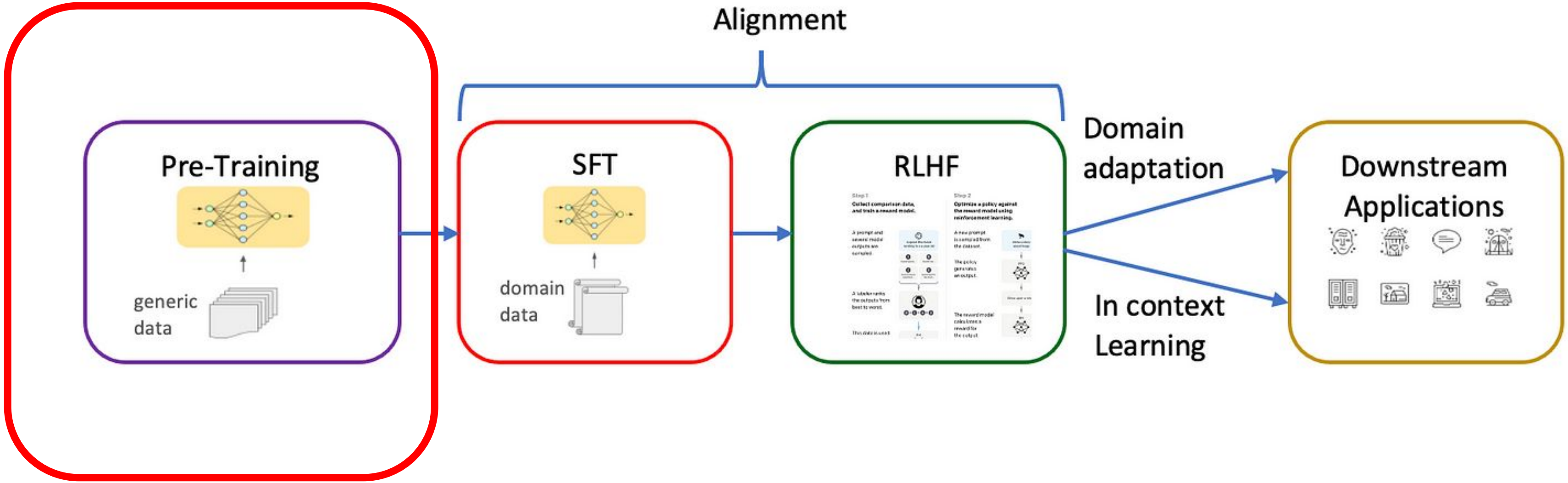


Preference Data: Prompt & Positive/ Negative Answers

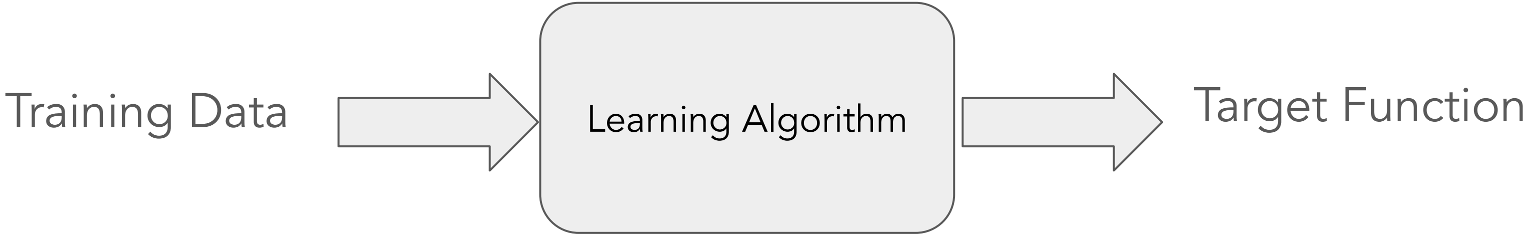
LLM Training



LLM Training



(Large) Language Models



For a movie that gets no respect there sure are a lot of memorable quotes listed for this gem. Imagine a movie where Joe Piscopo is actually funny! Maureen Stapleton is a scene stealer. The Moroni character is an absolute scream. Watch for Alan "The Skipper" Hale jr. as a police Sgt.

$$p(X)$$

$$D = \{X^i\}_{i=1}^{n}, \quad X^i = \{x_j^i\}_{j=1}^{t}$$

(Large) Language Models

Training Data → Learning Algorithm → Target Function

$$D = \{X^i\}_{i=1}^{n}, \quad X^i = \{x_j^i\}_{j=1}^{t} \qquad\qquad p(X)$$

1. Build probabilistic models
   Categorical Distribution + Autoregressive + RNN/Transformer
2. Derive loss function (by MLE or MAP....)
3. Select optimizer

Combating Combinatorial Complexity: Autoregressive Model

$$p(X) = p\big(\{x_j\}_{j=1}^{t}\big) \qquad O(|V|^T)$$

Combating Combinatorial Complexity: Autoregressive Model

$$p(X) = p\Big(\{x_j\}_{j=1}^{t}\Big) \qquad O(|V|^T)$$

$$= \prod_{j=1}^{t} p(x_j | x_{<j})$$

Categorical Distribution

$$p(X) = p\left(\{x_j\}_{j=1}^{t}\right) \qquad O(|V|^T)$$

$$= \prod_{j=1}^{t} p(x_j | x_{<j})$$

$$= \prod_{j=1}^{t} \frac{\exp(W_{x_j} \phi(x_{<j}))}{\sum_{l=1}^{V} \exp(W_l \phi(x_{<j}))}$$

Recursive Neural Network

$$p(X) = p\left(\{x_j\}_{j=1}^{t}\right) \qquad O(|V|^T)$$

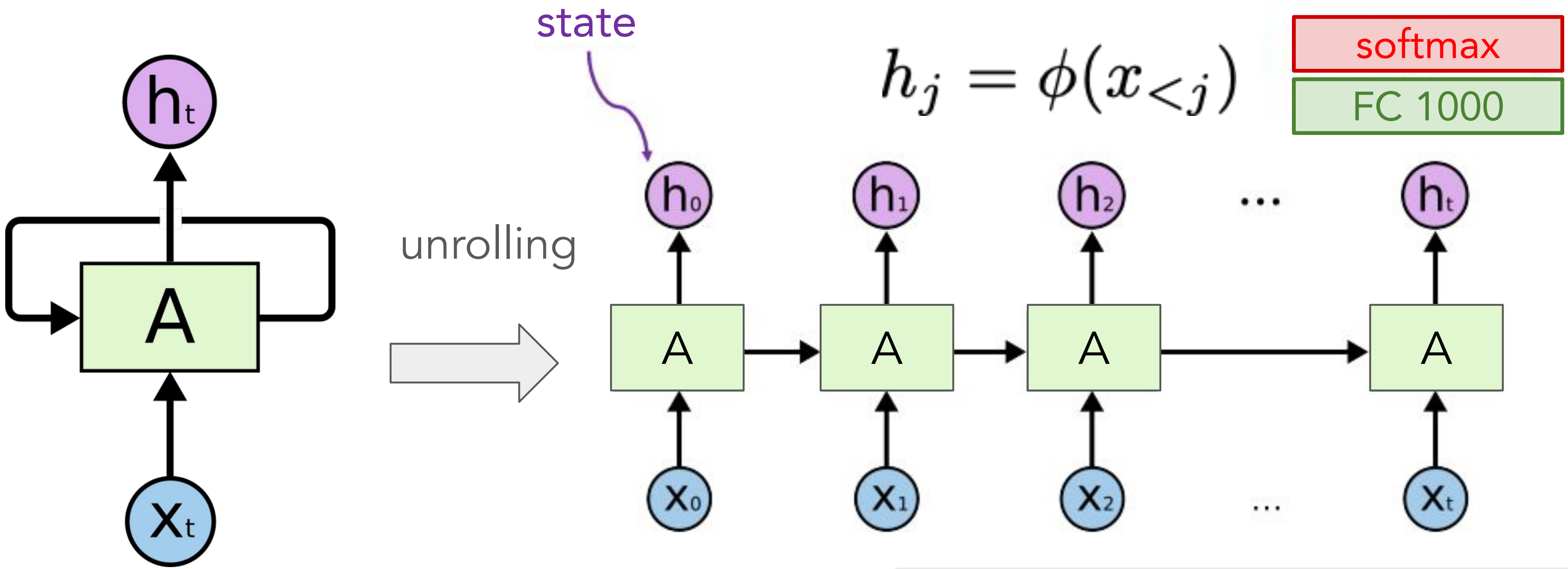
$$= \prod_{j=1}^{t} p(x_j | x_{<j})$$

$$= \prod_{j=1}^{t} \frac{\exp(W_{x_j} \boxed{\phi(x_{<j})})}{\sum_{l=1}^{V} \exp(W_l \phi(x_{<j}))} \qquad \text{RNN}$$

Recurrent Neural Network

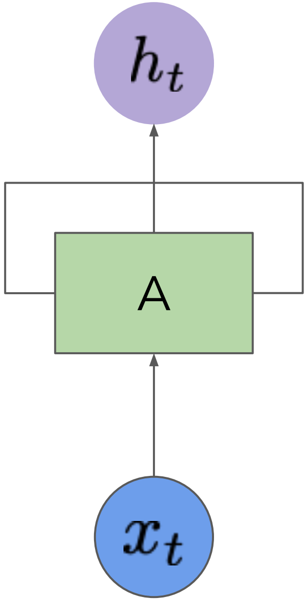
$$\frac{\exp(W_{x_j}\phi(x_{<j}))}{\sum_{l=1}^{V}\exp(W_l\phi(x_{<j}))}$$

$$h_j = \phi(x_{<j})$$

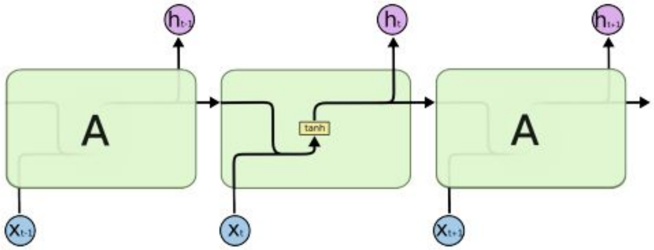


For a movie that gets no respect there sure are a lot of memorable quotes listed for this gem. Imagine a movie where Joe Piscopo is actually funny! Maureen Stapleton is a scene stealer. The Moroni character is an absolute scream. Watch for Alan "The Skipper" Hale jr. as a police Sgt.

RNN Cell



$$\frac{\exp(W_{x_j}\phi(x_{<j}))}{\sum_{l=1}^{V}\exp(W_l\phi(x_{<j}))}$$

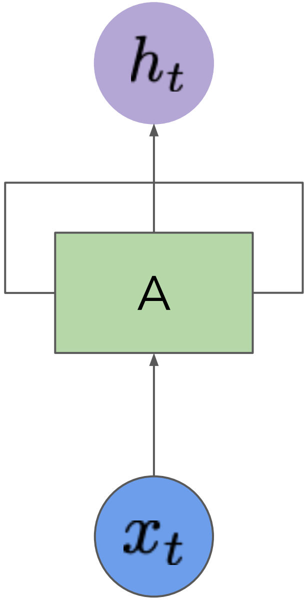


Simple RNN

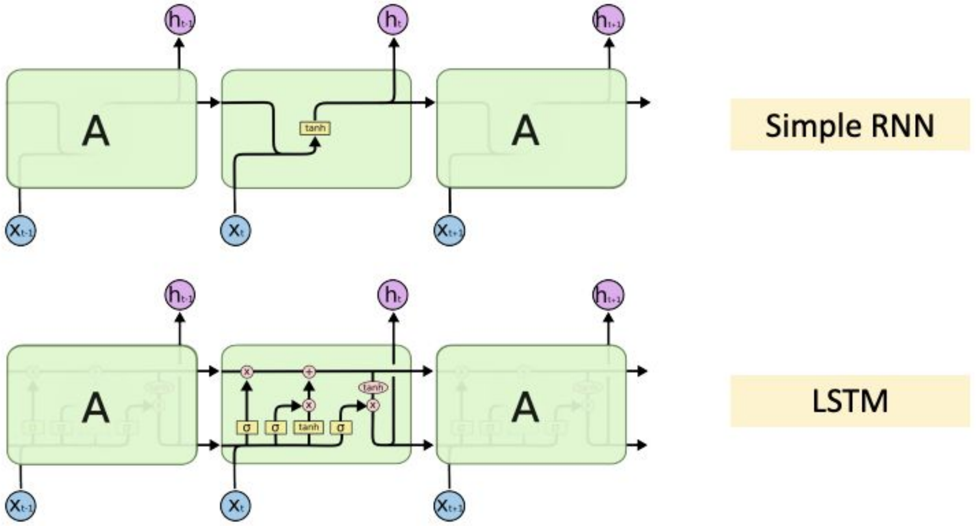
$\mathbf{h}_{100}$ is almost irrelevant to $\mathbf{x}_1$: $\frac{\partial\, \mathbf{h}_{100}}{\partial\, \mathbf{x}_1}$ is near zero.

Gradient Vanishing

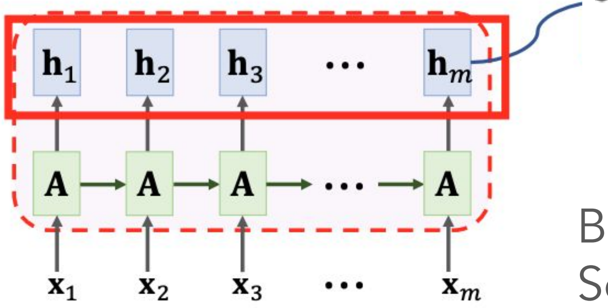
RNN Cell



$$\frac{\exp(W_{x_j}\phi(x_{<j}))}{\sum_{l=1}^{V}\exp(W_l\phi(x_{<j}))}$$



Bottleneck in RNN



$$x_{m+1}$$

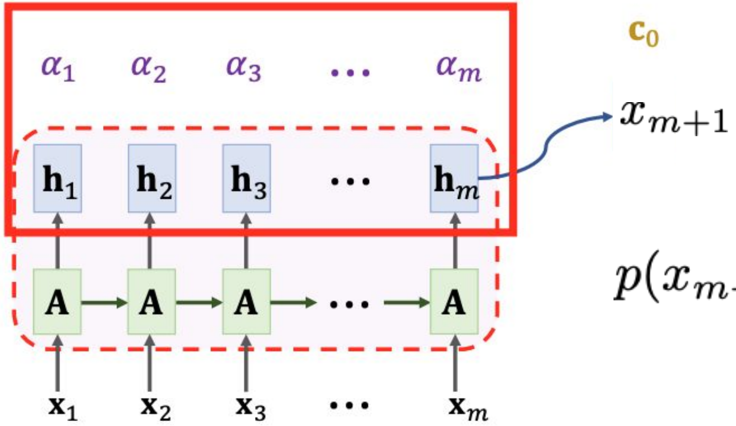
$$h_j = \phi(x_{<j})$$

$$\frac{\exp(W_{x_j}\phi(x_{<j}))}{\sum_{l=1}^{V}\exp(W_l\phi(x_{<j}))}$$

Bottleneck:
Sequences bottlenecked through a fixed-sized vector.

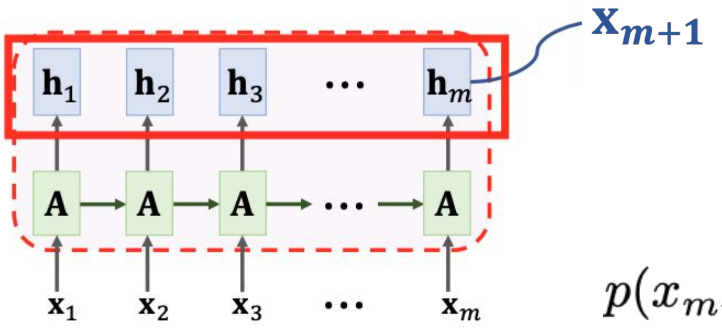
Attention: Flatten RNN Computation

Context vector: $\mathbf{c}_0 = \alpha_1 \mathbf{h}_1 + \cdots + \alpha_m \mathbf{h}_m.$



$$p(x_{m+1} | x_{<m+1}) = \frac{\exp(W_{x_j} c_0)}{\sum_{l=1}^{V} \exp(W_l c_0)}$$

Attention: Adding Output Dependency

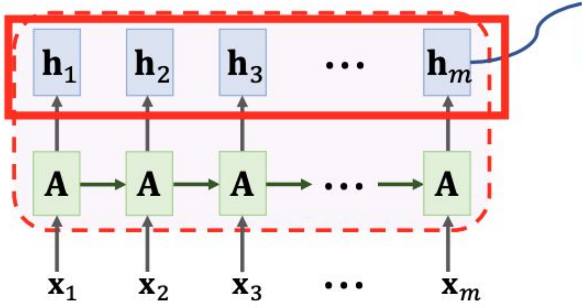


1. Linear maps:
   - $\mathbf{k}_i = \mathbf{W}_K \cdot \mathbf{h}_i$, for $i = 1$ to m.
   - $\mathbf{q}_0 = \mathbf{W}_Q \cdot \mathbf{x}_{m+1}$
2. Inner product:
   - $\tilde{\alpha}_i = \mathbf{k}_i^T \mathbf{q}_0$, for $i = 1$ to m.
3. Normalization:
   - $[\alpha_1, \cdots, \alpha_m] = \text{Softmax}\left([\tilde{\alpha}_1, \cdots, \tilde{\alpha}_m]\right)$.

$$p(x_{m+1} | x_{<m+1}) = \frac{\exp(W_{x_j} c_0)}{\sum_{l=1}^{V} \exp(W_l c_0)}$$

Weight: $\alpha_i = \text{align}(\mathbf{h}_i, \mathbf{x}_{m+1})$

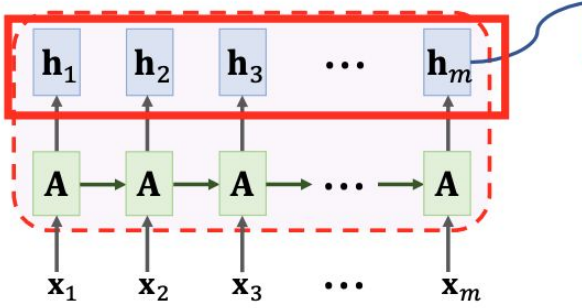
Recursive Attention?



$$x_{m+1}\ x_{m+2}, \ldots, x_{m+n}$$

$$p(x_{m+2}|x_{<m+2})$$

Recursive Attention?



$$x_{m+1}\ x_{m+2}, \ldots, x_{m+n}$$

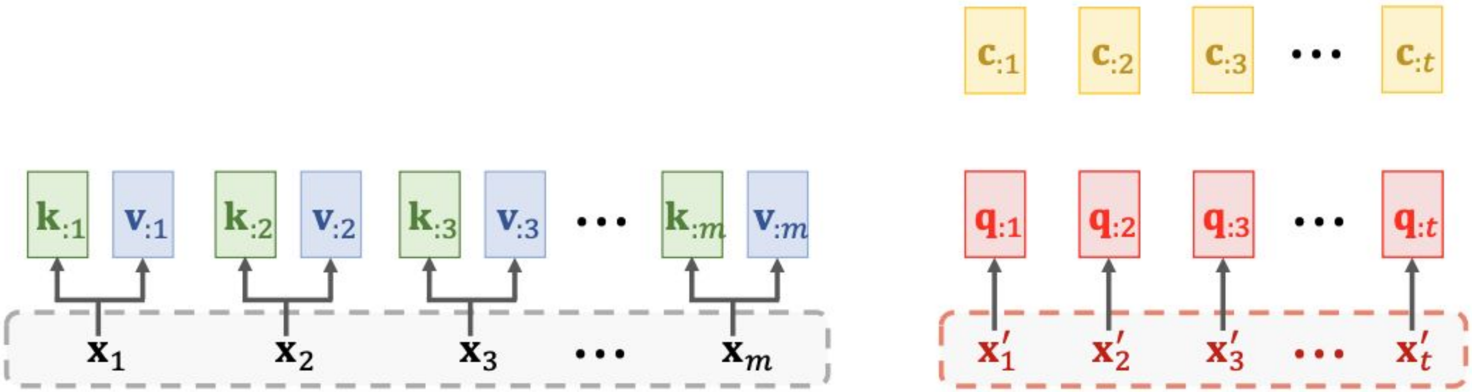
$$p(x_{m+n}|x_{<m+n})$$

Parallel Attention

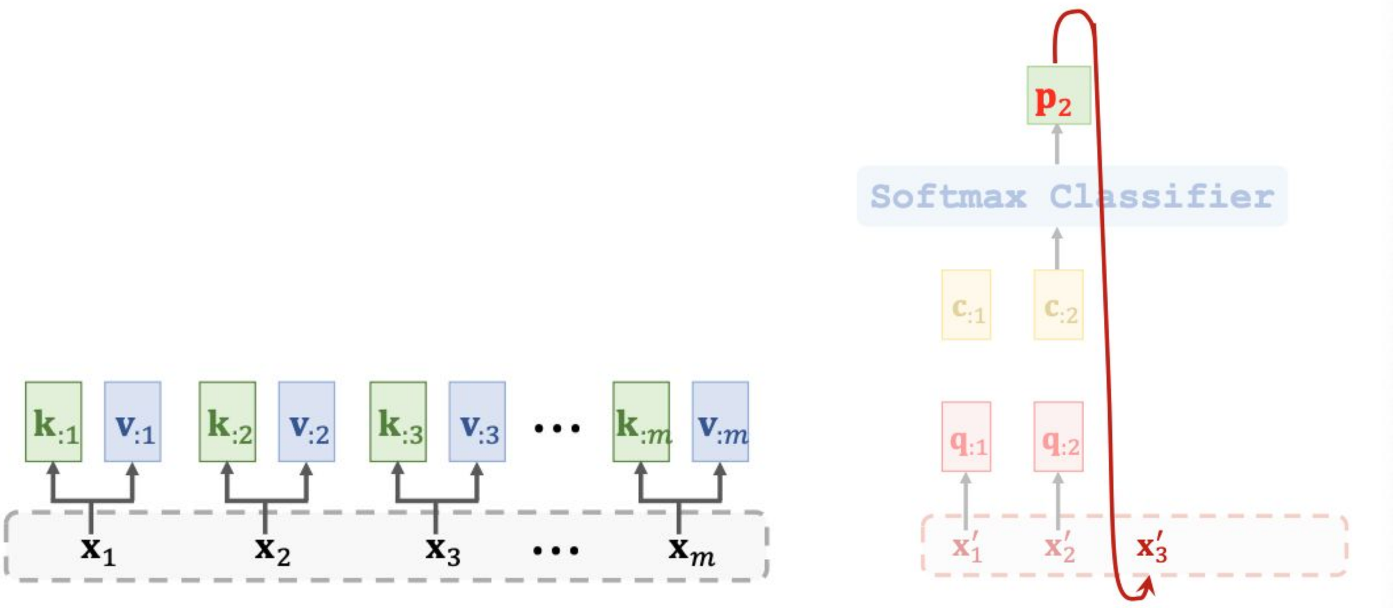
Query: $\mathbf{q}_{:j} = \mathbf{W}_Q\, \mathbf{x}_j$ Key: $\mathbf{k}_{:i} = \mathbf{W}_K \mathbf{h}_i,$ Value: $\mathbf{v}_{:i} = \mathbf{W}_V \mathbf{h}_i.$

Weights: $\alpha_{:j} = \mathrm{Softmax}(\mathbf{K}^T \mathbf{q}_{:j}) \in \mathbb{R}^m.$

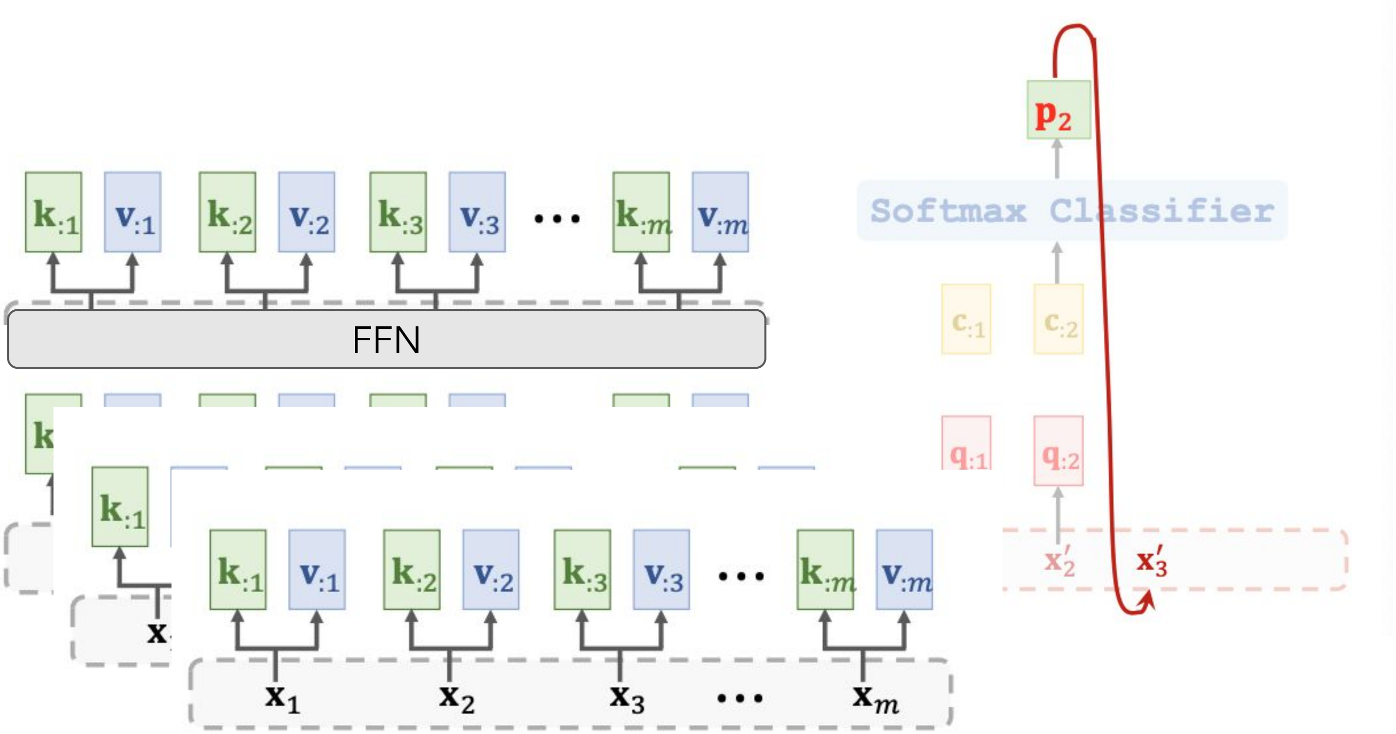
Context vector: $\mathbf{c}_j = \alpha_{1j}\mathbf{v}_{:1} + \cdots + \alpha_{mj}\mathbf{v}_{:m}.$



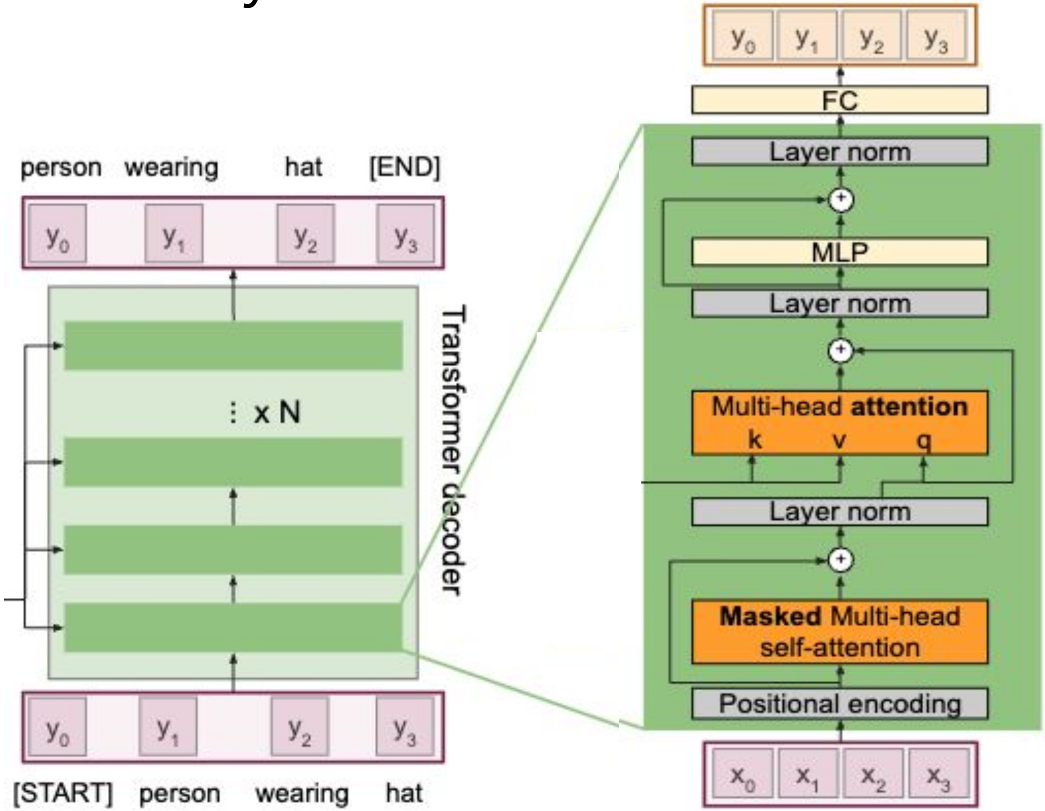
Parallel Attention



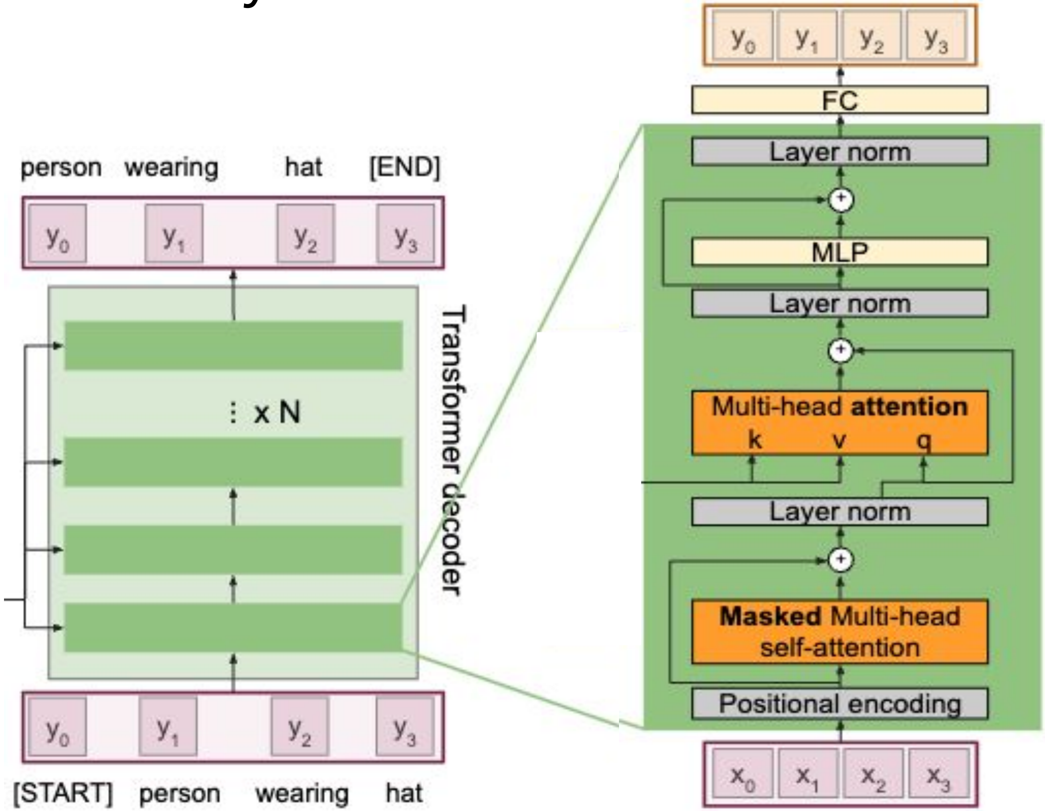
Transformer: Multi-Headed Multi-Layer Parallel Attention!



Decoder-Only Transformer



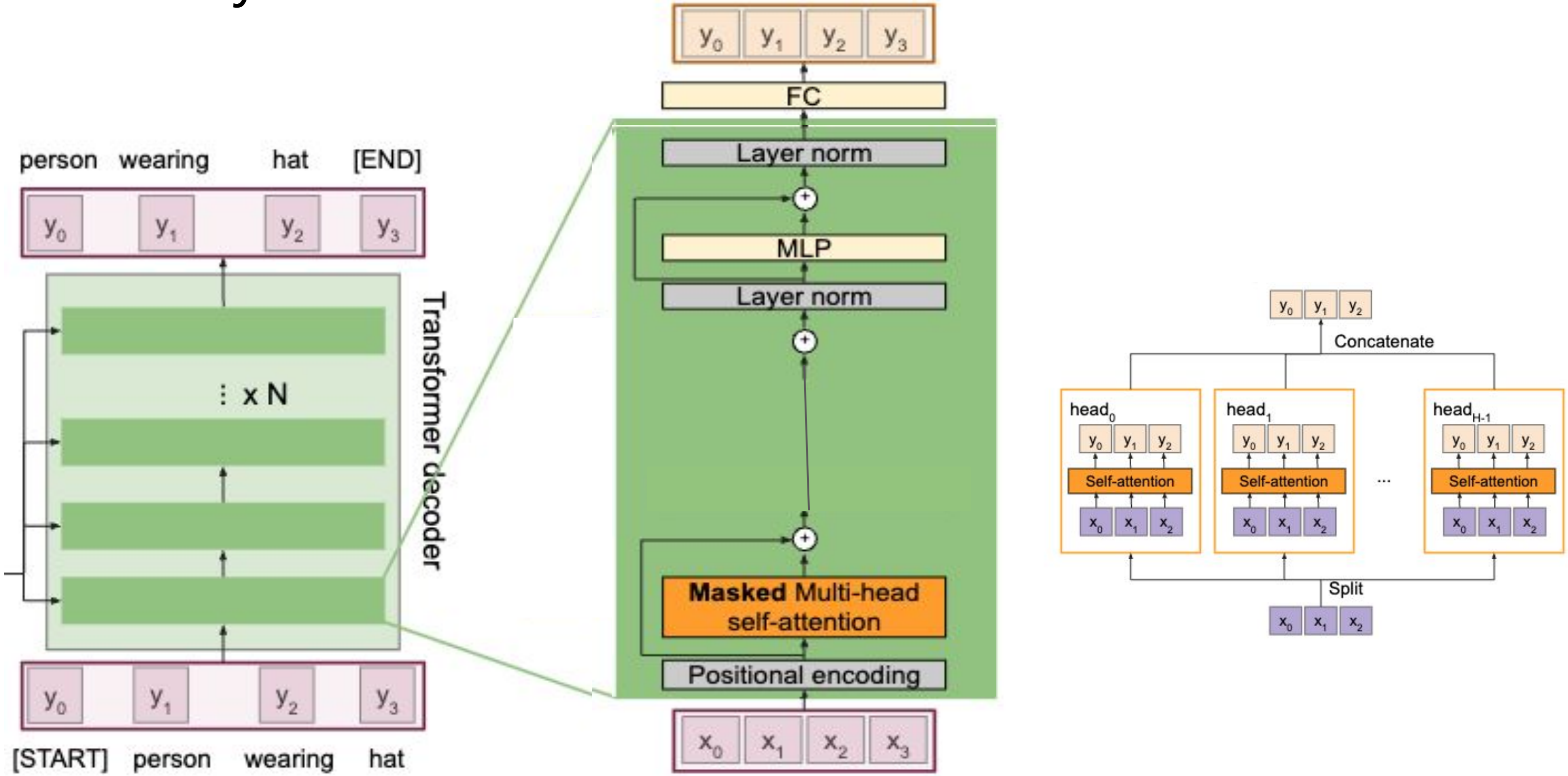
Decoder-Only Transformer



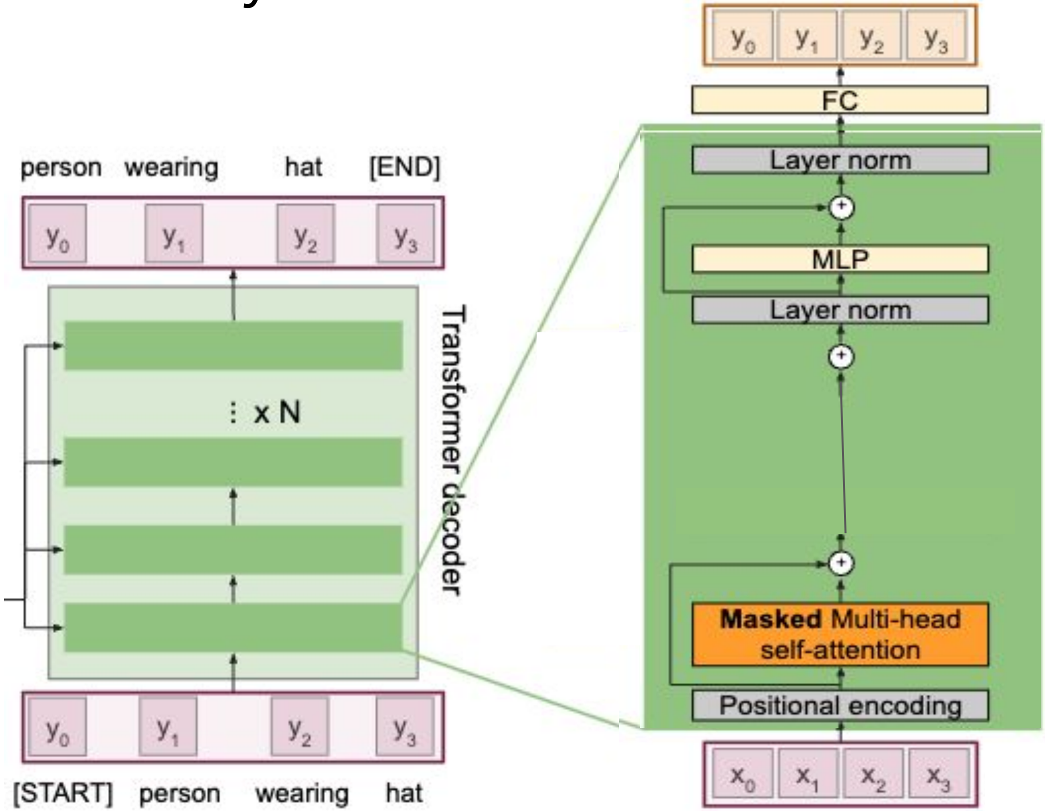
$$PE_{(pos,2i)} = sin(pos/10000^{2i/d_{model}})$$

$$PE_{(pos,2i+1)} = cos(pos/10000^{2i/d_{model}})$$

Decoder-Only Transformer



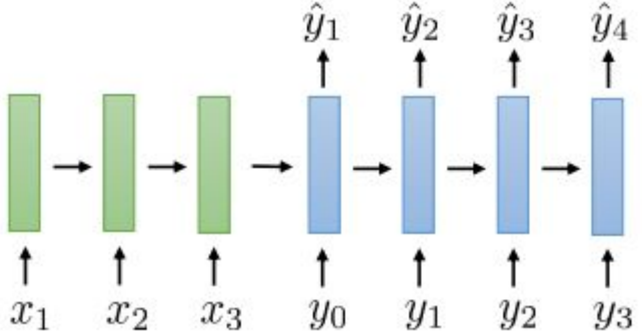
Decoder-Only Transformer

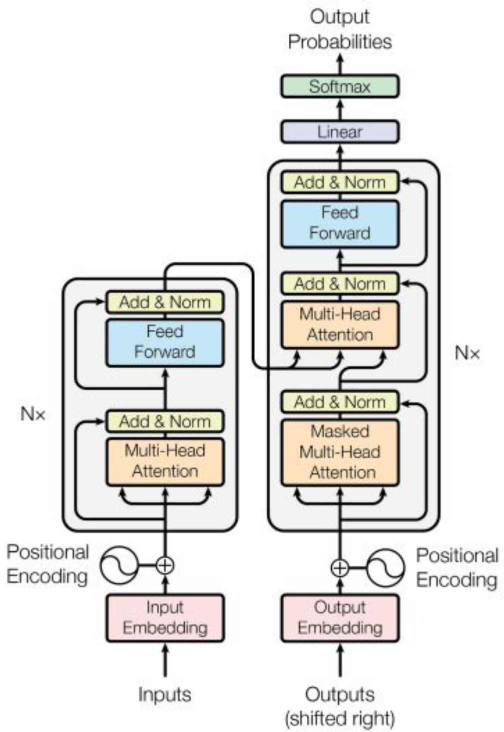


residual connect

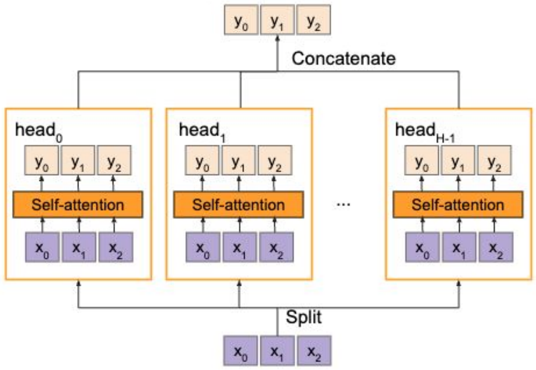
residual connect

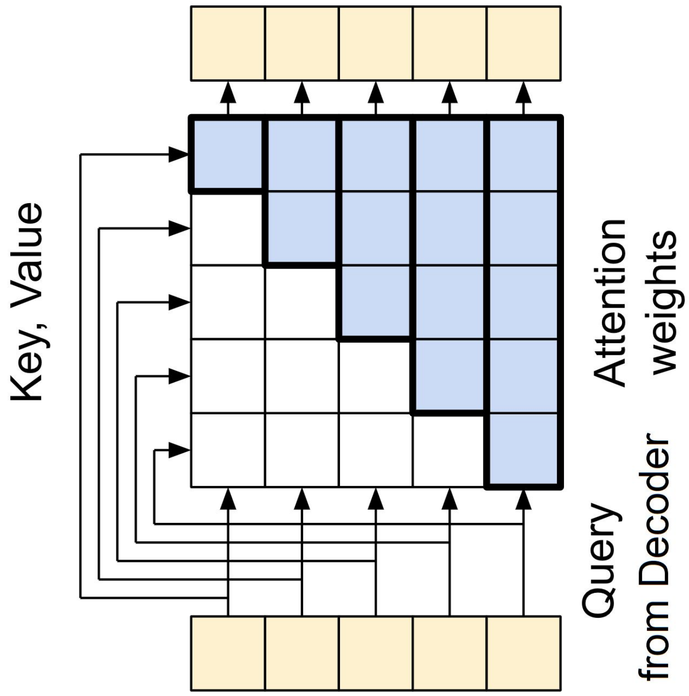
Original Transformer



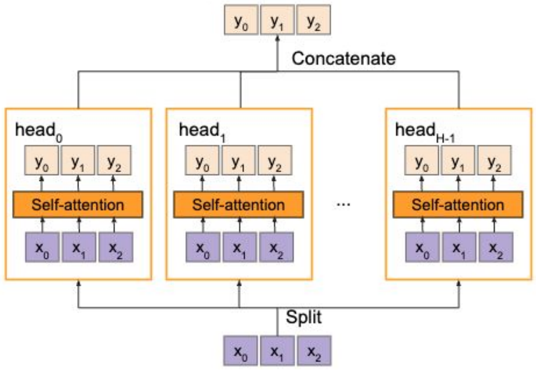


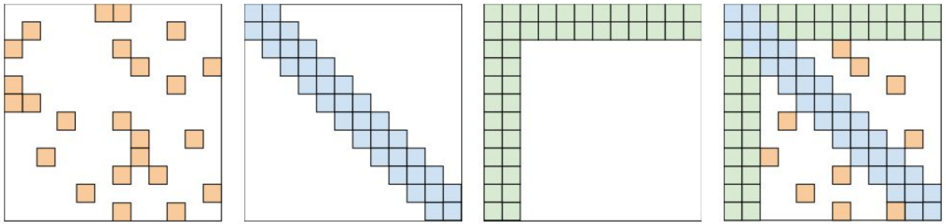
Limitations? Expensive Computation



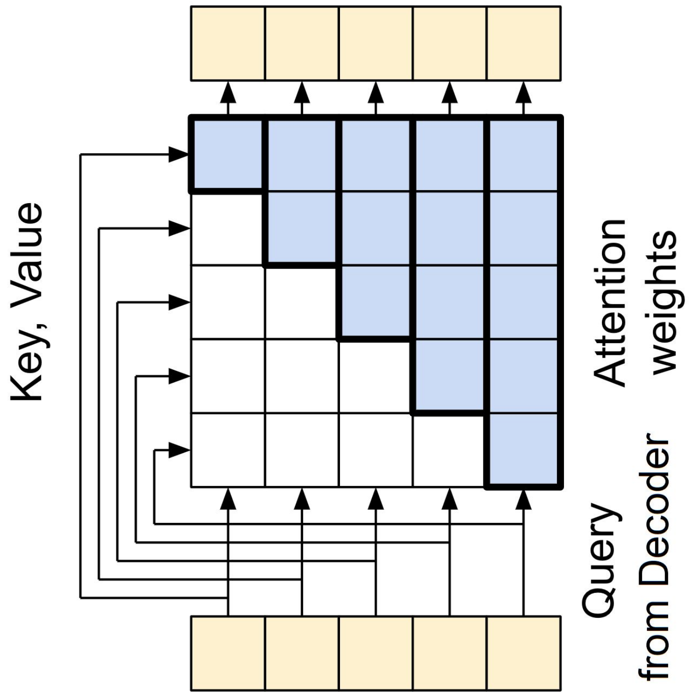


Limitations?



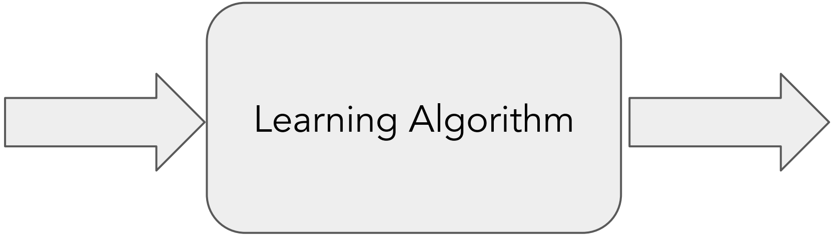


(a) Random attention (b) Window attention (c) Global Attention (d) BigBird



(Large) Language Models

Training Data



Target Function

$$D = \{X^i\}_{i=1}^n, \quad X^i = \{x_j^i\}_{j=1}^t$$

$$p(X)$$

1. Build probabilistic models
2. Derive loss function (by MLE or MAP….)
   MLE
3. Select optimizer

MLE

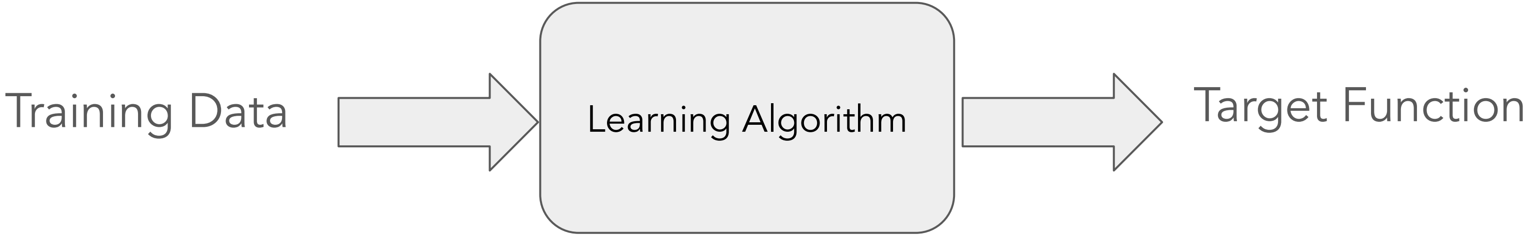
- Given all input data $D = \{X^i\}_{i=1}^n, \quad X^i = \{x_j^i\}_{j=1}^t$

$$p(X^i) = \prod_{j=1}^{t} \frac{\exp\left(W_{x_j^i}\phi\left(x_{<j}^i\right)\right)}{\sum_{l=1}^{V}\exp\left(W_l\phi\left(x_{<j}^i\right)\right)}$$

- Log-likelihood

$$\begin{aligned}
\ell(\phi) &= \sum_{i=1}^{n} p(X^i;\phi) = \sum_{i=1}^{n}\sum_{j=1}^{t}\log\left(x_j^i \mid x_{<j}^i;\phi\right) \\
&= \sum_{i=1}^{n}\sum_{j=1}^{t}\log\frac{\exp\left(W_{x_j^i}\phi\left(x_{<j}^i\right)\right)}{\sum_{l=1}^{V}\exp\left(W_l\phi\left(x_{<j}^i\right)\right)} \\
&= \sum_{i=1}^{n}\sum_{j=1}^{t} W_{x_j^i}\phi\left(x_{<j}^i\right) - \sum_{i=1}^{n}\sum_{j=1}^{t}\log\sum_{l=1}^{V}\exp\left(W_l\phi\left(x_{<j}^i\right)\right)
\end{aligned}$$

(Large) Language Models



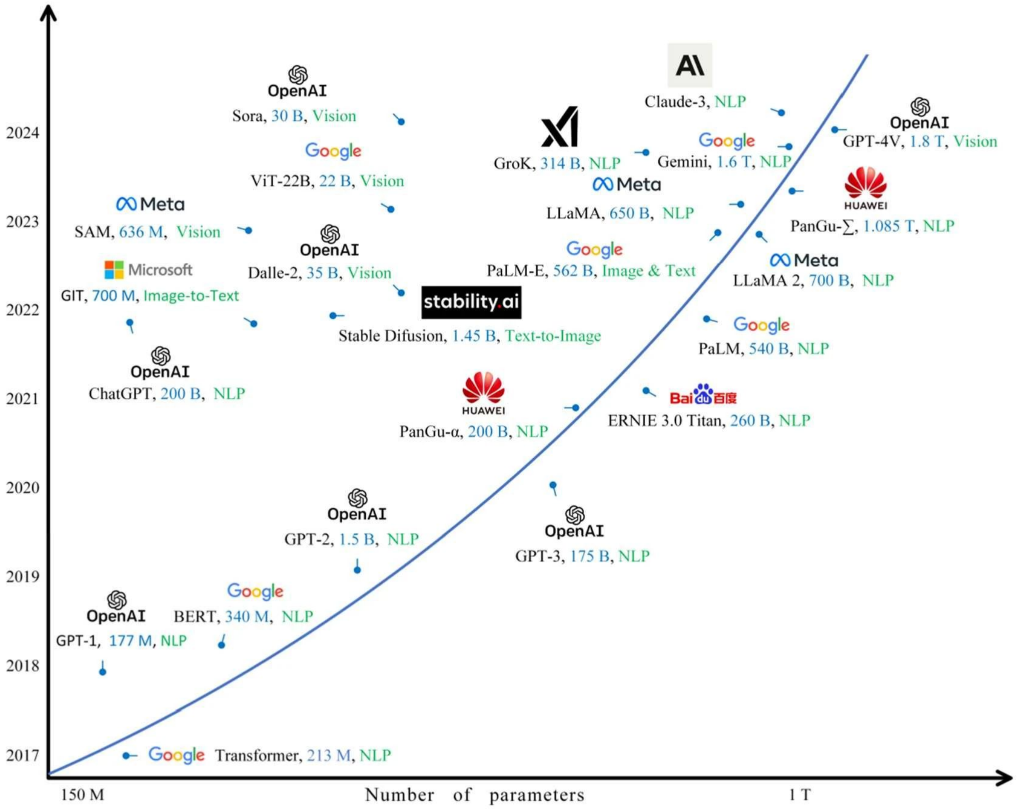
$$D = \{X^i\}_{i=1}^n, \quad X^i = \{x_j^i\}_{j=1}^t$$

$$p(X)$$

1. Build probabilistic models
2. Derive loss function (by MLE or MAP….)
3. Select optimizer

   Stochastic Gradient Descent

Large and Larger LMs



Tu, Xiaoguang, et al. "An overview of large AI models and their applications." *Visual Intelligence* 2.1 (2024): 1-22.

Summary

- Pretraining of LLM: MLE for unsupervised learning

- RNNs requires sequential processing
- RNNs have bottleneck, restricting the flexibility
- Attention is actually flattening RNN
- Transformer is deep attention

Q&A