

VISTEC
VIDYASIRIMEDHI
INSTITUTE OF SCIENCE AND TECHNOLOGY

WangchanX: Building Foundation Models for ThaiNLP

Can Udomcharoenchaikit

School of Information Science and Technology

VISTEC

12 Sep 2023

Outline

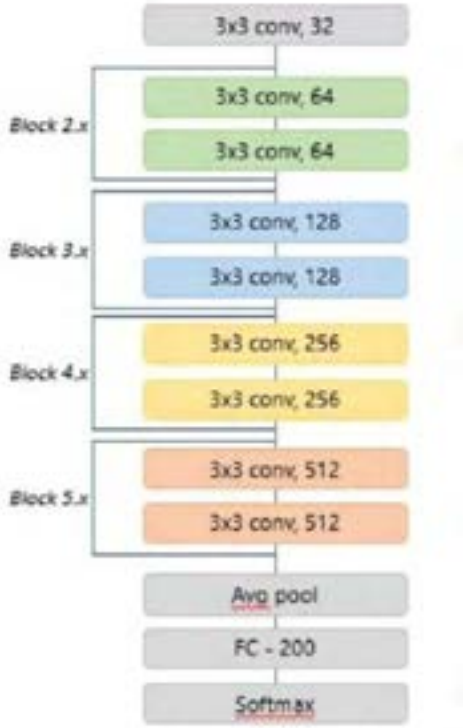
- Modern NLP Paradigm: Pretraining → Finetuning
- Transformers
- AIResearch 2023: Towards Thai Language Understanding and Generation

Modern NLP paradigm: Pretraining + Finetuning (Transfer Learning)

Pretrained images

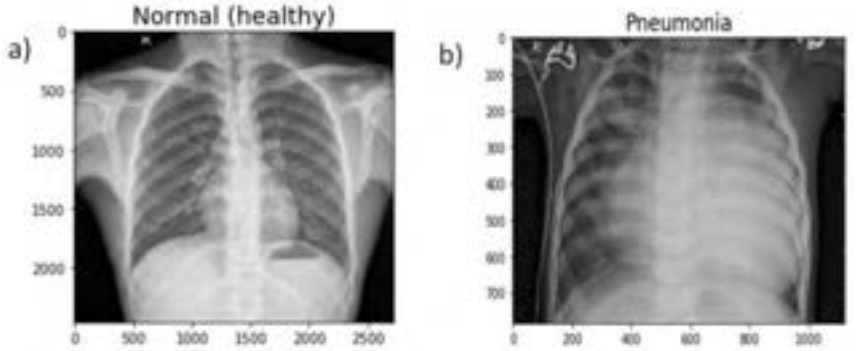
Sample Image from Tiny ImageNet (200 categories)

Figure from [Lei Sun]



ResNet

Images for target task



Chest Xray of (a) a healthy person and (b) a person suffering from pneumonia.

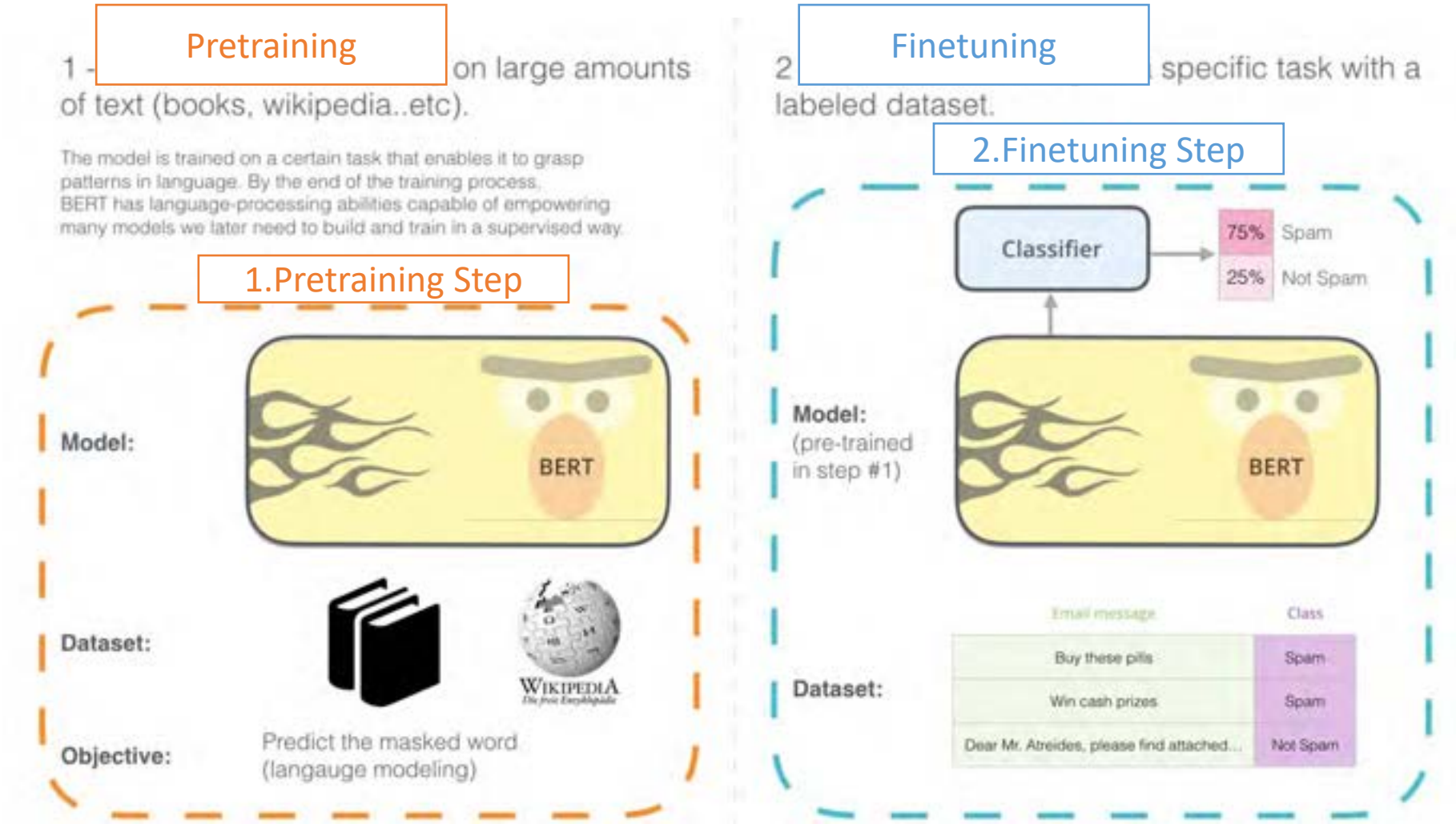
Figure from [Hashmi et al., 2020]

VISTEC-depa THAILAND AI RESEARCH INSTITUTE

Pretraining

Finetuning

Modern NLP paradigm: Pretraining + Finetuning



- Learn linguistic features from large text corpora (e.g., encyclopedia, new, spoken dialogs, web-crawled data) to solve downstream NLP tasks
- Pre-train model once and then finetune on downstream tasks (transfer learning)
- Reduce training time when finetuning on downstream tasks

[The Illustrated BERT, ELMo, and co. (How NLP Cracked Transfer Learning) – Jay Alammar – Visualizing machine learning one concept at a time. (jalammar.github.io)]

VISTEC-depa THAILAND AI RESEARCH INSTITUTE

NLP Model pretraining: Language Modeling

- A probabilistic model of a natural language
  - Input: a sequence of words
  - Output: probability of the next word (relationship between the next words and its **context**)

$$P(w_m \mid w_1, \ldots, w_{m-1})$$

  - John works in a hospital. He is a ____
    - Doctor: 10%
    - Nurse: 8%
    - Phatmacist: 5%
    - Janitor: 3%
    - ……

VISTEC-depa THAILAND AI RESEARCH INSTITUTE

Transformers: Language Models

- Most transformers are trained as language models.

flax-community/gpt2-base-thai

Text Generation · Transformers · PyTorch · JAX

Hosted inference API

Text Generation — Examples

ฉันชื่อแคนและฉันชอบ

Compute ⌘+Enter

JSON Output — Maximize

Dataset used to train flax-community/gpt2-base-thai

Hosted inference API

Text Generation — Examples

ฉันชื่อแคนและฉันชอบที่เธอจะได้จับมือทำเป็นรูปร่างที่น่ารักในรูปภาพของเธอ ฉันพยายามอย่างดีที่ส

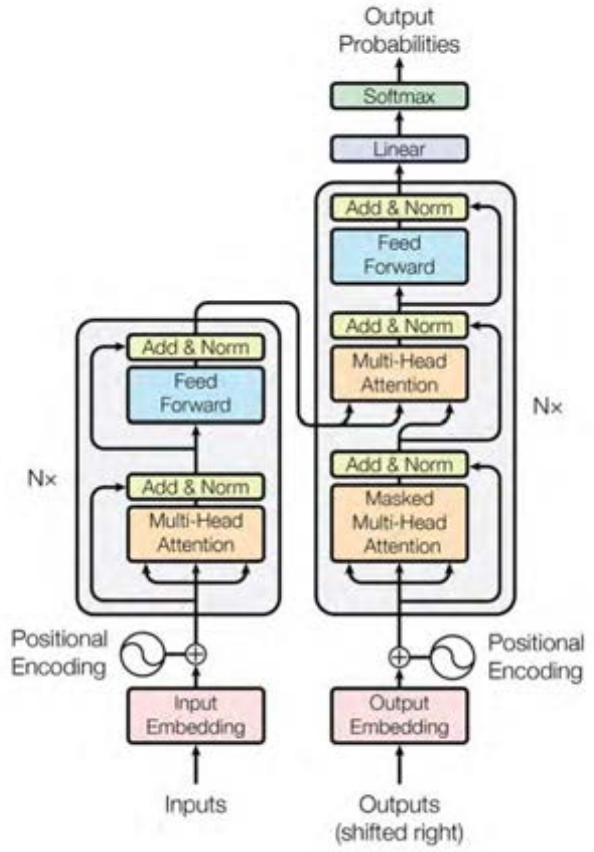
Compute ⌘+Enter

JSON Output — Maximize

Dataset used to train flax-community/gpt2-base-thai

$$P(X) = \prod_{t=1}^{n} P(x_t \mid x_{<t})$$

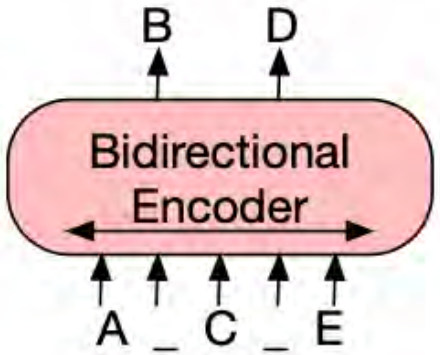
Next word prediction!



VISTEC-depa THAILAND AI RESEARCH INSTITUTE

Transformers: Language Models

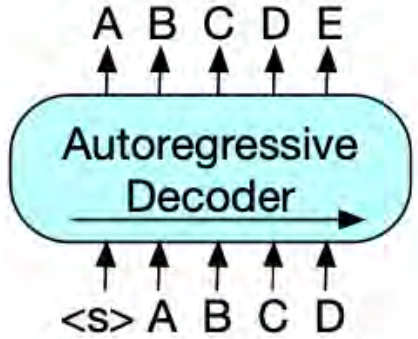
$$P(X) = \prod_{t=1}^{n} P(x_t \mid x_{\neq t})$$



BERT

(Encoder-only)

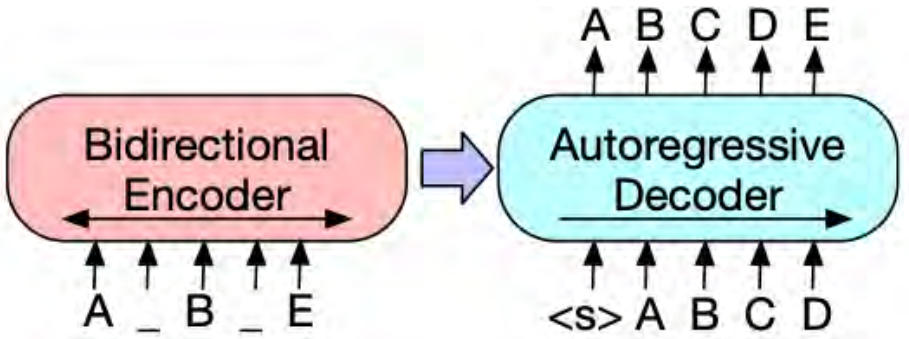
$$P(X) = \prod_{t=1}^{n} P(x_t \mid x_{<t})$$



GPT

(Decoder-only)

$$P_\theta(\boldsymbol{Y} \mid \boldsymbol{X}) = \prod_{t=1}^{m} P(y_t \mid y_{<t}, \boldsymbol{X}, \theta)$$



BART

(Encoder-Decoder)

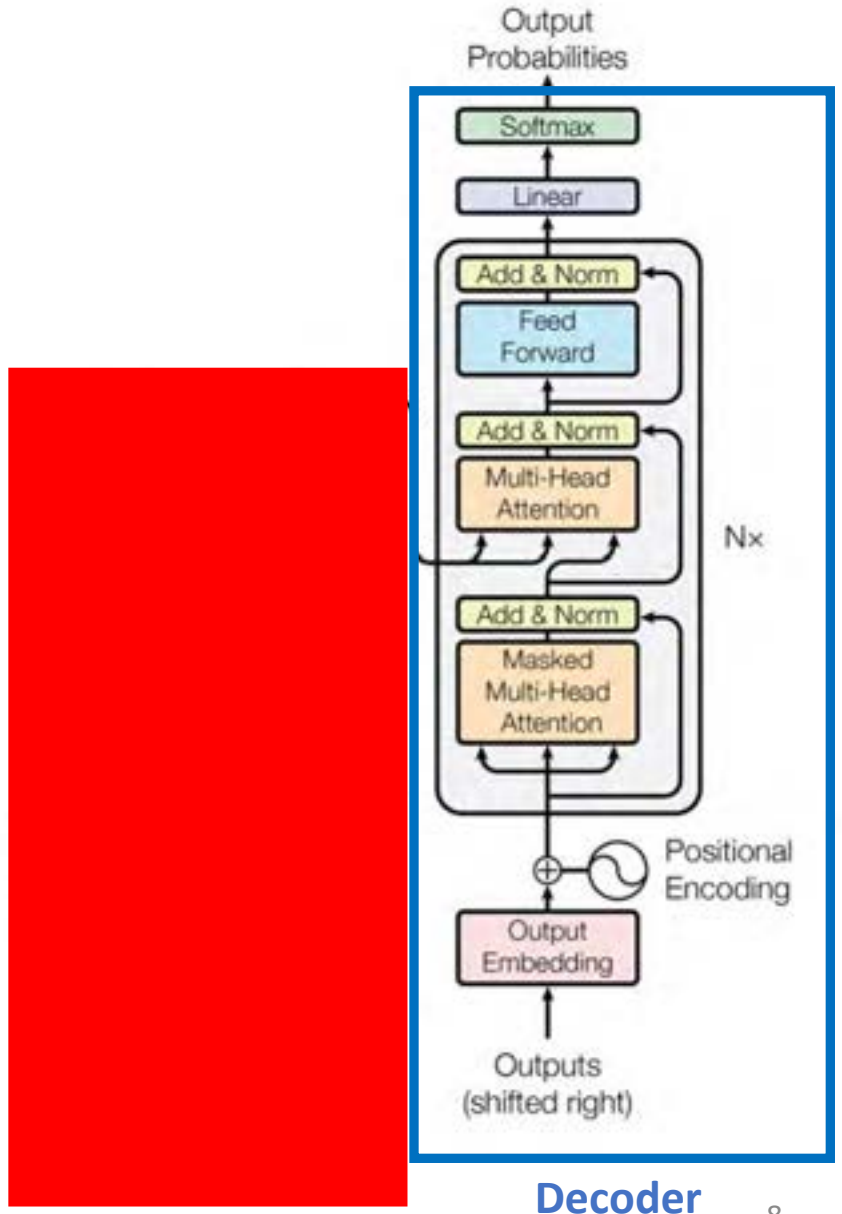
Image Source: [BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension](#) (Lewis et al., ACL 2020)

VISTEC-depa THAILAND AI RESEARCH INSTITUTE

Transformers: decoder-only LM

- Standard language modeling scheme
  - Next word prediction
  - unidirectional context/attention
  - Generate one token at a time based on previous tokens
  - John works in a hospital. He is a ____
- GPT models are decoder-only models
- Suitable for text generation

$$P(X) = \prod_{t=1}^{n} P(x_t \mid x_{<t})$$



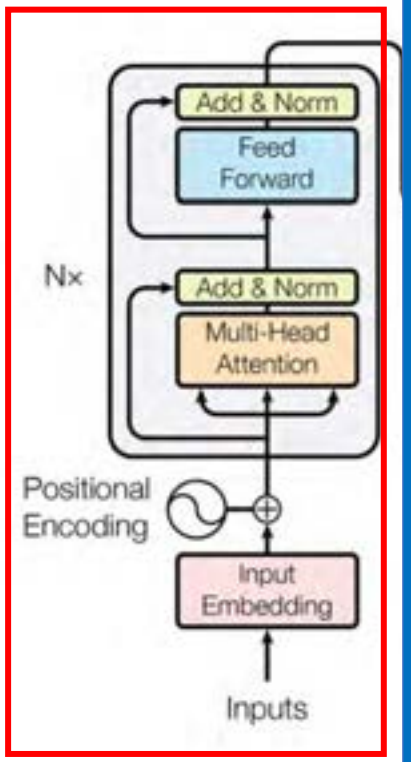
Decoder

Transformers: encoder-only LM

- “Fill-in-the-blank” Mask Language Modeling
  - Predict the masked word (“fill-in-the-blank”)
  - Bidirectional Context/Attention
  - Bob is a __ at Rangsit university.

$$P(X) = \prod_{t=1}^{n} P(x_t \mid x_{\neq t})$$

- Can encode a text input into a vector-representation
- BERT, RoBERTa, WangchanBERTa are encoder-only LMs.
- Suitable for natural language understanding tasks
  - Text classification
  - Text-pair classification
  - Etc.



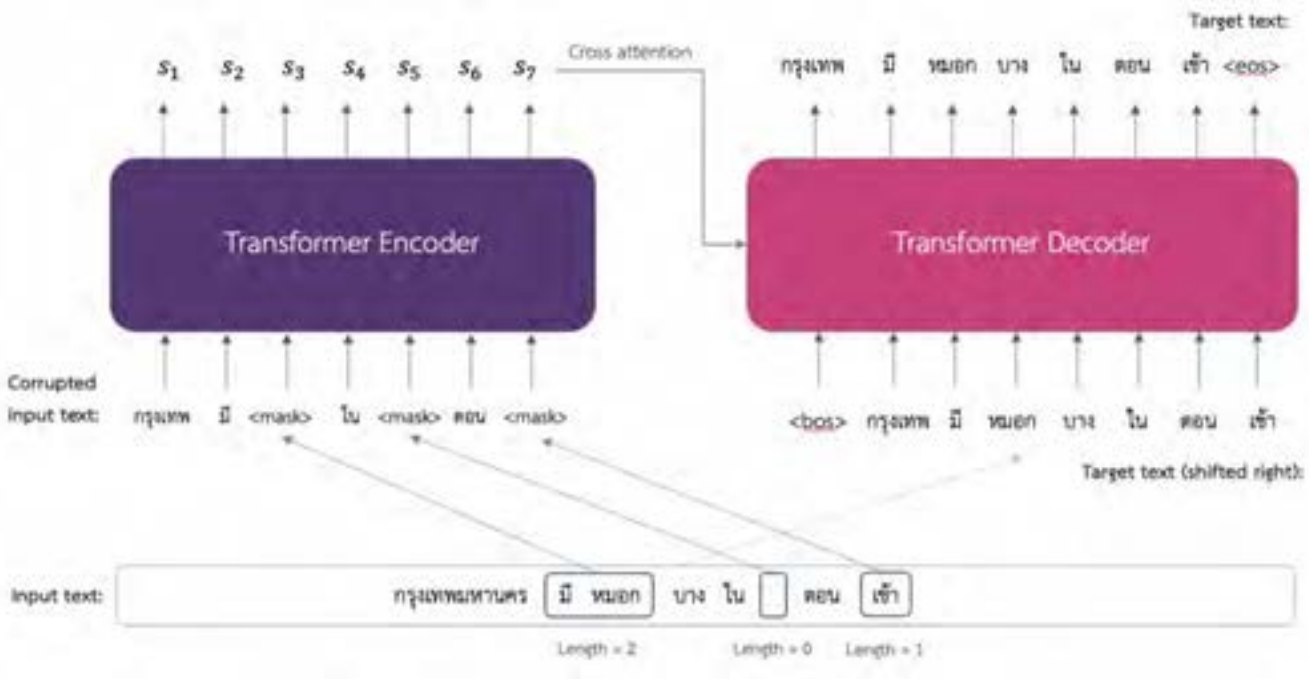
Encoder

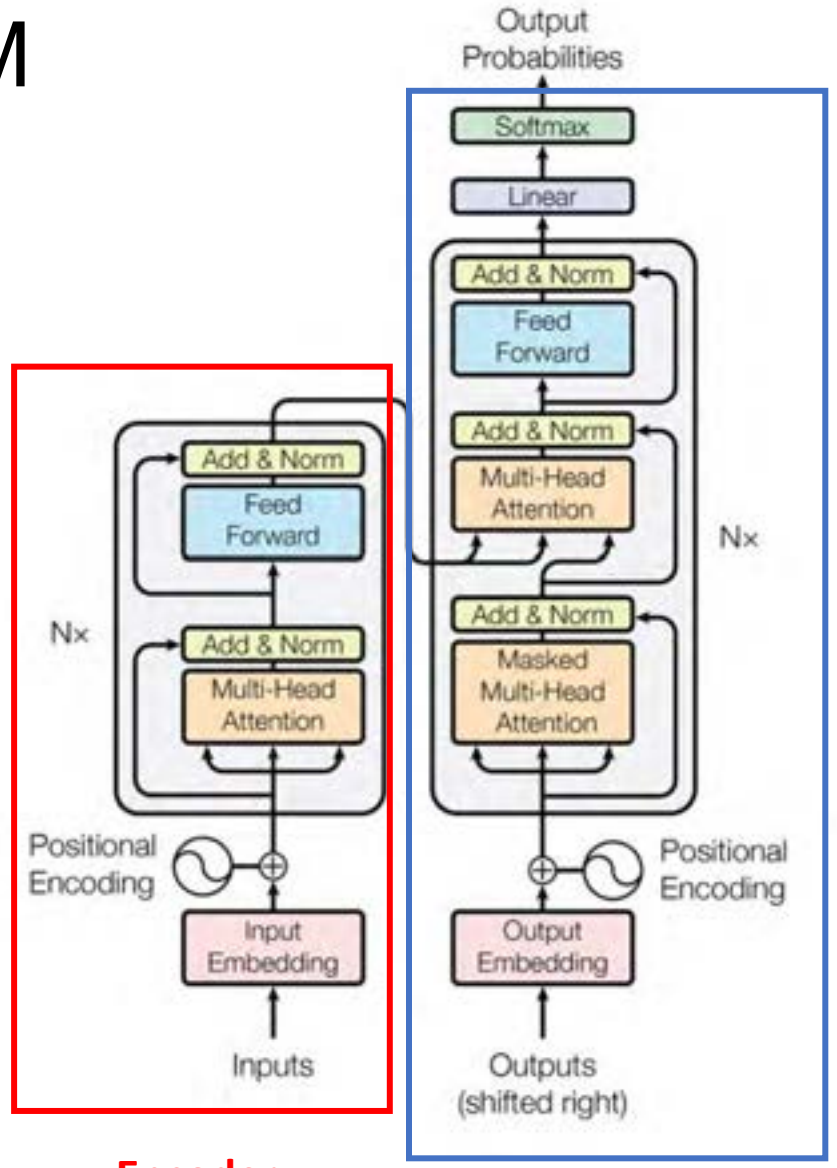
Decoder

VISTEC-depa THAILAND AI RESEARCH INSTITUTE

Transformers: encoder-decoder LM

- Corrupted Text Reconstruction

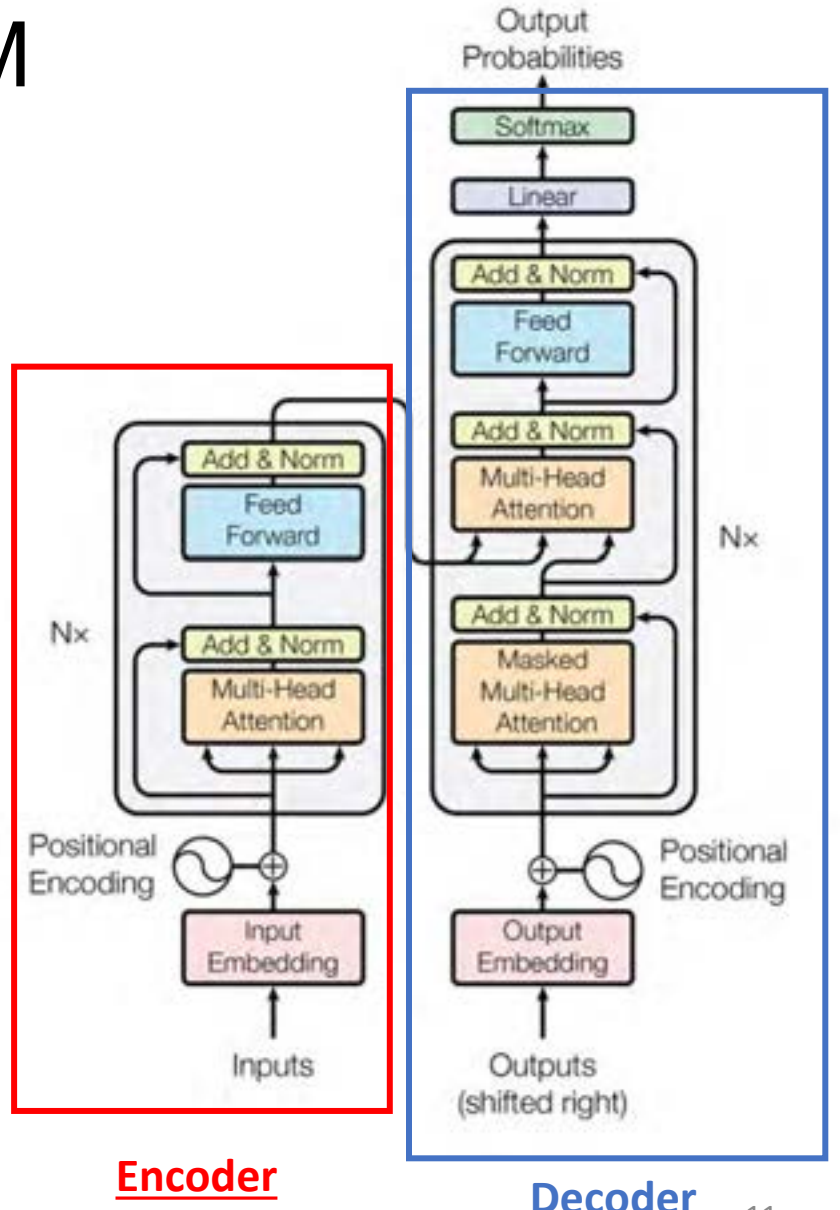




Transformers: encoder-decoder LM

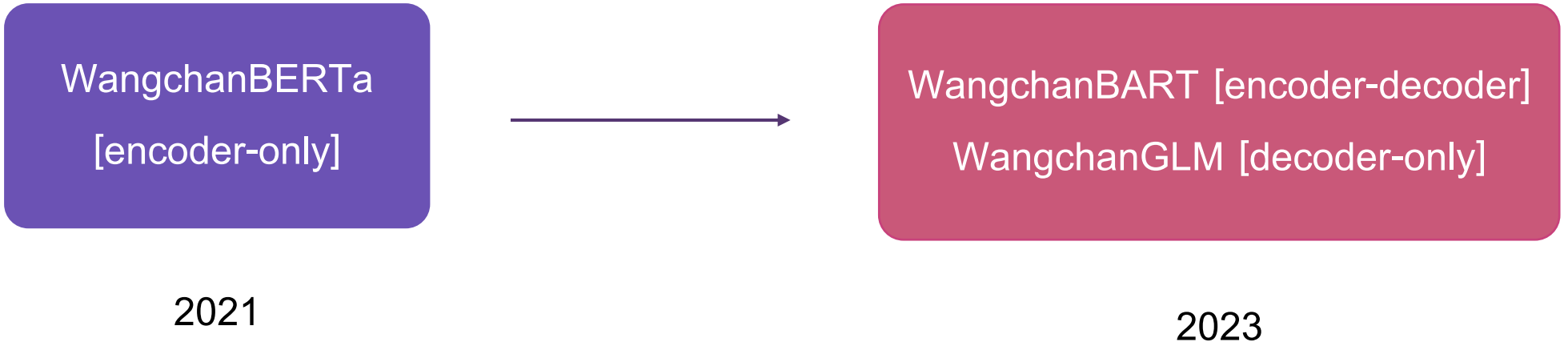
- Corrupted Text Reconstruction
- Best of both world:
  - Encoding: Language understanding
  - Decoding: Text generation
- BART and T5 are encoder-decoder models.
- Suitable for both generation and language understanding tasks

$$P_\theta(\boldsymbol{Y} \mid \boldsymbol{X}) = \prod_{t=1}^{m} P(y_t \mid y_{<t}, \boldsymbol{X}, \theta)$$



VISTEC-depa THAILAND AI RESEARCH INSTITUTE

Towards Thai Language Understanding and Generation



WangchanBART, a pre-trained model for language understanding and generation, is under development and planned to be released in Q4 2023.

VISTEC-depa THAILAND AI RESEARCH INSTITUTE

Towards Thai Language Understanding and Generation

- Provide building blocks for the Thai NLP community
- Encourage the development of applications integrated with Thai NLP
- Increase the capability of Thai NLP

VISTEC-depa THAILAND AI RESEARCH INSTITUTE

Our KIB structure comprises five academic labs, three Research & Development labs, and two deep-tech businesses.

Vertical Integration

B Business Benefits

Chatbots · Call Center Assistants · Personalized Healthcare Applications · Assistant Robots · ...

I Industrial Capability

Sense	Think	Act
Healthcare and Aging Assistant Tech	AI for Everyone	Human-Centric Automation

Wangchan Advanced Industrial Labs: Laboratories for Sense-Think-Act Machines

K Knowledge Foundation

Natural Language Processing · Machine Learning Algorithms · Computer Vision · Robotics · Biosignal Processing · AI Infrastructure · Optimization

Deep Tech Businesses

VISAI

SensAI

Research & Development Labs

INTERFACES INDUSTRIAL RESEARCH

VISTEC-depa THAILAND AI RESEARCH INSTITUTE

ARIA Advanced Robotics and Intelligent Automation

Academic Labs

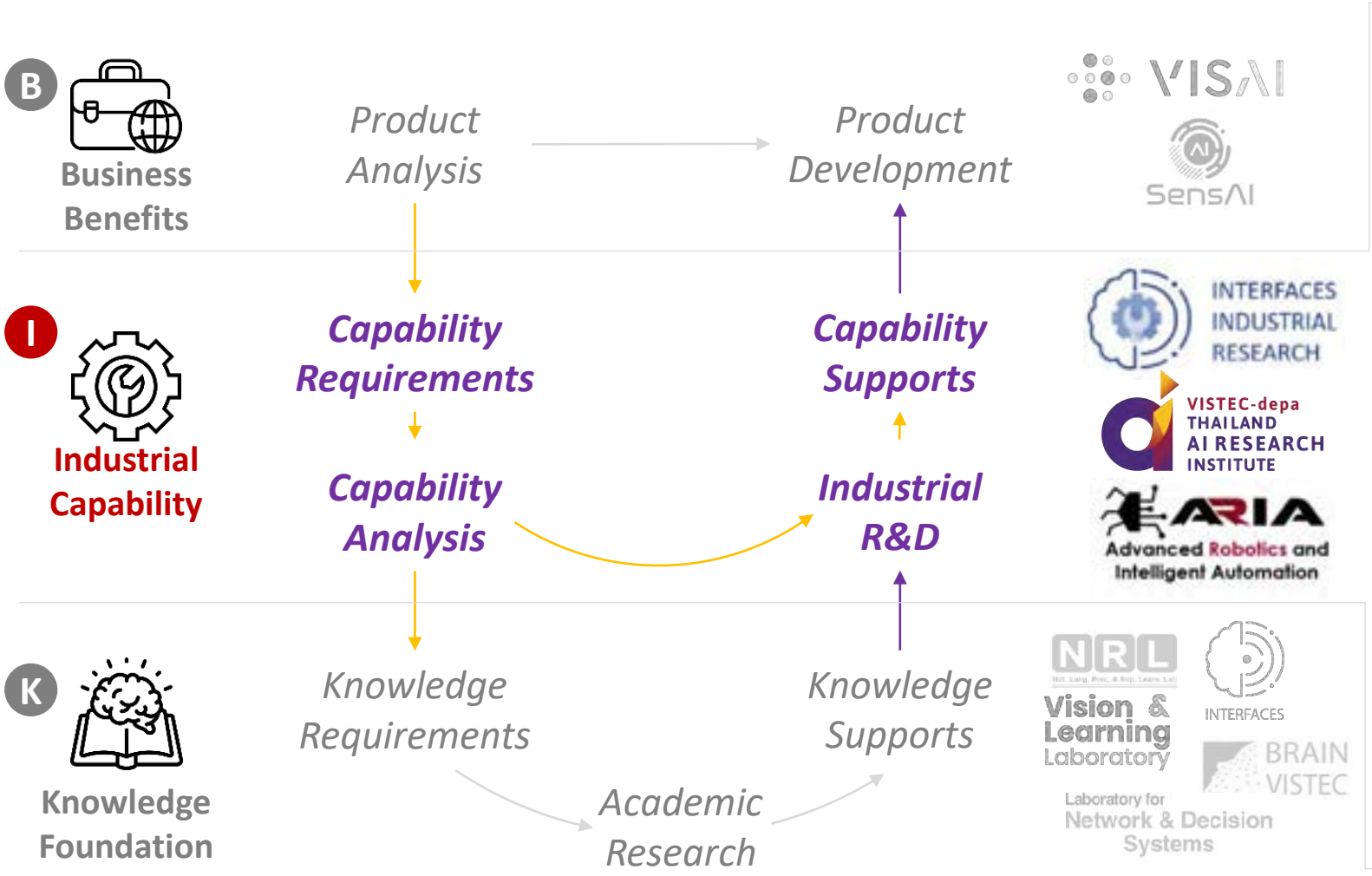
NRL Vision & Learning Laboratory

INTERFACES

BRAIN VISTEC

Laboratory for Network & Decision Systems

In 2022, we established Wangchan Advanced Industrial Labs as a cluster of R&D teams to bridge the gap between business and academic research and function as a platform for university spin-off enterprises.



Wangchan Advanced Industrial Labs

Laboratories for Sense-Think-Act Machines

Program S: Sense

Healthcare and Aging Assistant Tech: Developing wearable health monitoring devices and algorithms to process bio-signals.

Program T: Think

AI for Everyone: Building opensource natural language processing toolsets and datasets.

Program A: Act

Human-Centric Automation: Developing robots with the ability to navigate human environments and work with individuals safely.

We just concluded Phase 1 of the AI for Everyone program with four datasets and eight models. One of the models, *WangchanBERTa*, is a language model forming the basis for Phase 2.

Phase 1: 2019 - 2023

Open Datasets & Opensource Models

Item	Dataset	Model
Machine Translation (EN-TH)	✓	✓
Speech Emotion Recognition	✓	✓
Language Model: ***WangchanBERTa (1.2m Downloads)***		✓
Word Segmentation	✓	✓
Automatic Speech Recognition		✓
Sentence Embedding		✓
Nested Named Entity Recognition	✓	✓
Cross-lingual Search		✓

VISTEC-depa THAILAND AI RESEARCH INSTITUTE

Charin Polpanumas
PyThaiNLP

Lalita Lowphansirikul
Research Student

Weerayut Buaphet
Research Student

Peerat Limkonchotiwat
Research Student

Wannaphong Phatthiyaphaibun
Research Student

Ekapol Chuangsuwanich
Chulalongkorn University

Attapol Thamrongrattanarit
Chulalongkorn University

Can Udomcharoenchaikit
VISTEC

Sarana Nutanong
VISTEC

WangchanX offers a comprehensive set of features for developers to build NLP models to meet their specific needs.

Phase 2: 2023 - 2026

WangchanX

Opensource Foundation Models & Adaptation Toolsets

Revolutionize how we develop NLP applications in Thailand!

Developers no longer have to pay a premium for access and rely on off-the-shelf API services as their only option.

With WangchanX, developers can access our high-quality open-source foundation models and easy-to-use adaptation pipeline and benchmark datasets to create domain-specific and task-specific models.

Foundation Models	Adaptation Toolsets	Benchmarks
Pretrained Multilingual Large Language Models with enhanced Thai capabilities	Toolsets to adapt a foundation model to a specific domain or task	Comprehensive set of domain-specific benchmark datasets for various NLP tasks.

Available with no commercial restrictions (CC-BY-SA 4.0).

We launched Phase 2 in April. A trial version of *WangChanGLM* was released on April 29, 2023.

Phase 2: 2023 - 2026 *WangchanX*

Opensource Foundation Models & Adaptation Toolsets

Item	
Foundation Models	*WangchanBERTa [encoder-only]* *WangChanGLM [decoder-only]* *WangchanBART [encoder-decoder]* *WangchanUltima*
Adaptation Tools	Full/Partial Finetuning Adaptor Distillation Domain-Specific Models
Benchmarks & Datasets	General Domain Finance Medical Legal

VISTEC-depa THAILAND AI RESEARCH INSTITUTE


Charin Polpanumas
PyThaiNLP


Wannaphong Phatthiyaphaibun
Research Student


Lalita Lowphansirikul
Research Student


Weerayut Buaphet
Research Student


Pakin Siwatammarat
Research Student


Peerat Limkonchotiwat
Research Student


Patomporn Payoungkhamdee
Research Student


Titipat Achakulwisut
Mahidol University


Tara Sipo
Research Assistant


Rachata Ausavarungnirun
TGGS, KMUTNB


Ekapol Chuangsuwanich
Chulalongkorn University


Piyalitt Ittichaiwong
SIData+
Siriraj hospital

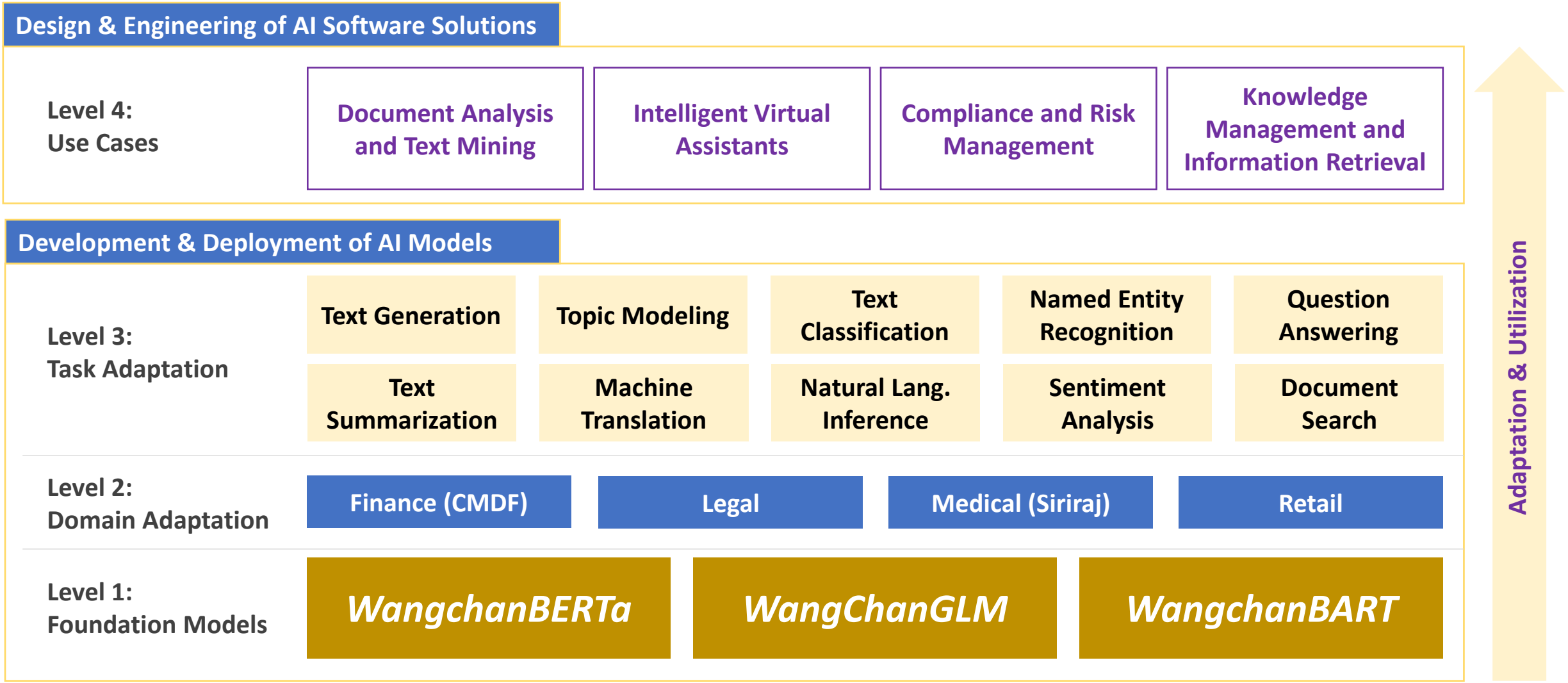

Can Udomcharoenchaikit
VISTEC


Putt Sakdhnagool
ThaiSC, NSTDA


Sarana Nutanong
VISTEC

Still hiring…. a lot!

Use cases are supported by three levels: foundation models, domain adaptation, and task adaptation.



VISTEC-depa
THAILAND
AI RESEARCH
INSTITUTE

What are the current status and our plan?

Foundation Models: WangChanGLM is a multilingual, instruction-finetuned Facebook XGLM-7.5B (released under CC-BY SA 4.0)

Instruction-following tasks: *reading comprehension*, *brainstorming*, and *creative writing*.

We provide

- the weights for a model finetuned on an English-only dataset (wangchanglm-7.5B-sft-en) and
- another checkpoint further finetuned on Google-Translated Thai dataset (wangchanglm-7.5B-sft-enth)

Options for the next step.

Finetune existing language models further with well-shaped subwords	This one is less resource-intensive, but there are many complications.
Revamp subword tokens of an existing model	This option is the least resource-intensive. However, it's also the riskiest, suitable for grad research.
Pretrain our language model with careful subword tokenization ***WangchanUltima***	This option is the most straightforward and most resource-intensive. Thanks for the HPC access from ThaiSC.

WangChanUltima is a planned multilingual language model with superior Thai monolingual and Thai-L2 cross-lingual performances trained on ThaiSC to be unveiled on VISTEC AI Day 2026

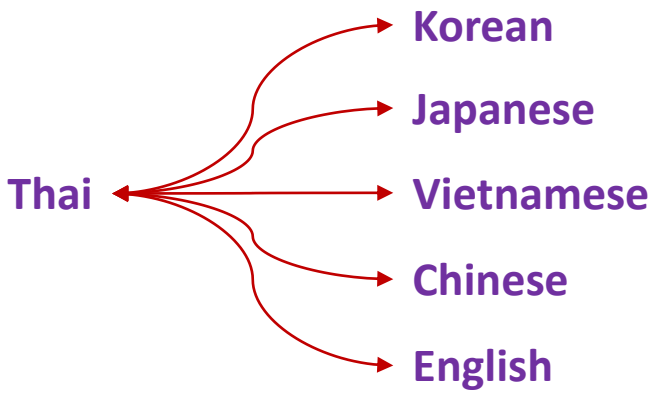
Achieving Superior Cross-Lingual Performance

Subword tokenization is critical to cross-lingual knowledge transfer.

Bio	logy	Psycho	logy
生物 shēngwù	学 xué	心理 xīnlǐ	学 xué
ชีว	วิทยา	จิต	วิทยา

WangchanUltima

By pretraining our own model with careful subword tokenization, we aim to achieve superior cross-lingual performances for the following language pairs.



This approach is also knowledge-intensive. In addition to full-time research staff, we have ***four research students*** looking into this problem for ***academic research***.

Foundation Models

Mini WangchanBERTa (August 2023)

Mini WangchanBERTa is a distilled model from WangchanBERTa

We reduce the model's parameter from 110M to 4M and 29M parameters.

The preliminary result demonstrated the faster running time (~6 times faster), and the performance gap between the original and distillation models is only +-6%

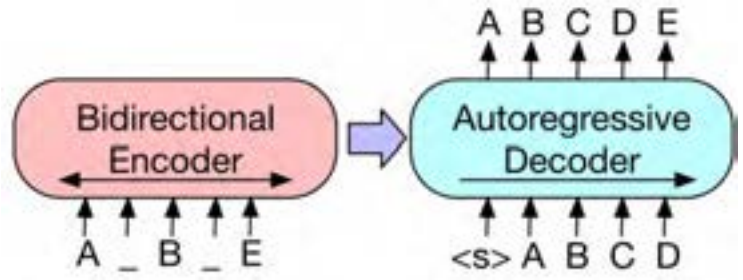
WangchanBART (December 2023)

WangchanBART is an encoder-decoder pretrained language model

The model is trained in the sequence-to-sequence manner.

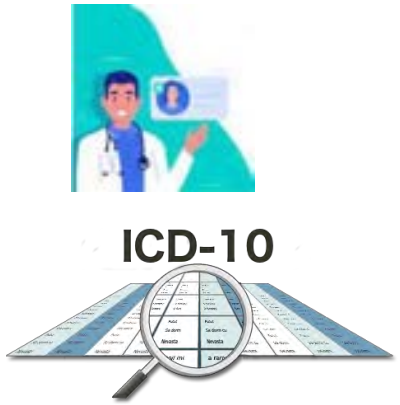
Thus, the model is good at sequence tasks (i.e., machine translation) and encoder tasks (i.e., text classification).

Also, the model can perform ChatGPT-like abilities but with less parameter (only 150M compared to 7B from LLaMA)



Benchmarks & Datasets

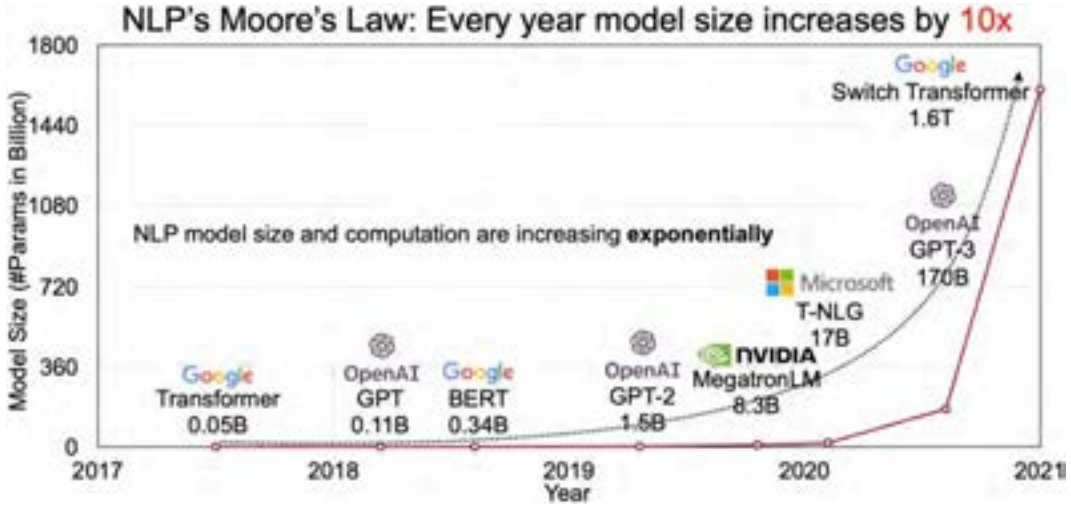
Dataset/Benchmark	Expected Released Date
Human-generated questionnaires for Vicuna-style evaluation	April 2023
Medical Text Classification dataset	May 2024
Financial NER dataset	June 2024
Financial Sentiment Analysis Dataset	June 2024
Legal Text Retrieval Benchmark	December 2024

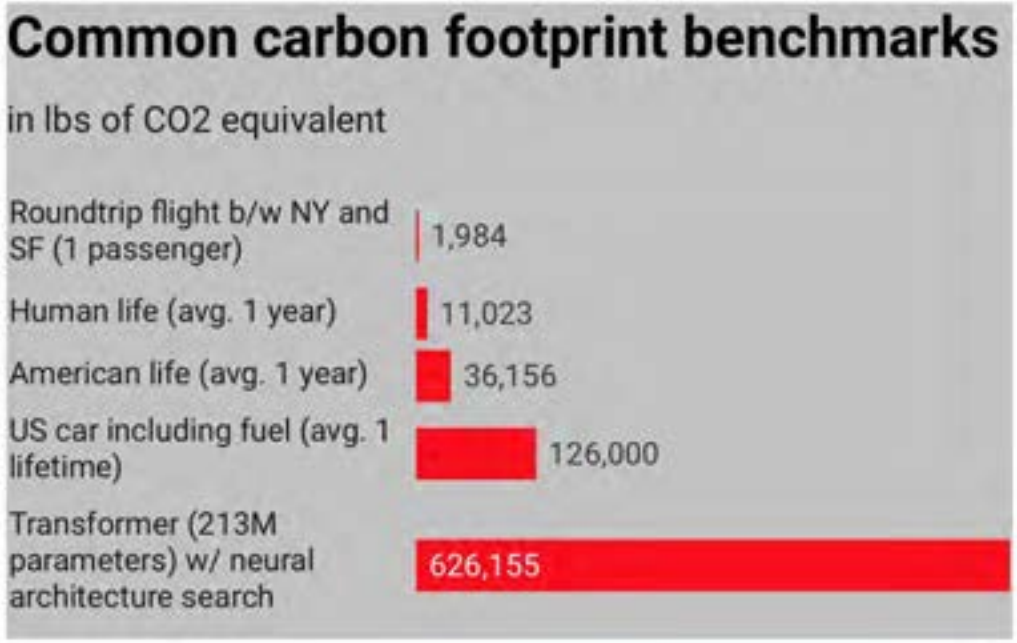


Milestones & Timeline

Milestone	Foundation Models	Adaptation Tools	Benchmarks
VISTEC AI Day 2023 Monday 22nd May	WangchanBERTa WangChanGLM	Adapter Code Release	4 Datasets
VISTEC AI Day 2024 Monday 20th May	Domain-Specific WangChanGLM (Legal, Medical, and Finance)	Full/Partial Finetune and Distillation Code Releases	3 benchmarks and 2 Thai instruction datasets
VISTEC AI Day 2025 Monday 19th May	WangChanGLM-V2 Mini WangChanGLM	Integration with PyThaiNLP	New LLM benchmark (10 tasks)
VISTEC AI Day 2026 Monday 18th May	WangchanUltima	Code Updates	A complete set of LLM benchmark (15 tasks 8 datasets)

Supercomputer and open-source projects





Pre-training to deployment

Pre-training

Finetuning

Amazon SageMaker

Deployment