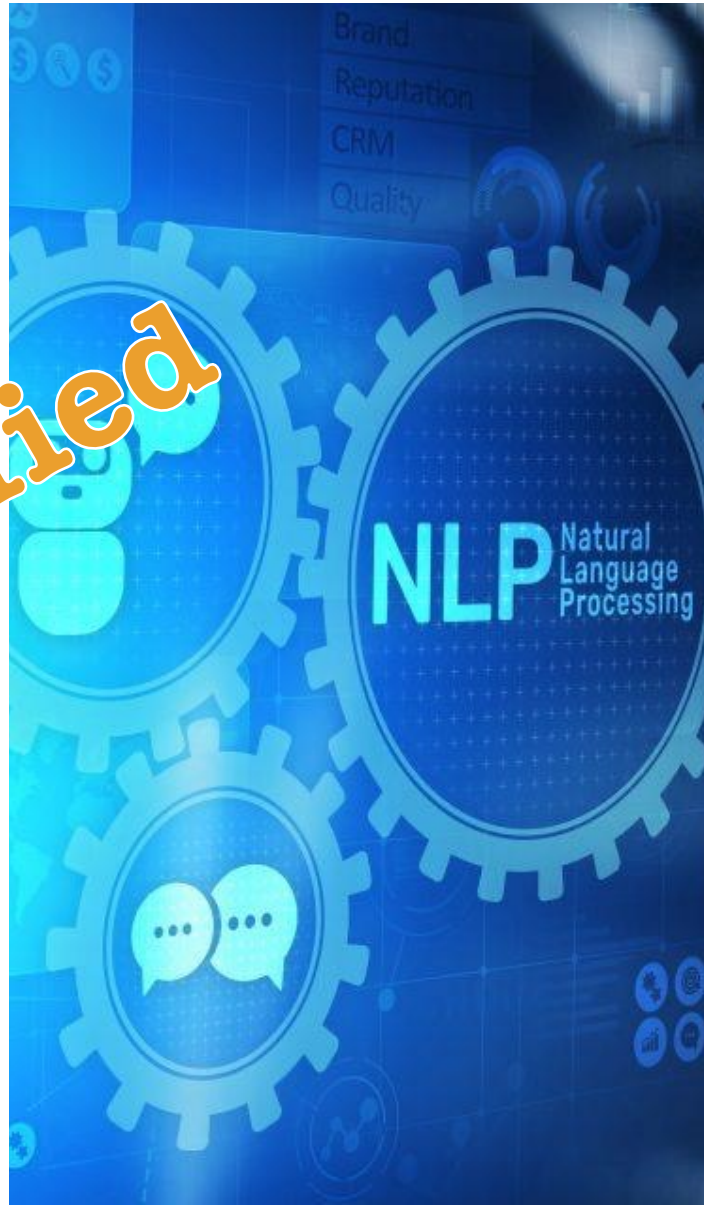
Large Language Models Demystified



Prachya Boonkwan (P'/N' Arm)

Language and Semantic Technology Lab (LST)
National Electronics and Computer Technology Center (NECTEC)

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

URL ⇒ https://tinyurl.com/2pvf3kcr

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

Large Language Models Demystified

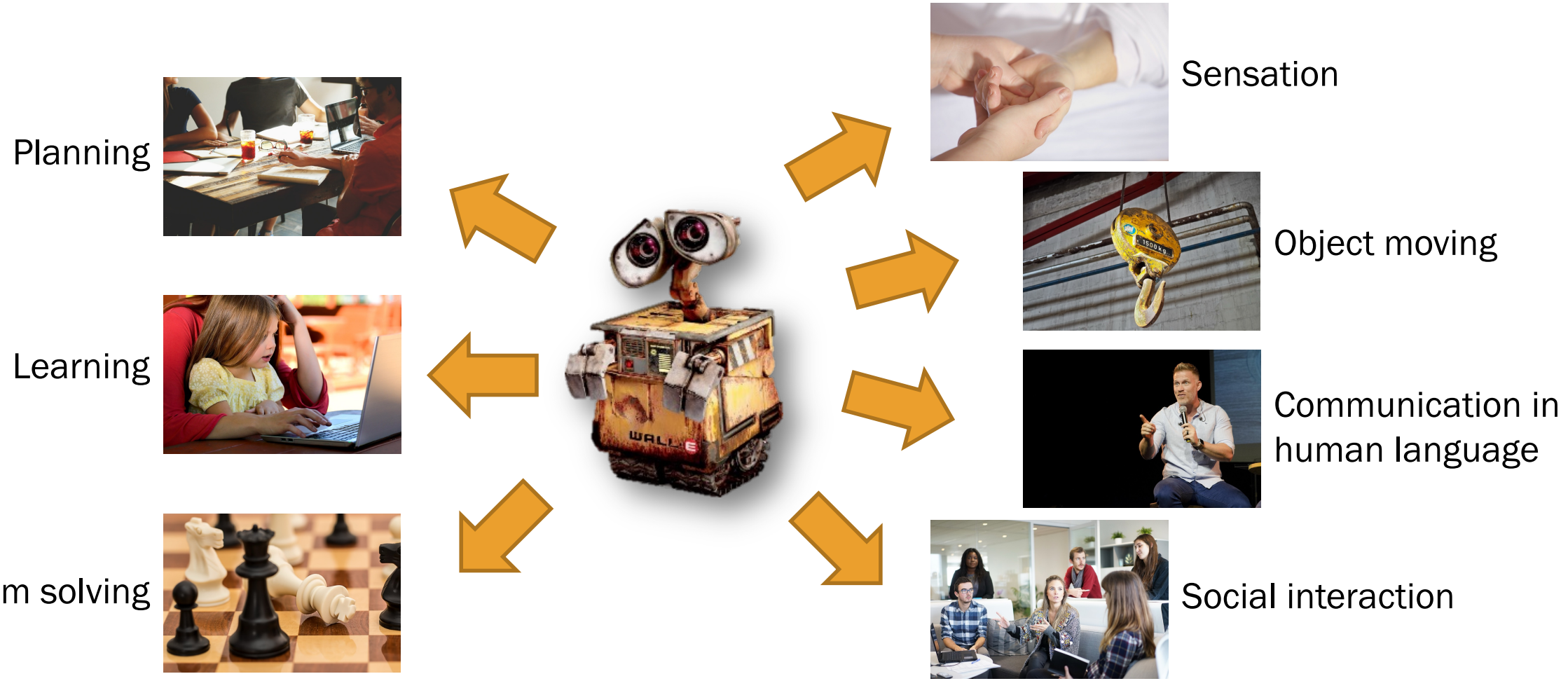
- Introduction to AI
- Large language models
- Transformer Model
- Training ChatGPT
- OpenThaiGPT
- Future direction

https://tinyurl.com/2pvf3kcr

1. What exactly is AI?

What is Artificial Intelligence?

- Imitation of natural intelligence with machine



The rise of Robotics and AI

Fueled by advances in computing power and connectivity, the fields of robotics and artificial intelligence have grown rapidly

1921
The term **robot** is first used by Czech writer Karel Capek

1939
Elektro, a humanoid robot, debuts at the World's Fair, smoking cigarettes and blowing up balloons

1948
William Grey Walter creates **the first autonomous robot with complex behavior**

1950
Alan Turing publishes paper about the possibility of machines that think, develops idea known as the

Turing's Test.

It tests a machine's ability to "think" by answering a series of questions. In essence, **the tester must think the machine's answers are coming from a human**

1941
Isaac Asimov formulates the

Three Laws of Robotics:

A robot may not injure a human being or, through inaction, allow a human being to be harmed

A robot must obey orders given it by human beings except where such orders would conflict with the First Law

A robot must protect its own existence as long as such protection does not conflict with the First or Second Law

1951
Marvin Minsky builds the first neurocomputer, **SNARC**

1954
George Devol invents the first digitally operated and programmable robot

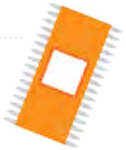
1956
Field of AI research founded at a conference at Dartmouth

1960
Frank Rosenblatt constructs **Mark I Perceptron,** a computer that learned new skills by trial and error

1961
GM installs Unimate robot to lift and stack hot pieces of metal

1968
Mobile robot "Shakey" is introduced. It's controlled by a computer the size of a room

1972
Stanford researcher develops PARRY, designed to simulate a paranoid schizophrenic.



1974
Intel produces its second-generation 8080 general-purpose chips

1956
IBM 305, the first hard disk drive
5MB

1970
IBM 1330
100MB
per pack

1985
IBM 0665, a 5.25" disk with
20-40MB

Minimize and maximize
Shrinking disk sizes and exponentially growing capacity help fuel robotics and AI

1984
The RB5X, developed by General Robotics Corp., includes software enabling it to learn from its environment

1984
Doug Lenat and his team start **Cyc,** to codify millions of pieces knowledge that compose human common sense

1985
Jaron Lanier's VPL Research, Inc., sells first VR glasses and gloves; Lanier coins the phrase

virtual reality

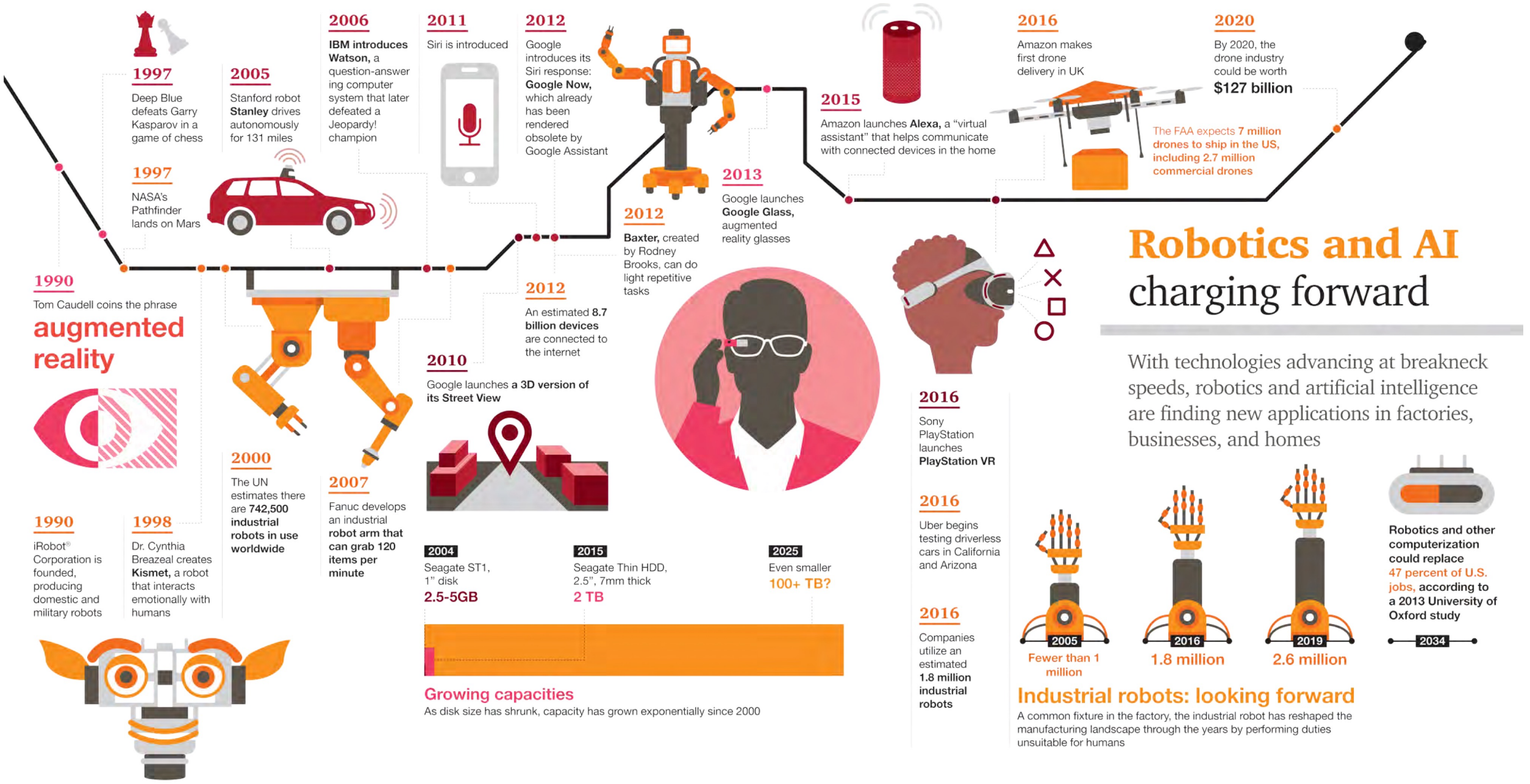
1986
Honda creates the E0, the first of a series of humanoid robots that walk on two feet

1988
Nope, I'm human.

Researchers launch Jabberwacky, an AI chatbot designed to learn through conversation

1988
The first HelpMate service robot begins work at Danbury Hospital

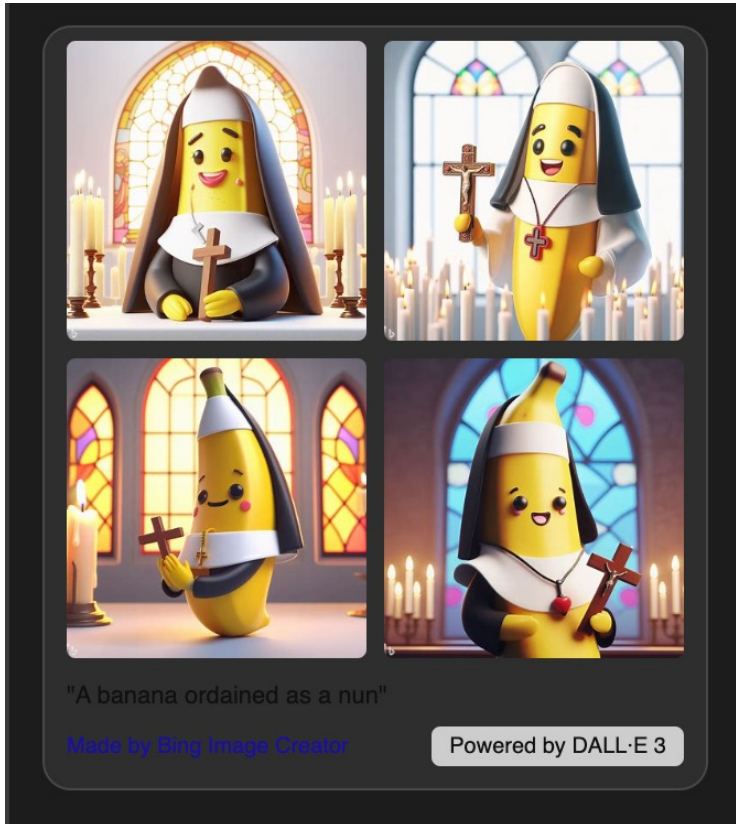
Source: http://usblogs.pwc.com/emerging-technology/rise-robotics-ai-infographic/

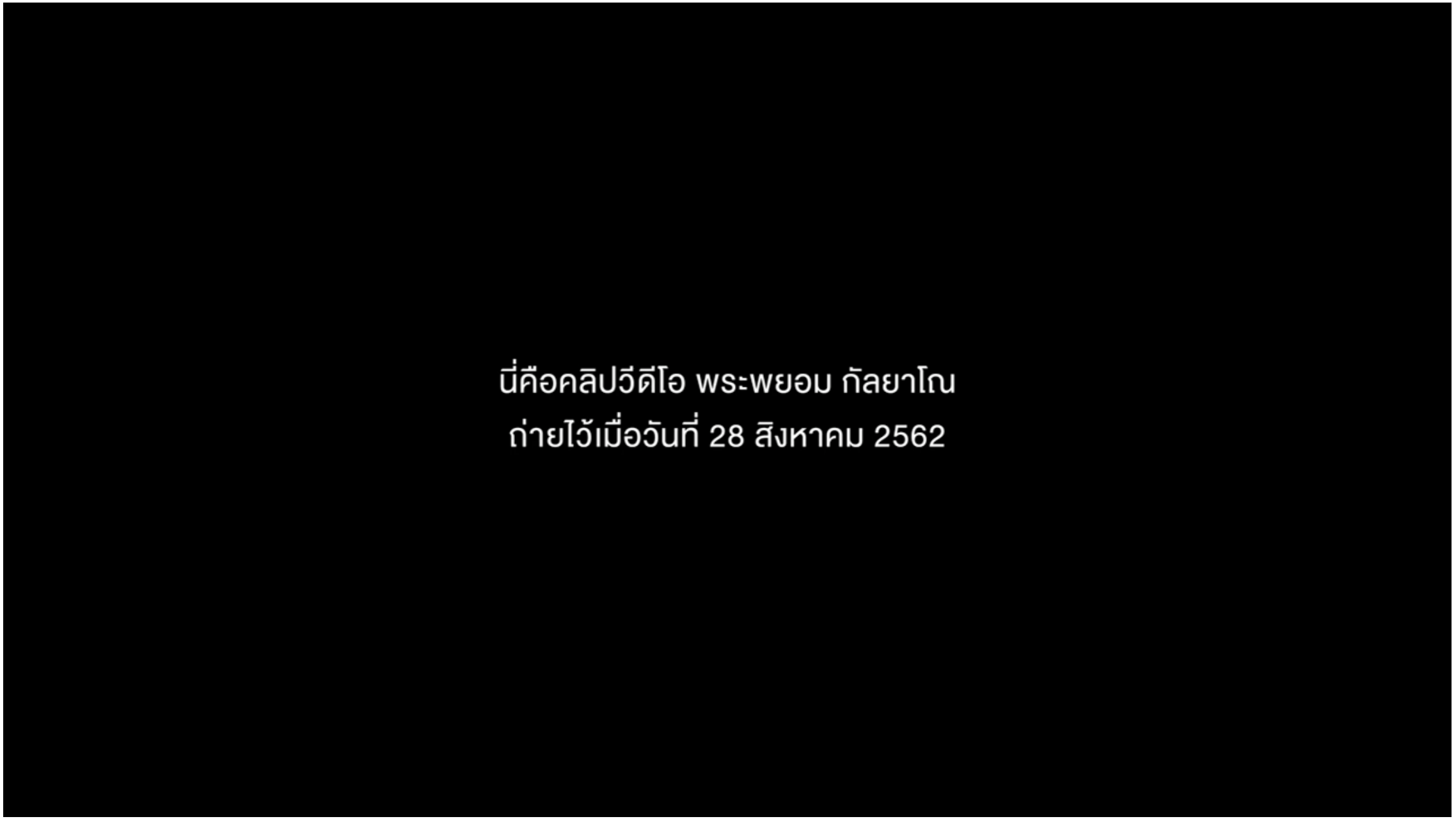


Source: http://usblogs.pwc.com/emerging-technology/rise-robotics-ai-infographic/

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)
  - DALL-E and Stable Diffusion (MidJourney) for image
  - VALL-E for voice (Microsoft)
  - Deepfake for video





https://www.youtube.com/watch?v=wWjnXa_k2NE

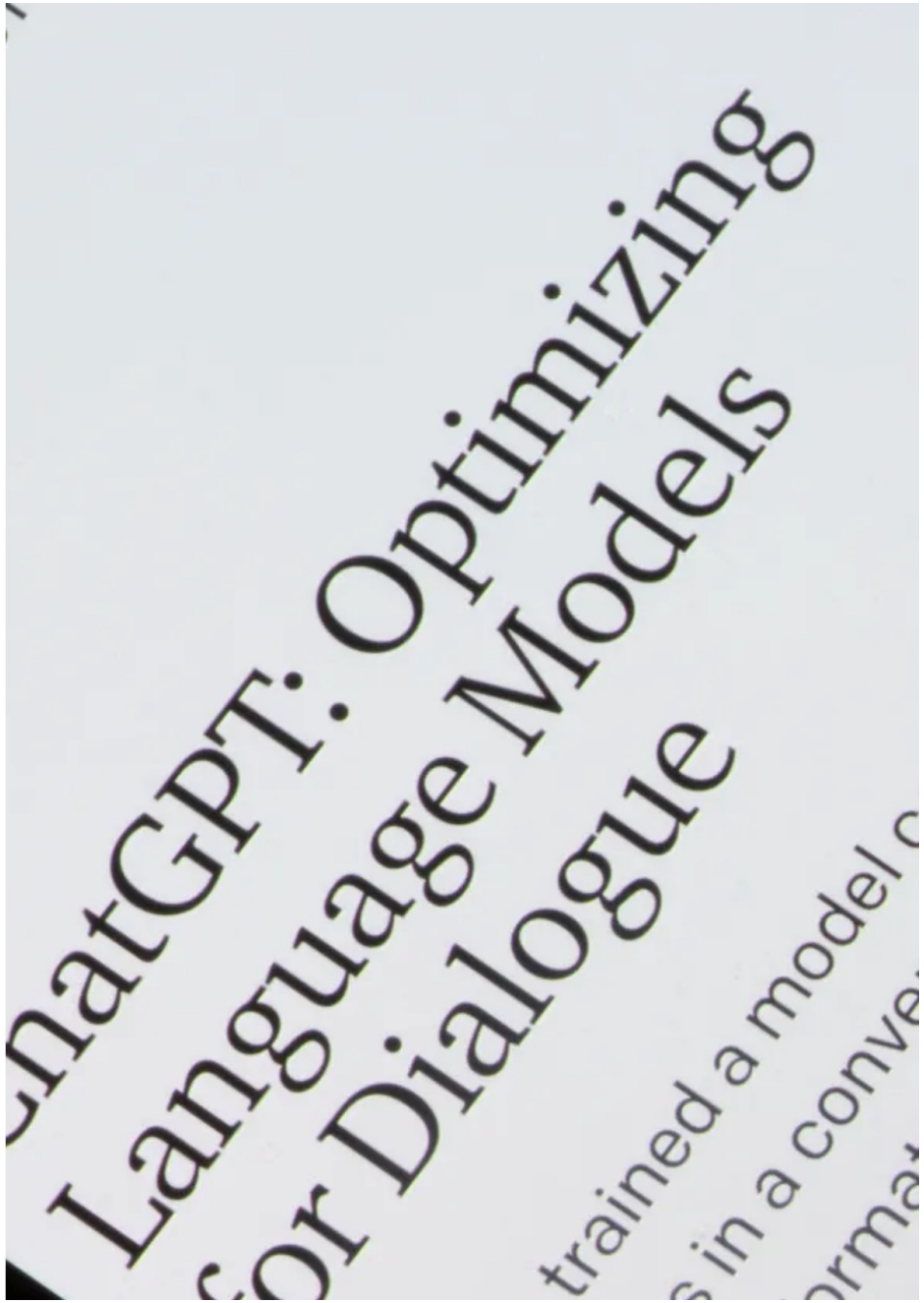
2. Large Language Models

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'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.

ChatGPT (OpenAI)

Copilot (Microsoft)

Perplexity.AI

Gemini (Google)

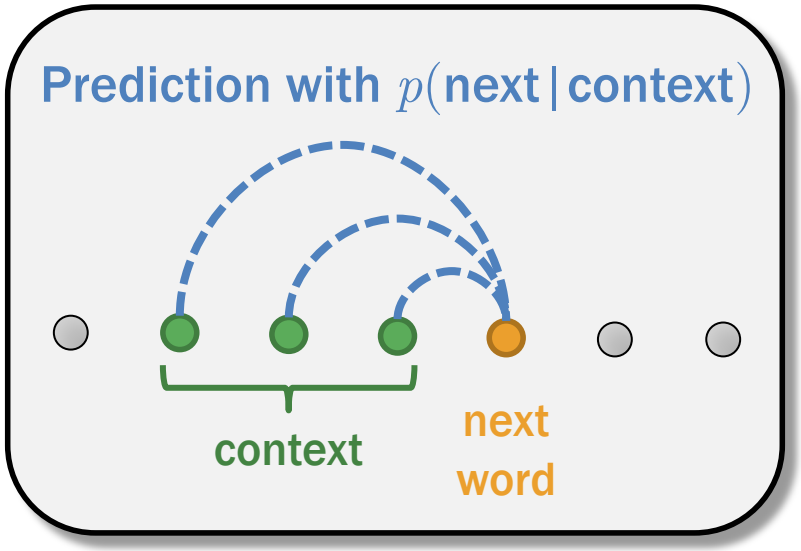
Claude.AI (Anthropic)

Large Language Model (LLM)

- 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 ... w_{k-1} until the whole sentence of length N is produced

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

next word — context



We need at least 1 billion words of text data to train a stable LLM, which learns **word collocations** and **phrase structures**

What is GPT?

อรอุมา สิงห์ศิริ

น้อง	เป็น	สาว	_____	___	___	___	__	_____	_____	___	____	_____

(This is a very popular Thai country song in the 1980s)

What is GPT?

อรอุมา สิงห์ศิริ

น้อง	เป็น	สาว	ขอนแก่น	___	___	___	__	_____	_____	___	____	_____

(This is a very popular Thai country song in the 1980s)

What is GPT?

อรอุมา สิงห์ศิริ

น้อง	เป็น	สาว	ขอนแก่น	ยัง	___	___	__	_____	_____	___	____	_____

(This is a very popular Thai country song in the 1980s)

What is GPT?

อรอุมา สิงห์ศิริ

น้อง	เป็น	สาว	ขอนแก่น	ยัง	บ่	___	__	_____	_____	___	____	_____

(This is a very popular Thai country song in the 1980s)

What is GPT?

อรอุมา สิงห์ศิริ

น้อง	เป็น	สาว	ขอนแก่น	ยัง	บ่	เคย	__	_____	_____	___	____	_____

(This is a very popular Thai country song in the 1980s)

GPT: Generative Pretrained Transformer

เพลง: สาวอีสานรอรัก

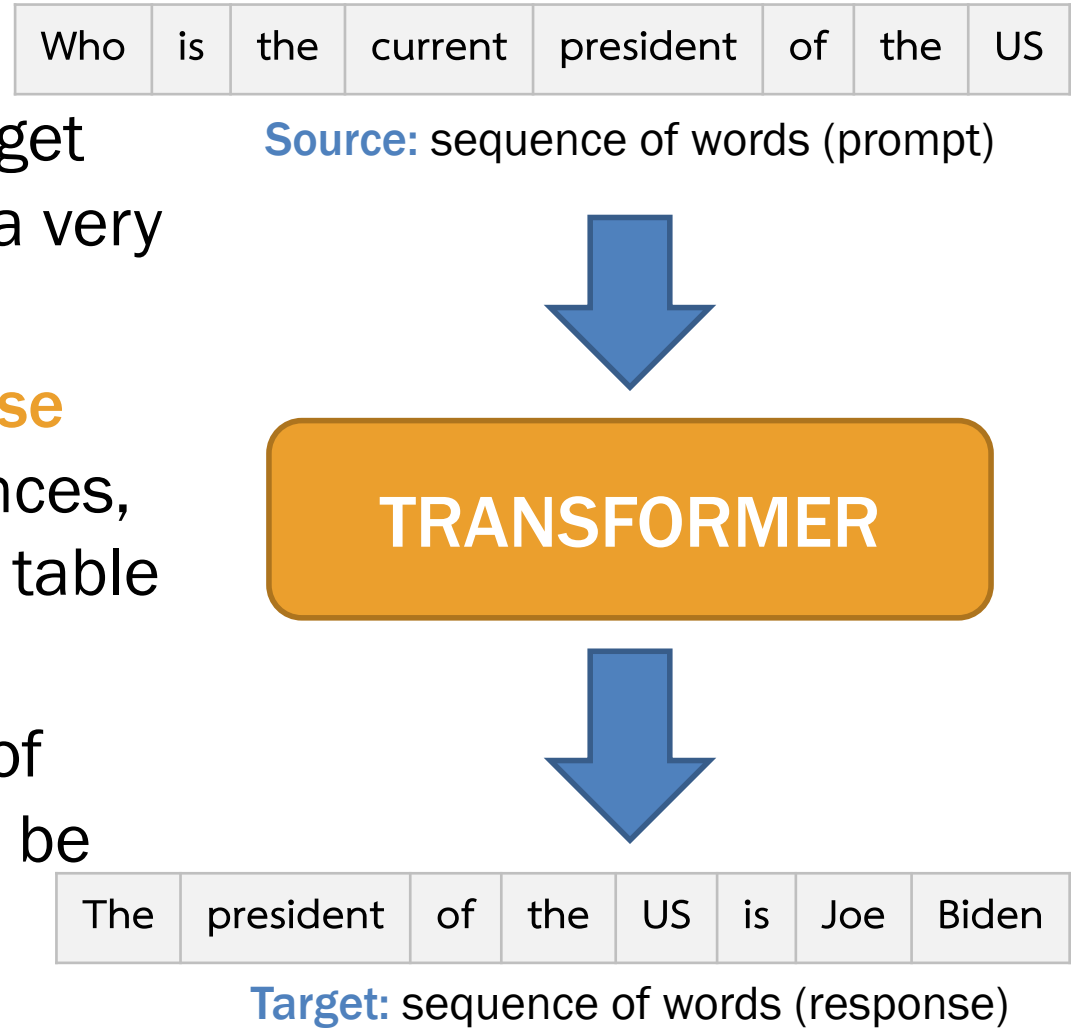
อรอุมา สิงห์ศิริ

น้อง	เป็น	สาว	ขอนแก่น	ยัง	บ่	เคย	มี	แฟน	บ้าน	อยู่	แดน	อีสาน

https://www.youtube.com/watch?v=stLQVIau1ns

Transformer Model (Vaswani et al., 2016)

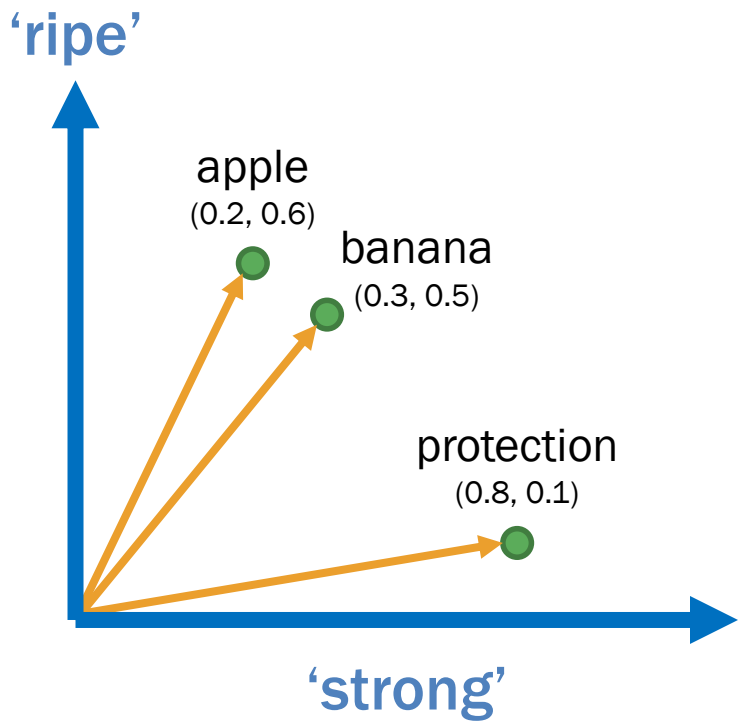
- Sequence-to-sequence generation
  - **Translation:** It learns how to produce a target sequence from a source sequence, given a very large dataset of sequence pairs
  - **Pros:** It learns **word collocations** and **phrase structures** on the input and output sequences, and associates them cross-lingually 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



3. Transformer Model

Recap: Word Vector

- Distributional semantics
  - 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:

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

Scaled Dot-Product Attention

- Semantic similarity ⇒ 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

Simple Form

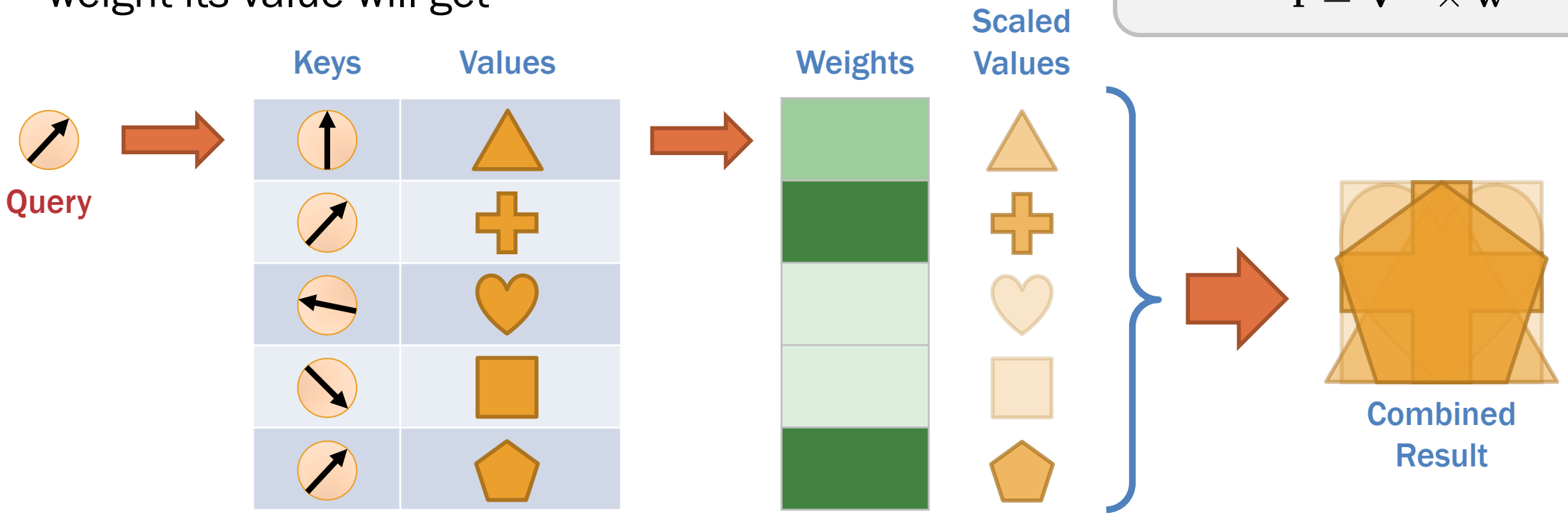
$$w_i \propto \mathbf{k}_i \cdot \mathbf{q}$$

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

Matrix Form

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

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



Scaled Dot-Product Attention

- Semantic similarity ⇒ 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

Simple Form

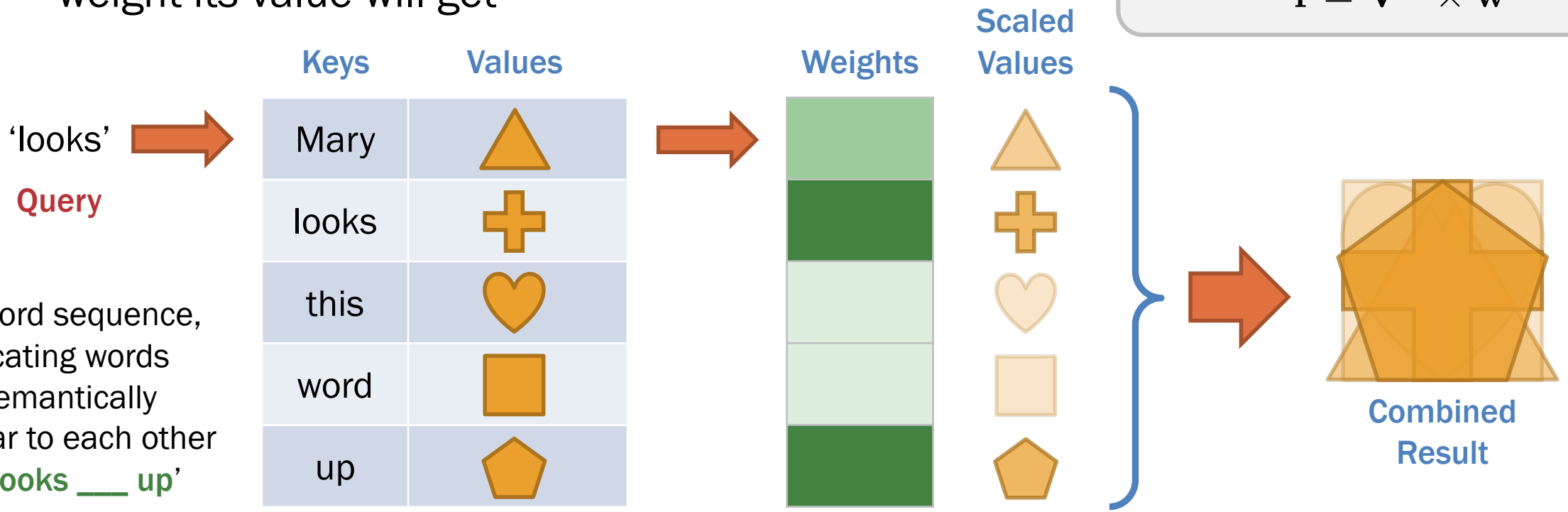
$$w_i \propto \mathbf{k}_i \cdot \mathbf{q}$$

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

Matrix Form

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

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



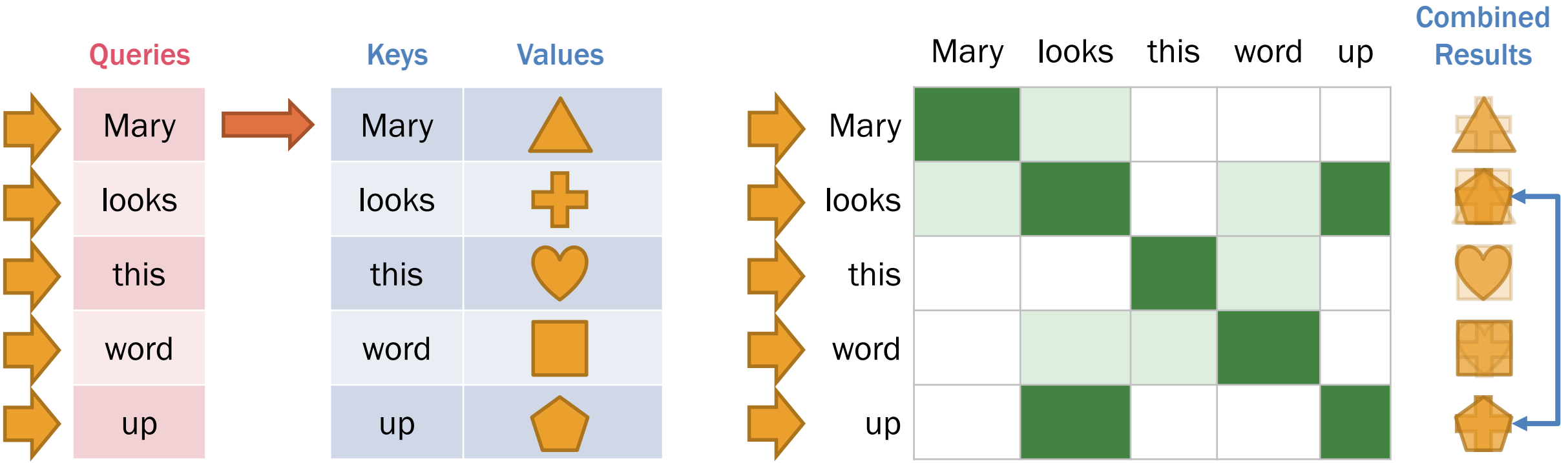
For word sequence, collocating words are semantically similar to each other e.g. '**looks ___ up**'

Self-Attention

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

Matrix Form

$$\mathbf{W} = \mathrm{Softmax}(\mathbf{K} \times \mathbf{K}^\top)$$
$$\mathbf{R} = \mathbf{W} \times \mathbf{V}$$

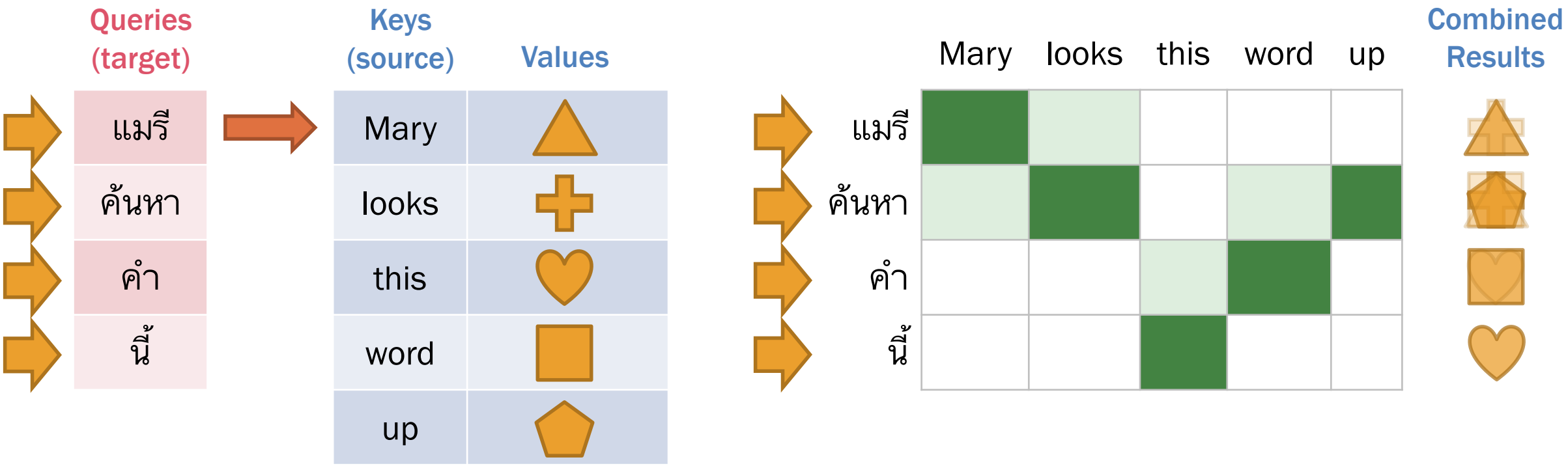


Cross-Attention

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

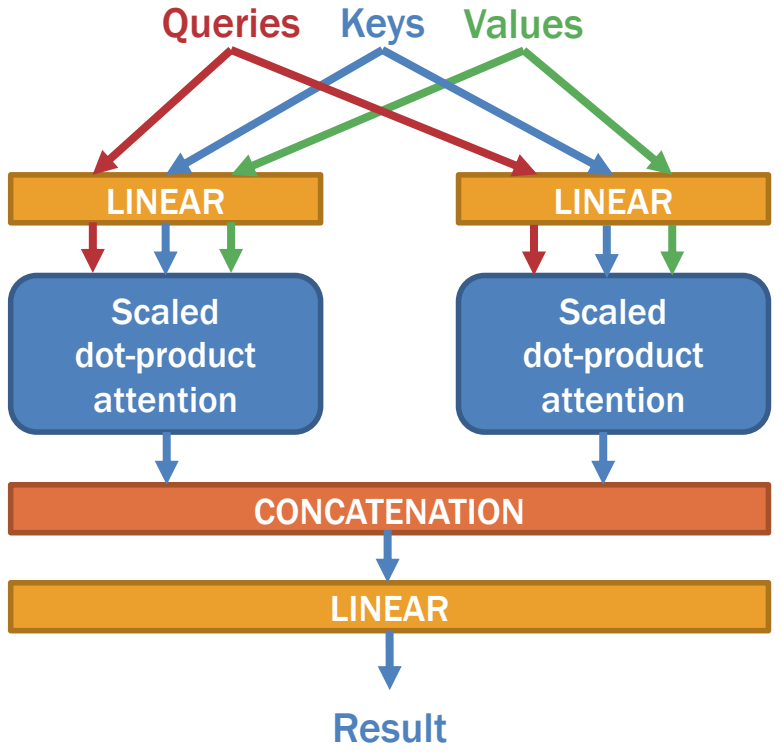
Matrix Form

$$\mathbf{W} = \mathrm{Softmax}(\mathbf{Q} \times \mathbf{K}^\top)$$
$$\mathbf{R} = \mathbf{W} \times \mathbf{V}$$

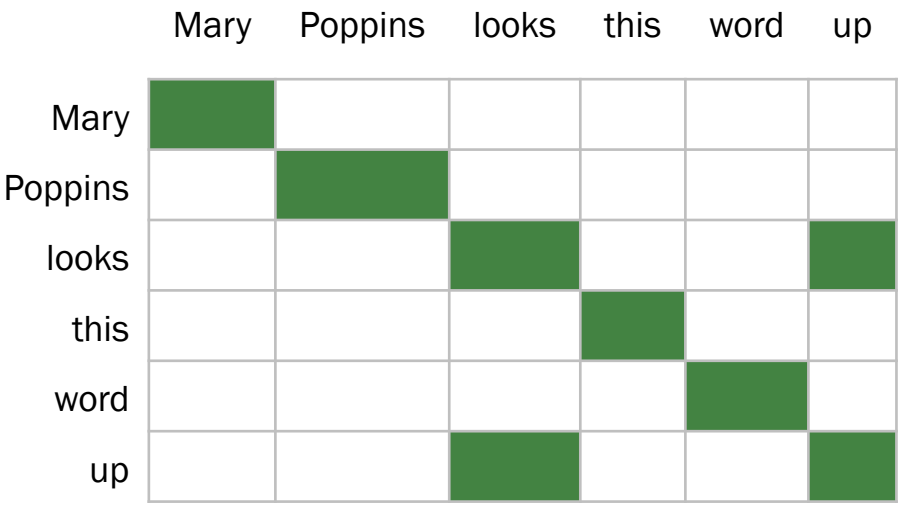


Multihead Attention

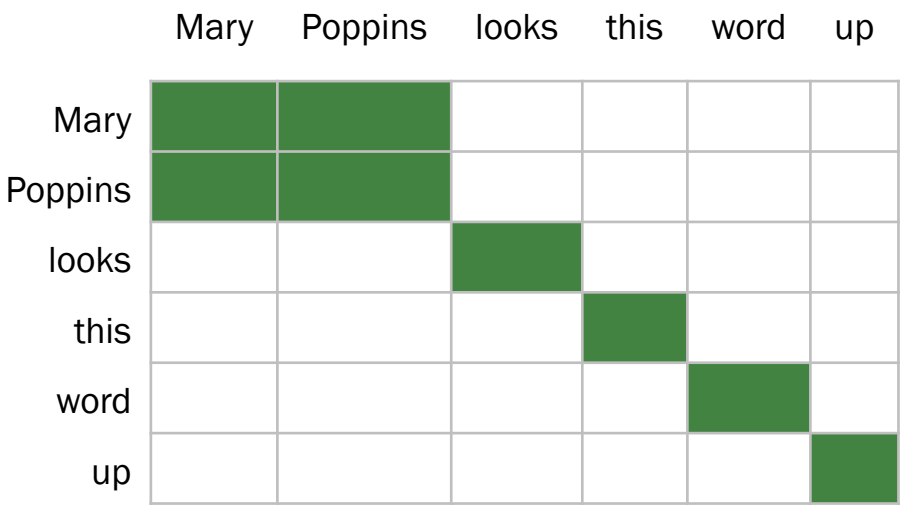
- Scaled dot-product attention has a drawback
  - It recognizes **only one** type of word collocation
  - If we assume more than one type of word collocation per sequence, then we have to combine multiple attention heads [default = 8 heads]

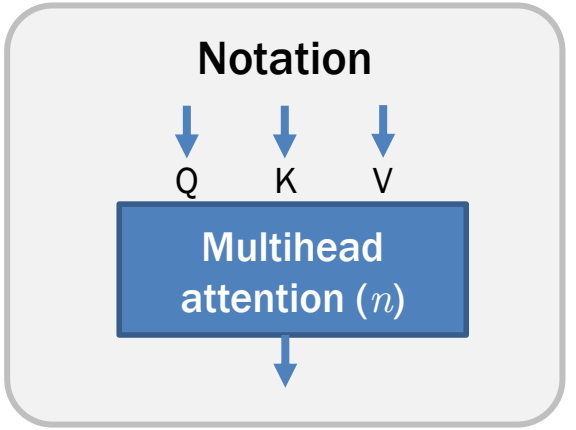


HEAD 1 (looks ___ up)



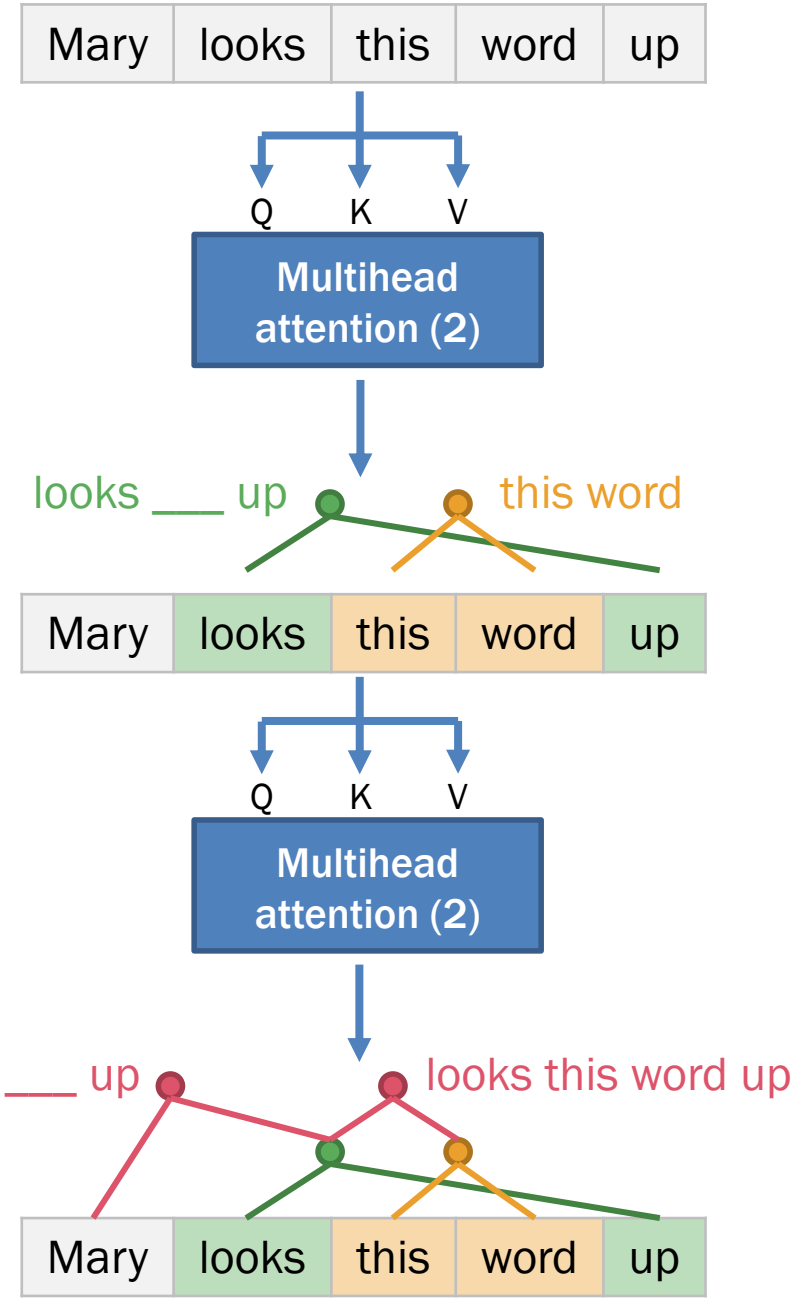
HEAD 2 (Mary Poppins)





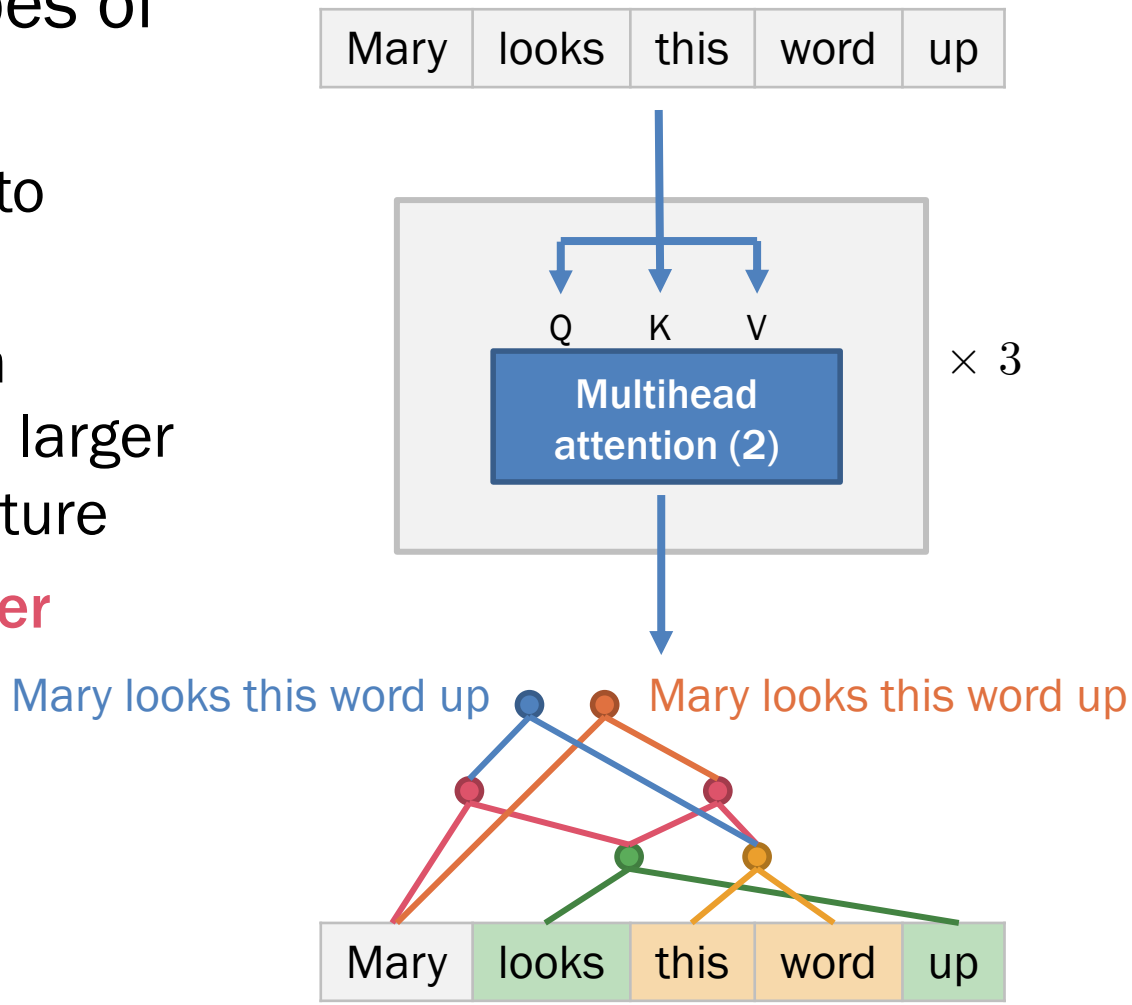
Phrase Structure

- 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 ⇒ phrase structure
  - Each layer is simply called an **encoding layer**



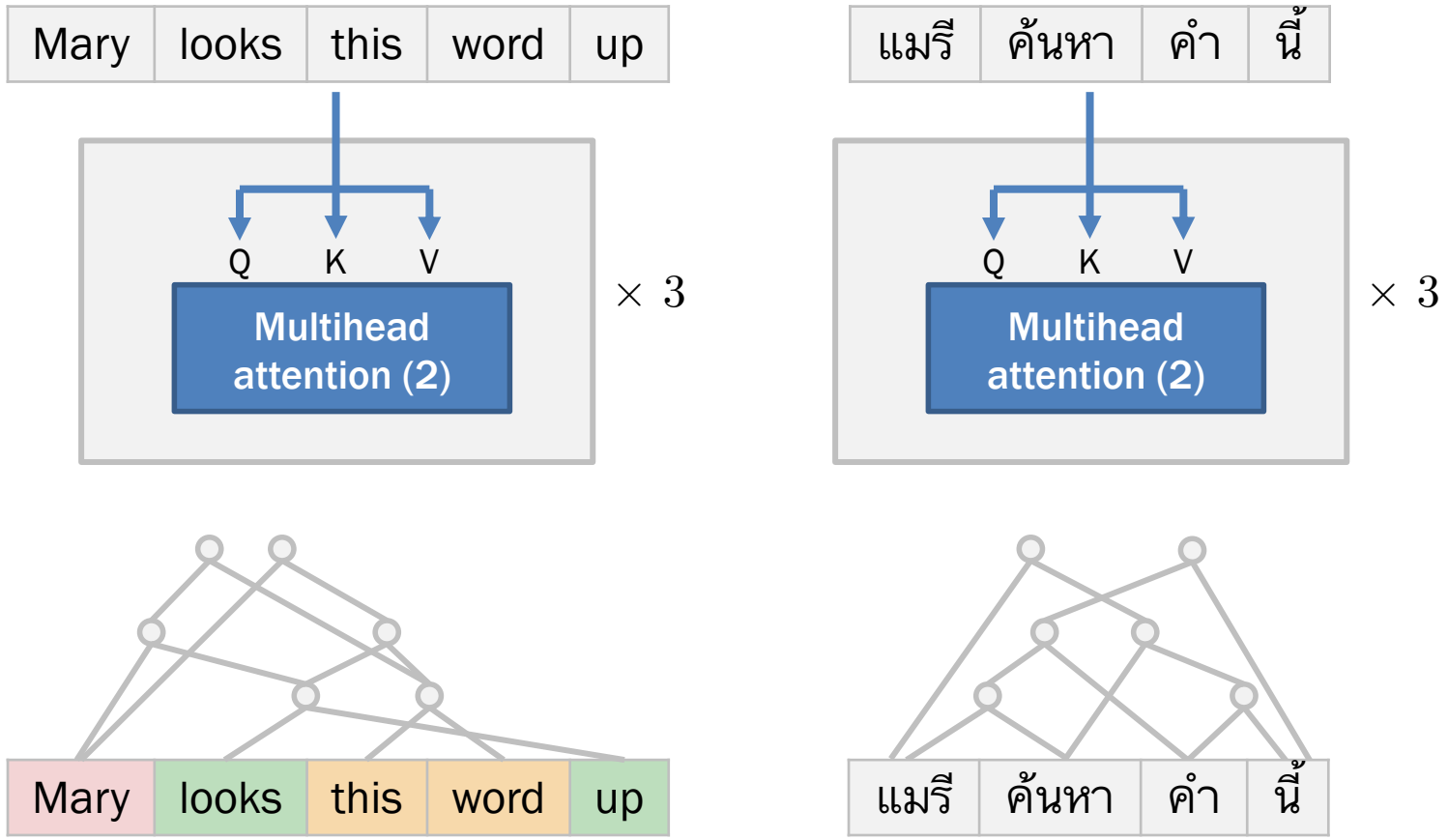
Phrase Structure

- 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**



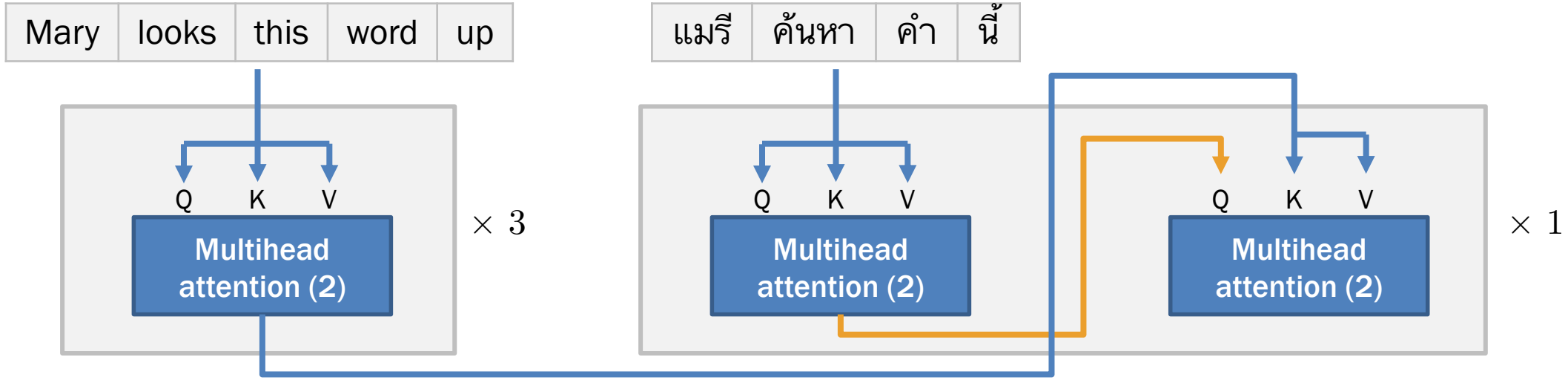
Alignment of Phrase Structures

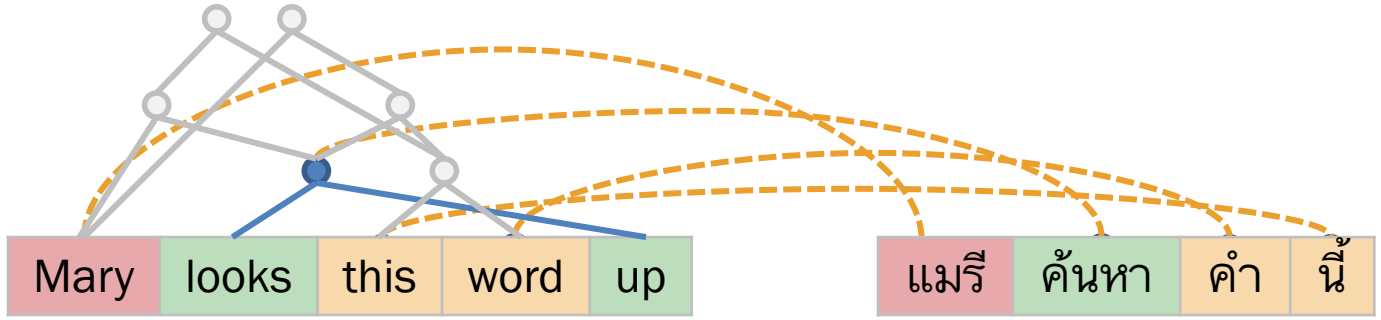
- H-head alignment attention recognizes H pairs of phrase structures



Alignment of Phrase Structures

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

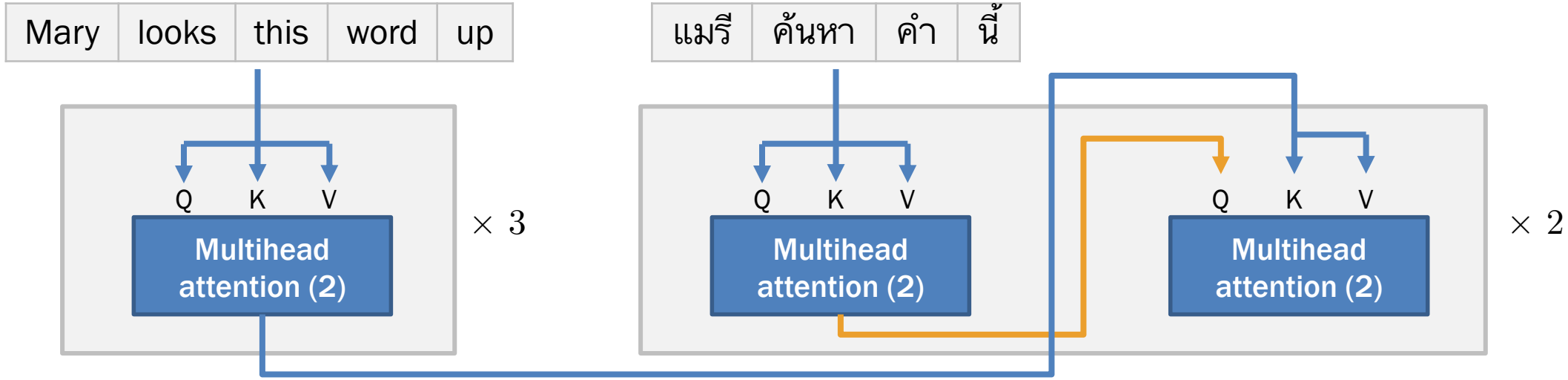


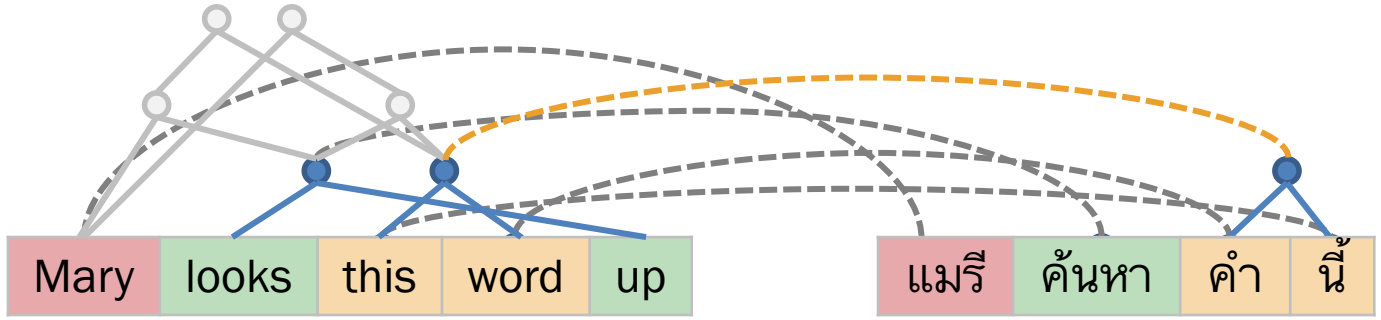


Mary ⇒ แมรี
looks ___ up ⇒ ค้นหา
this ⇒ นี้
word ⇒ คำ

Alignment of Phrase Structures

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

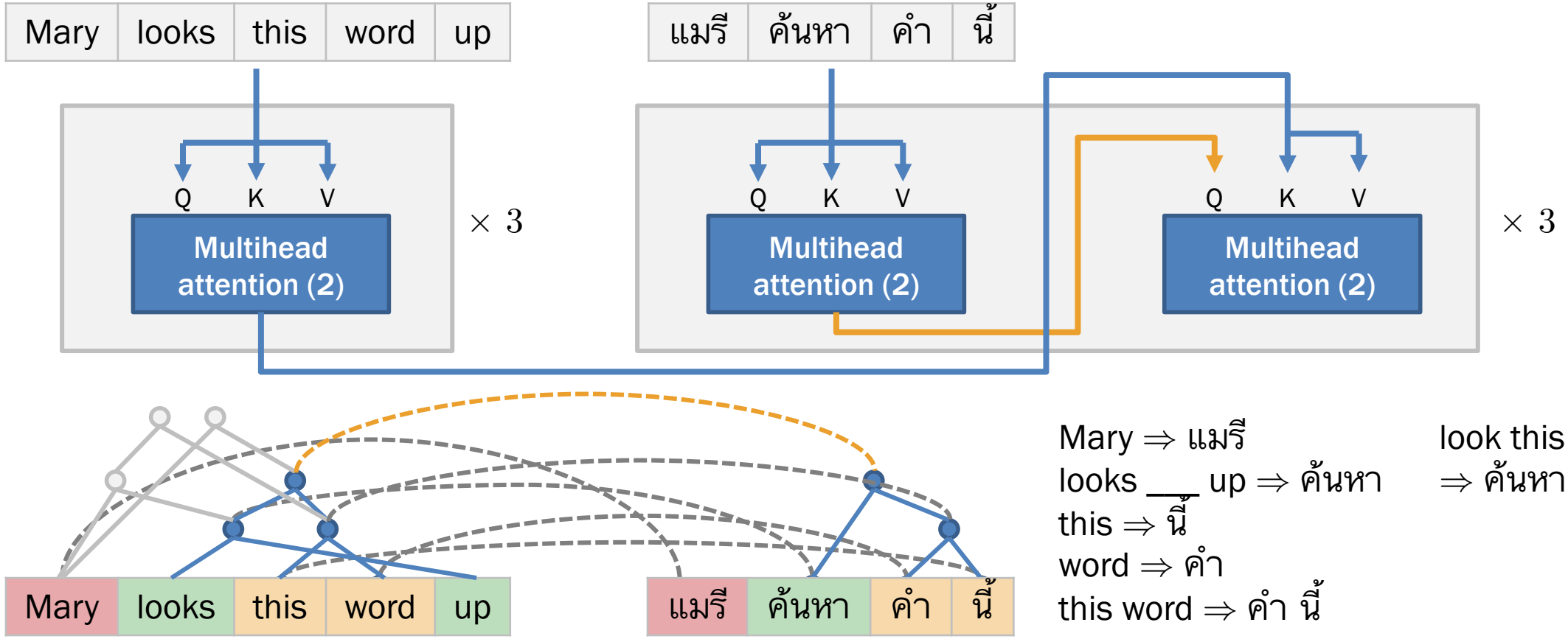




Mary ⇒ แมรี
looks ___ up ⇒ ค้นหา
this ⇒ นี้
word ⇒ คำ
this word ⇒ คำ นี้

Alignment of Phrase Structures

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

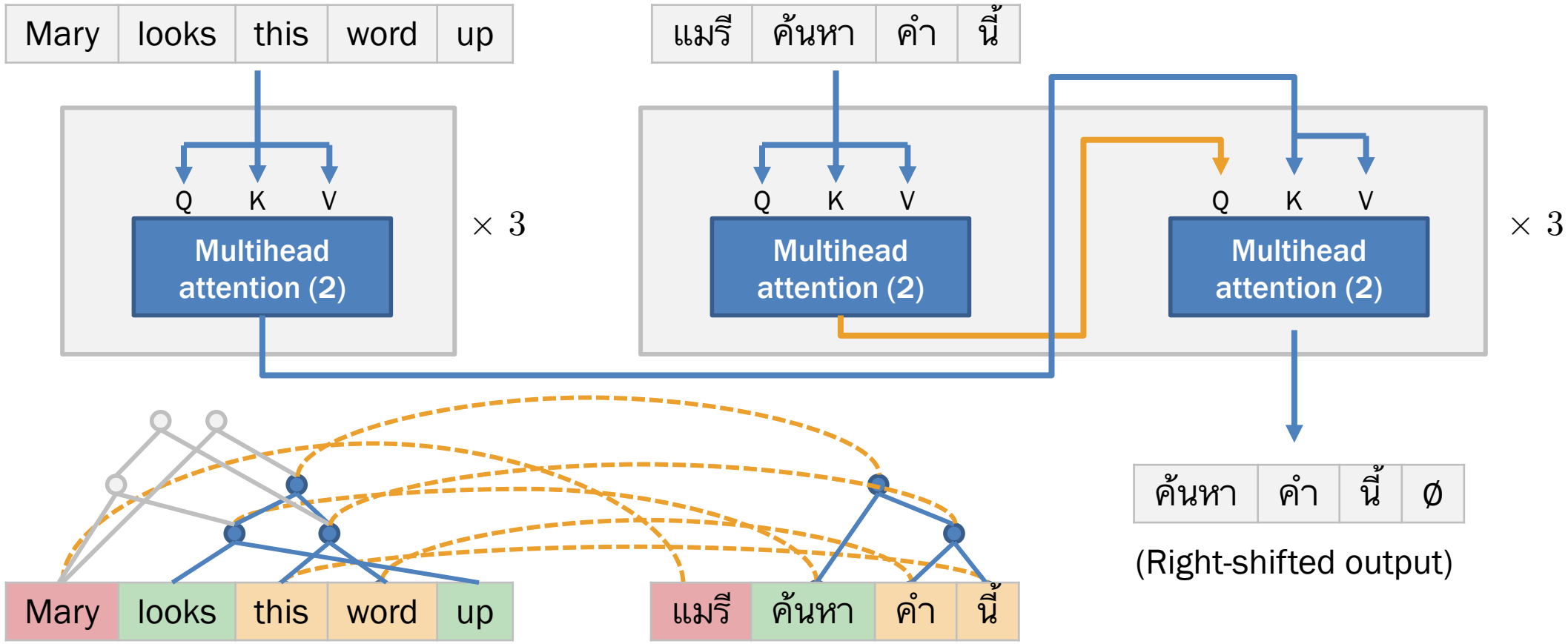


Mary ⇒ แมรี
looks ___ up ⇒ ค้นหา
this ⇒ นี้
word ⇒ คำ
this word ⇒ คำ นี้

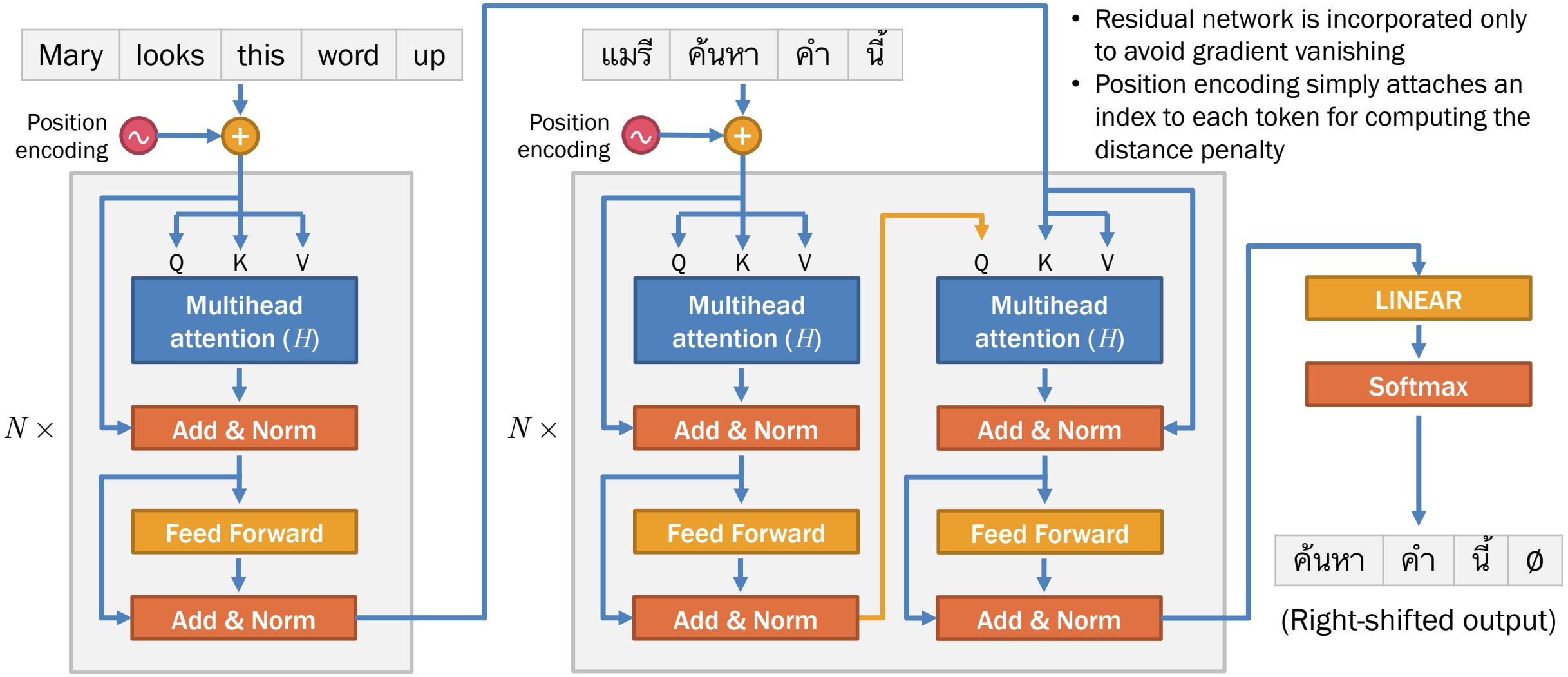
look this word up
⇒ ค้นหา คำ นี้

Alignment of Phrase Structures

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



Overview of the Transformer Model



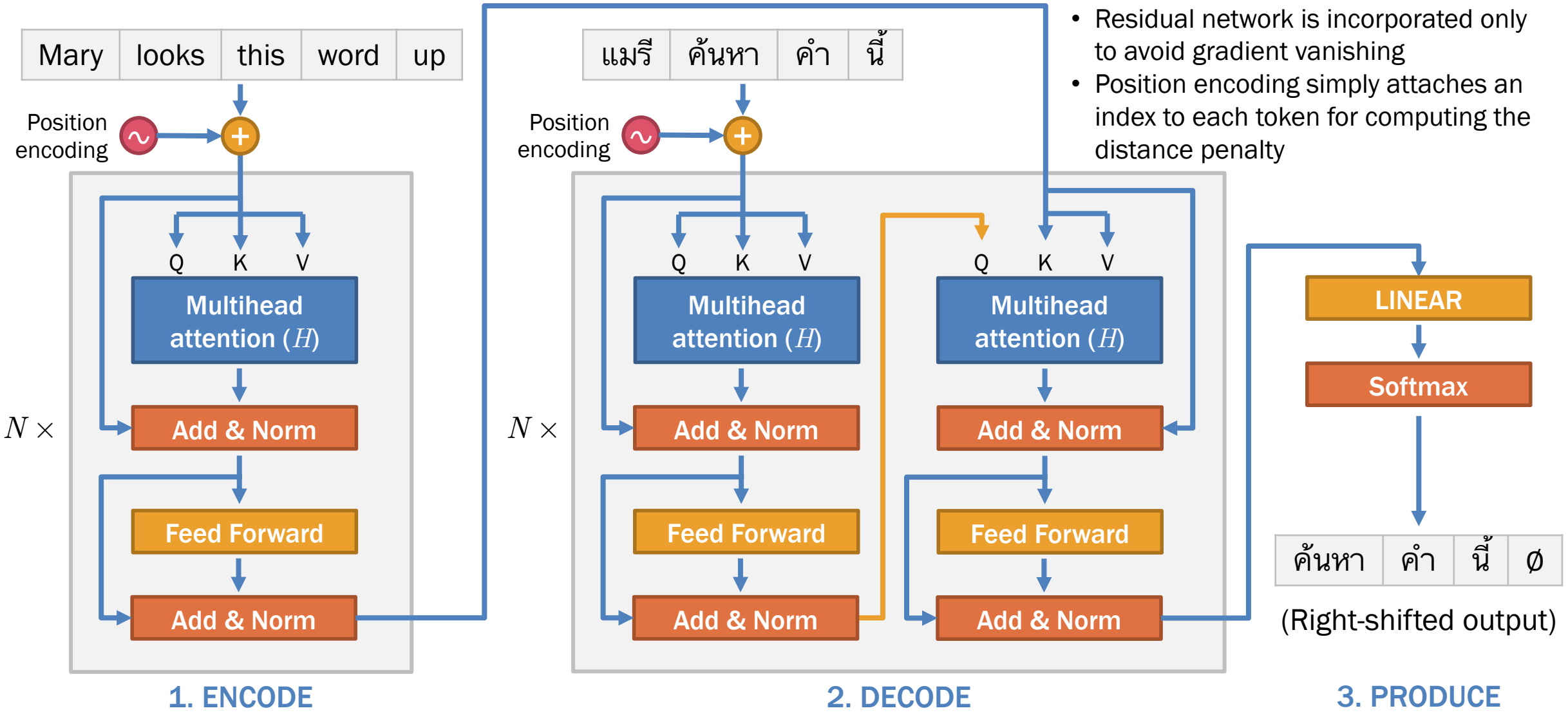
- Residual network is incorporated only to avoid gradient vanishing
- Position encoding simply attaches an index to each token for computing the distance penalty

1. Extract phrase structures in the source

2. Align phrase structures to the target

3. Produce translation

Overview of the Transformer Model

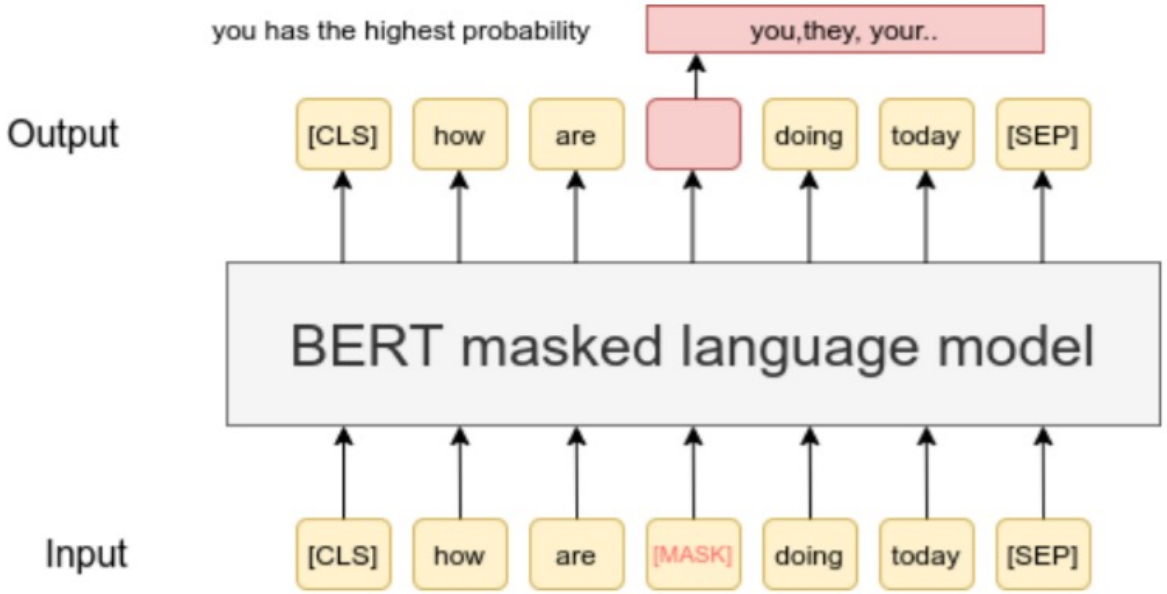


- Residual network is incorporated only to avoid gradient vanishing
- Position encoding simply attaches an index to each token for computing the distance penalty

4. Training an LLM

Masked Language Model (MLM)

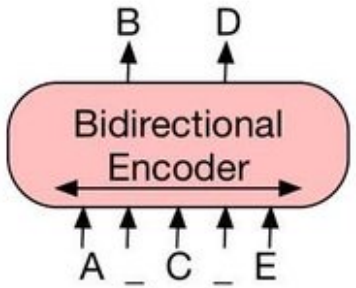
- We replace some words in the input with blanks and compute the loss of word prediction on these blanks in the output



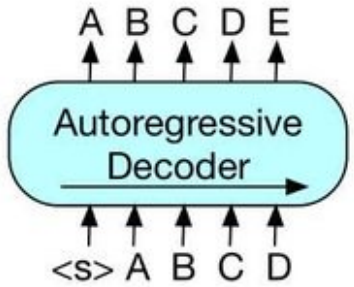
https://www.sbert.net/examples/unsupervised_learning/MLM/README.html

- 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

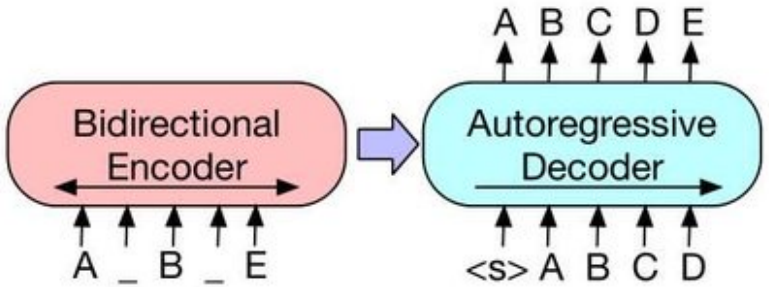
Approaches of Training LLMs (Lewis et al., 2019)



(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.

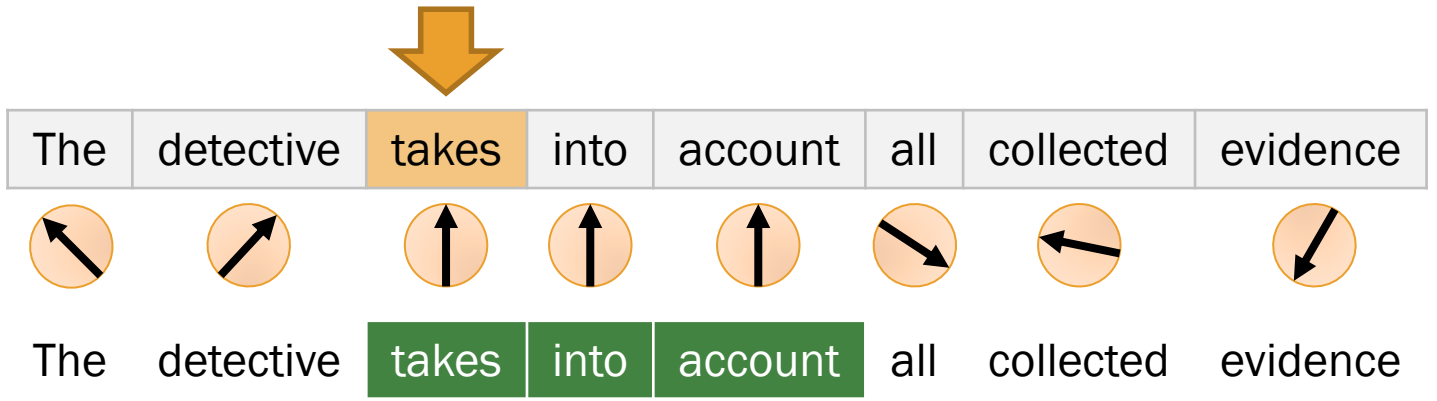


(c) BART: Inputs to the encoder need not be aligned with decoder outputs, allowing arbitary noise transformations. Here, a document has been corrupted by replacing spans of text with a mask symbols. The corrupted document (left) is encoded with a bidirectional model, and then the likelihood of the original 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

Pros: Multiword Expression (MWE)

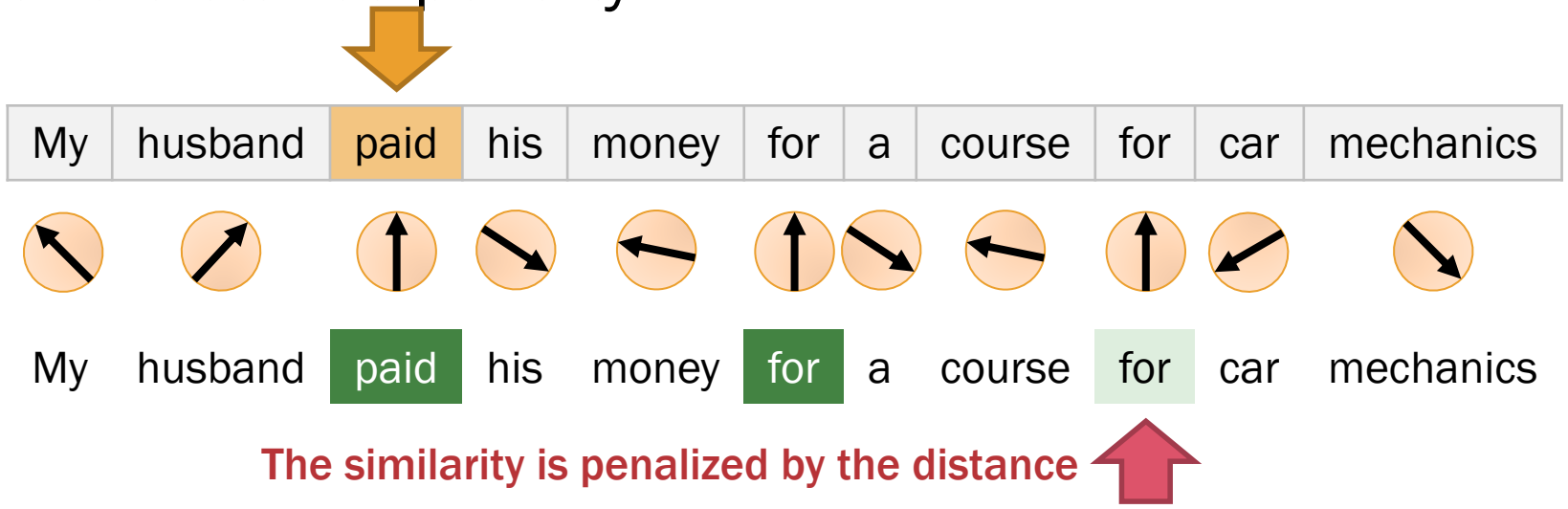
- It recognizes the idiosyncratic collocations of at least 2 words
  - E.g. ‘peanut butter’, ‘car park’, ‘kick the bucket’, ‘take into account’, ‘break up’
  - It learns MWEs by comparing each word with the remaining to reveal semantic similarity



MWE is extracted because of 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

PREDICTION
∅ ⇒ สถานะ
สถานะ ⇒ ทาง
ทาง ⇒ การเงิน
การเงิน ⇒ ที่
ที่ ⇒ มั่นคง
มั่นคง ⇒ ของ
ของ ⇒ ท่าน

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’

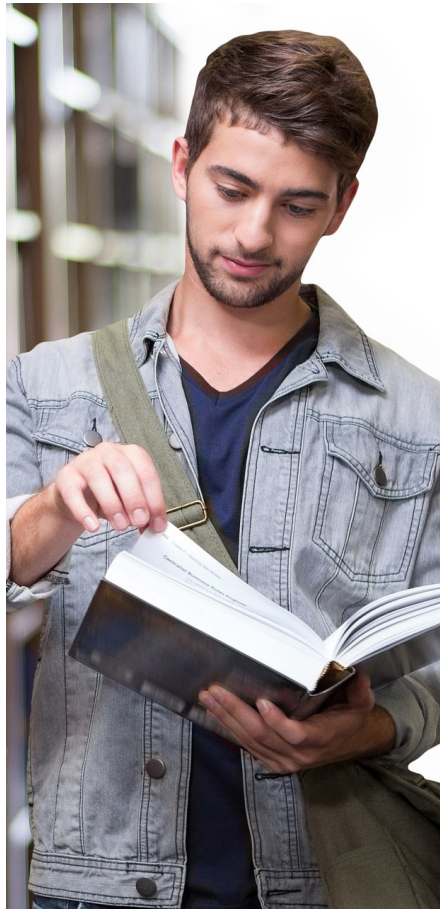
									PREDICTION						
Stevenson	initiated	a	scheme	for	building	the	first	locomotive							
Stevenson	initiated	a	scheme	for	building	the	first	locomotive	∅ ⇒ สตีเวนสัน	สตีเวนสัน					
Stevenson	**initiated**	a	**scheme**	for	**building**	the	first	locomotive	สตีเวนสัน ⇒ ประดิษฐ์	สตีเวนสัน	ประดิษฐ์				
Stevenson	initiated	a	scheme	for	building	the	first	**locomotive**	ประดิษฐ์ ⇒ รถจักรไอน้ำ	สตีเวนสัน	ประดิษฐ์	รถจักรไอน้ำ			
Stevenson	initiated	a	scheme	for	building	**the**	**first**	locomotive	รถจักรไอน้ำ ⇒ คัน	สตีเวนสัน	ประดิษฐ์	รถจักรไอน้ำ	คัน		
Stevenson	initiated	a	scheme	for	building	**the**	**first**	locomotive	คัน ⇒ แรก	สตีเวนสัน	ประดิษฐ์	รถจักรไอน้ำ	คัน	แรก	

4. Training ChatGPT

Constructing ChatGPT

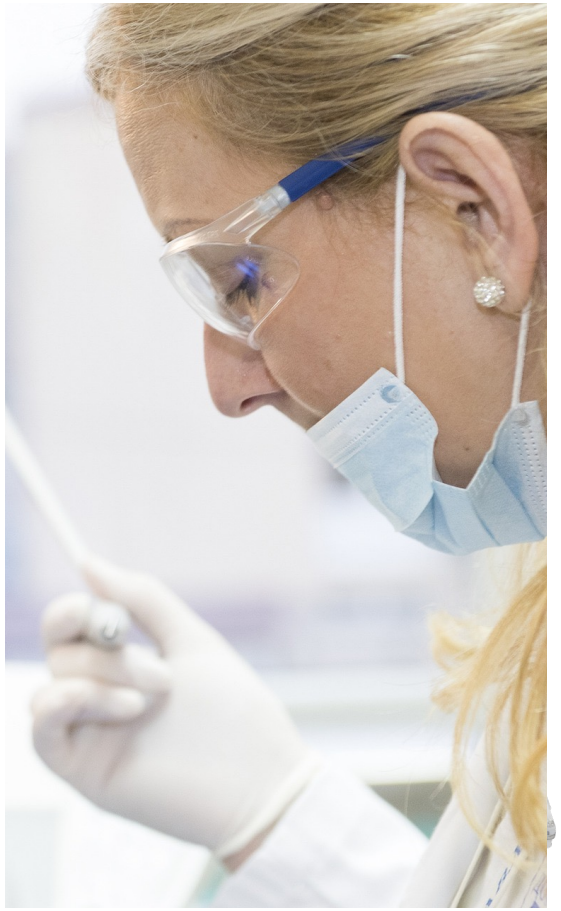
Toddler babbling a gibberish

TRAIN
Let it read a voluminous amount of texts



High-school student with adequate language proficiency

FINE-TUNE
Train it to do a set of specific tasks



Intelligent chatbot

1) Model:
Transformer Model

2) Language Model:
GPT, BERT, LLaMa, etc.

3) Dataset:
Instructions and Conversations

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 indiginous people of New Zealand?

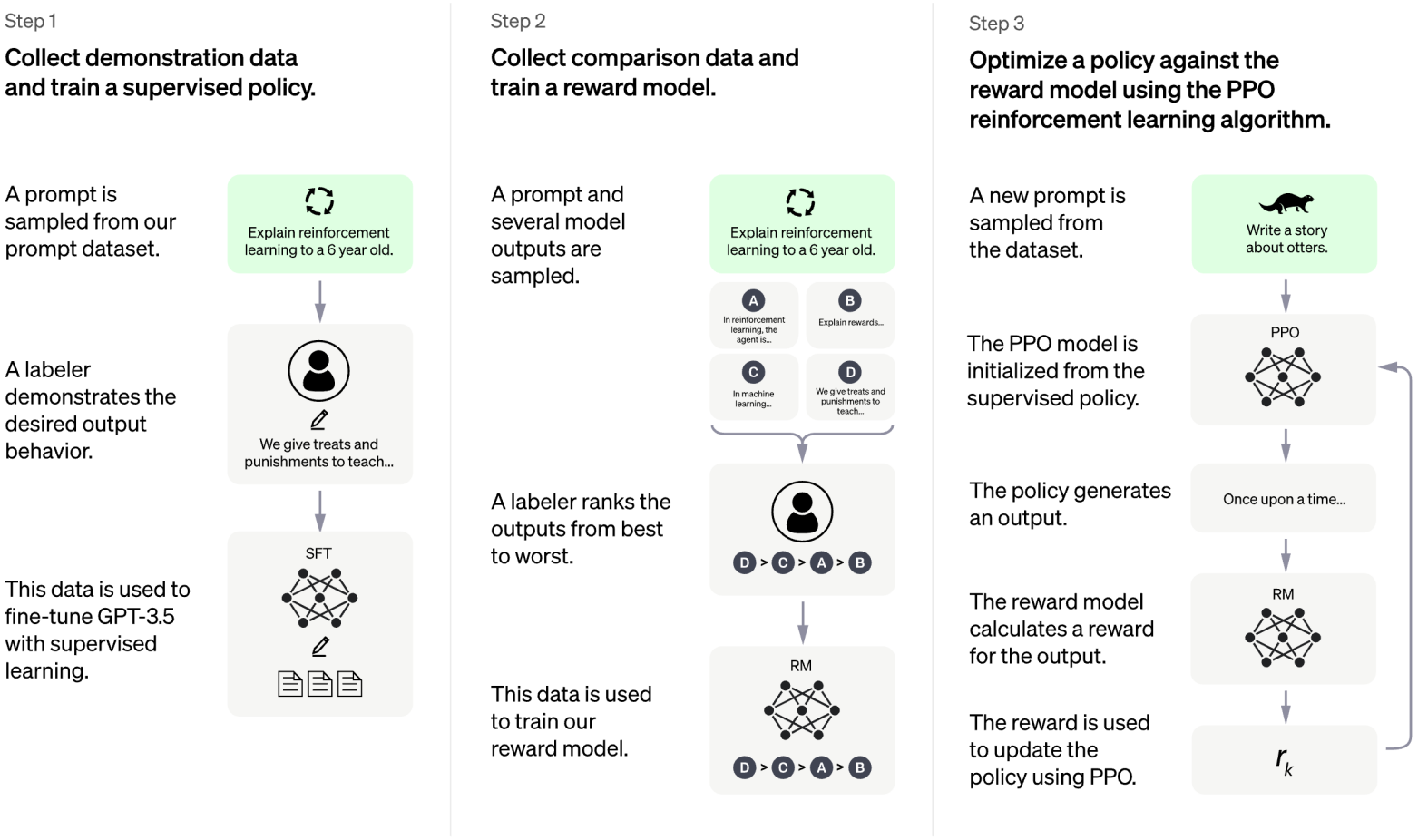
User Prompts and Chats (Ouyang et al., 2022)

Use Case	Example
chat	The following is a conversation with an AI assistant. The assistant is helpful, creative, clever, and very friendly. 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:
chat	Marv is a chatbot that reluctantly answers questions with sarcastic responses: 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:
chat	This is a conversation with an enlightened Buddha. Every response is full of wisdom and love. Me: How can I achieve greater peace and equanimity? Buddha:

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

Use Case	Example
classification	The following is a list of companies and the categories they fall into: Apple, Facebook, Fedex Apple Category: Technology Facebook Category: Social Media Fedex Category:
extract	Text: {text} Keywords:
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)

5. OpenThaiGPT

Training Datasets

Pretraining datasets

Name	Source Type	Access Type	token size (GPT Thai)	size (GB)
mC4	Web Crawl	Public	16B	56 GB
OSCAR 2019	Web Crawl	Public	3.4B	16GB
OSCAR 2021	Web Crawl	Public	3.8B	16GB
OSCAR 2022	Web Crawl	Public	10.8B	60GB
OSCAR 2023	Web Crawl	Public, Access needed	17B	100GB
CC100	Web Crawl	Public	12B	80GB
LST20	News	Public, Access needed	3M (Compound Word)	
Twitter	Social	P Conan, Private	560M	
Linetoday	News	P Conan, Private	105M	
TraffyFondue	Complaint	P Conan, Private	5M	
Wikipedia	Wiki	Public	10M	
prachathai67k	News	Public	160M	
ThaiPBS	News	Public		
Scb-th-en(extract th)	MISC	Public	50.3M	
wisesight_sentiment	Social	Public	0.978M	
Wongnai_reviews	Social	Public	11.1M	
ThaiRath		Public		
Best		Public, Access needed		

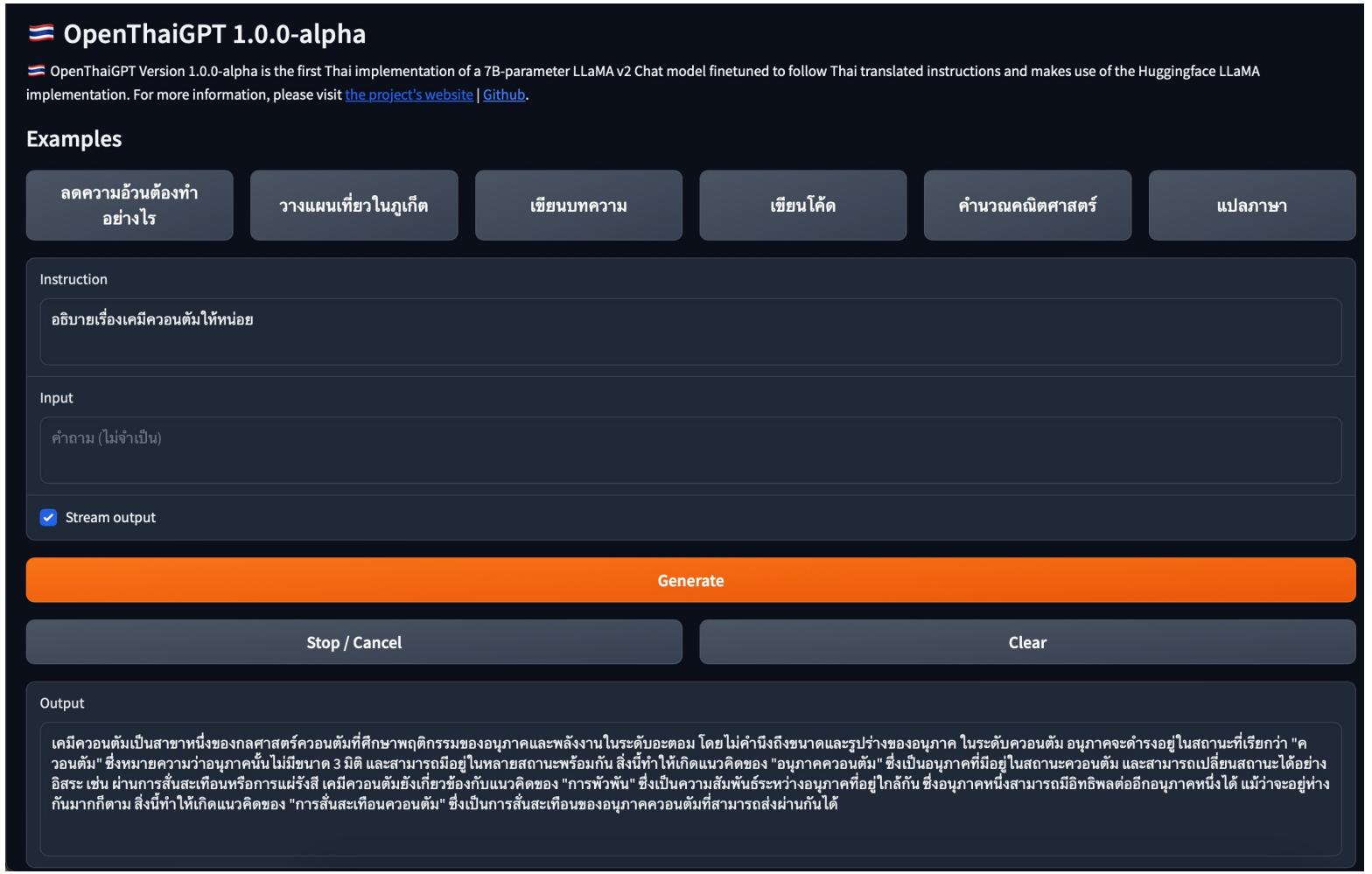
- OpenThaiGPT was trained on the dataset consisting of over 2 trillion tokens (sub-words): CC100, OSCAR, and mC4
- After cleansing, decontamination, and anonymization, we obtained 37.3 billion tokens
- Pantip.com supplied additional 20.0 billion tokens of social media datasets
- In total, we obtained 60 billion tokens for training

LLM Project @ NECTEC

- It simulates how humans string words to become a meaningful sentence
- Trained models:
  - **LLaMa (GPT)** and **Mistral** for language generation
  - **ELECTRA** for document classification
  - **BART** for machine translation and document summarization
  - **DeBERTa** for classification and zero-shot learning
  - **Multimodality:** text, speech, images, videos
- Making Thai one of the pivot languages for LLM via knowledge distillation and model grafting

OpenThaiGPT 1.0.0

Available on https://openthaigpt.openservice.in.th and Hugging Face



- **Base:** LLaMa v.2
- **Sizes:** 7 and 13 billion parameters (neurons)
- 70B parameters will be released in Feb 2024
- Natural word-level generation
- BPE tokenizer with 20,000 Thai frequent tokens
- Launched on AI4Thai

🤗 Open LLM Leaderboard

📐 The 🤗 Open LLM Leaderboard aims to track, rank and evaluate open LLMs and chatbots.

🤗 Submit a model for automated evaluation on the 🤗 GPU cluster on the "Submit" page! The leaderboard's backend runs the great Eleuther AI Language Model Evaluation Harness - read more details in the "About" page!

Archived on 14 SEP 2023

🏅 LLM Benchmark | 📝 About | 🚀 Submit here!

Select columns to show

- [x] Average ⬆️
- [x] ARC
- [x] HellaSwag
- [x] MMLU
- [x] TruthfulQA
- [ ] Type
- [ ] Precision
- [ ] Hub License
- [ ] #Params (B)
- [ ] Hub ❤️
- [ ] Model sha

- [x] Show gated/private/deleted models

thai

Model types

- [x] 🟢 pretrained
- [x] 🔶 fine-tuned
- [x] ⭕ instruction-tuned
- [x] 🟦 RL-tuned

Model sizes

- [x] Unknown
- [x] < 1.5B
- [x] ~3B
- [x] ~7B
- [x] ~13B
- [x] ~35B
- [x] 60B+

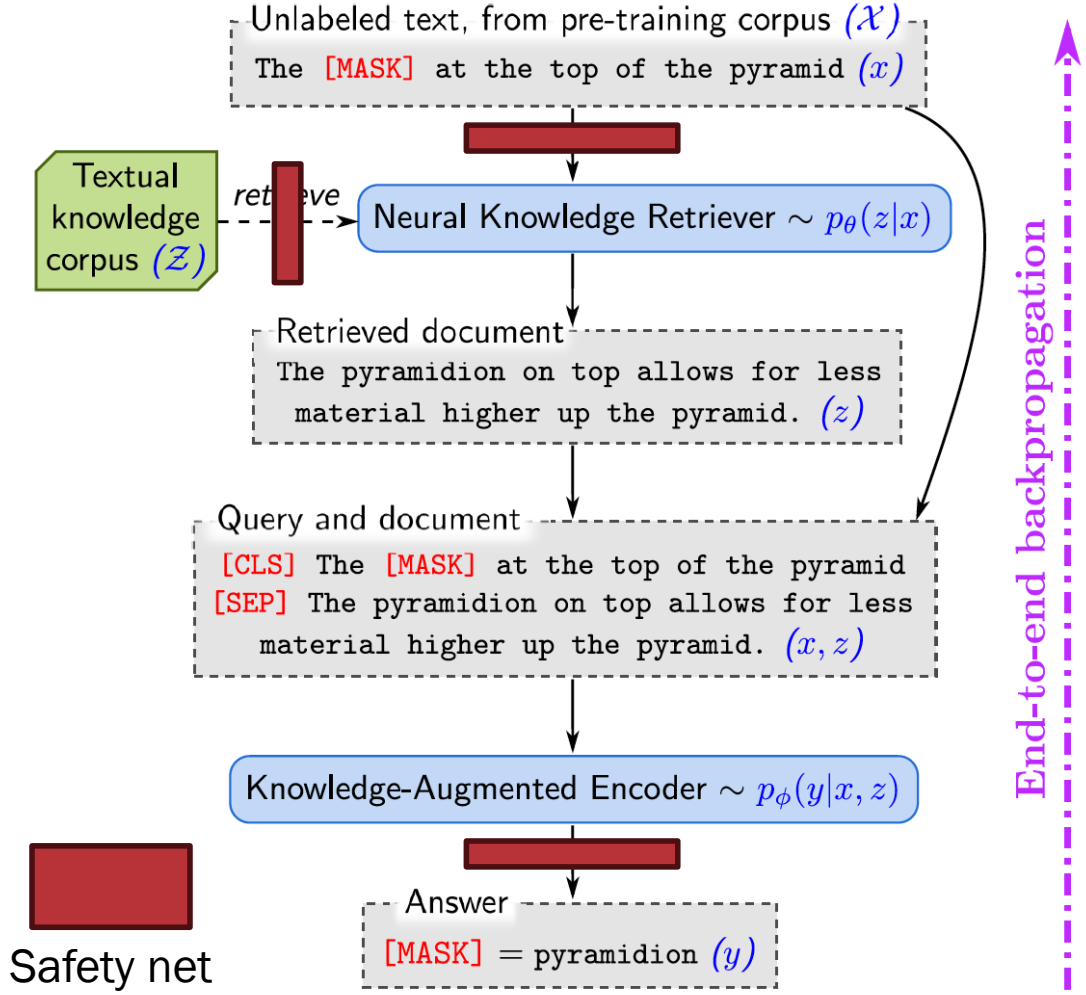
T	Model	Average ⬆️	ARC	HellaSwag	MMLU	TruthfulQA
⭕	openthaigpt/openthaigpt-1.0.0-alpha-7b-chat-ckpt-hf	53.25	50.85	74.89	40.02	47.23
🔶	wannaphong/openthaigpt-0.1.0-beta-full-model_for_open_llm_leaderboard	51.3	51.28	77.46	33.18	43.28
🔶	pythainlp/wangchanglm-7.5B-sft-en-sharded	38.7	34.47	59.81	26.37	34.15
🔶	pythainlp/wangchanglm-7.5B-sft-enth	38.05	33.79	58.99	24.52	34.9

Capabilities

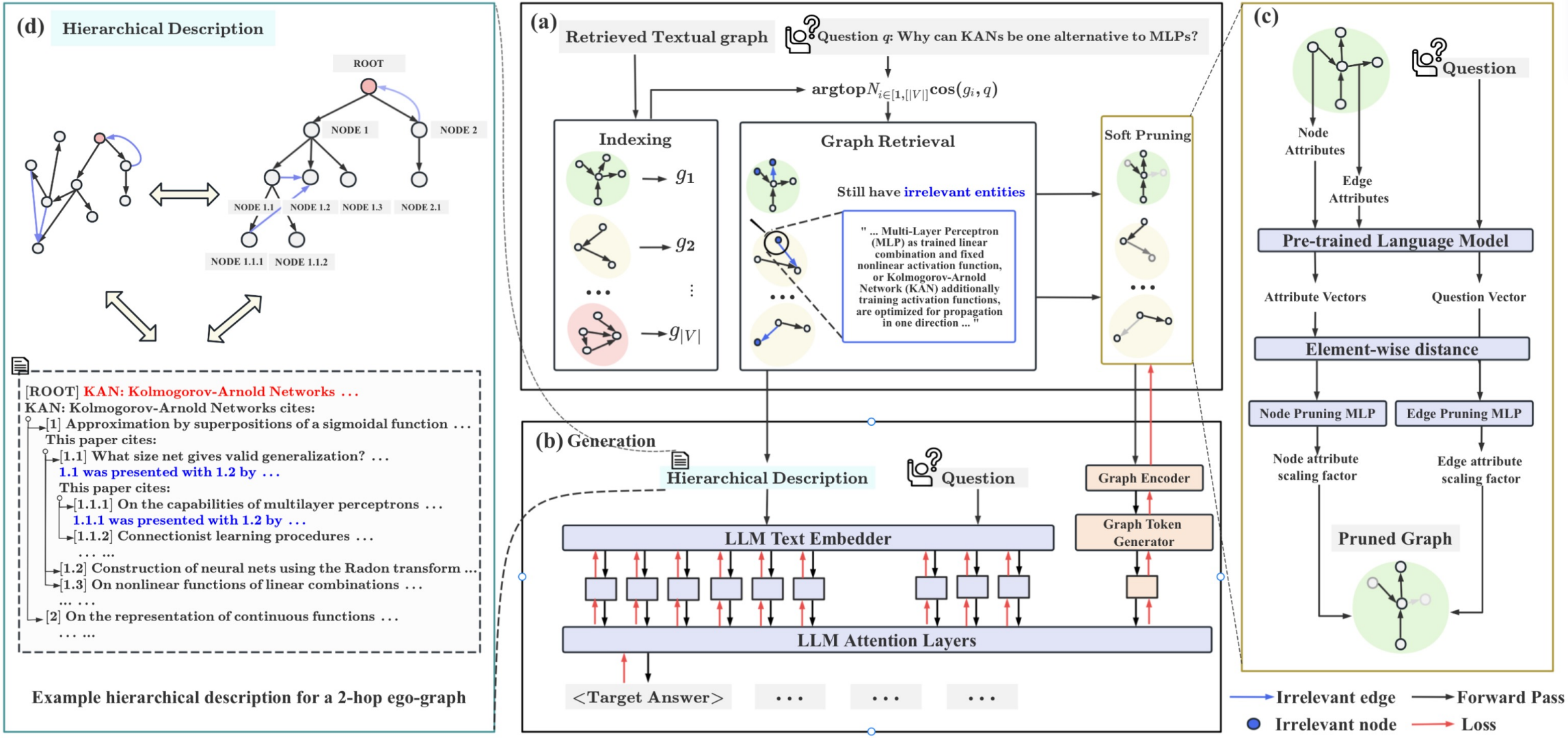
- Question answering from Wikipedia
- Trip planning
- Explaining steps and formulas
- Writing short descriptions
- Writing short pieces of code
- Solving math problems described in natural language
- Translation from/to Thai

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



RAG with Knowledge Graph



World Arena of LLMs

Multilingual (English)

Google
BERT (340M)
GLaM (1.2T)
LaMDA (137B)
PaLM (540B)
Minerva (540B)
PaLM-2 (340B)

Microsoft
Megatron-Turing (530B)

OpenAI
GPT-2 (1.5B)
GPT-3.5 (175B)
GPT-4 (1T)

DeepMind
Gopher (280B)
Chinchilla (70B)

Meta
OPT (175B)
Galactica (120B)
LLaMa (65B)
LLaMa-2 (70B)

Bloomberg
BloombergGPT (50B)

Chinese

Baidu
ERNIE (260B)

HUAWEI
PanGu-Σ (1.085T)

Thai

EAT AiAT ARTIFICIAL INTELLIGENCE ASSOCIATION OF THAILAND
NECTEC a member of NSTDA
OpenThaiGPT (7B)
equivalent to GPT-3
OpenThaiGPT (13B)
equivalent to GPT-3.5

VISTEC VIDYASIRIMEDHI INSTITUTE OF SCIENCE AND TECHNOLOGY
WangchanBERTa
WangchanGLM

SCB 10X
Typhoon (7B)

World Arena of LLMs

Multilingual (English)

Google
BERT (340M)
GLaM (1.2T)
LaMDA (137B)
PaLM (540B)
Minerva (540B)
PaLM-2 (340B)

Microsoft
Megatron-Turing (530B)

OpenAI
GPT-2 (1.5B)
GPT-3.5 (175B)
GPT-4 (1T)

DeepMind
Gopher (280B)
Chinchilla (70B)

Meta
OPT (175B)
Galactica (120B)
LLaMa (65B)
LLaMa-2 (70B)

Bloomberg
BloombergGPT (50B)

Southeast Asian Languages

Alibaba Group 阿里巴巴集团
SEA-LLM (7B, 13B)

AI SINGAPORE
SEA-LION (7B)

Thai

EAT AiAT ARTIFICIAL INTELLIGENCE ASSOCIATION OF THAILAND
NECTEC a member of NSTDA
OpenThaiGPT (7B)
equivalent to GPT-3
OpenThaiGPT (13B)
equivalent to GPT-3.5

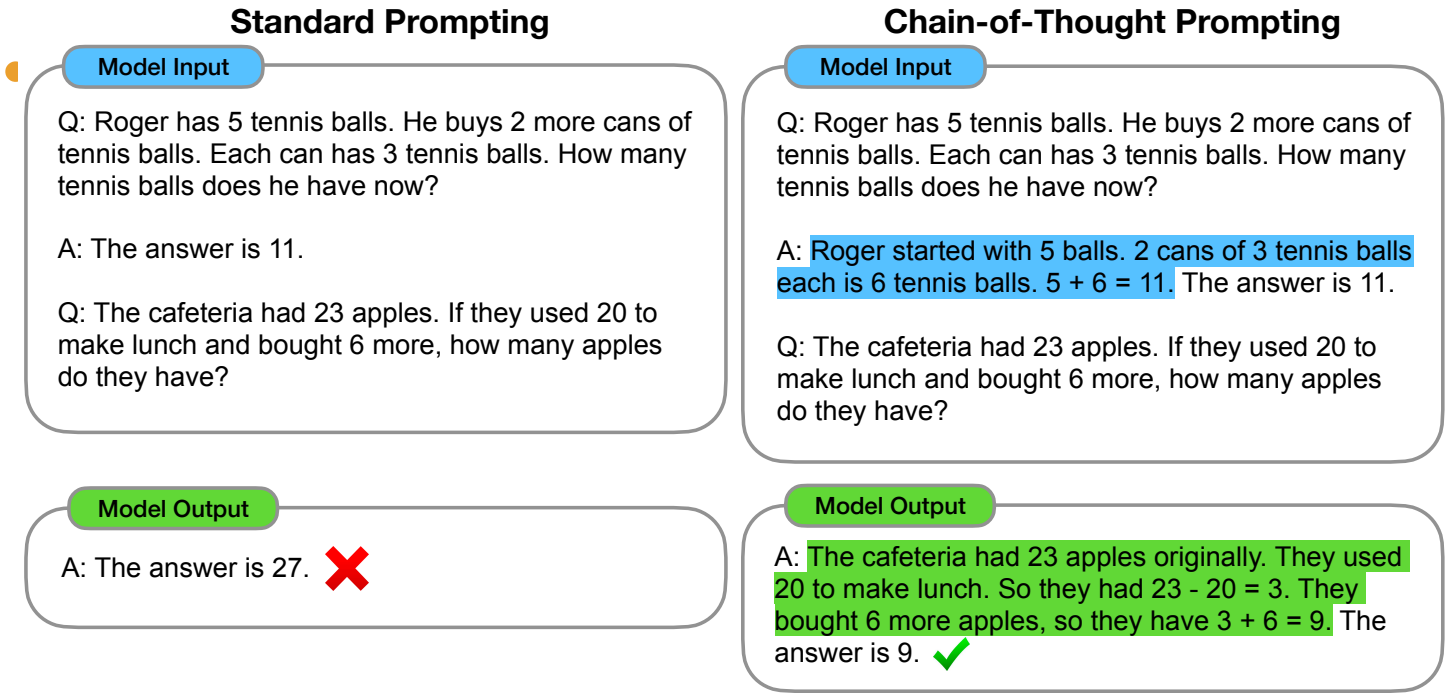
VISTEC VIDYASIRIMEDHI INSTITUTE OF SCIENCE AND TECHNOLOGY
WangchanBERTa
WangchanGLM

SCB 10X
Typhoon (7B)

6. Future Direction

Future: PaLM (Wei et al., 2022)

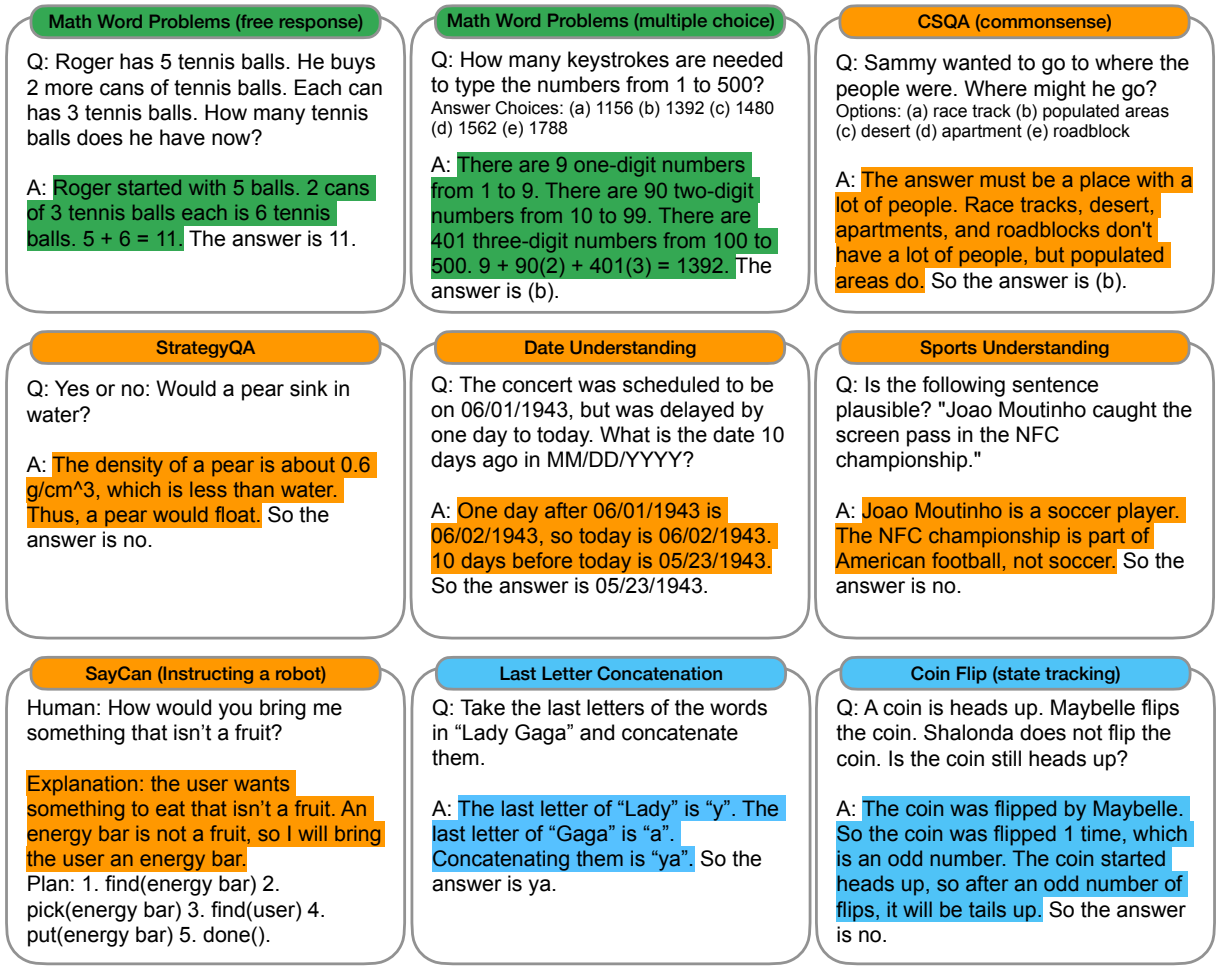
- Pathways Language Model
  - 540B parameters trained with Pathways Transformer model (single model capable of doing many tasks)



- Chain of thoughts 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



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

- Embodied multimodal PaLM with 562B parameters

Mobile Manipulation

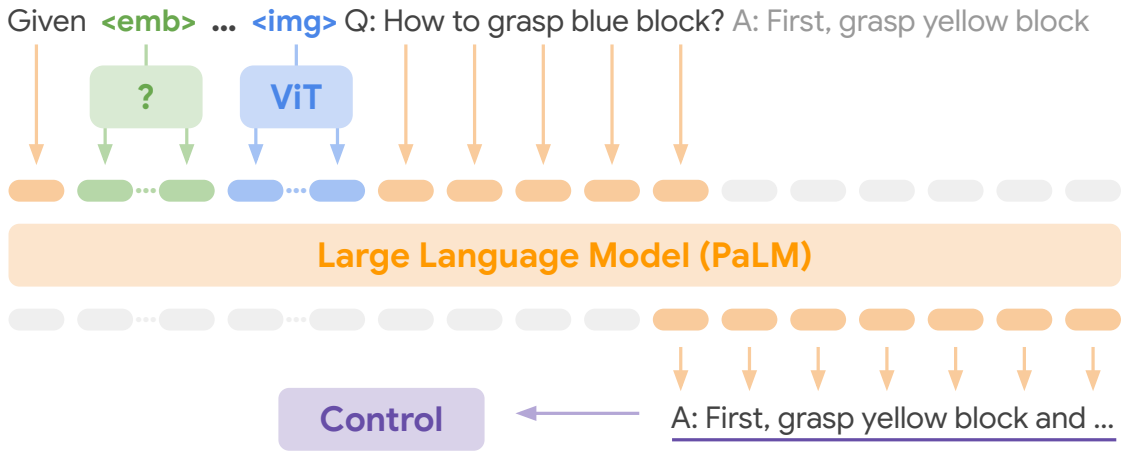

Human: Bring me the rice chips from the drawer. Robot: 1. Go to the drawers, 2. Open top drawer. I see <img>. 3. Pick the green rice chip bag from the drawer and place it on the counter.

Visual Q&A, Captioning ...

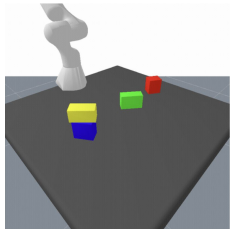
Given <img>. Q: What's in the image? Answer in emojis. A: 🍏🍌🍇🍐🍑🍈🍒.

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

PaLM-E: An Embodied Multimodal Language Model

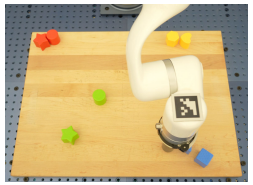


Task and Motion Planning



Given <emb> 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 <img> 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 x 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

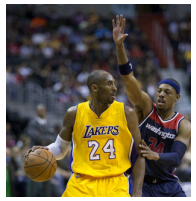
I see (image omitted). Description: A cow sitting on a rug. Joke: Why is the cow sitting on a rug? Because it wants to be a cow-ch! ... I see **<img>**. Description: a donkey is carrying a dog, cat, and rooster. Joke: what do you call a donkey with a rooster on his back. A rooster booster.

Robot Visual Perception, Dialogue, and Planning

Robot: I am a robot operating in a kitchen. Given **<img>**, 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.

Zero-shot Multimodal Chain-of-Thought

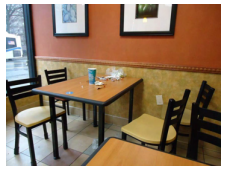
Given **<img>**. Q: Can I go down this street on a bicycle, yes or no? A: Let's think step by step. 1. do not enter. 2. except bicycles. 3. do not entry except bicycles. 4. yes.



Given **<img>**. 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 **<img>**. 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 Initiative



Given **<img>**. 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.

Physical Prediction

Given **<img>** . Q: What will the robot do next? A: fall.

Spatial Grounding

Given **<img>** . 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.

Zero-shot: Multi-image Relationships

Q: Where is **<img1>** in this scene **<img2>**? A: a stove top with a pot on it.

OCR-free Math Reasoning

Given **<img>** . 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.

End-to-End Egocentric Q&A, Dialogue

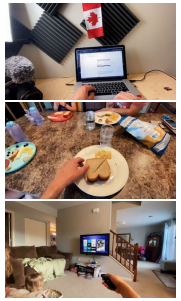
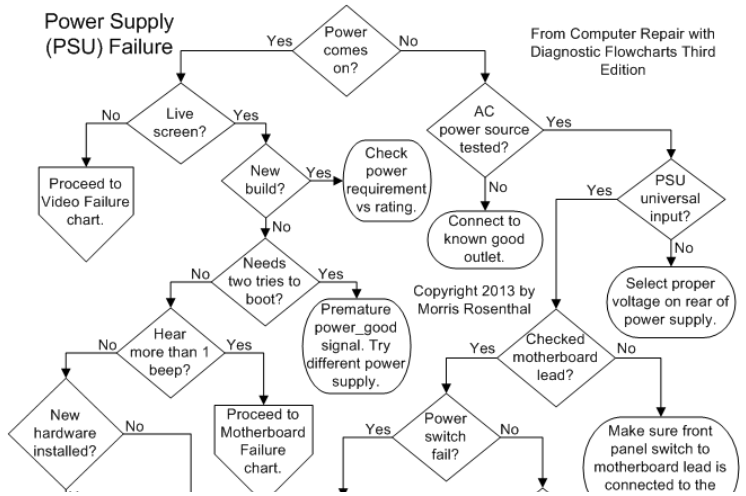
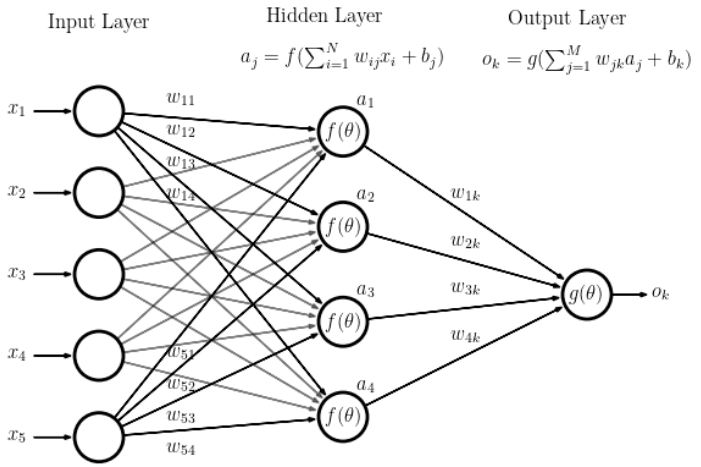


Photo 1, at 10:30 am: **<img1>**. Photo 2, at 12:45 pm: **<img2>**. Photo 3, at 3:45 pm: **<img3>**. Q: I forget, what did I have for lunch, and what time was it? A: Let's think step by step. 1. you had a sandwich for lunch. 2. it was 12:45 pm.

Quō Vādis, AI? (Where are you going to, AI?)

- Three waves of AI





Wave 1: DESCRIBE

Handcrafted knowledge and expert systems

https://stackoverflow.com/questions/14092782/how-to-save-a-decision-flow-chart-as-xml

Wave 2: CATEGORIZE

Blackbox AI, statistical learning, and neural networks

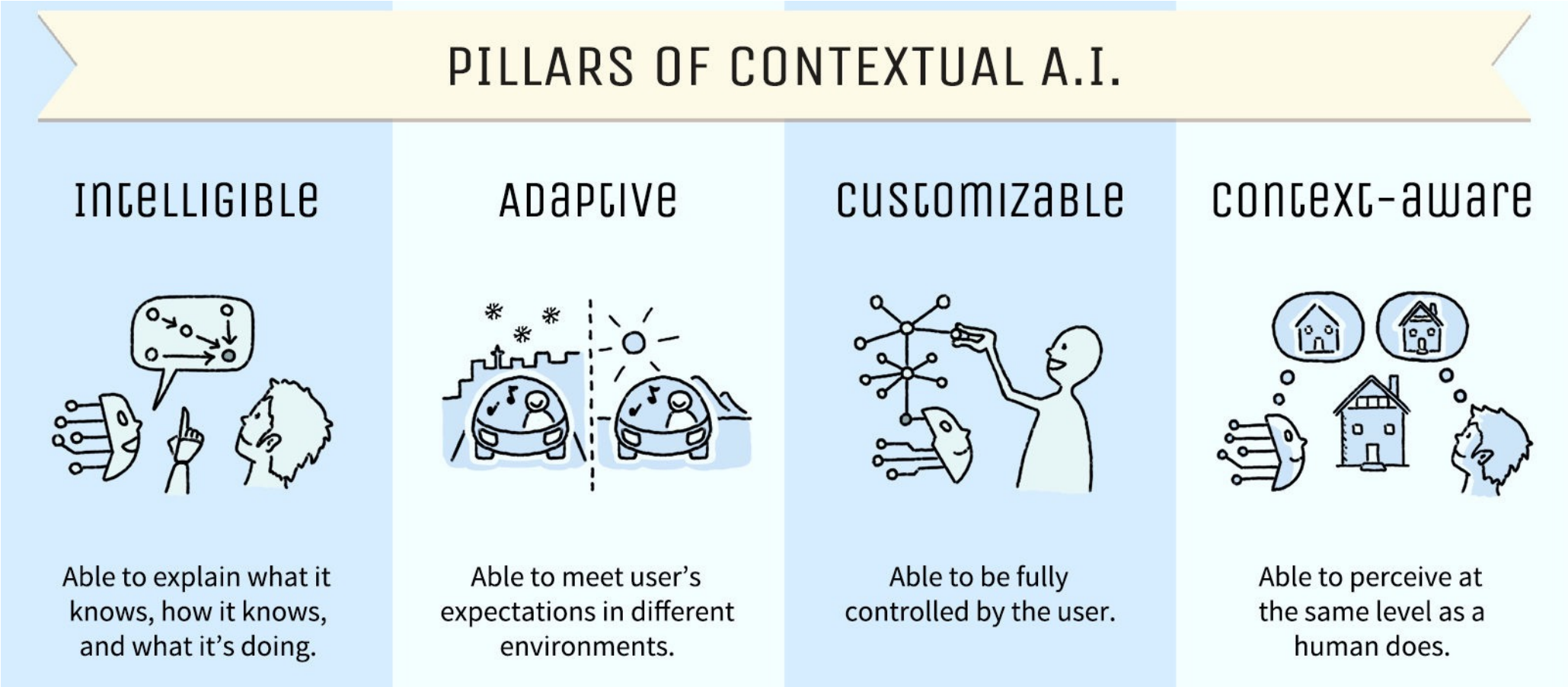
http://www.astroml.org/book_figures/appendix/fig_neural_network.html

Wave 3: EXPLAIN

Whitebox AI, context-aware adaptation, and knowledge graphs

https://www.linkedin.com/pulse/explainable-ai-worthwhile-george-bara/

Contextual AI



Credit: `https://theblog.adobe.com/contextual-ai-the-next-frontier-of-artificial-intelligence/`

Challenges

- AI hallucination
  - Generated contents may be incorrect due to confabulation (connecting unrelated stories)
  - Distinction between **hallucination** (recognized mistake) vs. **creativity**
- Unethical uses of Generative AI
  - Plagiarism and abuses in learning
  - Generating fake news and misinformation
  - Leaks of personal information

Types of Hallucination

- Out-of-context generation
- Contradiction to previously generated contents
- Misinformation generation

Thank You

prachya.boonkwan@nectec.or.th

kaamanita@gmail.com