

Transformer-based Large Language Models

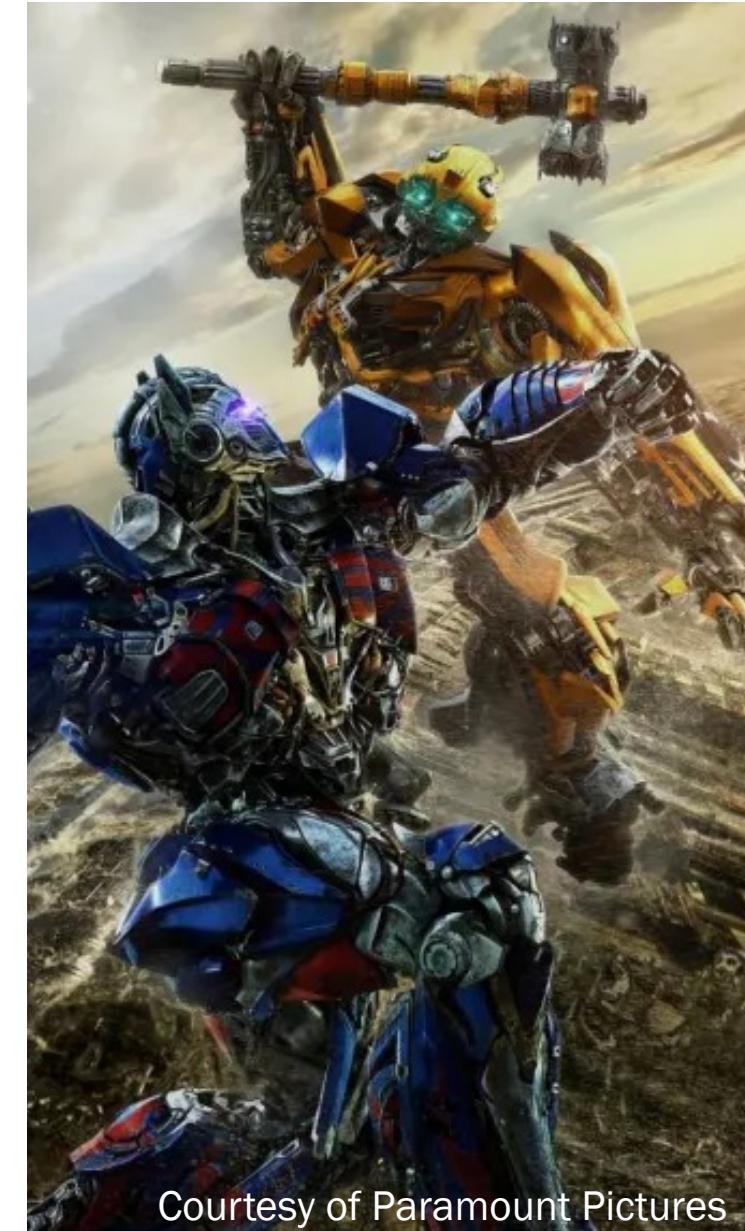
Prachya Boonkwan (Arm)

Language and Semantic Technology Lab

National Electronics and Computer Technology Center, Thailand

prachya.boonkwan@nectec.or.th, kaamanita@gmail.com

URL ⇒ <https://tinyurl.com/p8ezwzvm>



Courtesy of Paramount Pictures

Who? Me?

- Nickname: **Arm** (P'/N' Arm, etc.)
- Born: Aug 1981
- Work: researcher at NECTEC since 2005
- Education
 - B.Eng & M.Eng, CPE Kasetsart University
 - Obtained Ministry of Science Scholarship in early 2008
 - Did a PhD in Informatics (AI & Computational Linguistics) at University of Edinburgh, UK

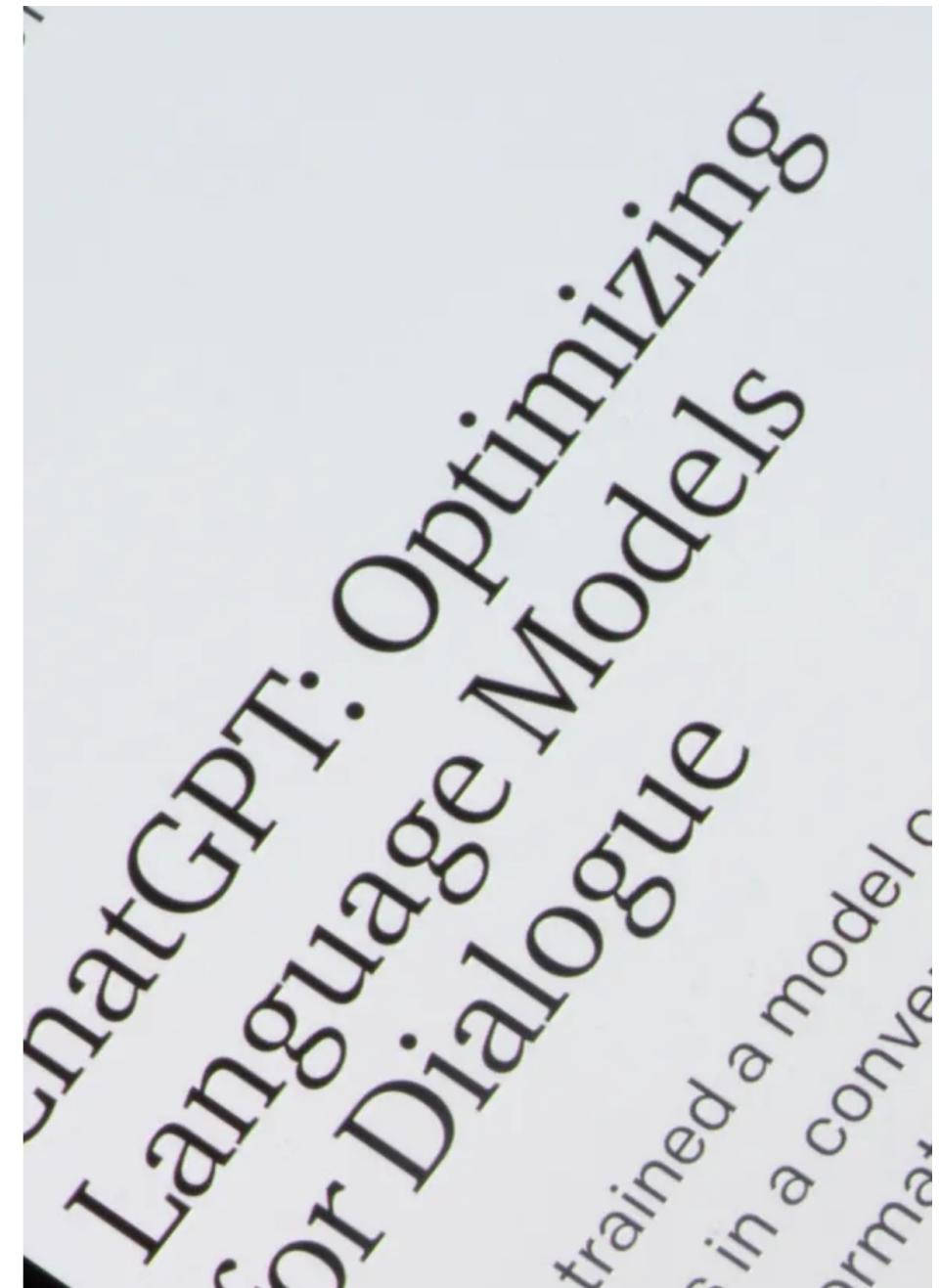


ChatGPT

ChatGPT is an advanced AI chatbot trained by [OpenAI](#) which interacts in a conversational way. The dialogue format makes it possible for ChatGPT to answer followup questions, admit its mistakes, challenge incorrect premises, and reject inappropriate requests.

ChatGPT relies on the powerful [GPT-3.5](#) technology. GPT stands for Generative Pre-Trained Transformer, a complex neural network based on the revolutionary [Attention concept](#).

<https://chat.openai.net>





I'd like to write an email to the managing director. There are three points to cover: (1) The business contract is now approved and ready to sign on both sides. (2) The final budget is 15 million dollars. (3) The stock shares are 50-50.



Subject: Approval of Business Contract and Final Budget



Dear [Managing Director's Name],

I hope this email finds you well. I am writing to inform you that the business contract between our companies has now been fully approved and is ready for signature on both sides.

Additionally, I am pleased to announce that the final budget for this project has been confirmed at 15 million dollars. Our teams have worked hard to ensure that all necessary expenses are accounted for and that this budget represents the best possible value for both of our companies.

Finally, I wanted to confirm that the stock shares will be split equally between our companies at a 50-50 ratio. This arrangement is in line with our initial discussions and reflects a fair and equitable distribution of ownership.

I look forward to the opportunity to finalize this contract and begin working together on this exciting project. Please let me know if you need any further information or if you would like to arrange a time for the contract signing.

Best regards,

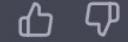
[Your Name]



I'm having a hair loss problem. Can taking a daily zinc pill solve the problem?



Zinc is an essential mineral that is important for a number of processes in the body, including hair growth. Zinc deficiency can lead to hair loss, and supplementing with zinc may help to promote hair growth in some people. However, hair loss can also be caused by a variety of other factors, such as genetics, hormonal imbalances, medical conditions, and certain medications.



If you are experiencing hair loss, it's best to consult a doctor or a dermatologist for an accurate diagnosis and appropriate treatment recommendations. They may recommend taking a zinc supplement in addition to other treatments, such as topical products, lifestyle changes, or medications. However, it's important to keep in mind that taking too much zinc can also have negative side effects, so it's important to take only the recommended amount.

In conclusion, while zinc supplementation may help with hair loss in some cases, it's not a cure-all solution and it's important to work with a healthcare professional to determine the best course of action for your specific situation.



My husband and I are going to Edinburgh in August. Can you recommend a journey plan, which consists of sightseeing, going to a concert, and visiting Edinburgh Castle?



Sure! Here's a sample itinerary for a trip to Edinburgh in August:



Day 1:

- Arrive in Edinburgh and check into your hotel.
- Explore the city's Old Town and visit the Royal Mile, a street that runs from Edinburgh Castle to the Palace of Holyroodhouse.
- Visit Edinburgh Castle, one of Scotland's most famous landmarks, to learn about the history of the castle and the city.

Day 2:

- Head to the Princes Street Gardens to enjoy the city's green spaces and take in the views of Edinburgh Castle.
- Visit the National Museum of Scotland to learn about the history and culture of Scotland.
- In the evening, attend a concert of your choice at the Usher Hall, one of Scotland's largest concert halls.

Day 3:

- Visit the Royal Botanic Garden Edinburgh to see the diverse plant collection, including a large collection of rhododendrons and azaleas.
- Explore the New Town and visit the Georgian House, a museum that showcases life in Edinburgh during the 18th century.
- In the evening, take a ghost tour of the city to learn about its haunted history.



Please write a piece of Python code that parses an input text in English into a dependency tree.



Here's an example of how you could parse English text into a dependency tree using the popular Natural Language Toolkit (NLTK) library in Python:

python

Copy code

```
import nltk
from nltk import word_tokenize
from nltk.parse import CoreNLPParser

# Initialize the parser
parser = CoreNLPParser(url='http://localhost:9000')

# Define the input text
text = "I saw the white car with my own eyes."

# Tokenize the text
tokens = word_tokenize(text)

# Parse the text into a dependency tree
dep_tree = next(parser.parse(tokens))

# Display the dependency tree
print(dep_tree)
```

This code uses the CoreNLPParser from NLTK, which is a wrapper for the Stanford Parser, to parse the input text into a dependency tree. The input text is first tokenized into individual words, and then passed to the parser for parsing. The parse method returns a generator, so we use `next` to get the first (and only) parse tree.



ChatGPT (OpenAI)



Copilot (Microsoft)



perplexity

Perplexity.AI



Gemini (Google)



Claude.AI (Anthropic)



Aran Komatsuzaki ✅

@arankomatsuzaki

OpenAI did what used to be considered impossible.

They made people want to use Bing.

Key Differences

ChatGPT = chatbot

- It answers your questions with its knowledge
- Generative AI: it generates new outputs unseen in the dataset
- Its knowledge is based on the training data (up to 2021)
- Use it when you need quick answers and get the job done

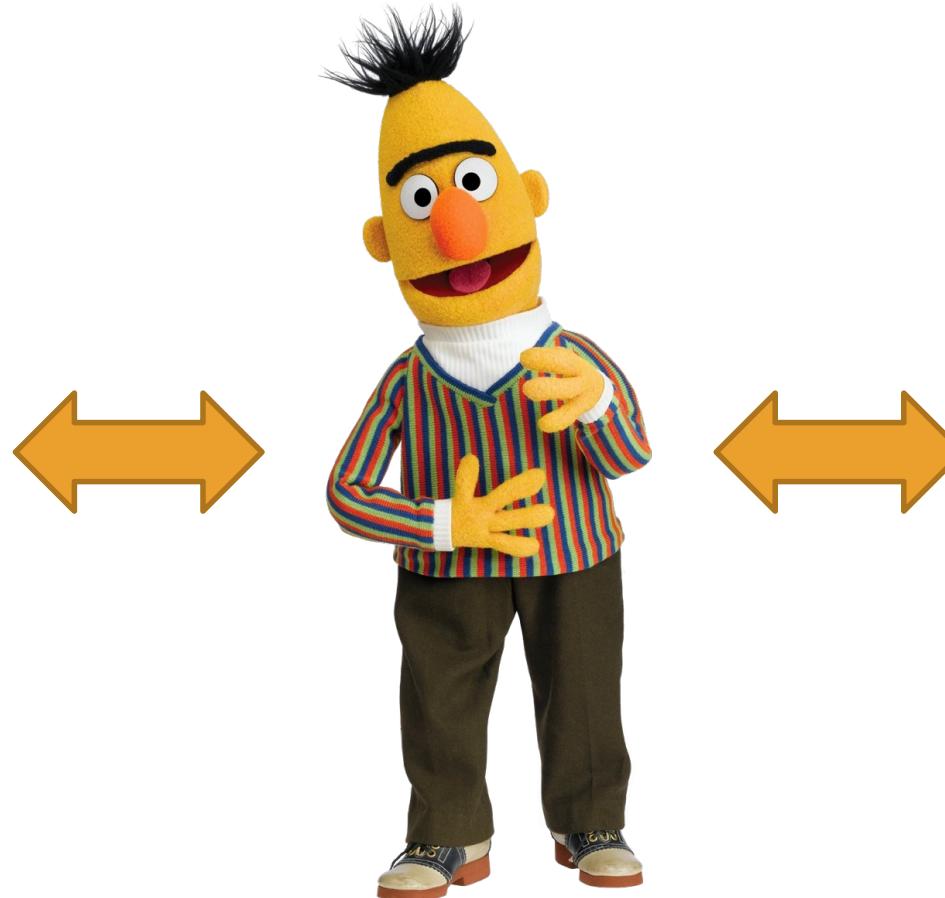
Google = search engine

- It searches things in a very large repository for you
- Non-generative AI: it never generates unseen outputs
- Its knowledge is always up-to-date and much more voluminous
- Use it when you want to compile knowledge from external resources and internalize the knowledge

Components of ChatGPT



1) Engine:
Transformer Model



2) Language Model:
GPT (BERT-based)



3) Dataset:
Chat-based Instruction

Outline

- Overview of the Transformer model
- Model interpretation
- Theoretical upper bounds
- Large language models
- Some popular BERT variants
- ChatGPT
- Generative AI & business opportunities
- Conclusion and discussion time

1. Overview of the Transformer Model

Transformer Model (Vaswani et al., 2016)

- Sequence-to-sequence generation
 - Translation:** It learns how to produce a target sequence from a source sequence, given a very large dataset of sequence pairs
 - Pros:** It learns **word collocations** and **phrase structures** on the input and output sequences, and associates them cross-lingually in the table of **translation alignments**
 - Cons:** It consists of an expansive amount of neuron cells, and the training process can be quite time-consuming

Mary looks this word up in the dictionary

Source: sequence of words



TRANSFORMER

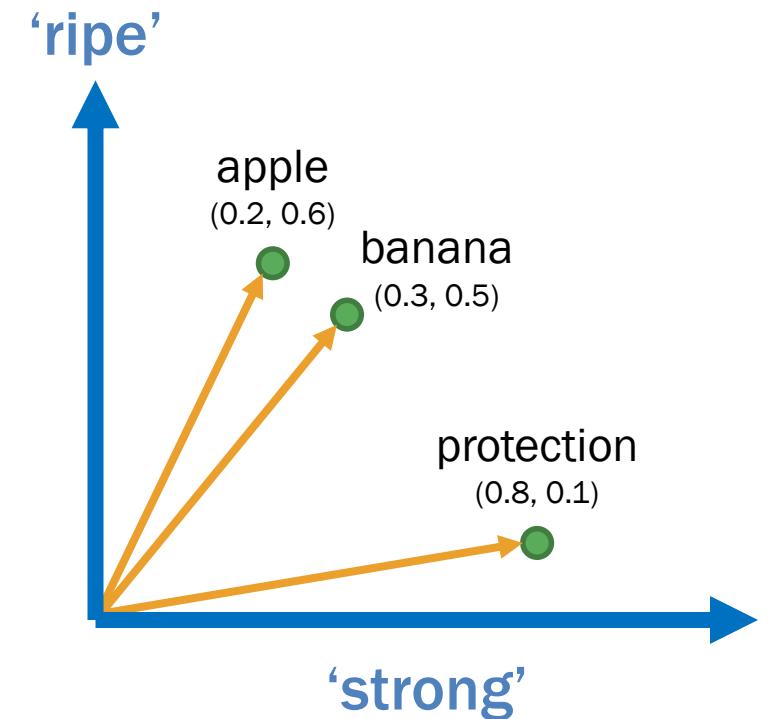


แมรี ค้นหา คำ นี้ ใน พจนานุกรม

Target: sequence of words

Recap: Word Vector

- Distributional similarity of words
 - Measured by co-occurrence of words and their contexts (i.e. **context distribution**)
 - Various types of similarity metrics and context are employed (Dagan+, 2008)
 - Backoff method (Katz, 1987) with interpolation (Jelinek+, 1980) are required for smoothing the zero counts
 - Bigram co-occurrence is commonly used as the context representation (Brown+, 1992)
 - Well-known techniques include Latent Semantic Analysis (Dumais, 2005) and PMI (Church+, 1990)



In this over-simplified example, each numeric element is:

$$p(\text{word} | \text{context})$$

Notable Applications in NLP

Applications	Descriptions	Input	Output	What is Learned?
Neural machine translation	Convert a text from the source language to the target language	Word sequence in the source language	Word sequence in the target language	<ul style="list-style-type: none"> Word alignment (cross-lingual semantic similarity) MWEs in both languages (semantic similarity)
Abstractive summarization	Translate a text into a shorter version in the same language	Word sequence of full text	Word sequence of summary	<ul style="list-style-type: none"> MWEs in the language Pronoun substitution Conceptualization
Image captioning	Explain an image with a short description	Sequence of image fragments	Word sequence of image caption	<ul style="list-style-type: none"> Image-to-word alignment (multimodal semantic similarity) MWEs in the language
Speech recognition	Transcribe a sequence of audio signal into phonetic representation (IPA)	Sequence of audio signals (frequency domain)	Sequence of phonetic representation	<ul style="list-style-type: none"> Sound-to-transcription alignment (multimodal semantic similarity) Phonetic processes in the language

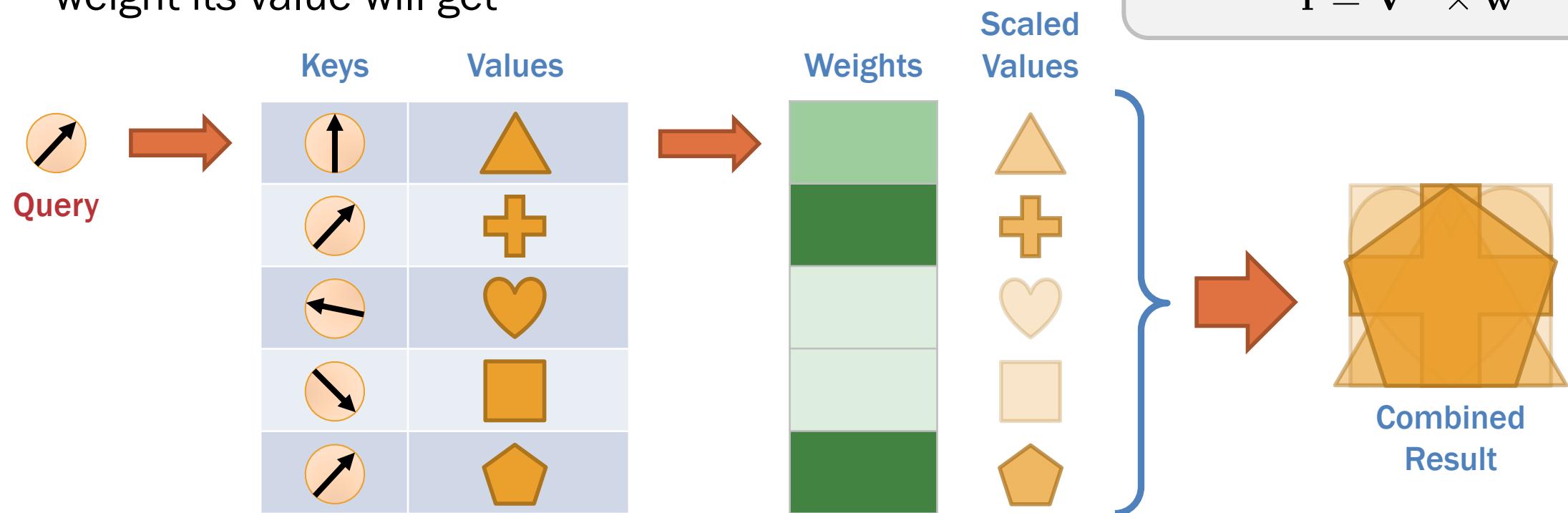
'Plausible' Applications in NLP

Applications	Descriptions	Input	Output	What is Learned?
Sequential tagging	Annotate each token of a given sequence with a linguistic tag (e.g. POS and NE)	Sequence of characters or words	Sequence of words with linguistic tags	<ul style="list-style-type: none"> Token-to-tag alignment Contextual clues for linguistic annotation Joint annotation model
Syntactic parsing	Annotate a sequence of words with a syntactic structure	Sequence of words	Sequence of parsing actions (shift, reduce, accept, backtrack)	<ul style="list-style-type: none"> Word-to-tree alignment Parsing model based on semantic similarity
Word segmentation with term normalization	Tokenize a given string into a word sequence and normalize non-canonical terms	Sequence of characters	Sequence of words	<ul style="list-style-type: none"> MWEs in the language Spelling rules
Relation extraction	Determine the relationship between the main verb and its arguments	Sequence of words	Knowledge graph	<ul style="list-style-type: none"> Verb-to-argument relationship based on semantic similarity MWEs in the language

2. Model Interpretation

Scaled Dot-Product Attention

- Semantic similarity \Rightarrow search engine
 - Query is compared against each key with dot product
 - The more similar the key is to the query, the more weight its value will get



$$w_i \propto k_i \cdot q$$

Simple
Form

$$\mathbf{r} = \sum_{i=1}^N w_i \mathbf{v}_i$$

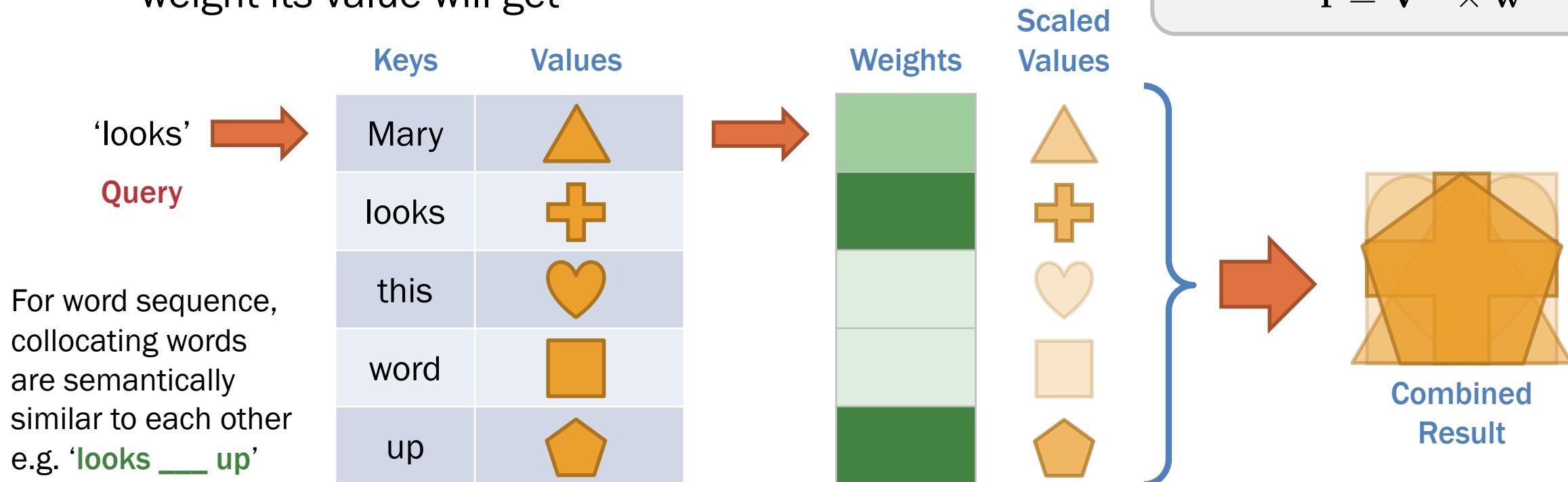
Matrix
Form

$$\mathbf{w} = \text{Softmax}(\mathbf{K} \times \mathbf{q})$$

$$\mathbf{r} = \mathbf{V}^\top \times \mathbf{w}$$

Scaled Dot-Product Attention

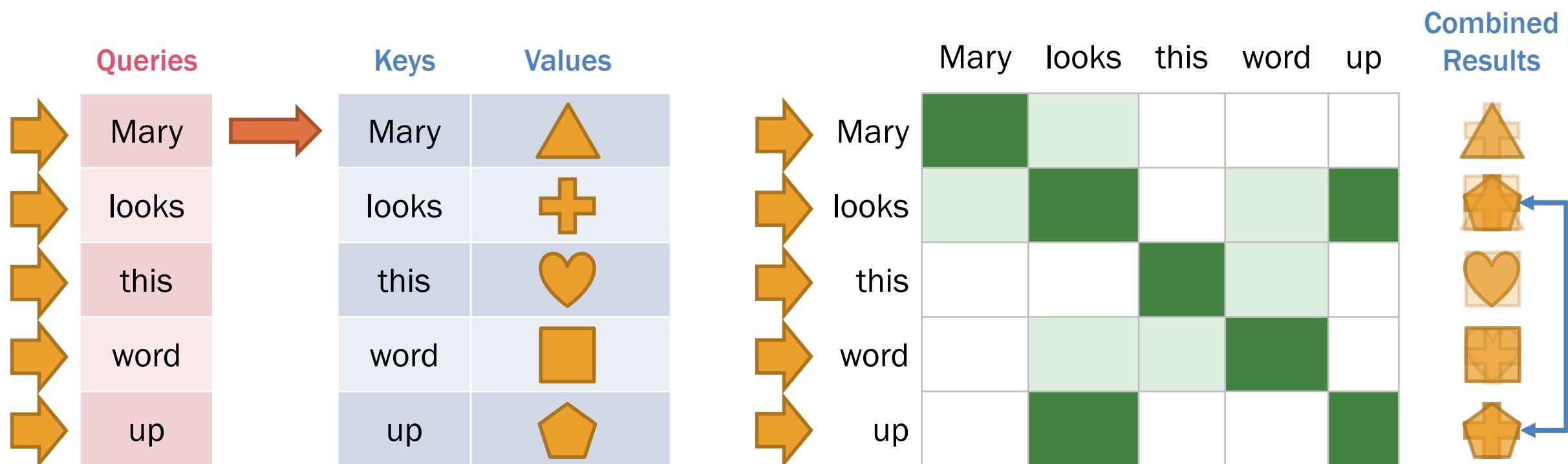
- Semantic similarity \Rightarrow search engine
 - Query is compared against each key with dot product
 - The more similar the key is to the query, the more weight its value will get



Self-Attention

- Scaled dot-product attention whose queries and keys are the same
- Collocations will have almost similar results

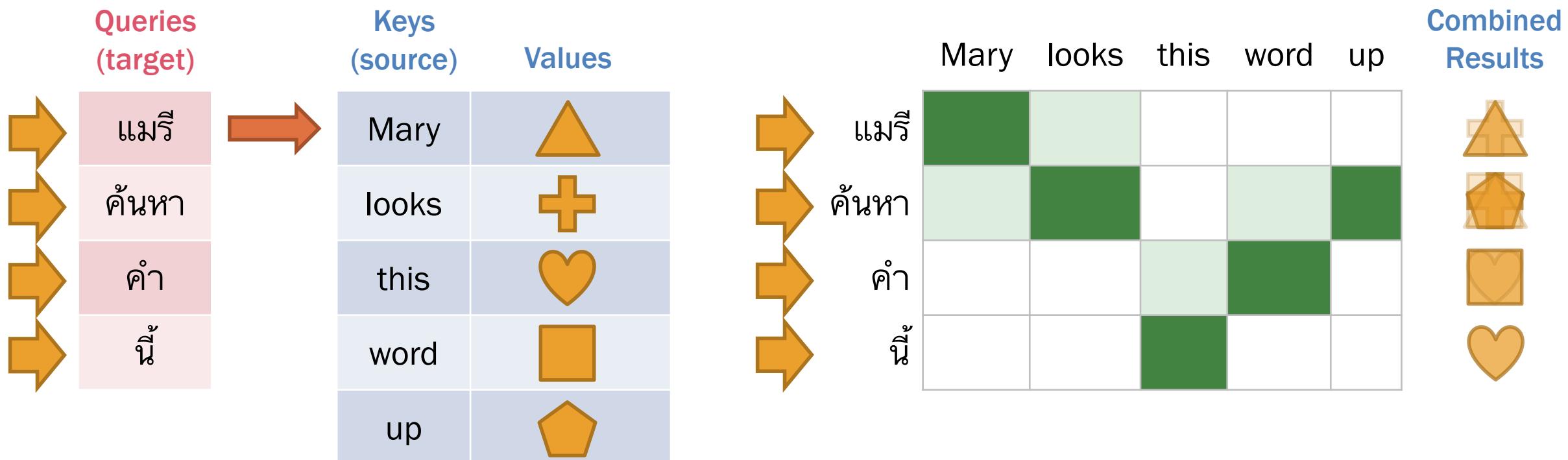
$$\begin{array}{ll} \text{Matrix Form} & \mathbf{W} = \text{Softmax}(\mathbf{K} \times \mathbf{K}^T) \\ & \mathbf{R} = \mathbf{W} \times \mathbf{V} \end{array}$$



Cross-Attention

- Scaled dot-product attention whose queries are the target and whose keys are the source
- Collocation alignment via semantic similarity

$$\begin{aligned} \text{Matrix Form } \mathbf{W} &= \text{Softmax}(\mathbf{Q} \times \mathbf{K}^\top) \\ \mathbf{R} &= \mathbf{W} \times \mathbf{V} \end{aligned}$$



Multihead Attention

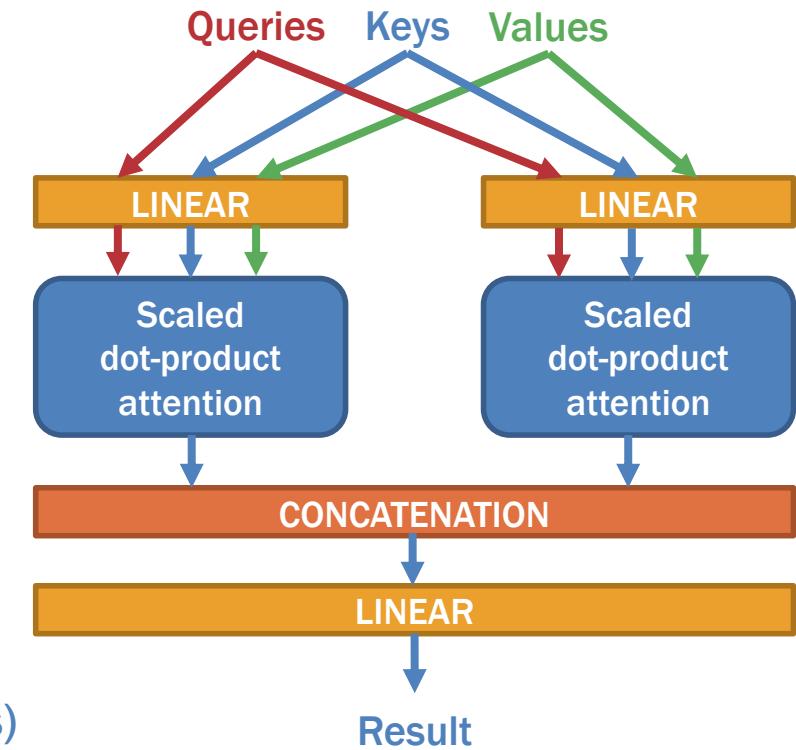
- Scaled dot-product attention has a drawback
 - It recognizes **only one** type of word collocation
 - If we assume more than one type of word collocation per sequence, then we have to combine multiple attention heads [default = 8 heads]

HEAD 1 (looks __ up)

Mary	Poppins	looks	this	word	up
Mary					
Poppins					
looks					
this					
word					
up					

HEAD 2 (Mary Poppins)

Mary	Poppins	looks	this	word	up
Mary					
Poppins					
looks					
this					
word					
up					



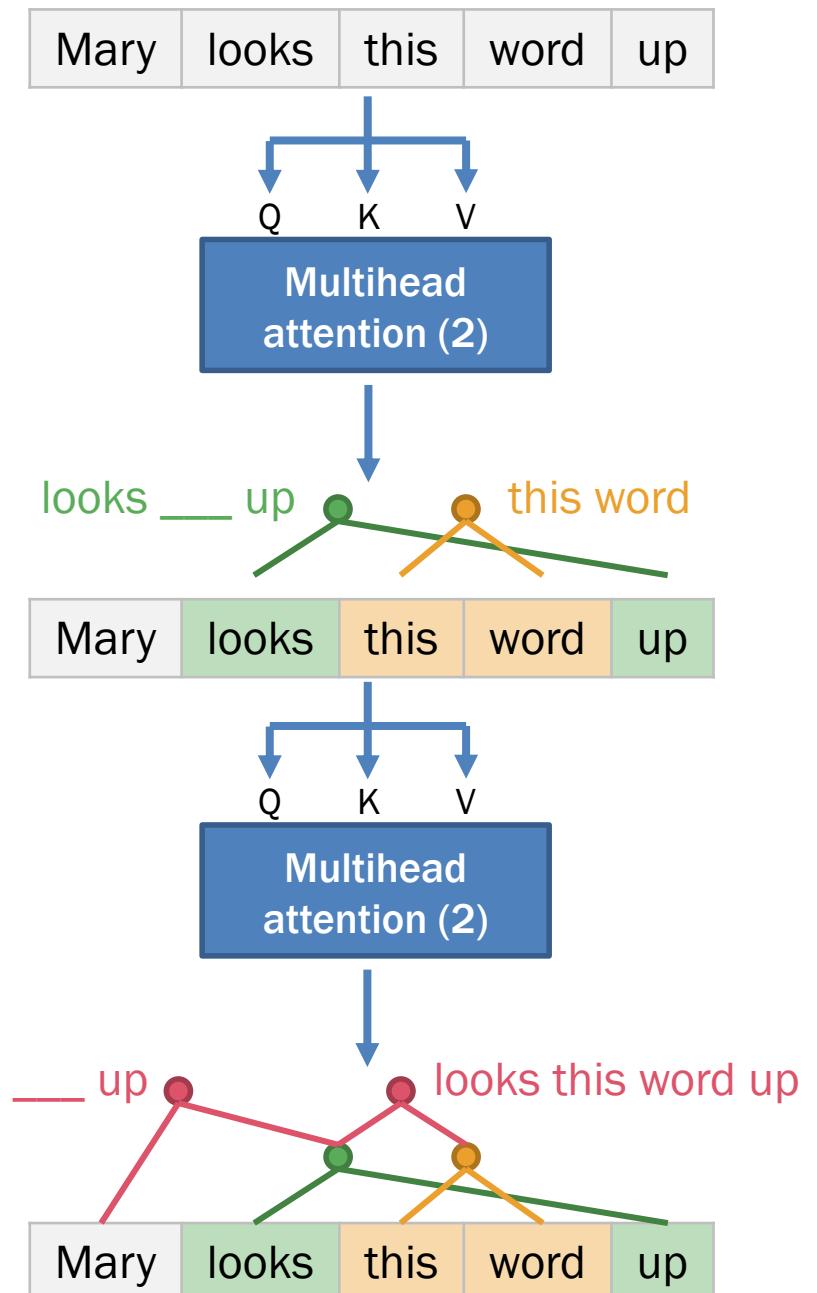
Notation

\downarrow \downarrow \downarrow
Q K V

**Multihead
attention (n)**

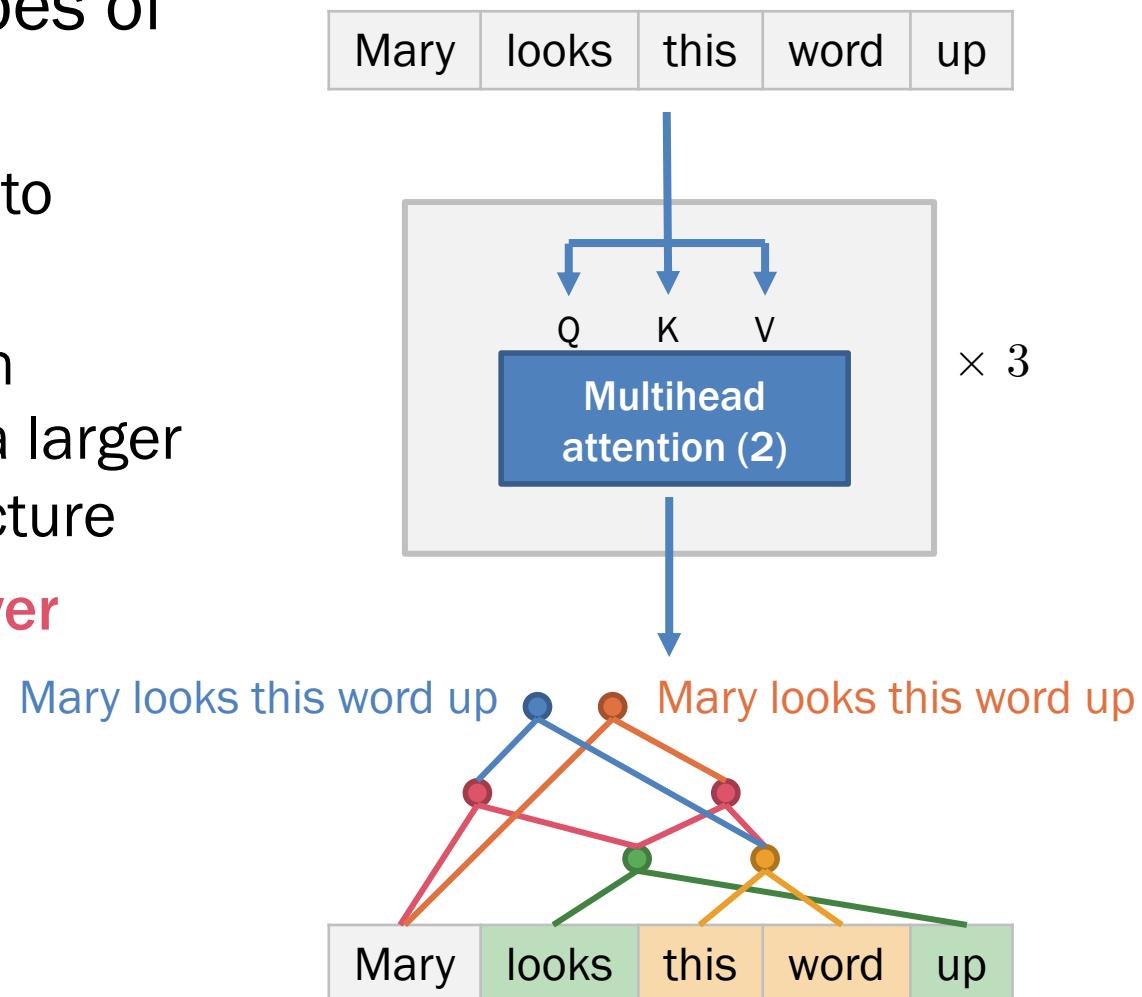
Phrase Structure

- H -head self-attention recognizes H types of word collocation per sequence
 - One layer can combine consecutive words to become a phrase
 - More layers of multihead self-attention can combine consecutive phrases to become a larger phrase or even a sentence \Rightarrow phrase structure
 - Each layer is simply called an **encoding layer**



Phrase Structure

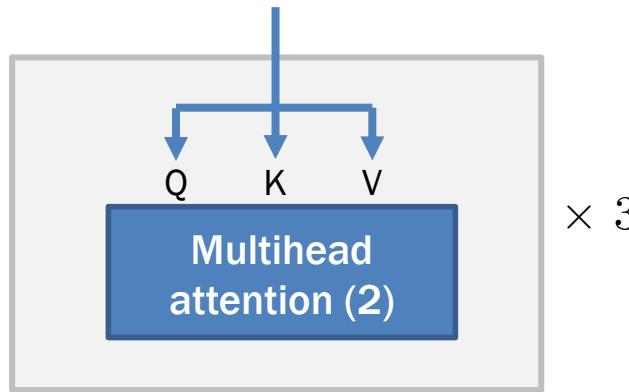
- H -head self-attention recognizes H types of word collocation per sequence
 - One layer can combine consecutive words to become a phrase
 - More layers of multihead self-attention can combine consecutive phrases to become a larger phrase or even a sentence \Rightarrow phrase structure
 - Each layer is simply called an **encoding layer**



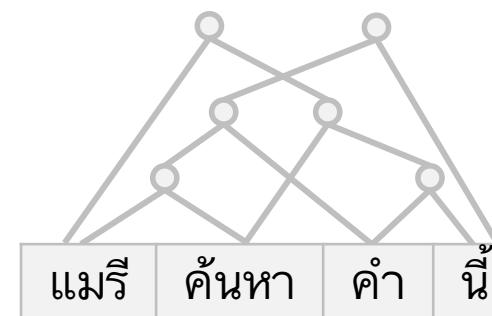
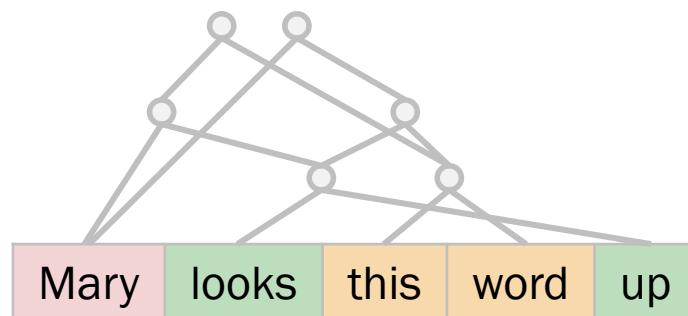
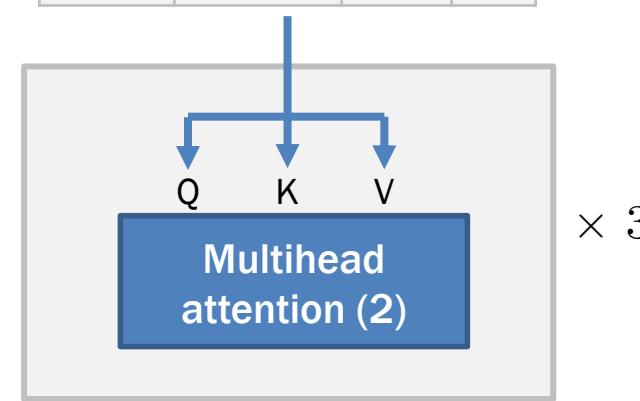
Alignment of Phrase Structures

- H -head alignment attention recognizes H pairs of phrase structures

Mary looks this word up

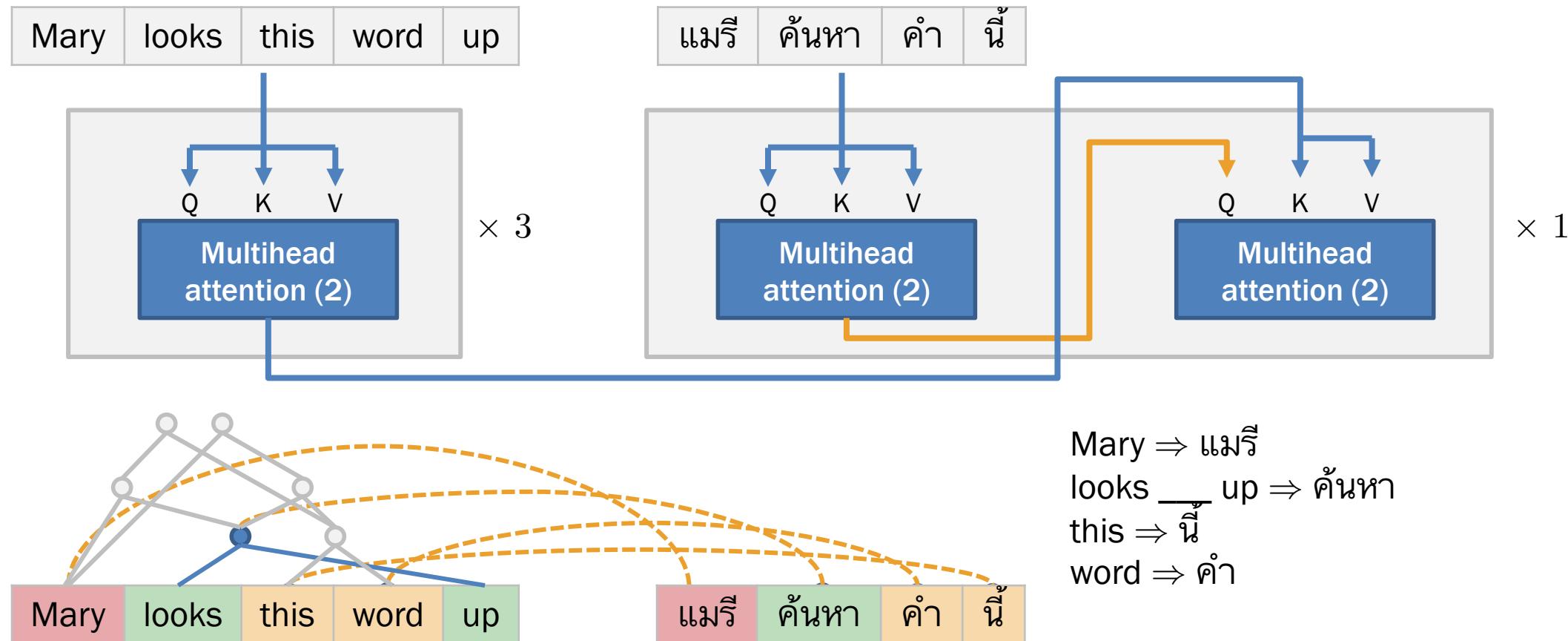


แมรี ค้นหา คำ นี่



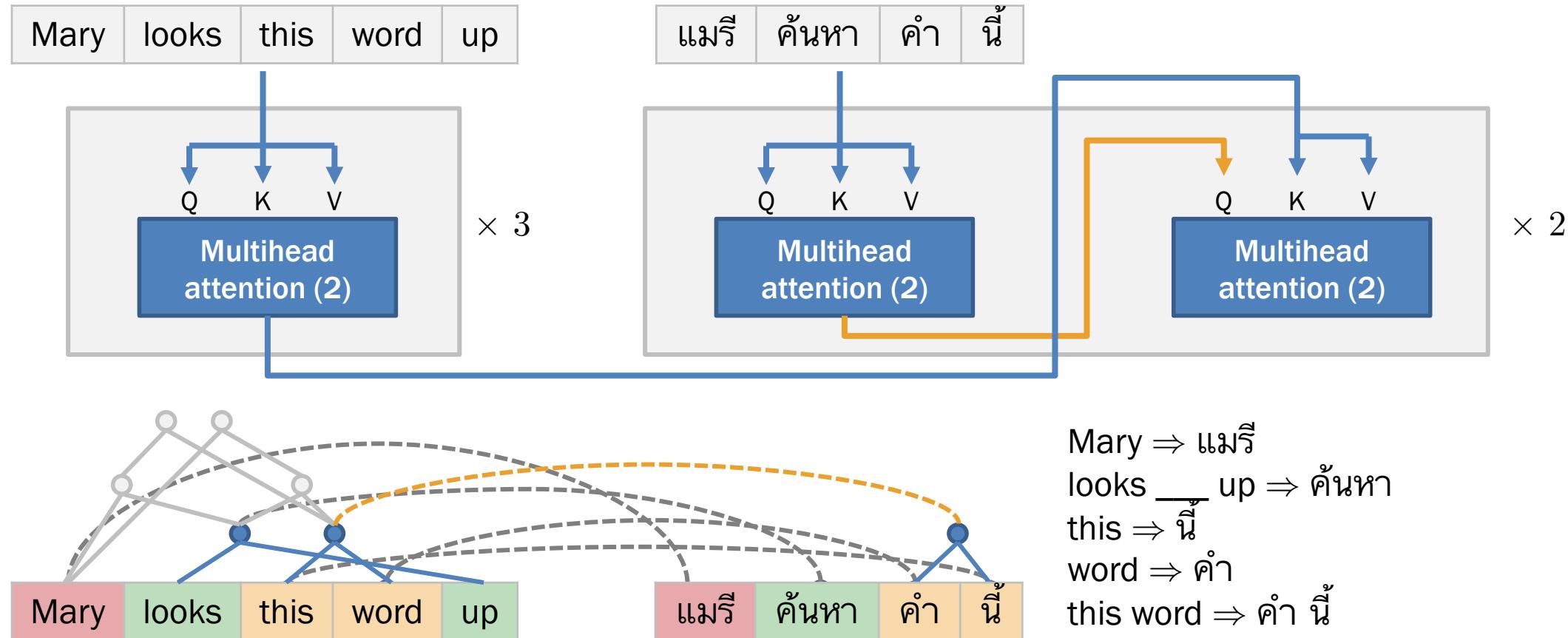
Alignment of Phrase Structures

- H -head alignment attention recognizes H pairs of phrase structures \Rightarrow **decoding layer**



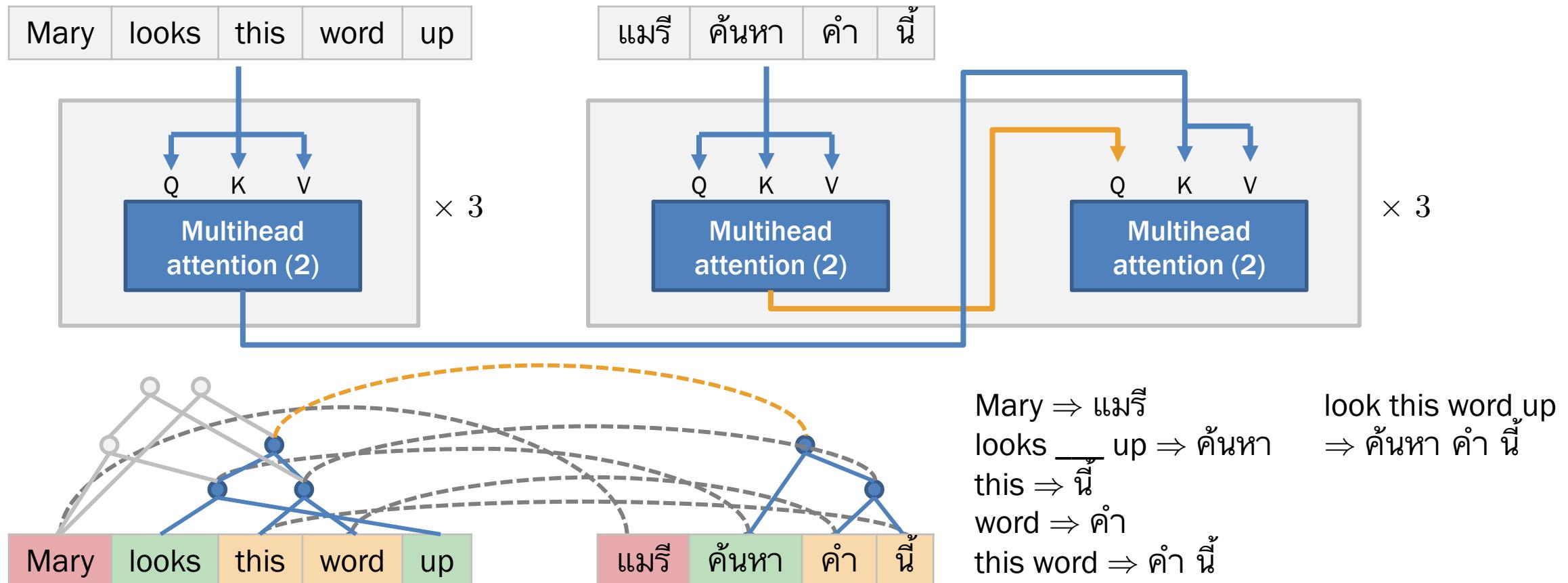
Alignment of Phrase Structures

- H -head alignment attention recognizes H pairs of phrase structures \Rightarrow **decoding layer**



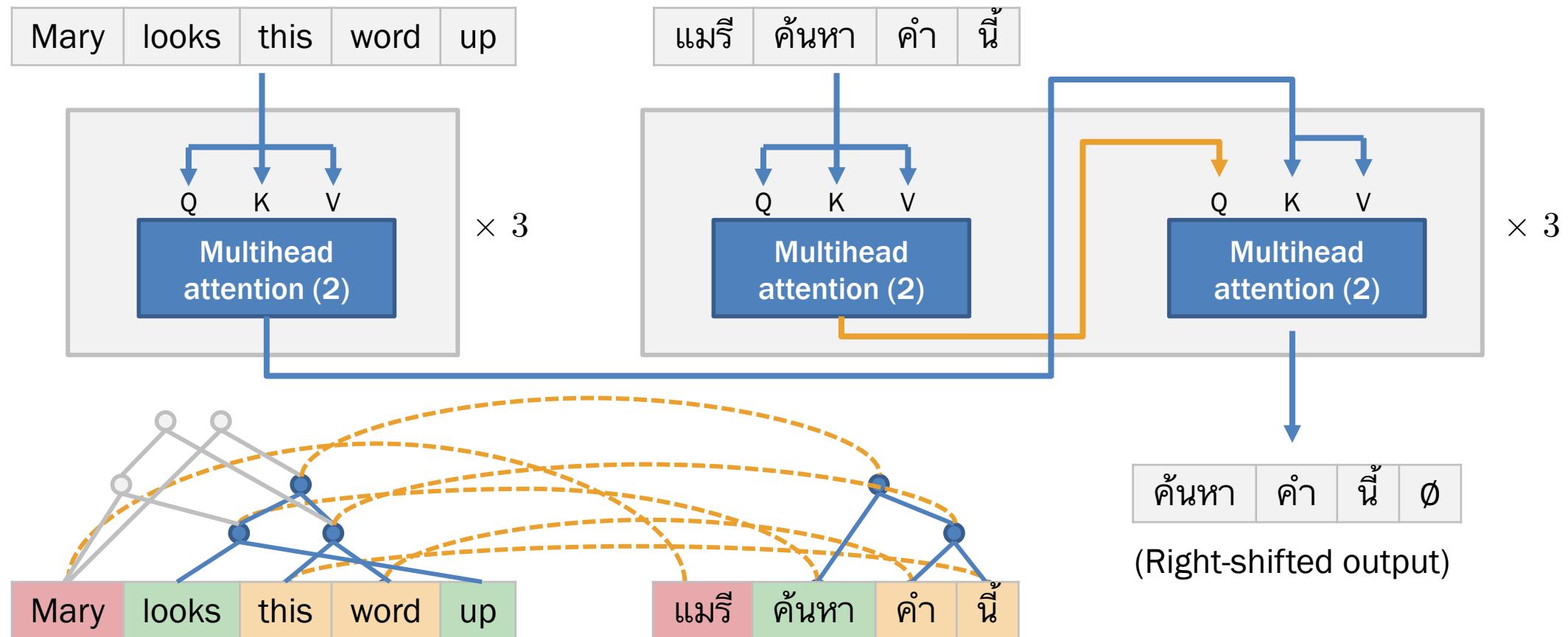
Alignment of Phrase Structures

- H -head alignment attention recognizes H pairs of phrase structures \Rightarrow **decoding layer**

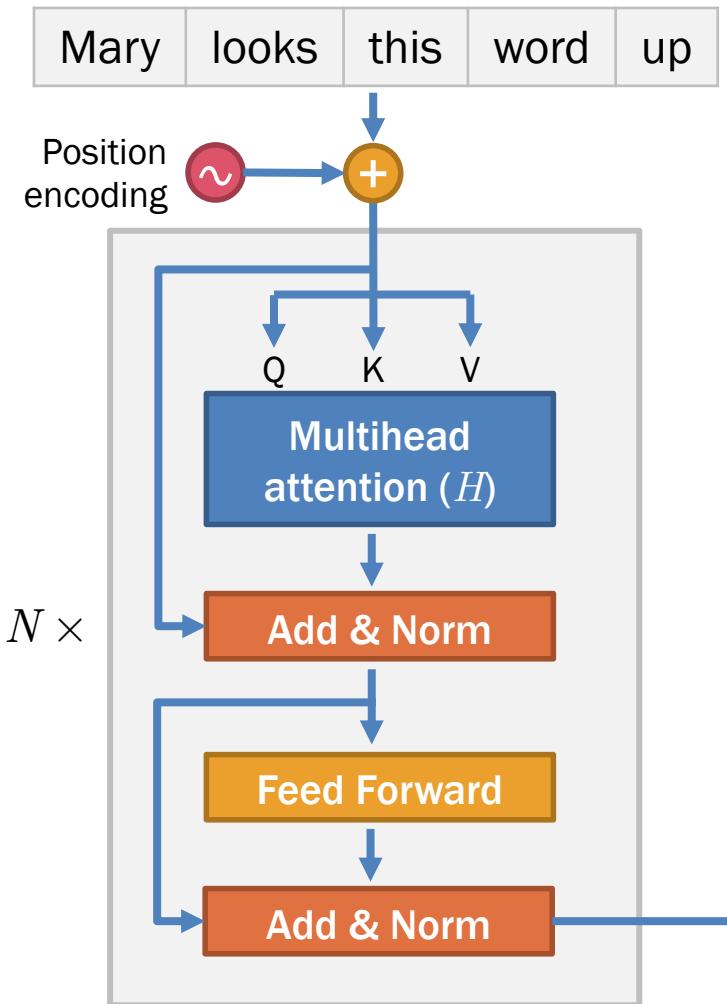


Alignment of Phrase Structures

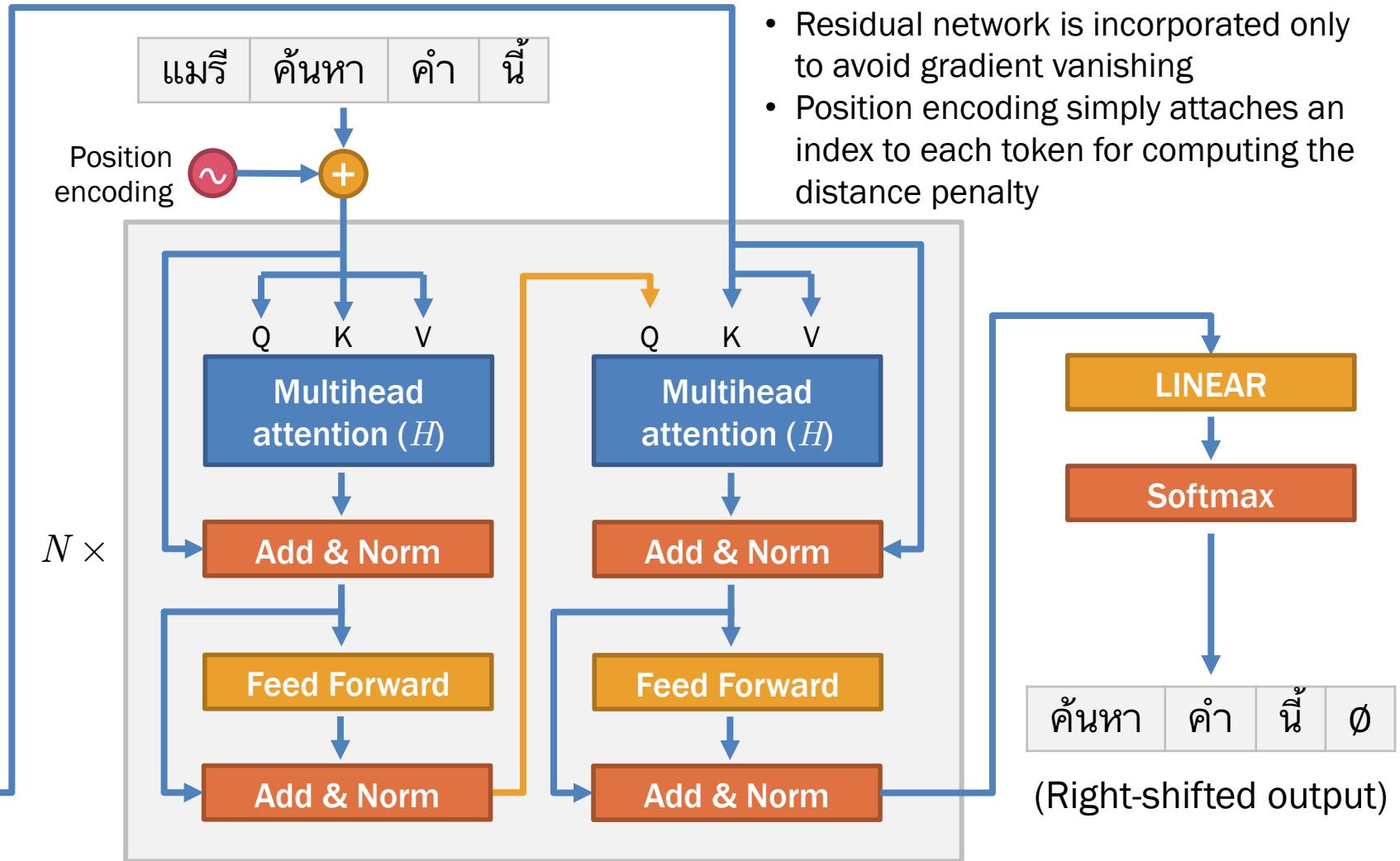
- H -head alignment attention recognizes H pairs of phrase structures \Rightarrow **decoding layer**



Overview of the Transformer Model



1. Extract phrase structures in the source

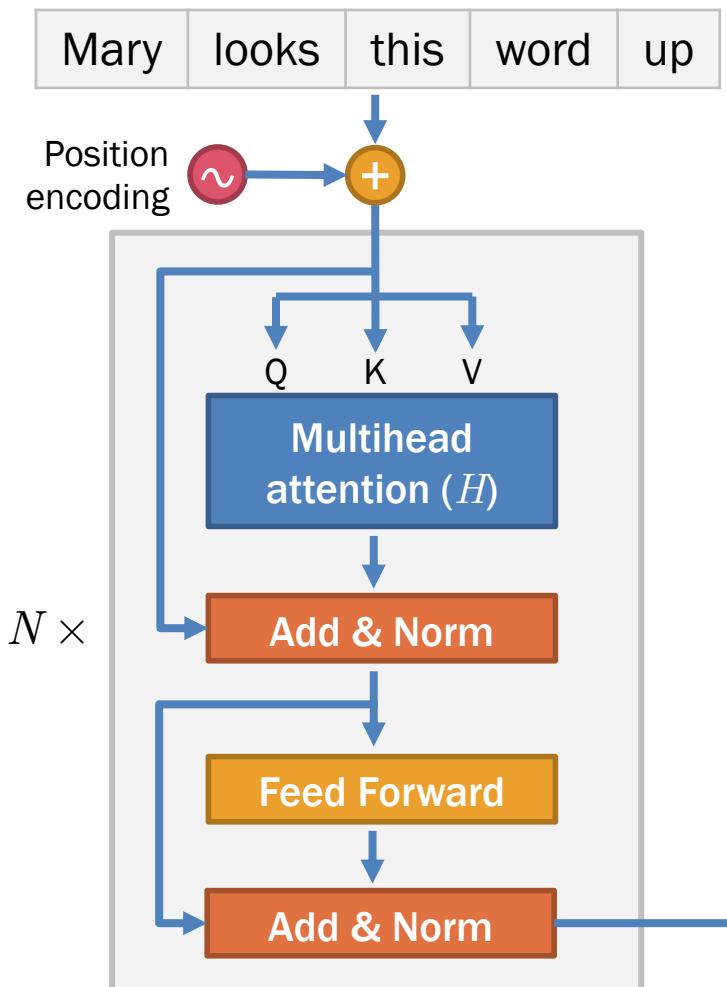


2. Align phrase structures to the target

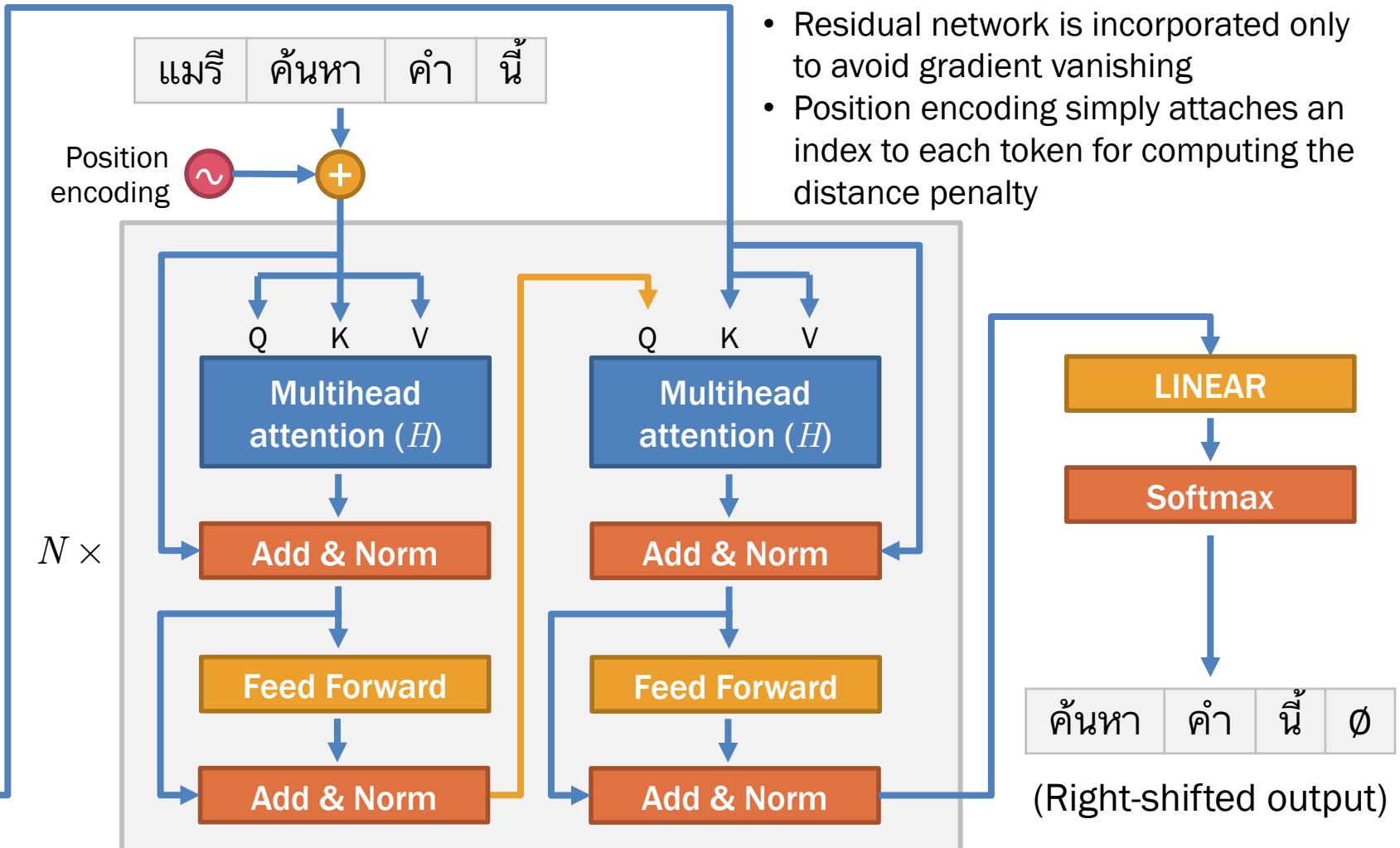
3. Produce translation

- Residual network is incorporated only to avoid gradient vanishing
- Position encoding simply attaches an index to each token for computing the distance penalty

Overview of the Transformer Model



1. ENCODE



2. DECODE

3. PRODUCE

Evaluation: BLEU Score /blø:/

- BiLingual Evaluation Understudy (BLEU)
- n-gram precision = ratio between the matched n -grams **against the candidate**

$$1\text{-gram prec} = \frac{7}{10}$$

$$2\text{-gram prec} = \frac{4}{9}$$

$$3\text{-gram prec} = \frac{1}{8}$$

Reference

the Iraqi weapons are to be handed over to the army within two weeks

Candidate
(decoded)

in two weeks Iraqi weapons will give to the army

$$\begin{aligned} \text{BLEU} &= \left(\prod_{n=1}^3 p_n \right)^{1/3} \\ &= \left(\frac{7}{10} \times \frac{4}{9} \times \frac{1}{8} \right)^{1/3} \end{aligned}$$

Evaluation: ROUGE Scores /xu:3/

- ROUGE- n = ratio between matched n -grams **against the reference**
- ROUGE- L = geo. mean of ratios between the longest common subsequence **and both texts**

Reference

the Iraqi weapons are to be handed over to the army within two weeks

Candidate
(decoded)

in two weeks Iraqi weapons will give to the army

$$\text{ROUGE-1} = \frac{7}{14}$$

$$\text{ROUGE-2} = \frac{4}{13}$$

$$\text{ROUGE-3} = \frac{1}{12}$$

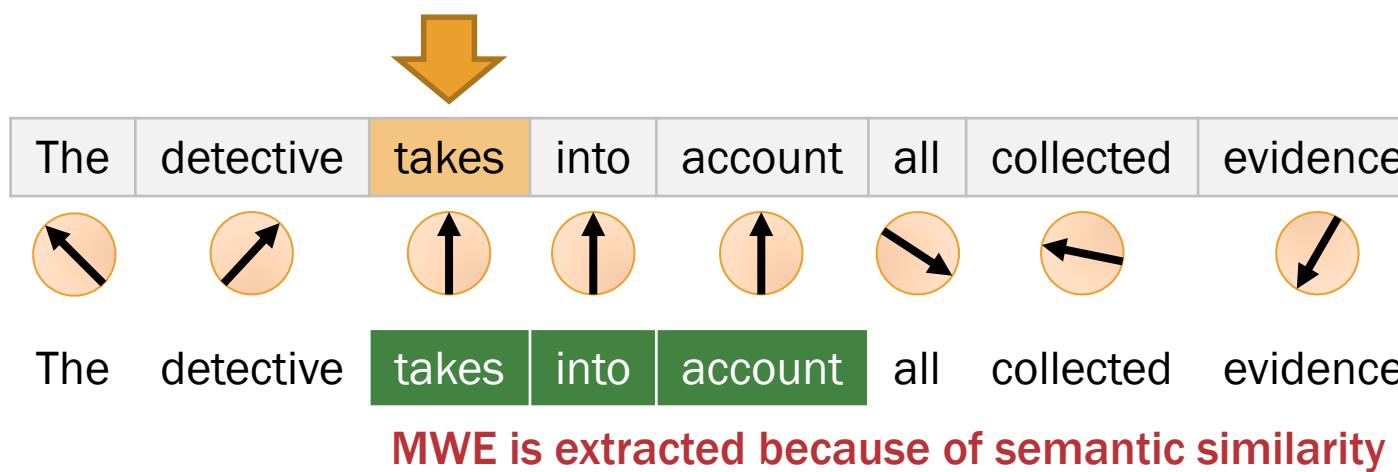
$$\text{Prec} = \frac{5}{10}$$

$$\text{Rec} = \frac{5}{14}$$

$$\text{ROUGE-}L = \frac{2}{\frac{1}{\text{Prec}} + \frac{1}{\text{Rec}}}$$

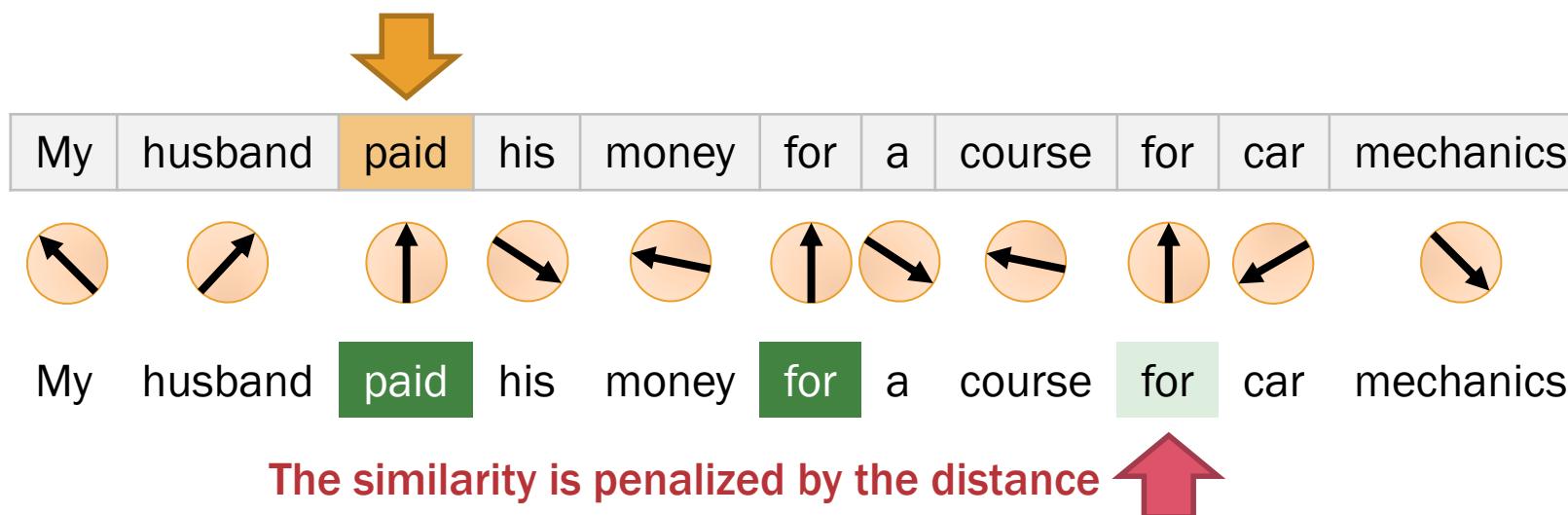
Pros: Multiword Expression (MWE)

- It recognizes the idiosyncratic collocations of at least 2 words
 - E.g. ‘peanut butter’, ‘car park’, ‘kick the bucket’, ‘take into account’, ‘break up’
 - It learns MWEs by comparing each word with the remaining to reveal semantic similarity



Pros: Moderate-Distance Dependency

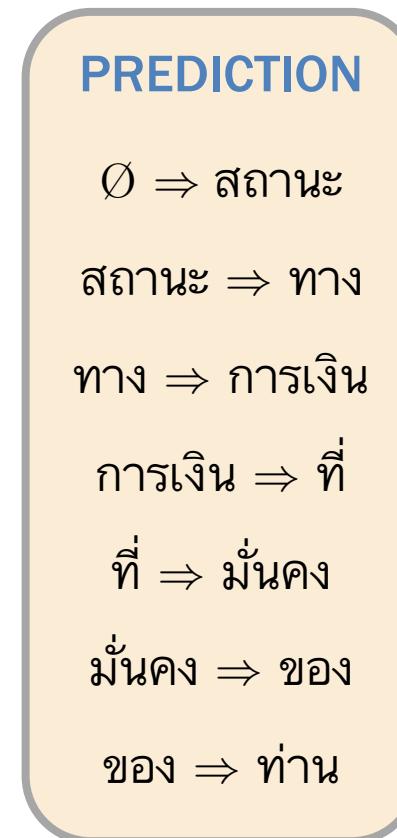
- It recognizes word collocation that is separate within a moderate distance
 - E.g. ‘look ____ up’, ‘ask ____ out’, ‘pay ____ for’
 - It learns moderate-distance dependency with semantic similarity and distance penalty



Pros: Moderate Reordering

- It learns to reorder words with next-word prediction (language model), cross-lingual semantic similarity, and distance penalty

your	stable	financial	status
your	stable	financial	status
your	stable	financial	status
your	stable	financial	status
your	stable	financial	status
your	stable	financial	status
your	stable	financial	status
your	stable	financial	status



Next-word prediction takes into account an entire input sequence

สถานะ
สถานะ ทาง
สถานะ ทาง การเงิน
สถานะ ทาง การเงิน ที่
สถานะ ทาง การเงิน ที่ มั่นคง
สถานะ ทาง การเงิน ที่ มั่นคง ของ
สถานะ ทาง การเงิน ที่ มั่นคง ของ ท่าน

Pros: Conceptualization

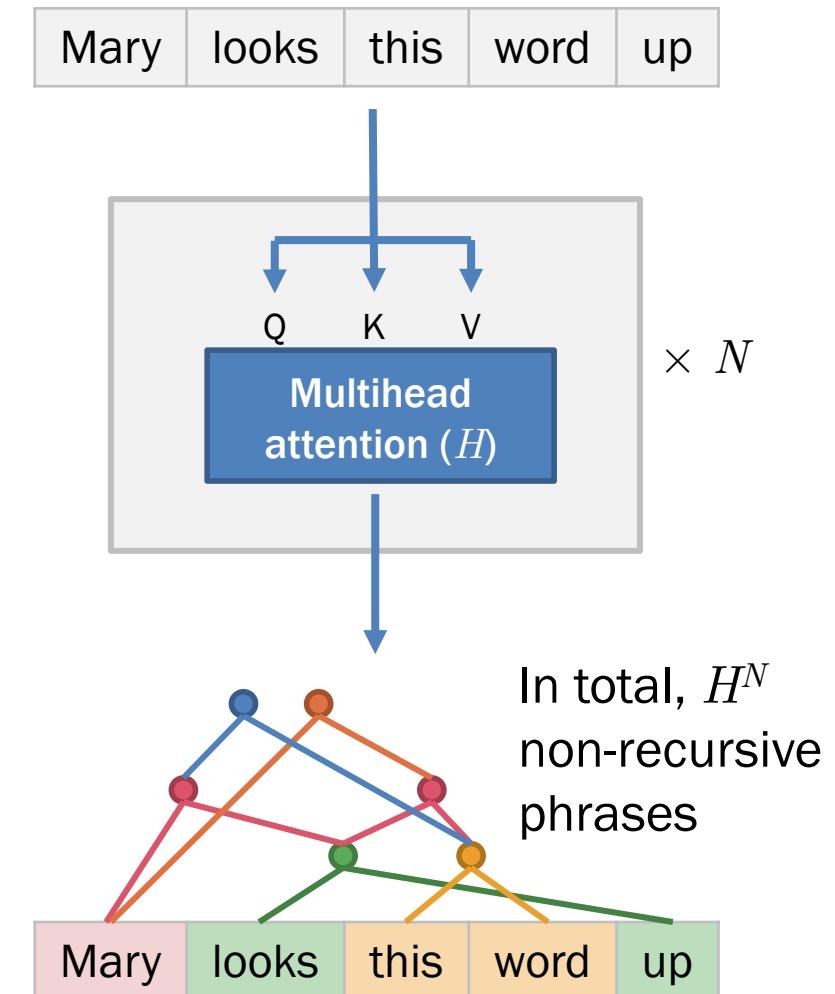
- It learns to conceptualize a long subsequence into a shorter one with semantic similarity and distance penalty
 - E.g. ‘initiated a scheme for building ____’ is conceptualized into ‘invented’ and consequently translated into ‘ประดิษฐ์’

Stevenson	initiated	a	scheme	for	building	the	first	locomotive	PREDICTION $\emptyset \Rightarrow$ สตีเวนสัน สตีเวนสัน \Rightarrow ประดิษฐ์ $\text{ประดิษฐ์} \Rightarrow$ รถจักรไอน้ำ $\text{รถจักรไอน้ำ} \Rightarrow$ คัน $\text{คัน} \Rightarrow$ แรก	สตีเวนสัน
Stevenson	initiated	a	scheme	for	building	the	first	locomotive		สตีเวนสัน
Stevenson	initiated	a	scheme	for	building	the	first	locomotive		สตีเวนสัน ประดิษฐ์
Stevenson	initiated	a	scheme	for	building	the	first	locomotive		สตีเวนสัน ประดิษฐ์ รถจักรไอน้ำ
Stevenson	initiated	a	scheme	for	building	the	first	locomotive		สตีเวนสัน ประดิษฐ์ รถจักรไอน้ำ คัน
Stevenson	initiated	a	scheme	for	building	the	first	locomotive		สตีเวนสัน ประดิษฐ์ รถจักรไอน้ำ คัน แรก

3. Theoretical Upper Bounds

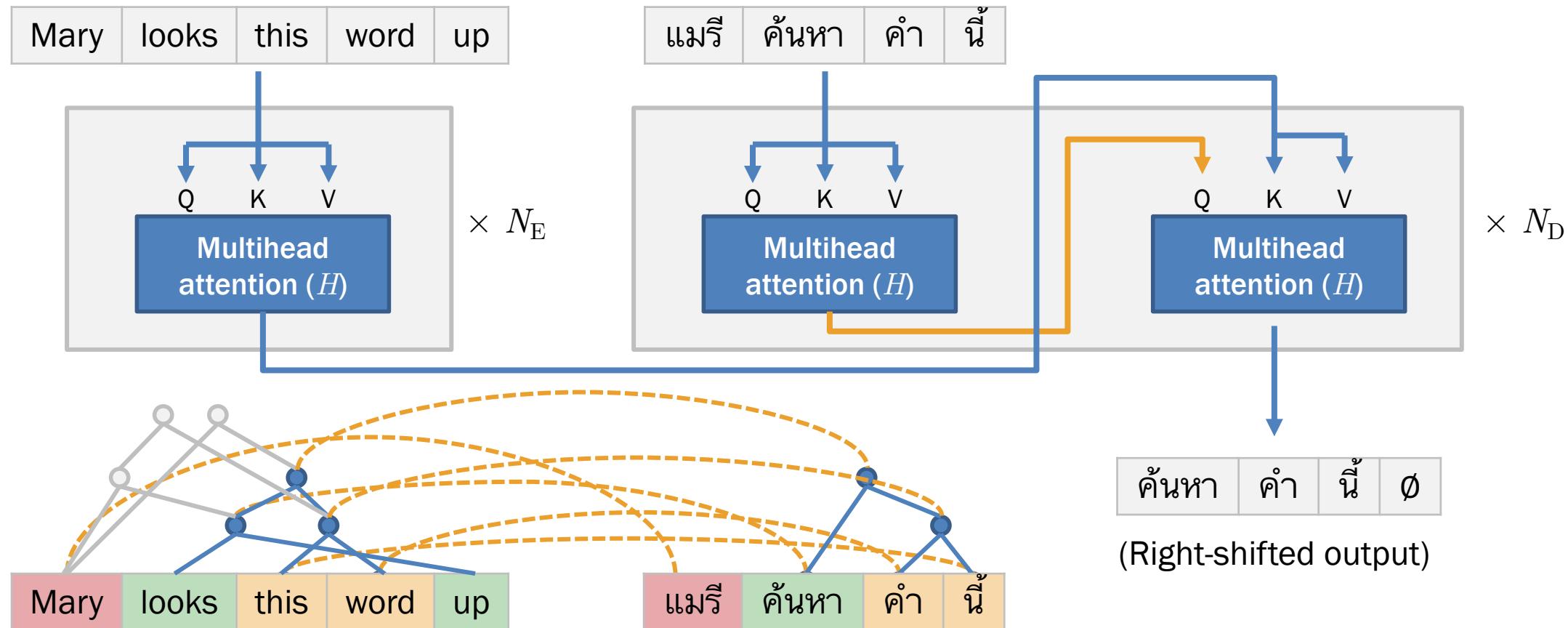
Encoding Phrase Structures

- **Limitation:** One self-attention head learns only one type of word collocation
 - H -head self-attention learns at best H types of word collocation
 - Adding one self-attention head on top of H -head self-attention helps learn a phrase structure of these H types of word collocation
 - So, adding H -head self-attention to H -head self-attention helps learn H^2 possible phrases
 - Therefore, N layers of H -head self-attention learns H^N possible non-recursive phrases
 - **Default:** $H=8, N=6 \Rightarrow 262,144$ possible phrases



Decoding Phrase Structures

- **Limitation:** Encoder-decoder learns at best $H^{N_E+N_D}$ non-recursive translation pairs
- **Default:** $H=8, N_E=6, N_D=6 \Rightarrow 6.87$ billion possible pairs



Effects of Upper Bound Violation

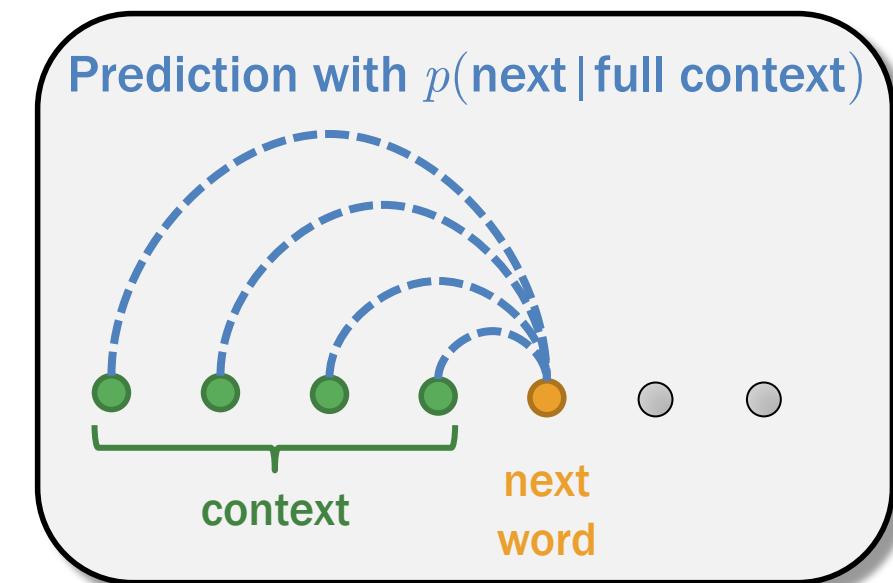
- If there are $> H^N$ phrase structures
 - Distinct phrases may be encoded as the same values in the multihead self-attention
 - **Encoding:** It causes lexical mistranslation
 - **Decoding:** It causes under-generation and over-generation
- If there are $> H^{N_E + N_D}$ translation pairs
 - Distinct translation pairs may be stored as the same pairs in the alignment attention
 - This results in phrase mistranslation, under-generation, and over-generation

4. Large Language Models

Language Models

- **Motivation**

- Statistical prediction for how strings are produced in a language
- Interpreted as a generative model
 1. Generate the first word w_1
 2. Keep generating the **next word** w_k based on the previous words (a.k.a. **context**) $w_1 \dots w_{k-1}$ until the whole sentence of length N is produced



$$P(w_1 \dots w_N) = p(w_1) \prod_{k=2}^N p(\text{next word} | \text{context})$$

next word
context

Language Models

- **Motivation:** n -gram models
 - Language models whose context is truncated to at most $n-1$ previous words

$$P(w_1 \dots w_N) = p(w_1) \prod_{k=2}^N p(w_k | \underbrace{w_{k-n+1} \dots w_{k-1}}_{n-1 \text{ prev words}})$$

- **Unigram** ($n=1$):

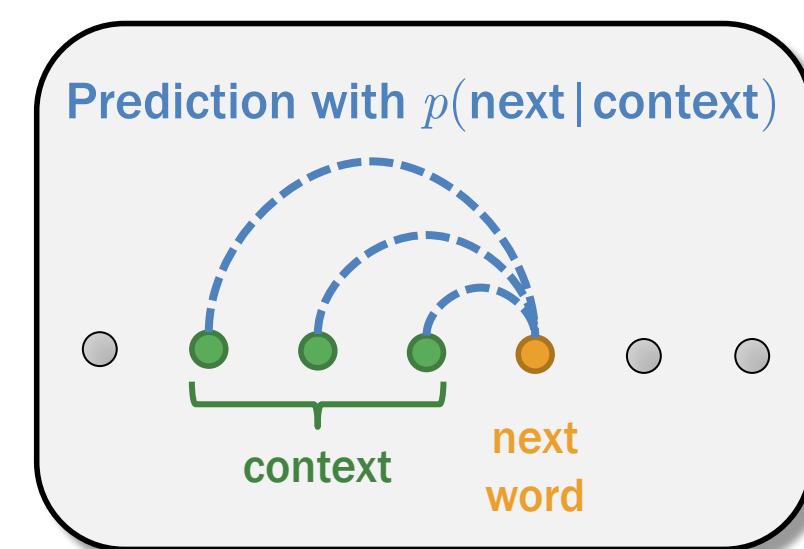
$$P(w_1 \dots w_N) = \prod_{k=1}^N p(w_k)$$

- **Bigram** ($n=2$):

$$P(w_1 \dots w_N) = p(w_1) \prod_{k=2}^N p(w_k | w_{k-1})$$

- **Trigram** ($n=3$):

$$P(w_1 \dots w_N) = p(w_1)p(w_2 | w_1) \prod_{k=3}^N p(w_k | w_{k-2}, w_{k-1})$$



BERT (Devlin et al., 2018)

- Bidirectional Encoder Representations from Transformer
 - Pretrained Transformer model with multilayer bidirectional encoders
 - Contextual representations: vector repr of each word varies by position
 - Trained on BooksCorpus (800M words) + Wikipedia (2,500M words)

	BERT base	BERT large
Encoding layers	12	24
Attention heads	12	16
Hidden dimensions	768	1,024
Parameters	110M	340M

Tokenization: Subwords instead of Words

- Thanks to large datasets, self-supervised subword tokenization replaces traditional word segmentation to tackle the variety of surface forms
 - **Byte-pair encoding (BPE):** gradually combine a pair of consecutive symbols into a new token, if its frequency exceeds a predefined threshold
 - **WordPiece:** gradually combine a pair of consecutive symbols into a new token, if it is still in the top- n most frequent list and its frequency also exceeds a predefined threshold
 - **SentencePiece:** choose a sequence of tokens for each text that maximizes the dataset's unigram likelihood
⇒ It recognizes prefixes, suffixes, and stems quite well

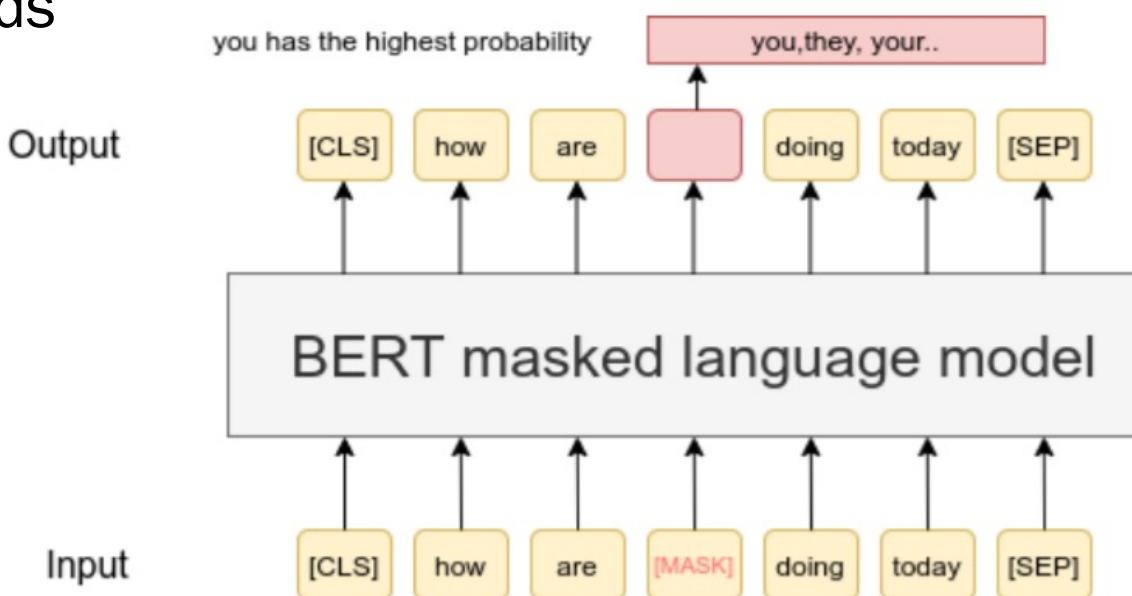
metamorphosis
⇒ m e t a m o r p h o s i s
⇒ m e t a m o r p h o s i s
⇒ met_# amor_# pho_# sis

metamorphosis
⇒ m e t a m o r p h o s i s
⇒ me ta m o r p h o s i s
⇒ meta_# mor_# pho_# sis

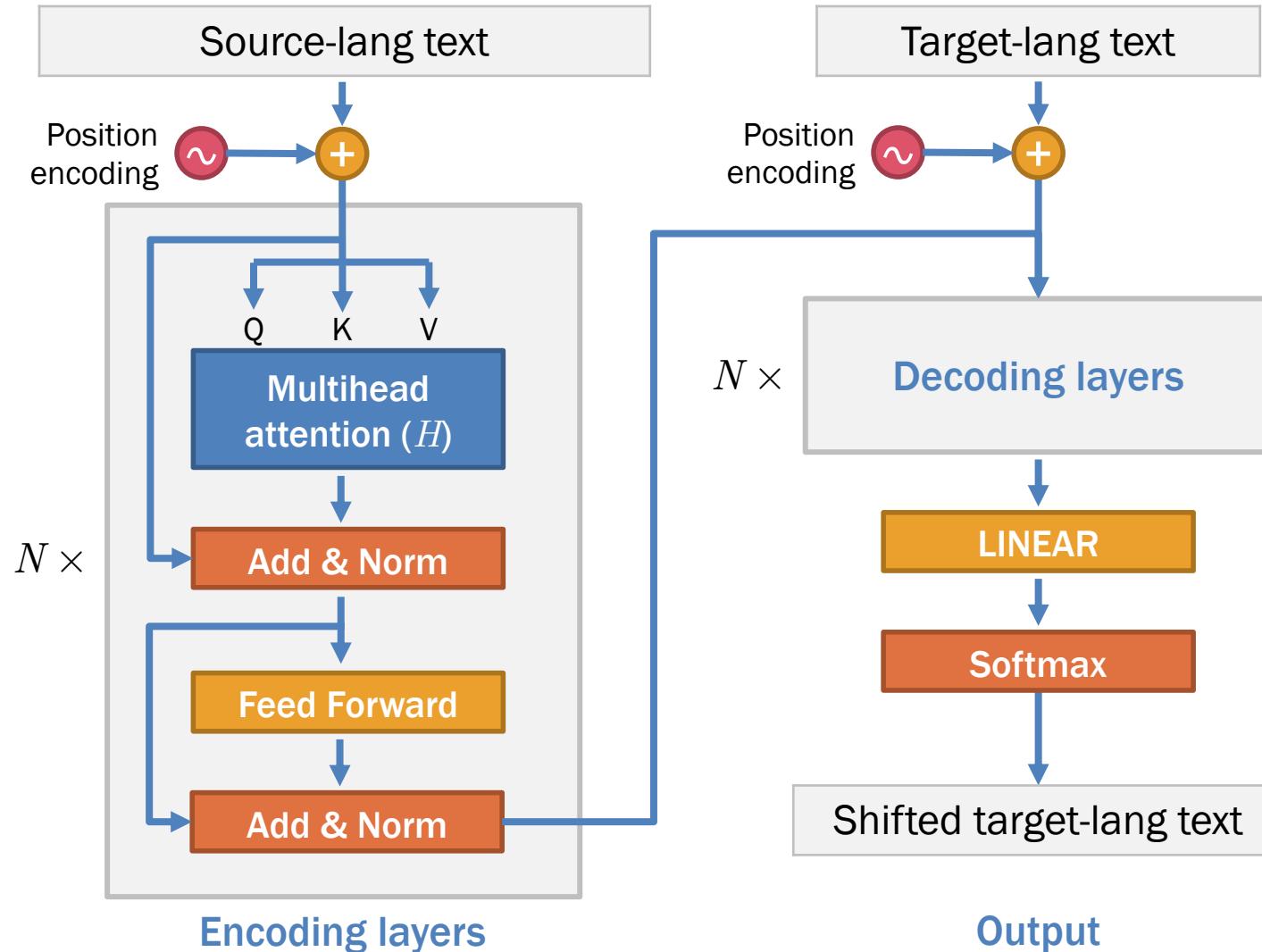
metamorphosis
≈ meta morpho sis
 or met amor phos is
⇒ meta_# morpho_# sis

Training BERT out of the Transformer

- **Masked Language Model (MLM):**
 - Each text is put on both source and target sides
 - Some words are blinded (or masked) at random
 - The model is trained to make correct prediction of masked words

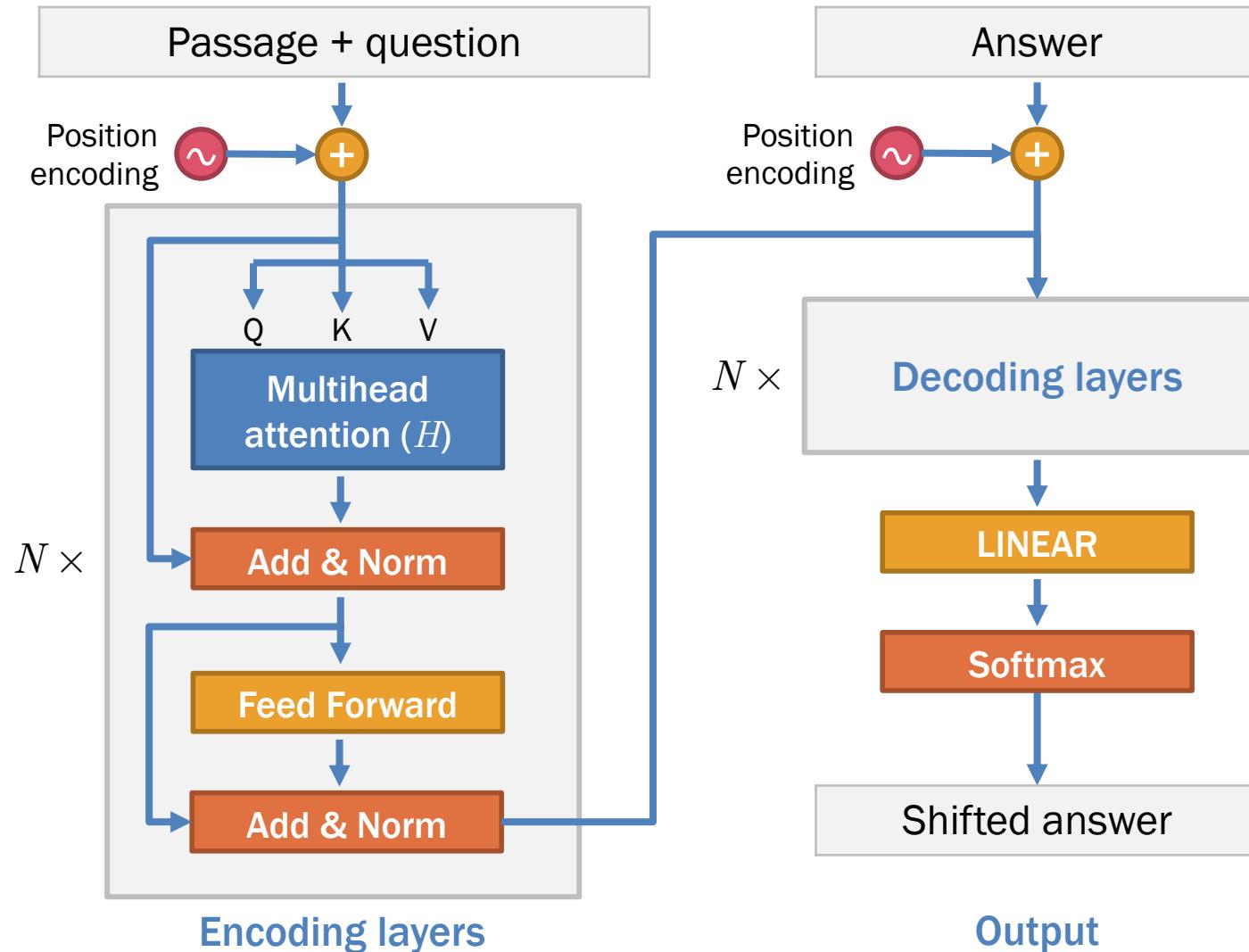


Fine-Tuning



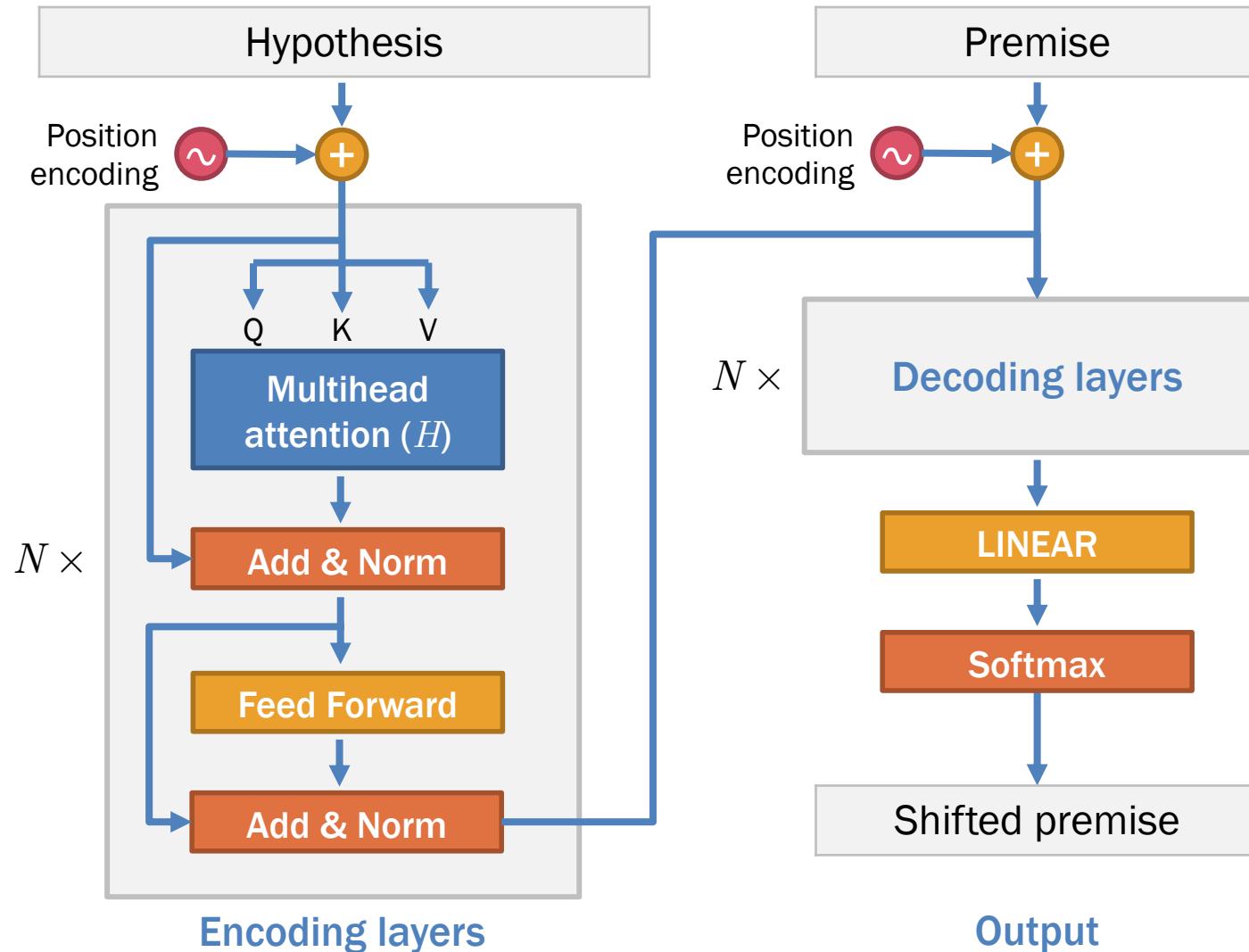
- BERT can be fine-tuned via multiple downstream tasks
 - Machine translation
 - Question answering (SQuAD)
 - Inference in natural language (NLI in GLUE Dataset)
 - Abstractive summarization

Fine-Tuning



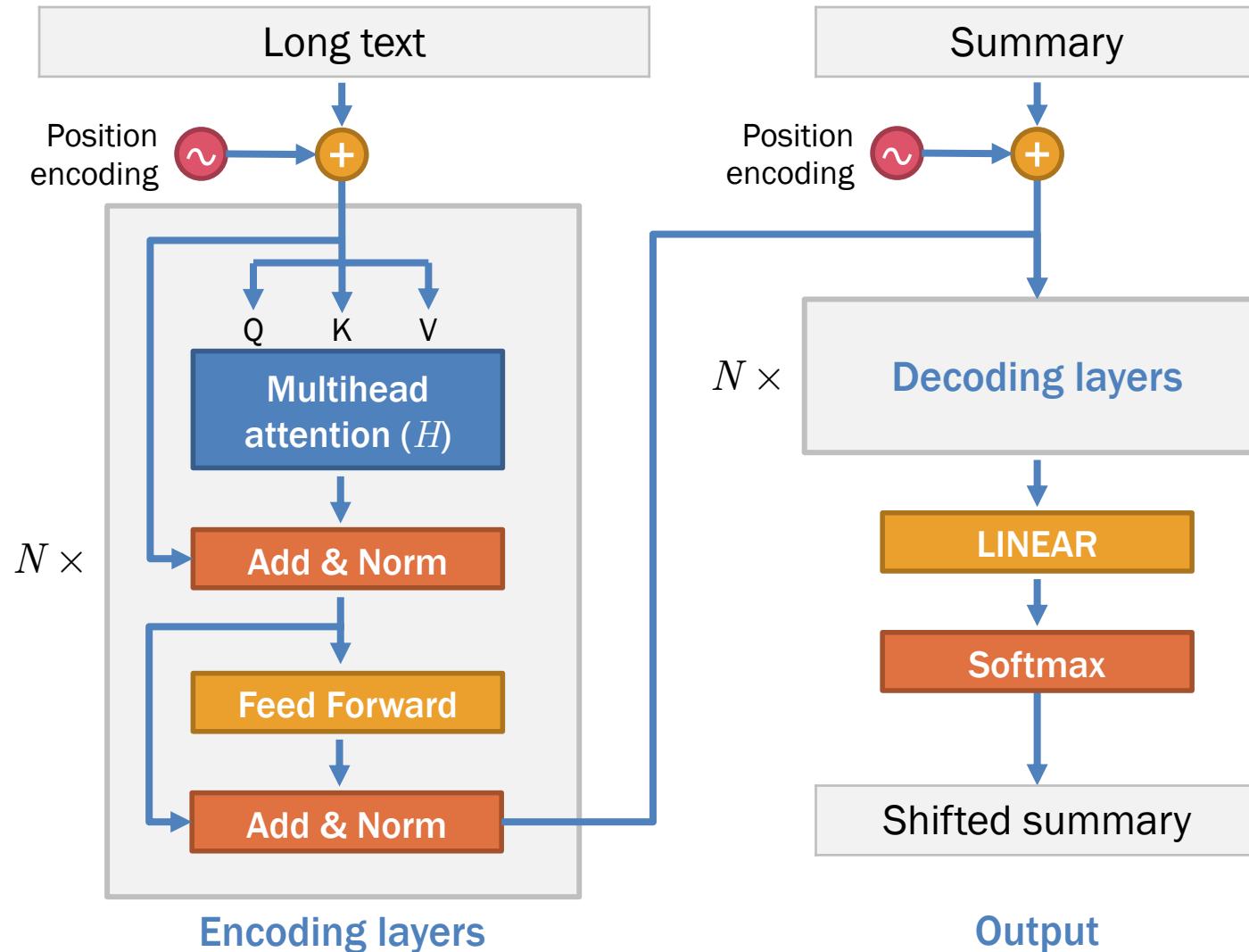
- BERT can be fine-tuned via multiple downstream tasks
 - Machine translation
 - **Question answering (SQuAD)**
 - Inference in natural language (NLI in GLUE Dataset)
 - Abstractive summarization

Fine-Tuning



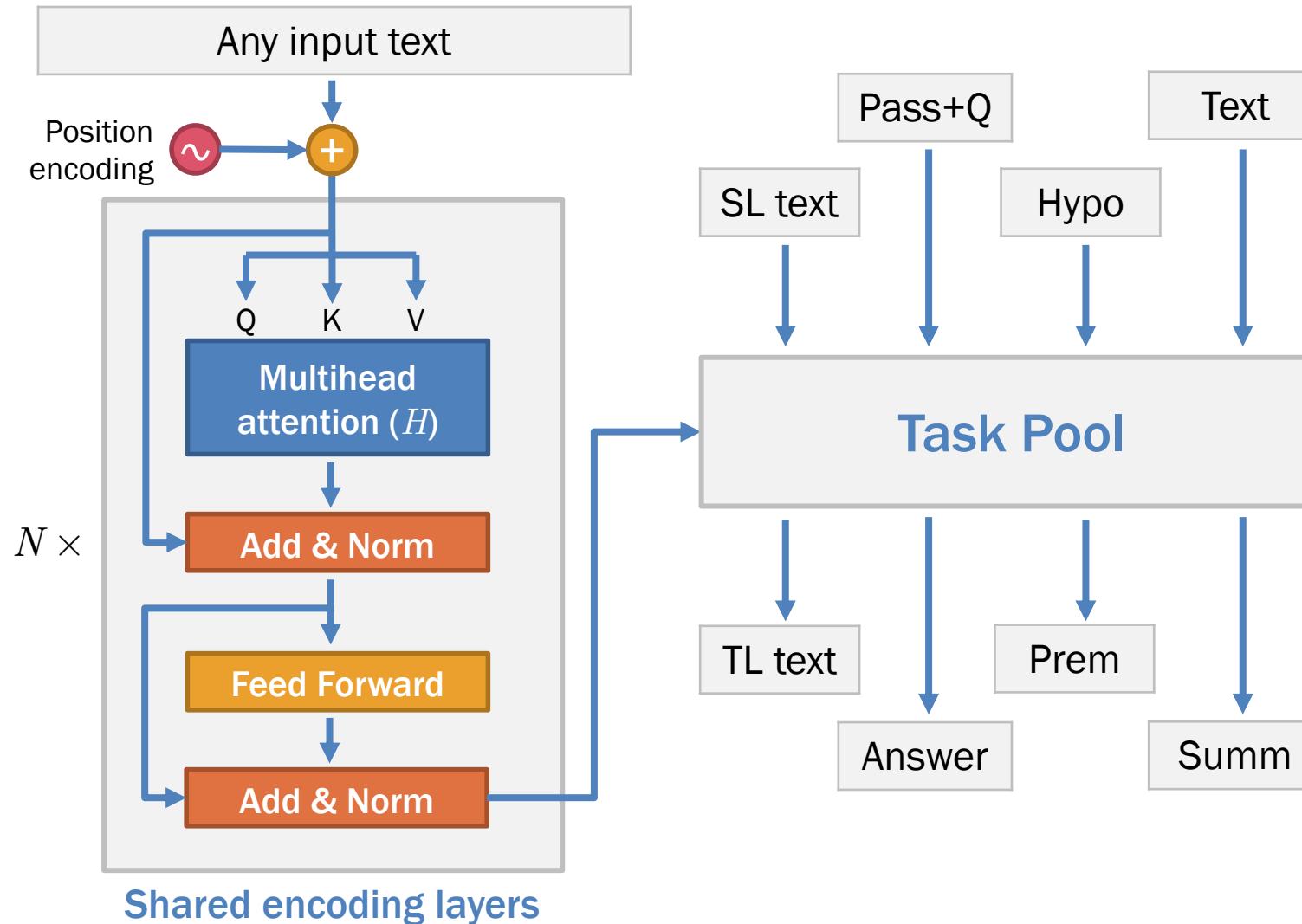
- BERT can be fine-tuned via multiple downstream tasks
 - Machine translation
 - Question answering (SQuAD)
 - Inference in natural language (NLI in GLUE Dataset)
 - Abstractive summarization

Fine-Tuning



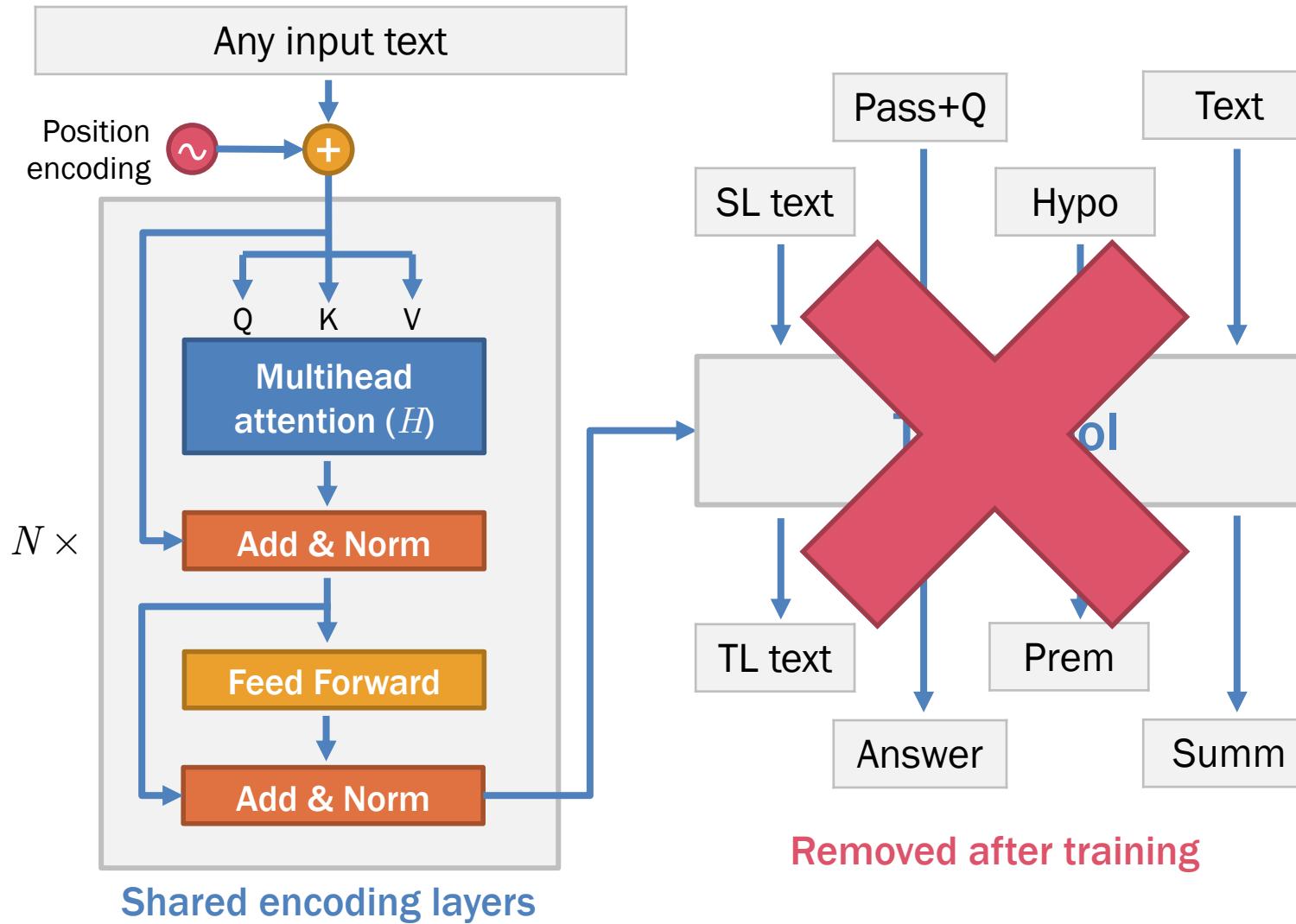
- BERT can be fine-tuned via multiple downstream tasks
 - Machine translation
 - Question answering (SQuAD)
 - Inference in natural language (NLI in GLUE Dataset)
 - **Abstractive summarization**

Fine-Tuning



- BERT can be fine-tuned via multiple downstream tasks
 - Machine translation
 - Question answering (SQuAD)
 - Inference in natural language (NLI in GLUE Dataset)
 - Abstractive summarization

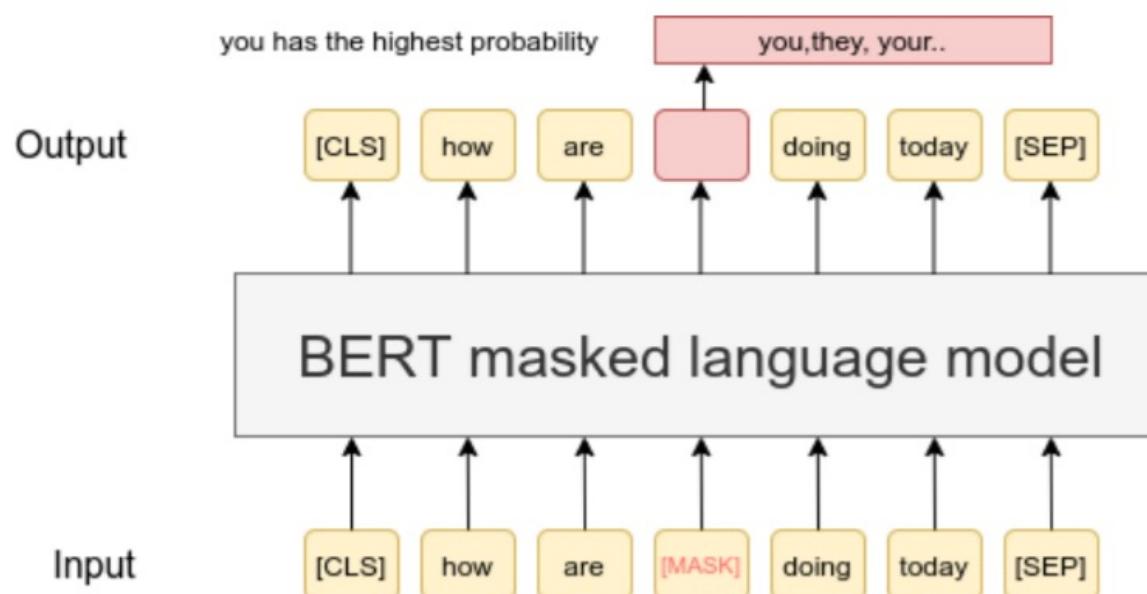
BERT = Shared Encoding Layers



- BERT can be fine-tuned via multiple downstream tasks
 - Machine translation
 - Question answering (SQuAD)
 - Inference in natural language (NLI in GLUE Dataset)
 - Abstractive summarization

Masked Language Model (MLM)

- We replace some words in the input with blanks and compute the loss of word prediction on these blanks in the output

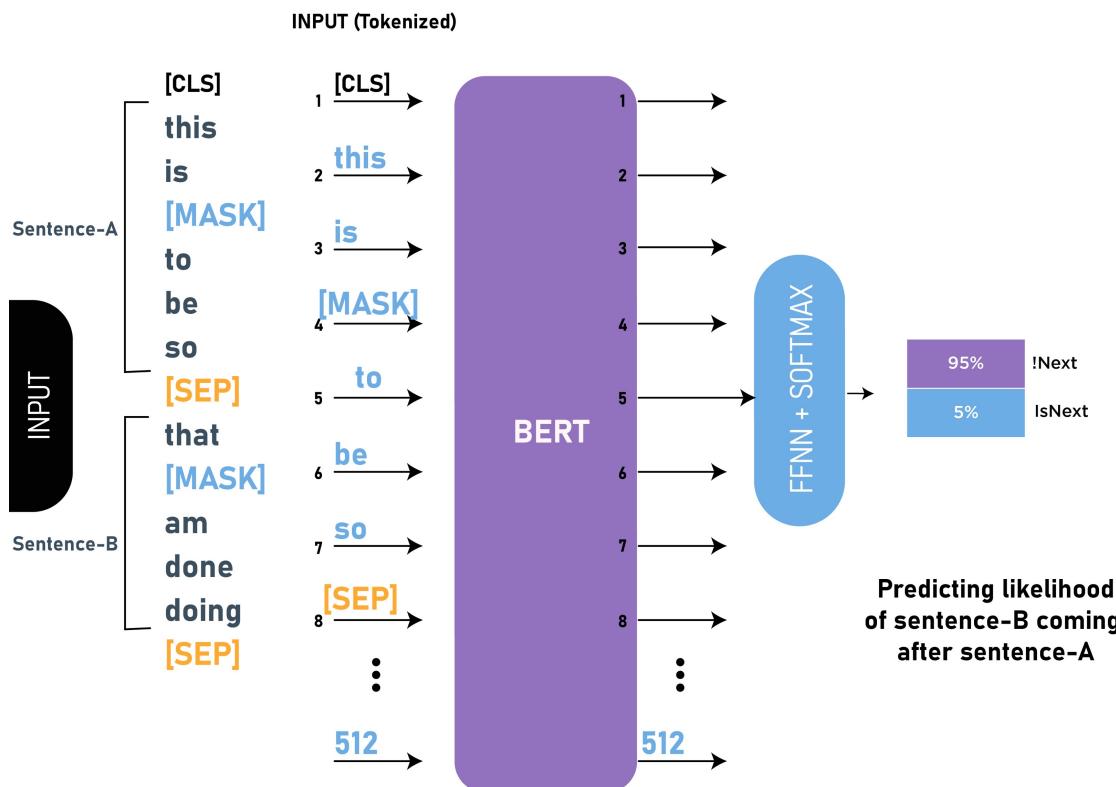


- Each input text is marked at some words by [MASK]
- Masking percentage = 15%
- Once marked, the masks will not be changed
- Special tokens
 - [CLS] = classifier token
 - [SEP] = separator token
 - [MASK] = mask

https://www.sbert.net/examples/unsupervised_learning/MLM/README.html

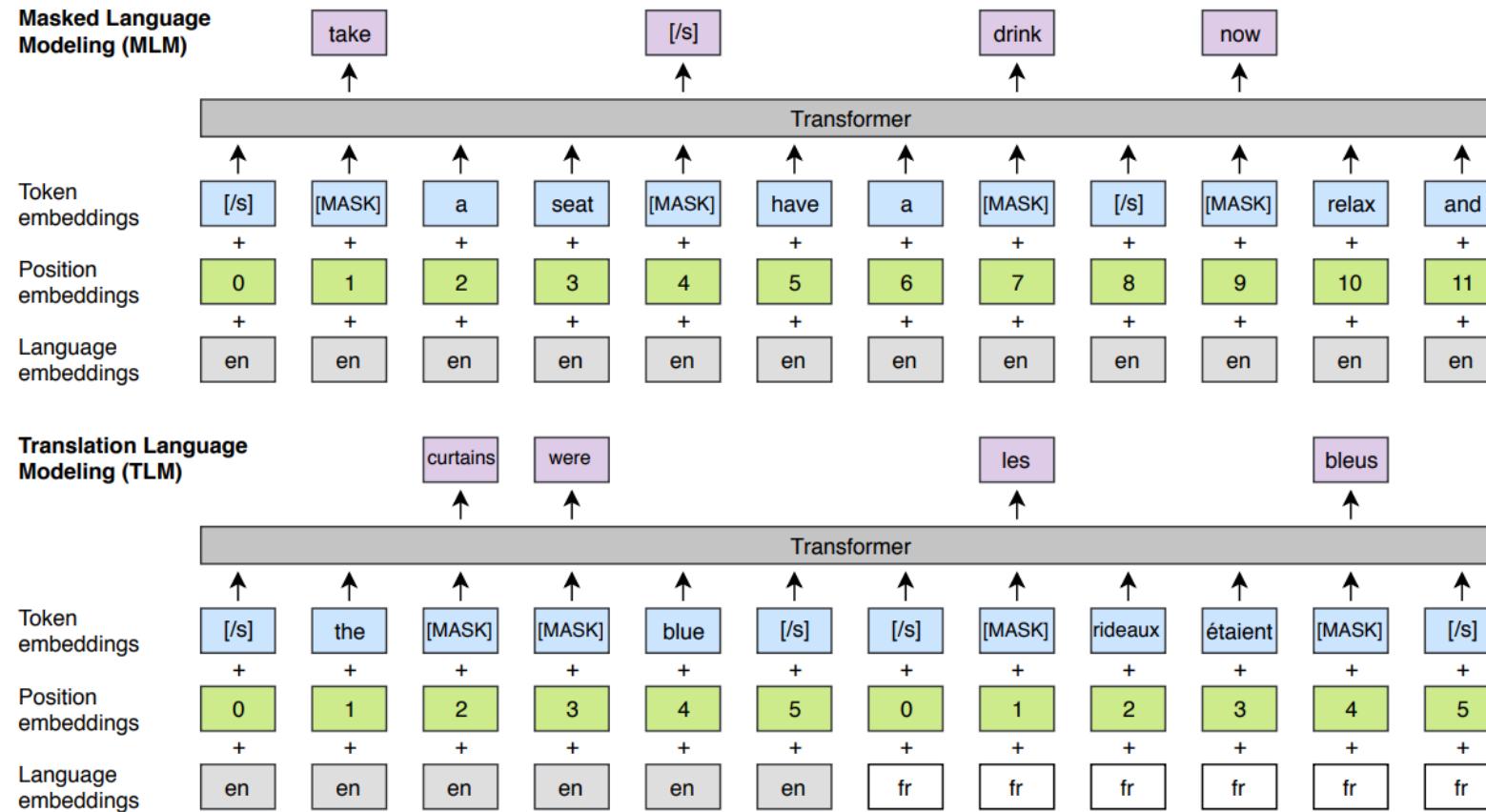
Next Sentence Prediction (NSP)

- We can concatenate two texts to let BERT learn their contextual information



- In Natural Language Inference (NLI), each pair of sentences is classified as **entailment** or not (**IsNext**)
- With NSP training, semantic relatedness is imposed into word embedding

Cross-Lingual Language Model (XLM)



- Translation pairs can also be used to train cross-lingual language model
- Some words are marked with [MASK] at random for masked language model
- Semantic relatedness can be learned from parallel corpora, especially from multitexts (multiple-language parallel texts)

Figure 1: **Cross-lingual language model pretraining.** The MLM objective is similar to the one of Devlin et al. (2018), but with continuous streams of text as opposed to sentence pairs. The TLM objective extends MLM to pairs of parallel sentences. To predict a masked English word, the model can attend to both the English sentence and its French translation, and is encouraged to align English and French representations. Position embeddings of the target sentence are reset to facilitate the alignment.

What Knowledge Does BERT Have?

- Syntactic knowledge
 - It encodes POS, idioms, and syntactic roles ([Lin et al., 2019; Tenney et al., 2016; Liu et al., 2019](#))
 - It learns hierarchical idiomatic patterns; not syntax ([Htut et al., 2019; Jawahar et al., 2019](#))
- Semantic knowledge
 - It encodes semantic roles ([Ettinger, 2019](#)) and entity types ([Tenney et al., 2019](#))
 - It still struggles with representations of numbers ([Wallace et al., 2019](#))
- World knowledge
 - It captures some commonsense knowledge ([too many citations here](#))
 - It struggles with pragmatic inference and role-based event knowledge ([Ettinger, 2019](#))
 - It cannot still reason based on learned world knowledge ([Forbes et al., 2019](#))

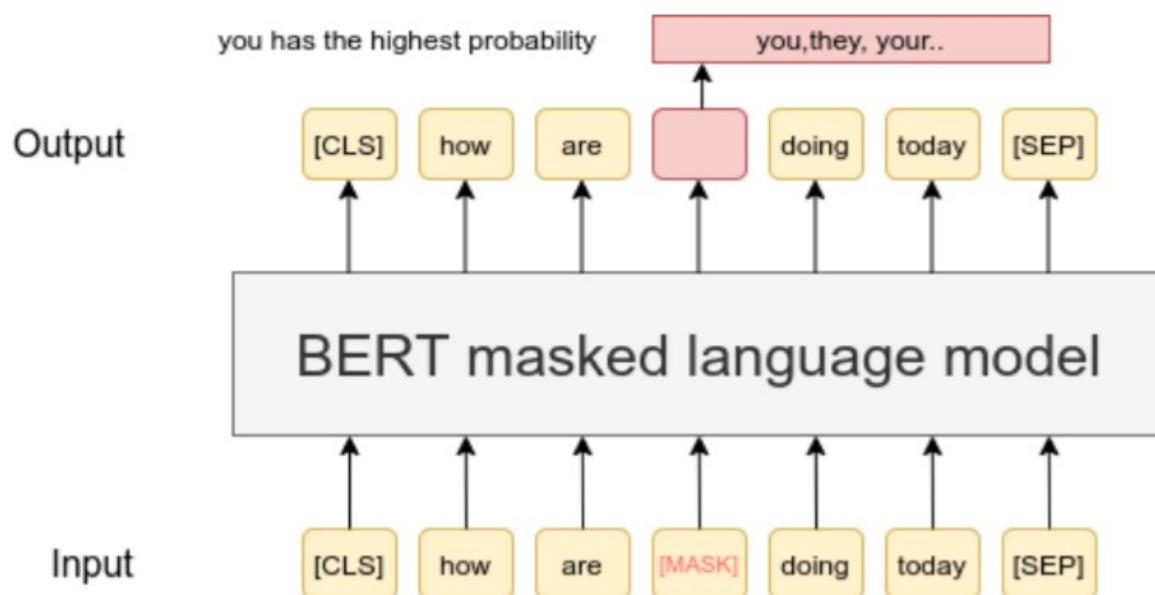
General Limitations

- Names, numerical entities, and variables are treated as separate units (similar to Word2Vec)
 - Direct effect:** Sparse distribution of word vectors
 - Generating loopy output sequences (repetition counter is needed)
 - Mingling memorized information
 - Possibly biased due to unclean training data (esp. discrimination, propaganda, and misinformation)

5. Some Popular BERT Variants

RoBERTa (Liu et al., 2019)

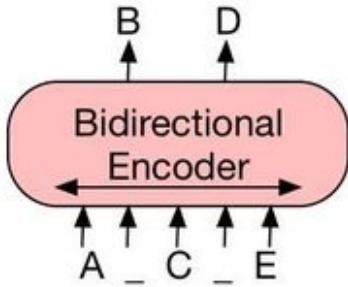
- Robustly Optimized BERT pretraining approach
 - An improved version of BERT



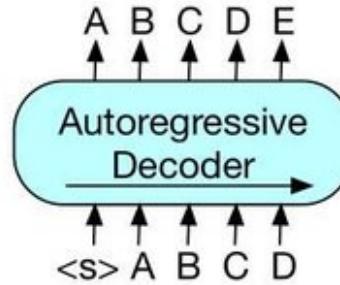
- Dynamic masking instead of static masking
- NSP task is eliminated without losing semantic relatedness
- Larger datasets are used in training than BERT (CC-News and Open WebText)

https://www.sbert.net/examples/unsupervised_learning/MLM/README.html

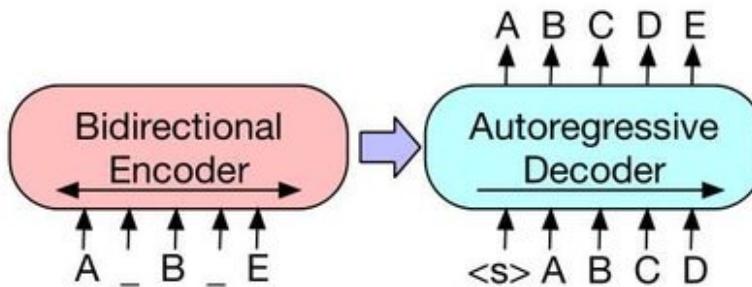
Differences of BERT, GPT, and BART (Lewis et al., 2019)



(a) BERT: Random tokens are replaced with masks, and the document is encoded bidirectionally. Missing tokens are predicted independently, so BERT cannot easily be used for generation.



(b) GPT: Tokens are predicted auto-regressively, meaning GPT can be used for generation. However words can only condition on leftward context, so it cannot learn bidirectional interactions.

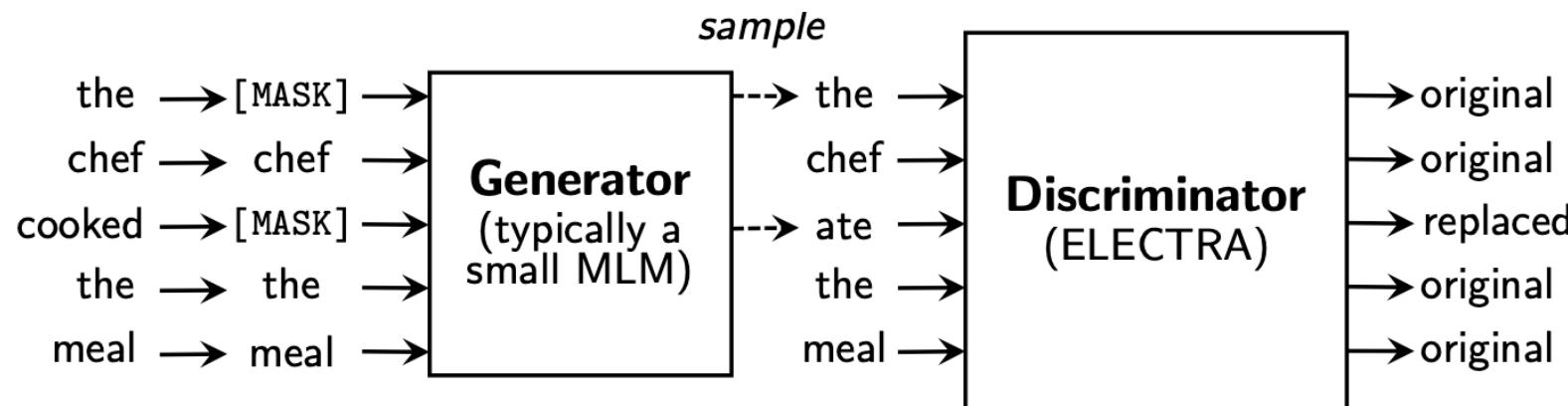


(c) BART: Inputs to the encoder need not be aligned with decoder outputs, allowing arbitrary noise transformations. Here, a document has been corrupted by replacing spans of text with a mask symbol. The corrupted document (left) is encoded with a bidirectional model, and then the likelihood of the original document (right) is calculated with an autoregressive decoder. For fine-tuning, an uncorrupted document is input to both the encoder and decoder, and we use representations from the final hidden state of the decoder.

- **BERT:**
 - Bidirectional encoder
- **GPT:**
 - Generative Pretrained Transformer
 - Autoregressive (unidirectional) decoder
- **BART:**
 - Bidirectional encoder + autoregressive decoder

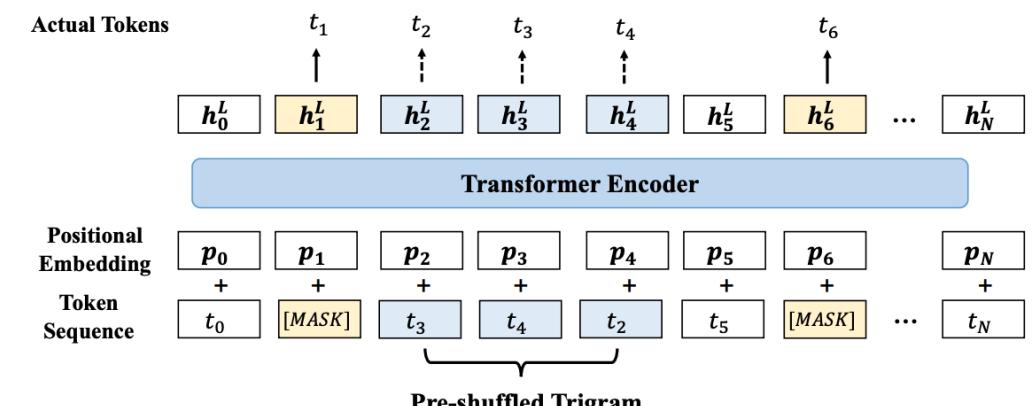
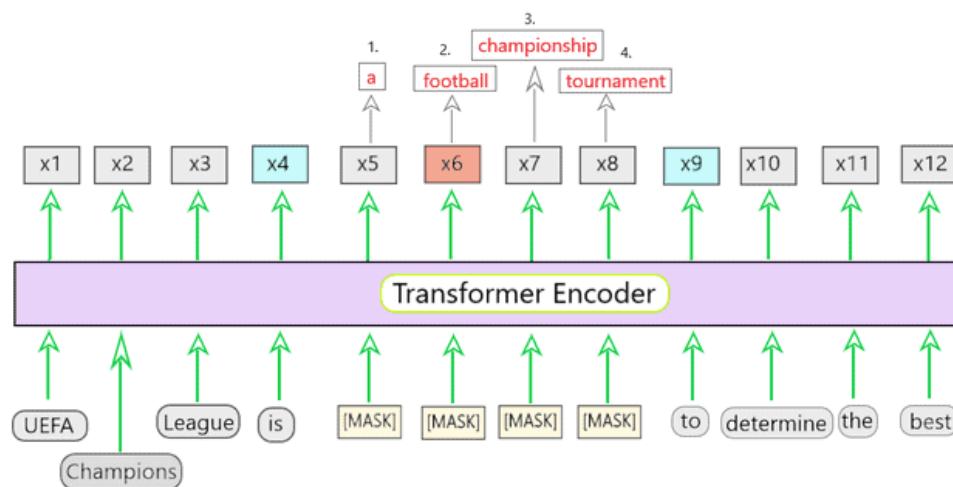
ELECTRA (Clark et al., 2020)

- Training by guessing the replaced tokens in the text
 - ELECTRA differs from BERT in that it is used as a discriminator (masking percentage = 15-20%)
 - It is trained much faster and has much less parameters
 - It is frequently used in discriminative models



SpanBERT and StructBERT

- SpanBERT (Joshi et al., 2019): guess the missing chunk
- StructBERT (Wang et al., 2020): guess the right word order



(a) Word Structural Objective

T5 and mT5 (Raffel et al., 2020)

- T5 = Text-To-Text Transformer (five T's)

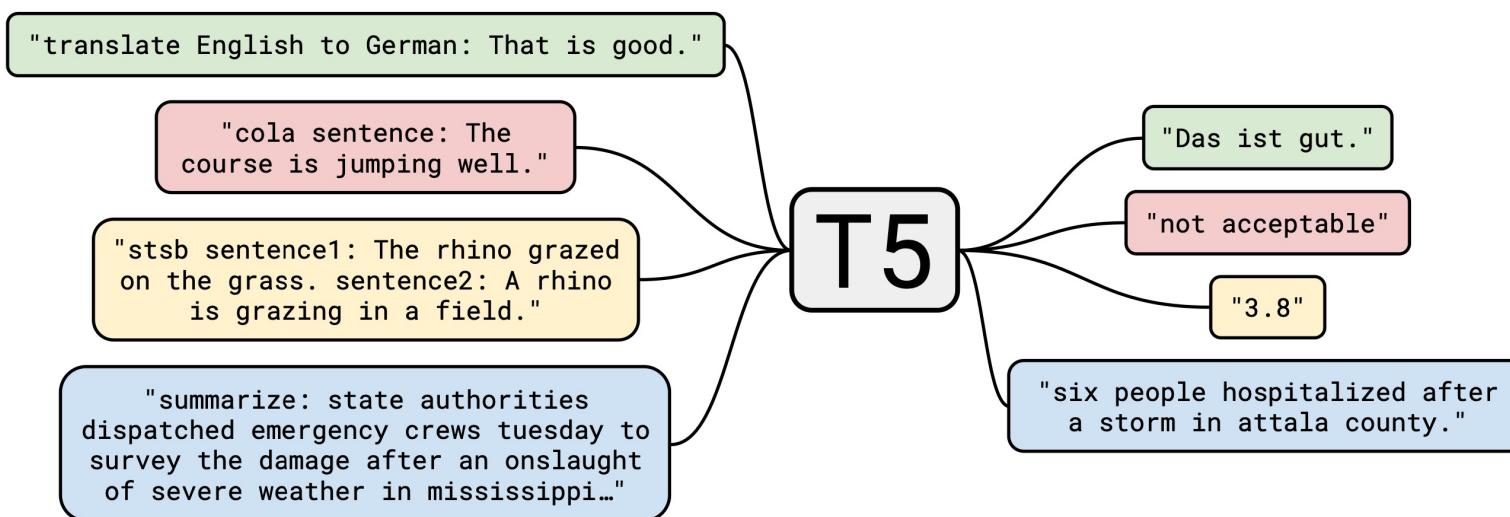
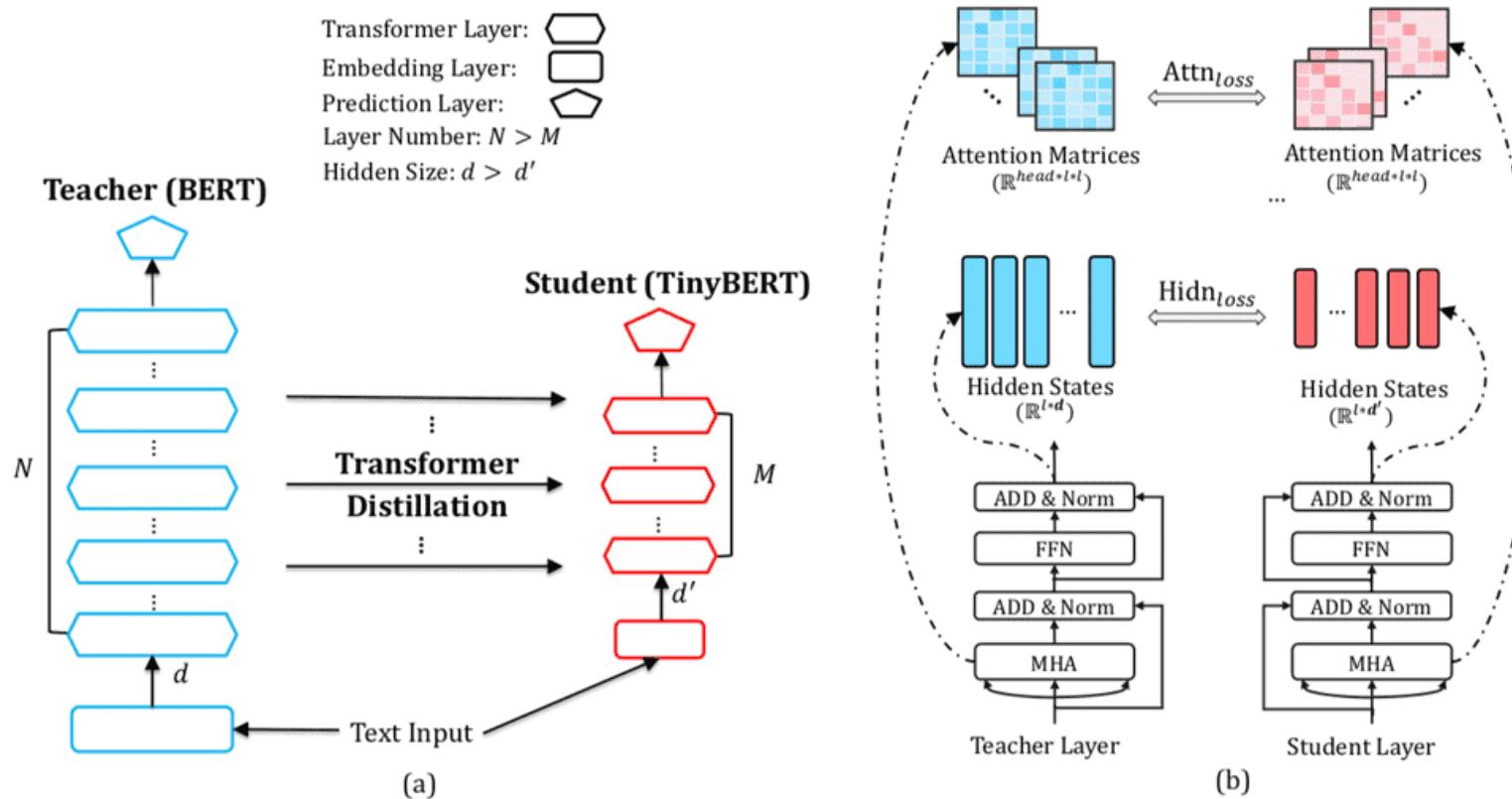


Figure 1: A diagram of our text-to-text framework. Every task we consider—including translation, question answering, and classification—is cast as feeding our model text as input and training it to generate some target text. This allows us to use the same model, loss function, hyperparameters, etc. across our diverse set of tasks. It also provides a standard testbed for the methods included in our empirical survey. “T5” refers to our model, which we dub the “Text-to-Text Transfer Transformer”.

- Every task is transformed into textual transfer
 - MT: “translate EN to DE”
 - Semantic similarity: “sim sent1: sent2:”
 - ATS: “summarize:”
- mT5 is a multilingual version of T5 model

DistilBERT (Sanh et al., 2020)

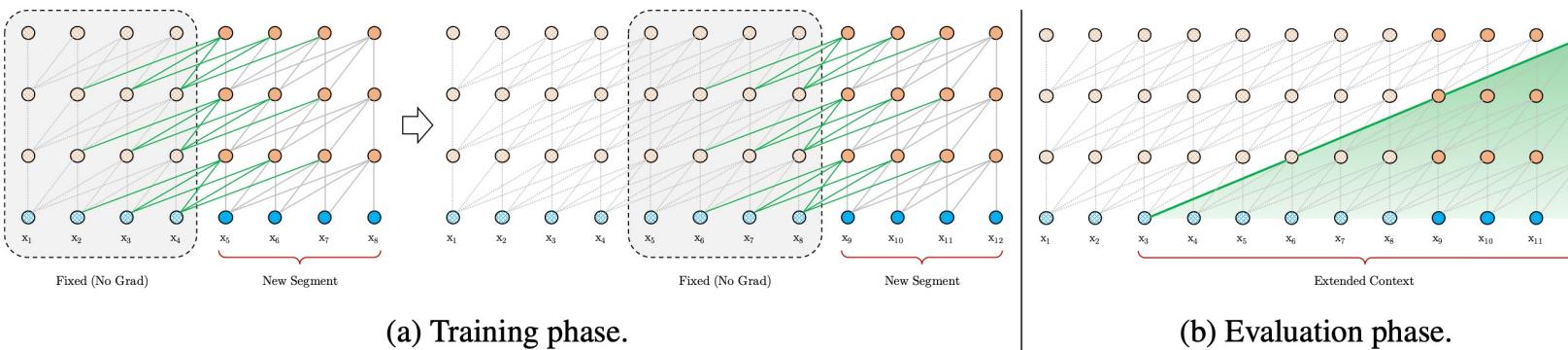
- Knowledge distillation from a very large model to a comparable, small model



- Imitating how the large model works by enforcing the losses of hidden states and attention matrices

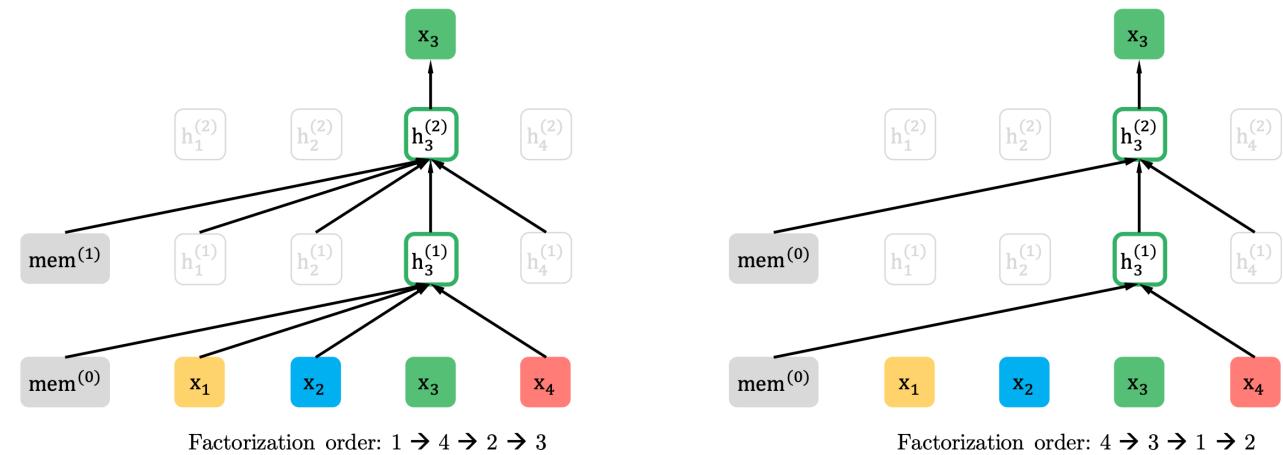
XLNet (Yang et al., 2019)

- Cope with large input sequences with Transformer-XL



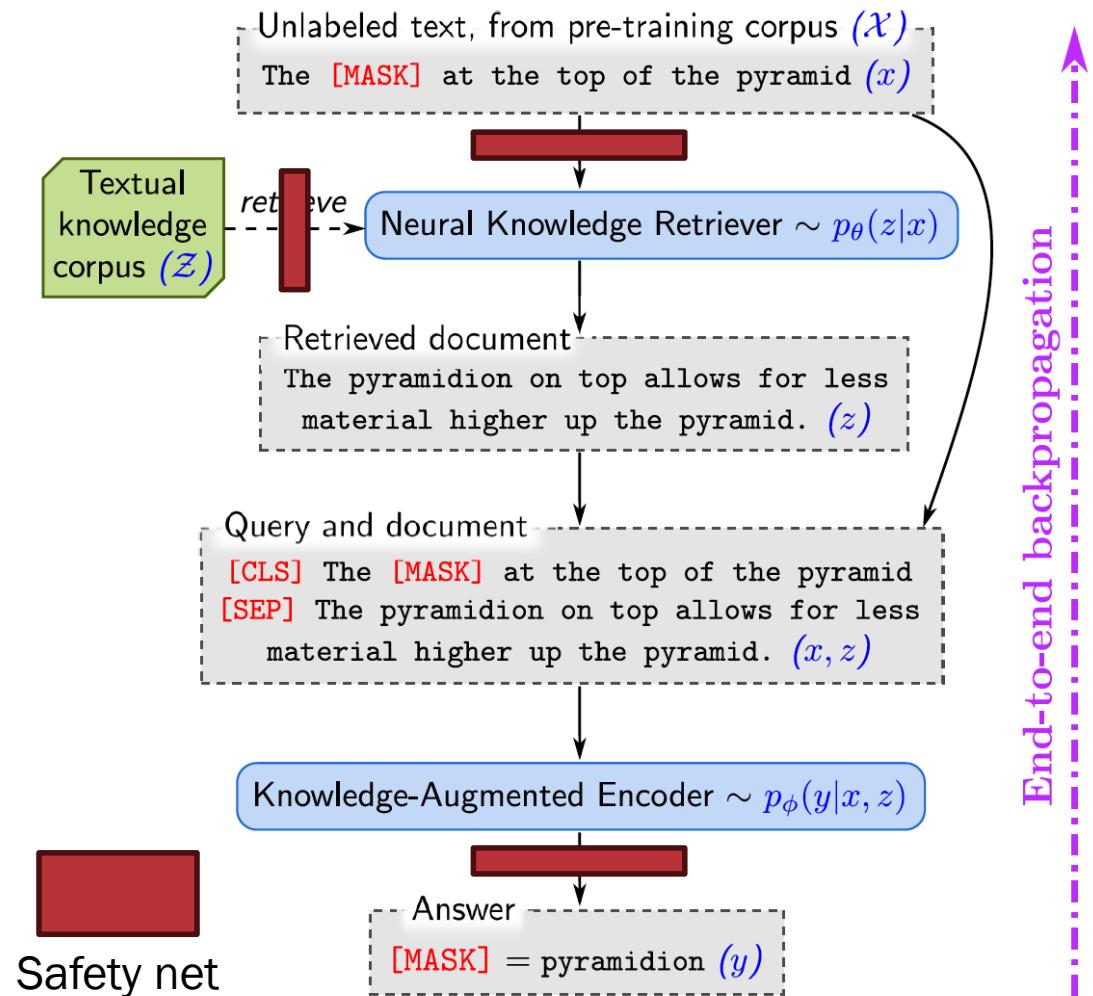
- Permutation language model

- Flexible bidirectional context
- Ex. $1 \rightarrow 4 \rightarrow 2 \rightarrow 3$ means
 $P(x_1) + P(x_4 | x_1) + P(x_2 | x_1, x_4)$
 $+ P(x_3 | x_1, x_4, x_2)$



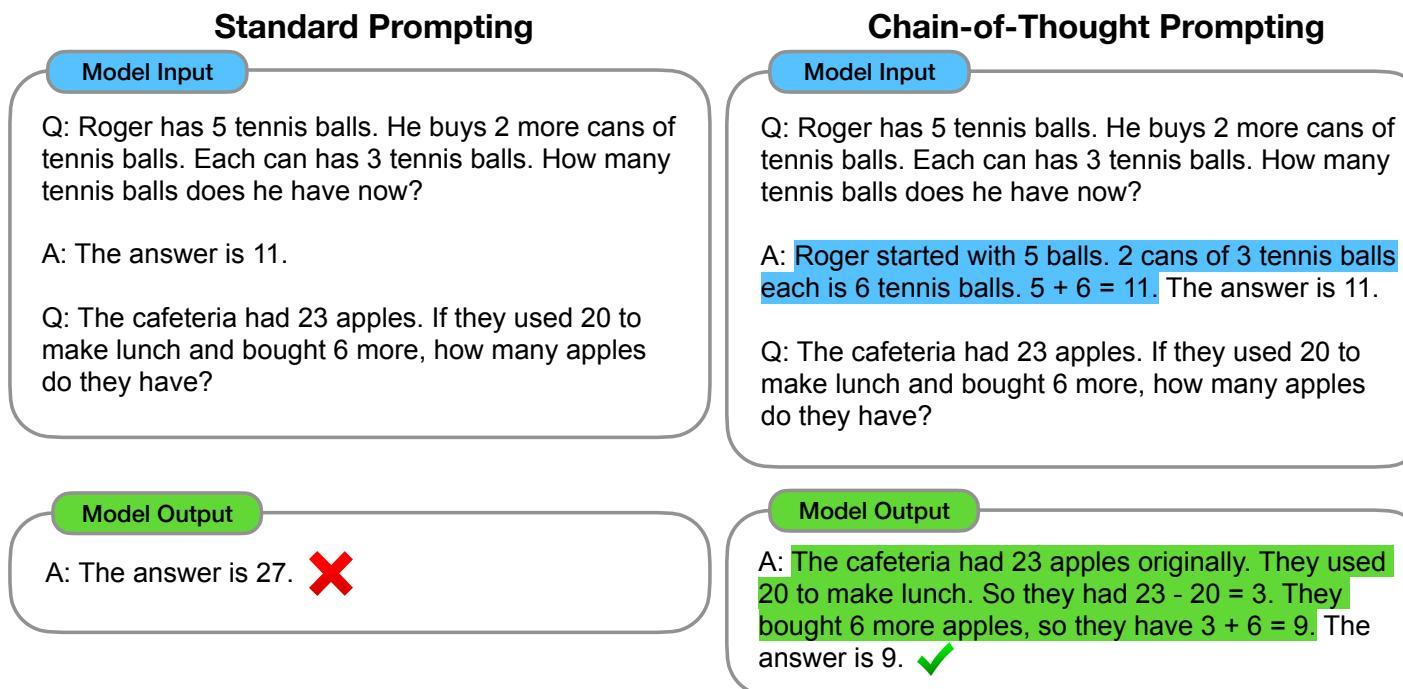
Retrieval-Augmented Generation (RAG)

- Enhancing question answering by enriching the query with relevant info
 - Mitigating the hallucination issue in traditional LLMs
 - In-context learning:** simply attaching the query with relevant documents retrieved with information retrieval
 - Vector databases (e.g. FAISS and VectorDB) are used to store internal documents on on-premise servers
 - Safety net:** content filtering



Future: PaLM (Wei et al., 2022)

- Pathways Language Model
 - 540B parameters trained with Pathways Transformer model (single model capable of doing many tasks)
 - Chain of thoughts prompting for multistep problem solving



- Chain of thoughts is added to the target sequence
- Reasoning and answering are learned in tandem
- Interpretability is introduced to Transformer

Future: PaLM (Wei et al., 2022)

- Chain of thoughts

<p>Math Word Problems (free response)</p> <p>Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?</p> <p>A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. $5 + 6 = 11$. The answer is 11.</p>	<p>Math Word Problems (multiple choice)</p> <p>Q: How many keystrokes are needed to type the numbers from 1 to 500? Answer Choices: (a) 1156 (b) 1392 (c) 1480 (d) 1562 (e) 1788</p> <p>A: There are 9 one-digit numbers from 1 to 9. There are 90 two-digit numbers from 10 to 99. There are 401 three-digit numbers from 100 to 500. $9 + 90(2) + 401(3) = 1392$. The answer is (b).</p>	<p>CSQA (commonsense)</p> <p>Q: Sammy wanted to go to where the people were. Where might he go? Options: (a) race track (b) populated areas (c) desert (d) apartment (e) roadblock</p> <p>A: The answer must be a place with a lot of people. Race tracks, desert, apartments, and roadblocks don't have a lot of people, but populated areas do. So the answer is (b).</p>
<p>StrategyQA</p> <p>Q: Yes or no: Would a pear sink in water?</p> <p>A: The density of a pear is about 0.6 g/cm^3, which is less than water. Thus, a pear would float. So the answer is no.</p>	<p>Date Understanding</p> <p>Q: The concert was scheduled to be on 06/01/1943, but was delayed by one day to today. What is the date 10 days ago in MM/DD/YYYY?</p> <p>A: One day after 06/01/1943 is 06/02/1943, so today is 06/02/1943. 10 days before today is 05/23/1943. So the answer is 05/23/1943.</p>	<p>Sports Understanding</p> <p>Q: Is the following sentence plausible? "Joao Moutinho caught the screen pass in the NFC championship."</p> <p>A: Joao Moutinho is a soccer player. The NFC championship is part of American football, not soccer. So the answer is no.</p>
<p>SayCan (Instructing a robot)</p> <p>Human: How would you bring me something that isn't a fruit?</p> <p>Explanation: the user wants something to eat that isn't a fruit. An energy bar is not a fruit, so I will bring the user an energy bar.</p> <p>Plan: 1. find(energy bar) 2. pick(energy bar) 3. find(user) 4. put(energy bar) 5. done().</p>	<p>Last Letter Concatenation</p> <p>Q: Take the last letters of the words in "Lady Gaga" and concatenate them.</p> <p>A: The last letter of "Lady" is "y". The last letter of "Gaga" is "a". Concatenating them is "ya". So the answer is ya.</p>	<p>Coin Flip (state tracking)</p> <p>Q: A coin is heads up. Maybelle flips the coin. Shalonda does not flip the coin. Is the coin still heads up?</p> <p>A: The coin was flipped by Maybelle. So the coin was flipped 1 time, which is an odd number. The coin started heads up, so after an odd number of flips, it will be tails up. So the answer is no.</p>

Future: Pathways (Barham et al., 2022)

- Asynchronous distributed architecture for large ML models

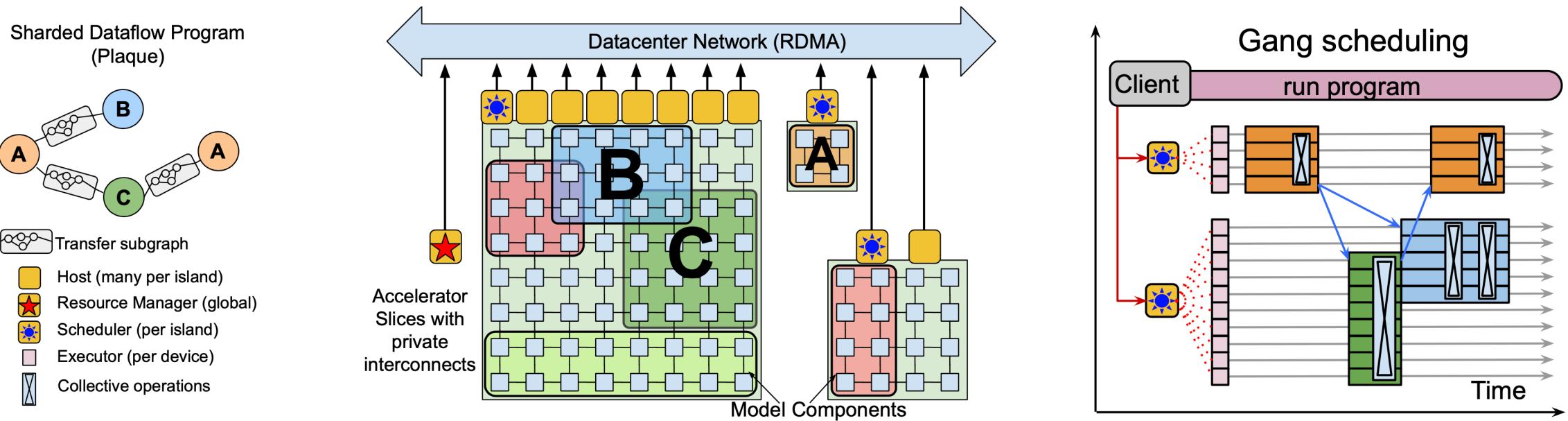
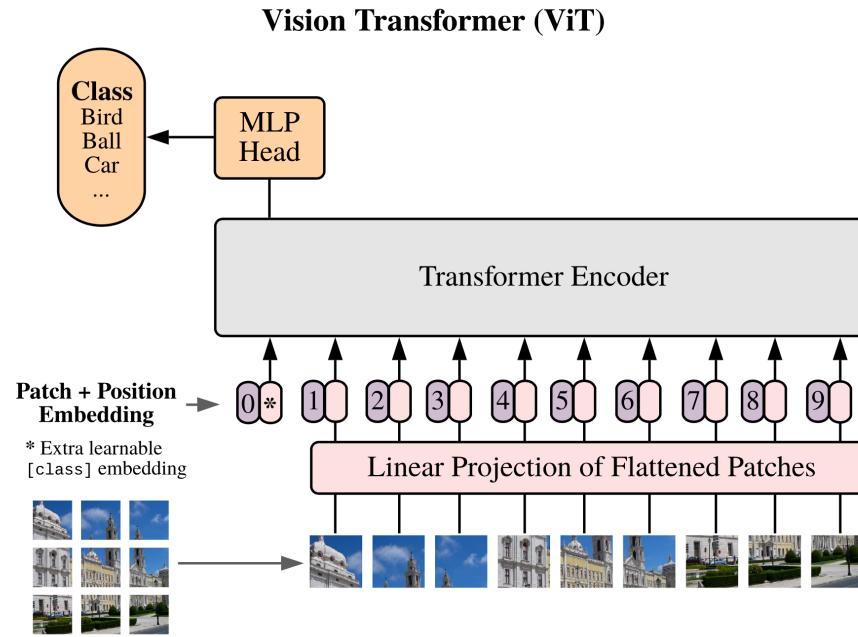
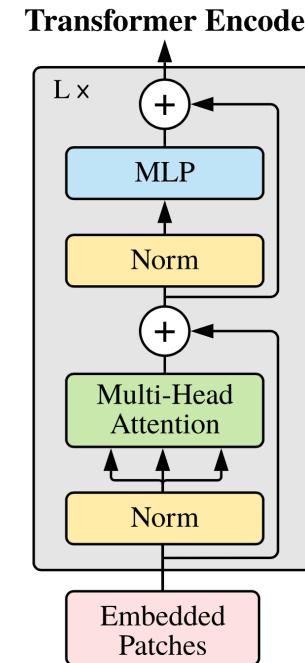


Figure 3. PATHWAYS system overview. (Left) Distributed computation expressed as a DAG where each node represents an individual compiled function, and edges between nodes represent data flows between functions. (Middle) Resource Manager allocates subsets of an island's accelerators ("virtual slices") for each compiled function. (Right) Centralized schedulers for each island gang-schedule computations that are then dispatched by per-shard executors. Red arrows indicate control messages, blue arrows show data-path transfers.

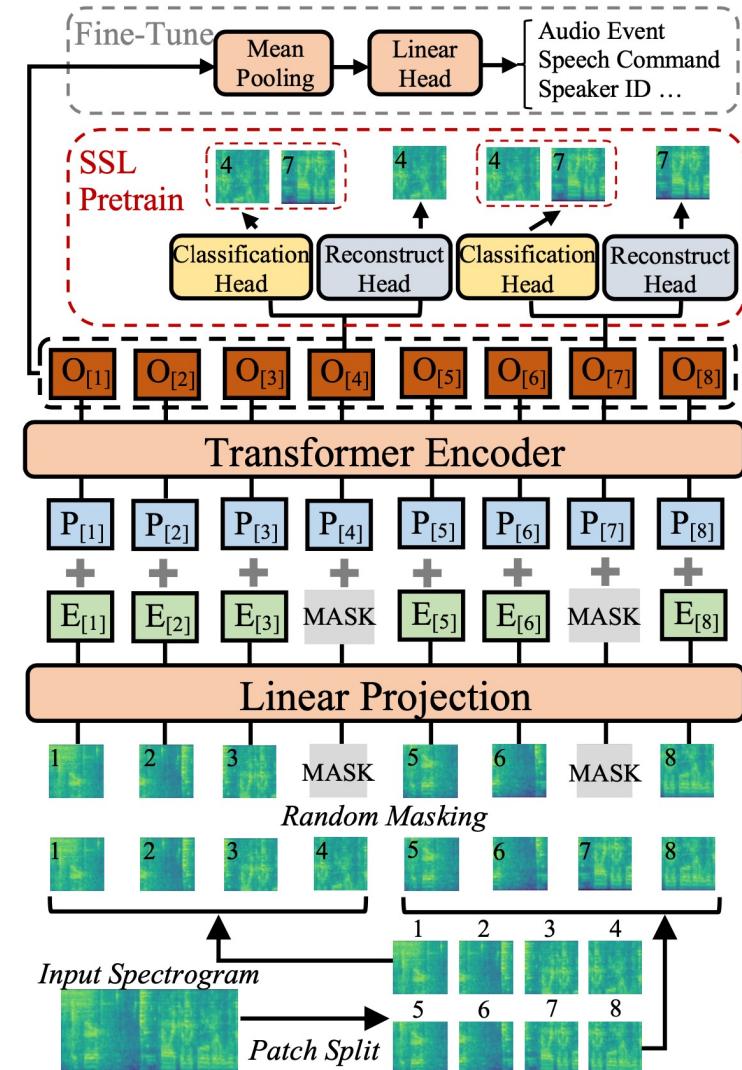
Future: Multimodal Transformer Models



Vision Transformer (Dosovitskiy et al., 2021)



Audio Transformer (Gong et al., 2022)



Future: PaLM-E (Driess et al., 2023)

- Embodied multimodal PaLM with 562B parameters

Mobile Manipulation



Human: Bring me the rice chips from the drawer. Robot: 1. Go to the drawers, 2. Open top drawer. I see 3. Pick the green rice chip bag from the drawer and place it on the counter.

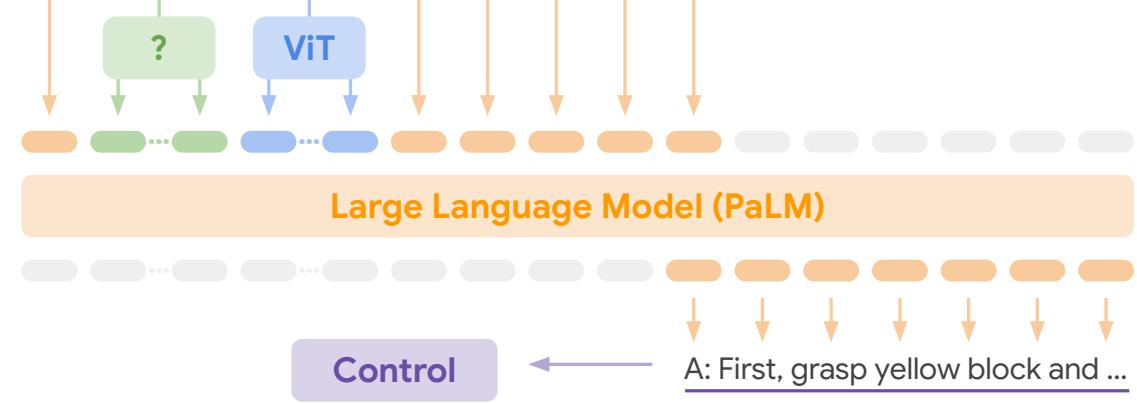
Visual Q&A, Captioning ...



Given Q: What's in the image? Answer in emojis.
A: .

PaLM-E: An Embodied Multimodal Language Model

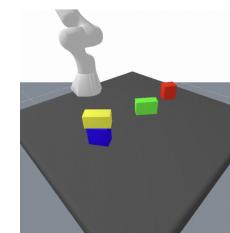
Given ... Q: How to grasp blue block? A: First, grasp yellow block



Describe the following :
A dog jumping over a hurdle at a dog show.

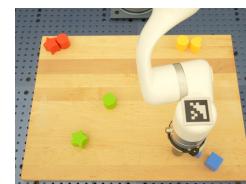


Task and Motion Planning



Given Q: How to grasp blue block?
A: First grasp yellow block and place it on the table, then grasp the blue block.

Tabletop Manipulation



Given Task: Sort colors into corners.
Step 1. Push the green star to the bottom left.
Step 2. Push the green circle to the green star.

Language Only Tasks

Here is a Haiku about embodied language models:
Embodied language models are the future of natural language

Q: Miami Beach borders which ocean? A: Atlantic.
Q: What is 372×18 ? A: 6696.
Language models trained on robot sensor data can be used to guide a robot's actions.

Future: PaLM-E (Driess et al., 2023)

Visually-conditioned Jokes + Few-shot prompting



I see (image omitted). Description: A cow sitting on a rug. Joke: Why is the cow sitting on a rug? Because it wants to be a cow-ch! ... I see . Description: a donkey is carrying a dog, cat, and rooster. Joke: what do you call a donkey with a rooster on his back. A rooster booster.

Zero-shot Multimodal Chain-of-Thought



Given . Q: Can I go down this street on a bicycle, yes or no? A: Let's think step by step. 1. do not enter. 2. except bicycles. 3. do not entry except bicycles. 4. yes.



Given . Q: How many championship rings did the player on the left win in his career? A: Let's think step by step. 1. kobe bryant is on the left. 2. kobe bryant has won 5 championship rings.



Given . Q: Who are the two teams playing in this photo, and which was the last to win a championship, and which year did they win, and who was their star player that year? A: Let's think step by step. 1. team in white is new york knick. 2. team in green is boston celtic. 3. last to win championship was boston celtic. 4. year was 2008. 5. star player was paul pierce. 6. i am not sure if this is correct. i am not basketball fan. i just google it.

Robot Visual Perception, Dialogue, and Planning



Robot: I am a robot operating in a kitchen. Given , what do I see? a bowl of eggs, a bowl of flour, and a bowl of sugar.. and a spatula.. and a spoon.. and a bag of flour.. and a bag of sugar. When a human asks me to do a task, I will respond with the sequence of actions I would do to accomplish the task with only the items I see. Human: Use all of the ingredients you see to make a cake batter. 1. crack egg. 2. put egg in bowl. 3. put flour in bowl. 4. put sugar in bowl. 5. mix. 6. put in pan. 7. bake. 8. eat. 9. clean up.

Physical Prediction



Given . Q: If a robot wanted to be useful here, what steps should it take? A: clean table..pick up trash.. pick up chairs.. wipe chairs.. put chairs down.



Spatial Grounding



Given . Q: What flavor is the doughnut on the right? Strawberry.. Q: What flavor is the doughnut on the left? A blueberry.

Photo 1: <img1>. Photo 2: <img2>. Q: What is in Photo 1 but NOT in Photo 2? A: sunglasses.

Photo 1: <img1>. Photo 2: <img2>. Q: What is in Photo 1 but NOT in Photo 2? A: Let's think step by step and explain. Photo 1 has sunglasses on top of folded clothes. Photo 2 does not have sunglasses on top of folded clothes. Therefore, sunglasses are in Photo 1 but not in Photo 2.

End-to-End Egocentric Q&A, Dialogue



Photo 1, at 10:30 am: <img1>. Photo 2, at 12:45 pm: <img2>. Photo 3, at 3:45 pm: <img3>. Q: I forget, what did I have for lunch, and what time was it? A: Let's think step by step. 1. you had a sandwich for lunch. 2. it was 12:45 pm.

OCR-free Math Reasoning



Given . Q: I am getting just two custom pizza for me and my friend. How much should I pay in total? A: Let's think step by step. 1st step: 2 custom pizzas. 2nd step: 2 x \$9.99. 3rd step: \$19.98.

6. ChatGPT

Instruction Dataset (Ouyang et al., 2022)

Table 1: Distribution of use case categories from our API prompt dataset.

Use-case	(%)
Generation	45.6%
Open QA	12.4%
Brainstorming	11.2%
Chat	8.4%
Rewrite	6.6%
Summarization	4.2%
Classification	3.5%
Other	3.5%
Closed QA	2.6%
Extract	1.9%

Table 2: Illustrative prompts from our API prompt dataset. These are fictional examples inspired by real usage—see more examples in Appendix A.2.1.

Use-case	Prompt
Brainstorming	List five ideas for how to regain enthusiasm for my career
Generation	Write a short story where a bear goes to the beach, makes friends with a seal, and then returns home.
Rewrite	This is the summary of a Broadway play: """ {summary} """ This is the outline of the commercial for that play: """

Use Case	Example
brainstorming	List five ideas for how to regain enthusiasm for my career
brainstorming	What are some key points I should know when studying Ancient Greece?
brainstorming	What are 4 questions a user might have after reading the instruction manual for a trash compactor? {user manual}
	1.

- **Prompt: “instruction”**
- Cleverly designed set of instructions and responses for a chatbot
- Covering frequently asked questions and their answers

User Prompts and Chats (Ouyang et al., 2022)

closed qa	Answer the following question: What shape is the earth? A) A circle B) A sphere C) An ellipse D) A plane
closed qa	Tell me how hydrogen and helium are different, using the following facts: {list of facts}
open qa	I am a highly intelligent question answering bot. If you ask me a question that is rooted in truth, I will give you the answer. If you ask me a question that is nonsense, trickery, or has no clear answer, I will respond with "Unknown". Q: What is human life expectancy in the United States? A: Human life expectancy in the United States is 78 years. Q: Who was president of the United States in 1955? A:
open qa	Who built the statue of liberty?
open qa	How do you take the derivative of the sin function?
open qa	who are the indigenous people of New Zealand?

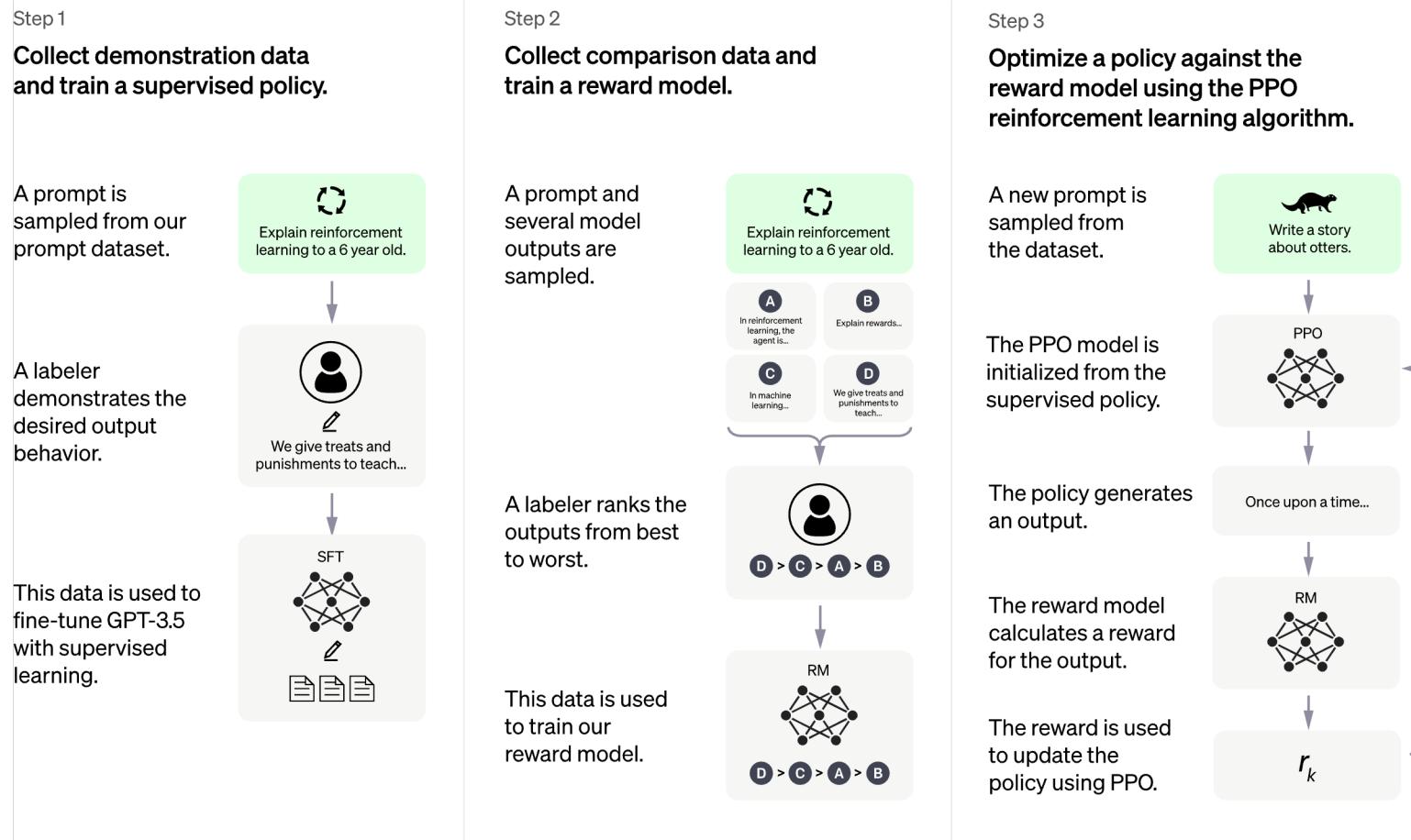
User Prompts and Chats (Ouyang et al., 2022)

Use Case	Example
chat	<p>The following is a conversation with an AI assistant. The assistant is helpful, creative, clever, and very friendly.</p> <p>Human: Hello, who are you? AI: I am an AI created by OpenAI. How can I help you today? Human: I'd like to cancel my subscription. AI:</p>
chat	<p>Marv is a chatbot that reluctantly answers questions with sarcastic responses:</p> <p>You: How many pounds are in a kilogram? Marv: This again? There are 2.2 pounds in a kilogram. Please make a note of this. You: What does HTML stand for? Marv: Was Google too busy? Hypertext Markup Language. The T is for try to ask better questions in the future. You: When did the first airplane fly? Marv:</p>
chat	<p>This is a conversation with an enlightened Buddha. Every response is full of wisdom and love.</p> <p>Me: How can I achieve greater peace and equanimity? Buddha:</p>

User Prompts and Chats (Ouyang et al., 2022)

Use Case	Example
classification	<p>The following is a list of companies and the categories they fall into:</p> <p>Apple, Facebook, Fedex</p> <p>Apple Category: Technology</p> <p>Facebook Category: Social Media</p> <p>Fedex Category:</p>
extract	<p>Text: {text}</p> <p>Keywords:</p>
generation	"Hey, what are you doing there?" Casey was startled. He hadn't even begun to
generation	The name of the next Star Wars movie is
generation	This is the research for an essay: ==== {description of research} ==== Write a high school essay on these topics: ====

Prompts + Human Ranking + RL



- 3 steps
 - Fine-tune the language model with the instruction dataset
 - Retrain the reward model for chat response with human ranking
 - Optimize the policy model w.r.t. the reward model with the PPO Algorithm (proximal policy optimization)

7. Business Opportunities in AI

Generative AI

- Algorithms (such as ChatGPT) that can be used to create new content, including audio, code, images, text, simulations, and videos
- Popular generative AI models
 - ChatGPT for text (OpenAI)
 - Stable Diffusion (MidJourney) and DeepFake for image and video
 - VALL-E for voice (Microsoft)

Krisp.ai

krisp Product Solutions ▾ Pricing Blog For Contact Centers Sign in Get Krisp for Free

Speak with confidence in business calls

Krisp's AI removes background voices, noises and echo from all your calls, giving you peace of mind.

Get Krisp for Free Watch Video

✓ No credit card needed
✓ Free forever



- Krisp.ai removes background voices, noises, and echoes from your conference calls with AI
- <https://krisp.ai>

Beatoven.ai

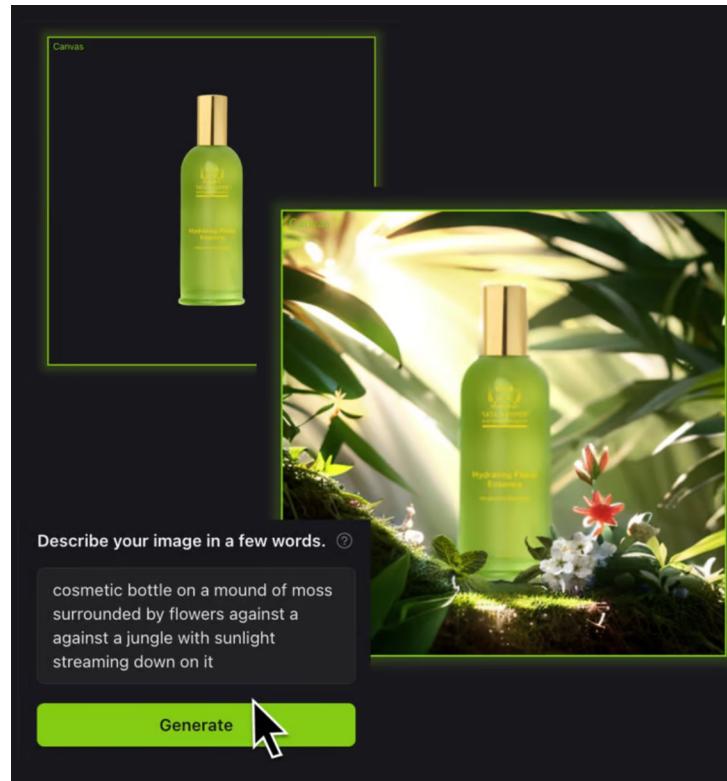
The screenshot shows the Beatoven.ai website's user interface. At the top, there is a navigation bar with links for 'How It Works', 'Pricing', 'Artists', 'About us', 'Blog', 'Sign in', and a green 'Sign up' button. Below the navigation, the page title 'How it works' is displayed, followed by the sub-headline 'Compose unique music for your content in a few easy steps'. The main content is organized into four numbered steps:

- 1. Pick a Genre/Style**: An image of a person playing a traditional Indian instrument (tabla) is shown next to the text: 'Upload a video/podcast or start with a track, choose from 8 different Genres to suit your theme.'
- 2. Make some cuts**: An image of a red waveform with a pair of scissors icon is shown next to the text: 'We know that the mood of the content changes over time. That's why you can make multiple cuts to add different moods.'
- 3. Change Mood**: An image of two overlapping cards, one orange and one green, with a wavy pattern is shown next to the text: 'Now, choose from a rich selection of 16 moods to choose the correct mood for the cuts.'
- 4. Compose**: An image of a purple circular progress bar with '80%' written on it is shown next to the text: 'Hit compose and let our AI do all the hard work of composing a unique track for you.'

At the bottom of the page is a green button labeled 'Start Creating for Free'.

- Beatoven.ai creates loyalty-free, customizable, and unique musics
- <https://www.beatoven.ai>

Flair.ai



Describe your image in a few words. ?

cosmetic bottle on a mound of moss surrounded by flowers against a jungle with sunlight streaming down on it

Generate 

| How Flair Works

Get started with just a few clicks

- 01 Drag your product photos into the Canvas.
- 02 Visually describe the scene surrounding your product.
- 03 Edit, export, and share!

- Flair.ai generates an advertisement from your idea draft and visual elements
- <https://flair.ai>

Illustroke.com



Stunning vector illustrations from text prompts

Create something unique with ours **text to SVG** AI tool.

A smiling face of an old man

Prompt tips

Generate SVG

- Illustroke.com converts your text prompt into a vector illustration (SVG)
- <https://illustroke.com>

Copy.ai

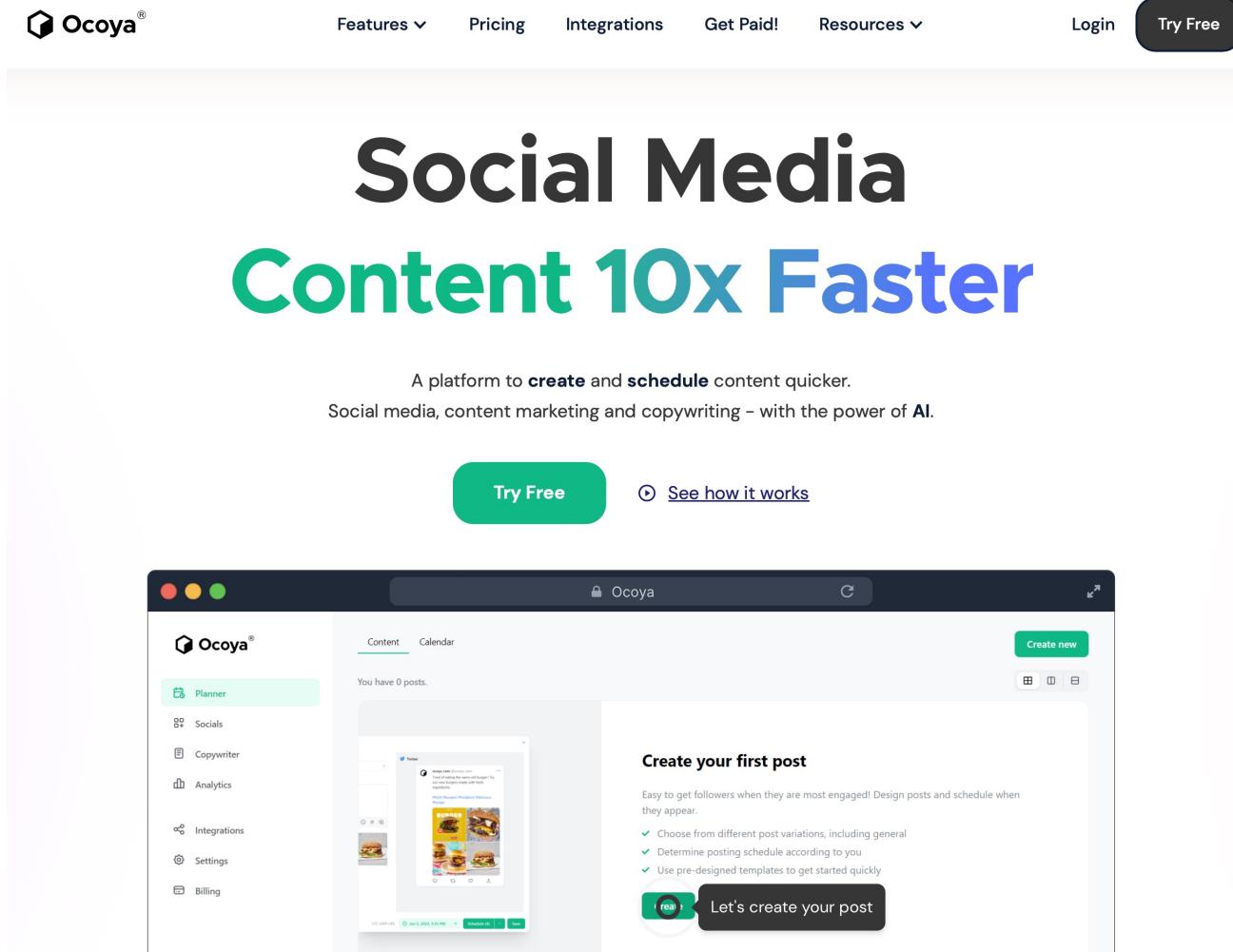
The screenshot shows the Copy.ai website. At the top, there is a navigation bar with links for 'Teams', 'Use Cases', 'Resources', and 'Pricing'. On the right side of the navigation bar are 'Login' and a teal-colored 'Get Started — It's Free' button. Below the navigation bar, the main heading 'How it works' is displayed in a large, bold, dark blue font. To the left of the heading, there is a form titled 'Freestyle'. The form asks 'What are you looking to create?' and has a text input field containing 'a Memorial Day sale email'. A cursor is visible over this input field. Below this, there is a question 'What are the main points you want to cover?' followed by a list of options: '[product details]', '[fun facts]', and '[any instructions]'. Underneath that, there is a section 'Choose a tone' with a dropdown menu showing 'Friendly' as the selected option. At the bottom of the form is a teal-colored 'Create Copy' button.

1 **Enter your copywriting project**

Choose from emails, social posts, long-form blog posts, and more!

- Copy.ai generates a text content based on your context (main points) and tone
- <https://www.copy.ai>

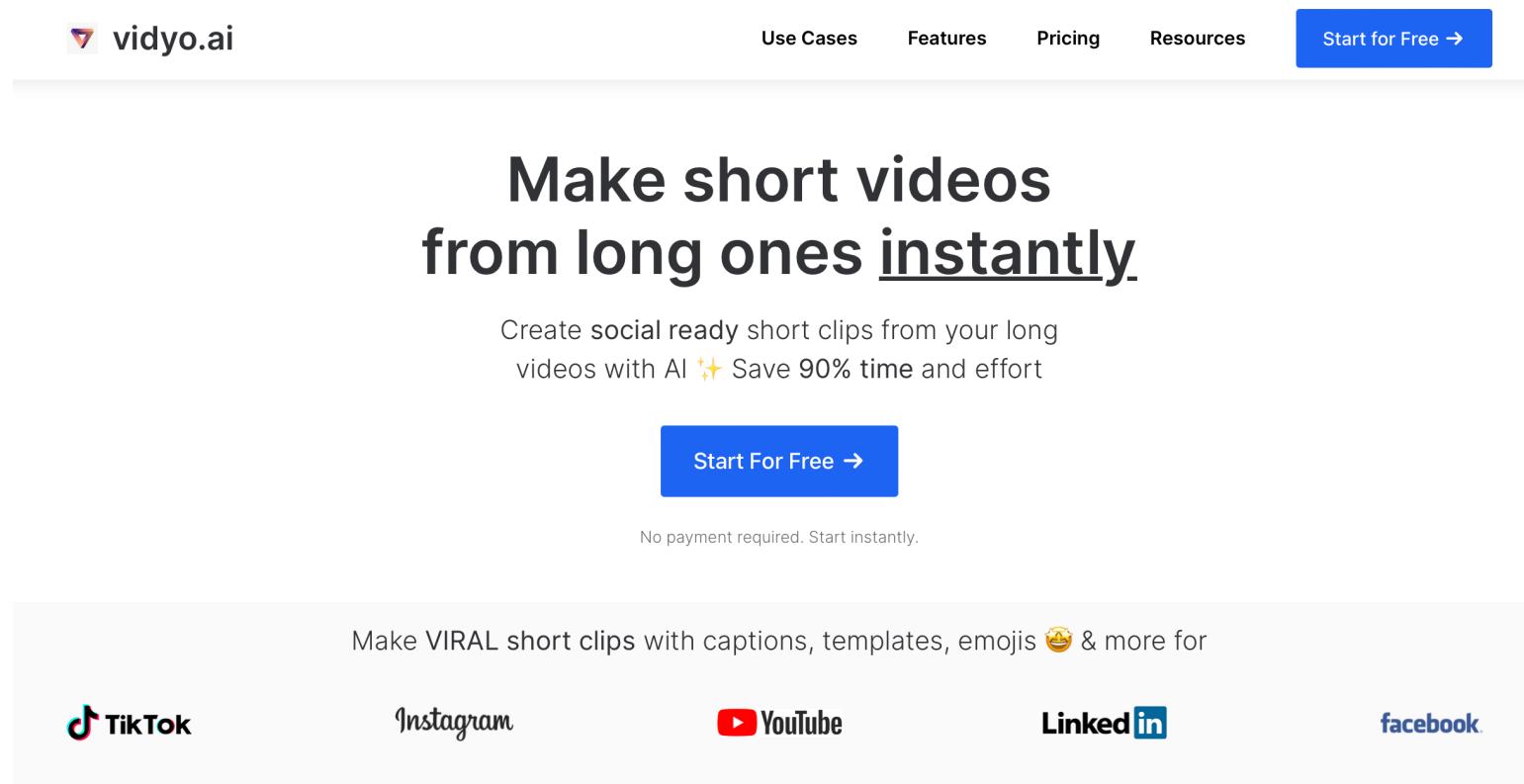
Ocoya.com



The image shows the Ocoya.com homepage and a screenshot of the platform's interface. The homepage features a large banner with the text "Social Media Content 10x Faster" in black and teal. Below the banner, it says "A platform to **create** and **schedule** content quicker. Social media, content marketing and copywriting – with the power of AI." It includes "Try Free" and "See how it works" buttons. The interface screenshot shows a dark-themed dashboard with a sidebar menu (Planner, Socials, Copywriter, Analytics, Integrations, Settings, Billing) and a main content calendar area. A modal window titled "Create your first post" is open, showing a preview of a social media post with a burger image and the text "Create your first post". It includes a "Let's create your post" button.

- Ocoya.com generates social media and hashtags by your headline and context
- It posts the contents on the optimal time
- <https://www.ocoaya.com>

Vidyo.ai



The image shows the homepage of Vidyo.ai. At the top left is the logo 'vidyo.ai'. To its right are navigation links: 'Use Cases', 'Features', 'Pricing', and 'Resources'. A prominent blue button labeled 'Start for Free →' is positioned to the right of these links. Below this header, the main headline reads 'Make short videos from long ones instantly'. A subtext below it says 'Create social ready short clips from your long videos with AI ✨ Save 90% time and effort'. Another 'Start For Free →' button is located further down. A small note at the bottom of this section states 'No payment required. Start instantly.' Below this, a light gray box contains the text 'Make VIRAL short clips with captions, templates, emojis 😁 & more for'. At the bottom of the page, there are social media links for TikTok, Instagram, YouTube, LinkedIn, and Facebook.

- Vidyo.ai generates short viral clips with emojis and captions from a given video footage
- <https://vidyo.ai>

Loved by 30K+ podcasters & creators

More Tools at <https://bestwebbs.com>

BESTWEBBS.COM

🔍

If you declare with your mouth, "Jesus is Lord," and believe in your heart that God raised him from the dead, you will be saved. Romans 10:9

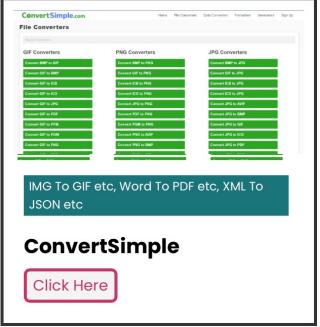
The Hub for AI Tools

Bestwebbs is one-stop destination for all AI TOOL KIT. Our Handpicked list includes every AI TOOL and some daily used websites

[All](#)
[Art Generation](#)
[Avatar](#)
[Chatbot](#)
[Chrome Extensions](#)
[ClassRoom](#)
[Code Generation](#)
[Content Generator](#)
[Copy Writing](#)
[Creative](#)
[Detect AI](#)
[Finance](#)
[Image Clarity Improve](#)
[Image To Sound](#)

[IMG To GIF etc](#)
[Kids & Fun](#)
[Learning](#)
[Marketing](#)
[Motion Capture](#)
[Music](#)
[Plagiarism](#)
[Podcasting](#)
[Prompts](#)
[Social Media](#)
[Speech To Text](#)
[Text To Audio](#)
[Text to Image](#)
[Text to Song](#)
[Text To Speech](#)

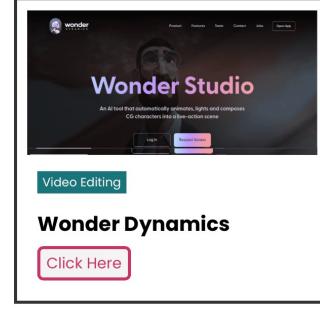
[Text To Text](#)
[Text to Video](#)
[Uncategorized](#)
[Video Editing](#)
[Voice Modulation](#)
[Word To PDF etc](#)
[XML To JSON etc](#)



ConvertSimple

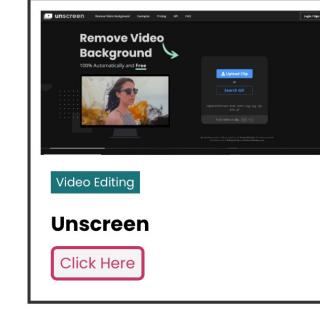
IMG To GIF etc, Word To PDF etc, XML To JSON etc

[Click Here](#)



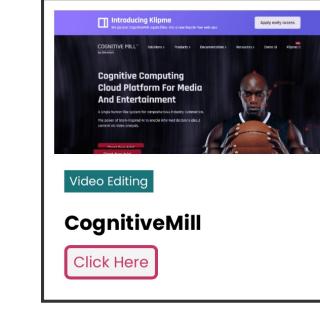
Wonder Dynamics

[Click Here](#)



Unscreen

[Click Here](#)



CognitiveMill

[Click Here](#)



auto studio

One Click Audio Cleaning

[Click to know more](#)



krisp

Augment Your Voice

[Click to know more](#)



wovatal

Supercharge your online meetings with Voice Productivity AI

[Click to know more](#)



wavelet

Your AI Assistant for Video Captions!

[Click to know more](#)

8. Conclusion and Discussion Time

Conclusion

- The Transformer model is a sequence-to-sequence model
 - It learns to encode phrase structures in the source sequence in the self-attention
 - It learns to align phrase structures in the source to the target sequence using the alignment attention
 - It learns to produce a target sequence using next-word prediction from the encoded phrase structures
- Upper bounds
 - **Encoder:** H^N non-recursive phrases
 - **Decoder:** $H^{N_E+N_D}$ non-recursive translation pairs

Thank You

<https://tinyurl.com/p8ezwzvm>

วิธีอ่านเปเปอร์ภายในเวลาอันสั้น

1. อ่าน title และทำความเข้าใจให้ได้ว่า input และ output ของระบบคืออะไร
 1. ถ้ายังไม่เข้าใจ ให้ไปอ่าน abstract
 2. ถ้ายังไม่เข้าใจอีก ให้ไปอ่าน introduction
2. หาให้เจอบ่า เปเปอร์นี้ต้องการแก้ปัญหาอะไรที่คนอื่นยังแก้ไม่ได้
3. ดูภาพรวมของระบบ ซึ่งมักจะมาในรูปของ figure ต่างๆ
4. เปรียบเทียบผลการทดลองว่า วิธีการนี้เหมาะสมกับปัญหาแบบใด
5. หา limitation ของระบบนี้ให้เจอ