IN-DEPTH LOOK AT TRANSFORMER BASED MODELS

Training Objectives & Architectures

1

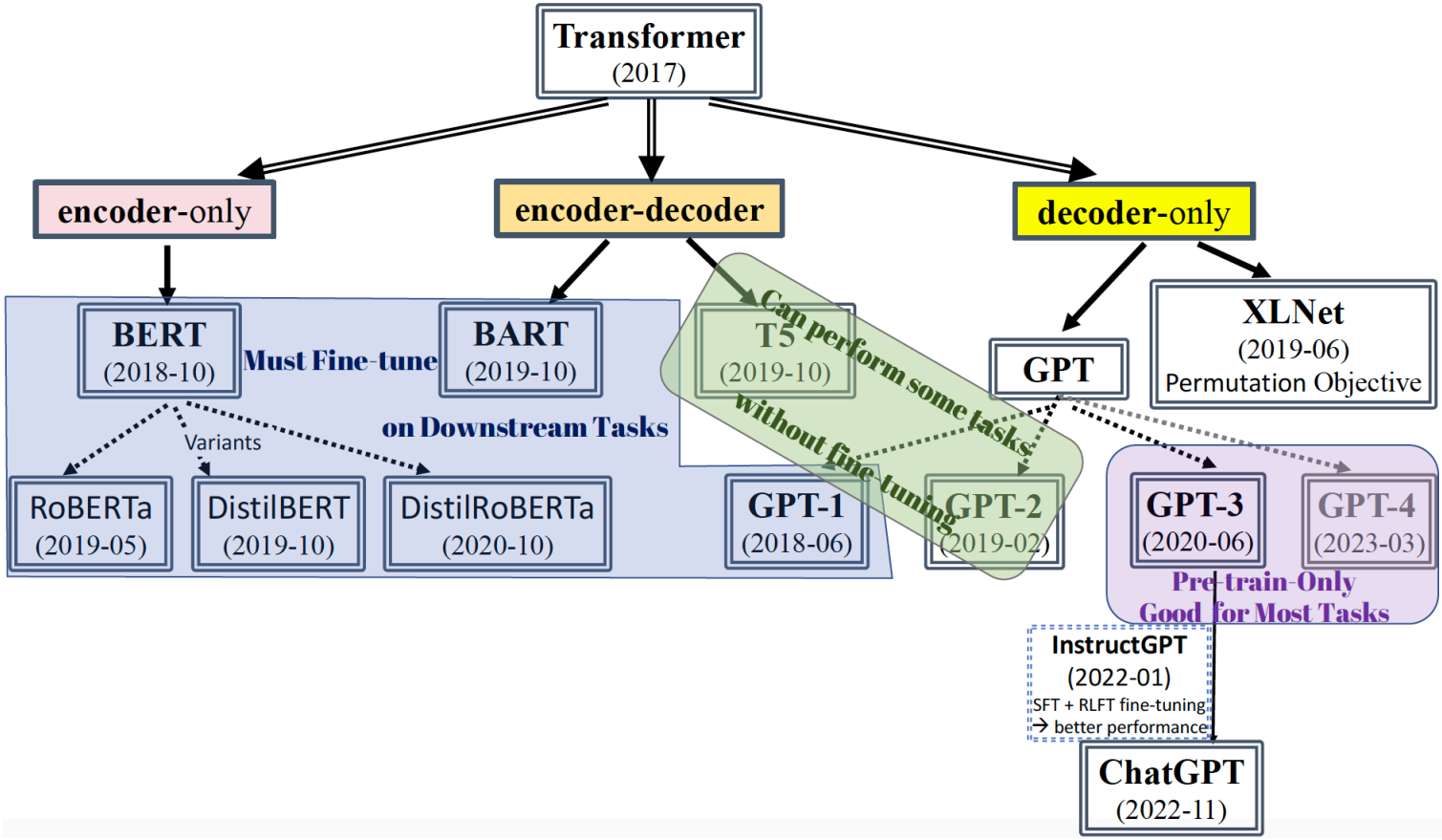
BERT, GPT, T5, BART & XLNet: Comprehensively Compared

Presenter: Yule Wang, PhD

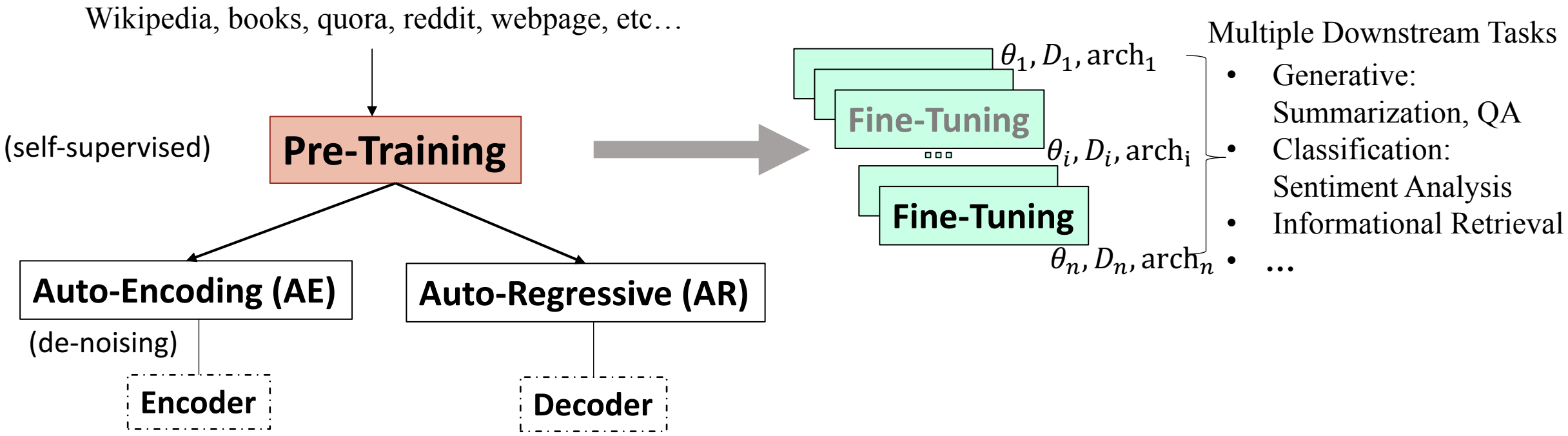
My Relavant blog

OUTLINE

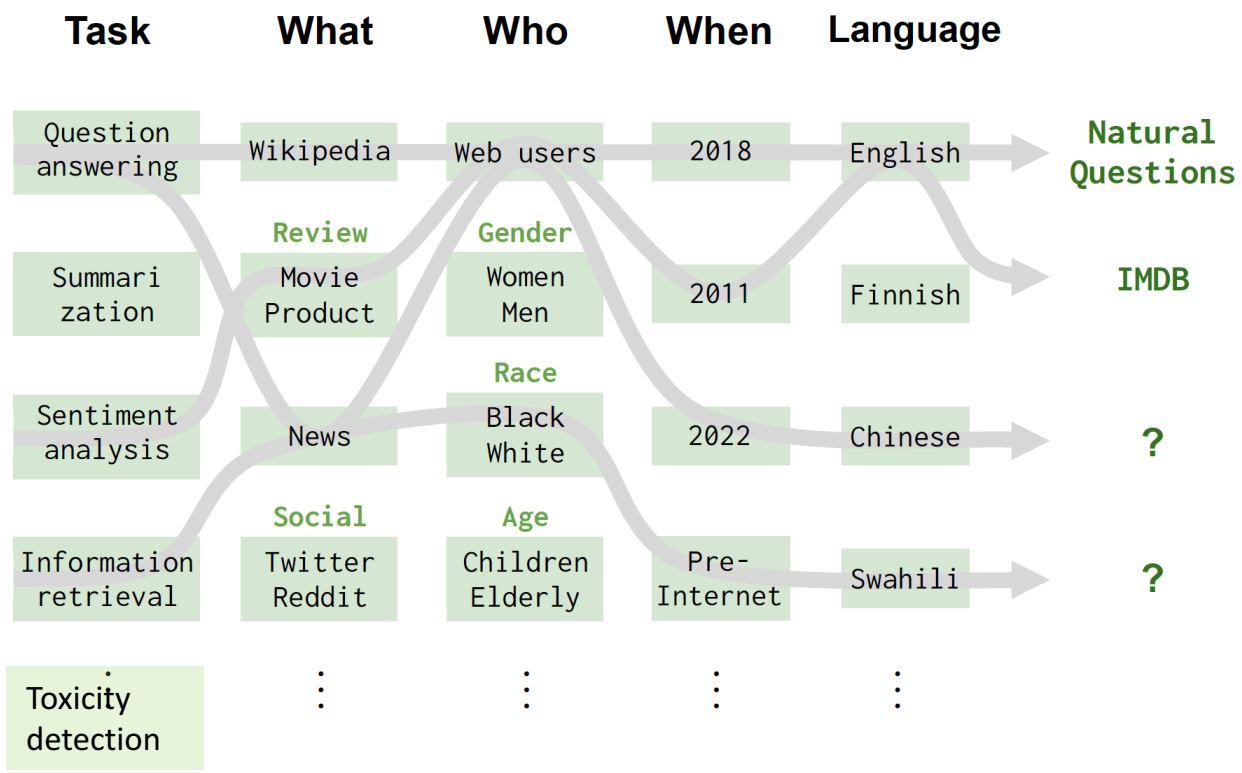
- Pre-training, Fine-tuning
- Transformer
  – encoder, decoder
- Pre-training Objectives
  -- AutoEncoding, AutoRegressive
- Emergence of In-Context Learning
- Unifying Multi-Tasks?
- Fine-tuning by InstructGPT



TRADITIONAL: PRE-TRAINING → FINE-TUNING



MULTI-TASKS, DOMAINS



User-Facing Tasks

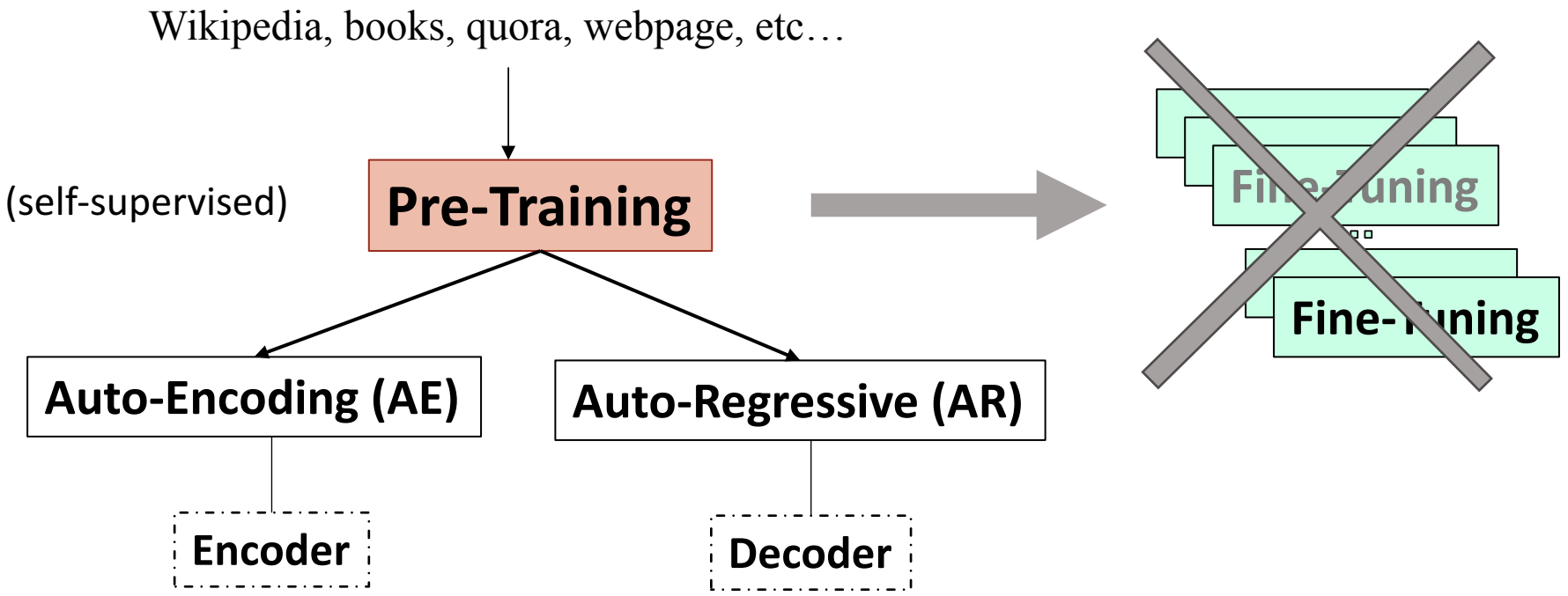
Dialogue

Named Entity Recognition
Part-of-Speech Tagging
Syntactic Parsing
information extraction
...

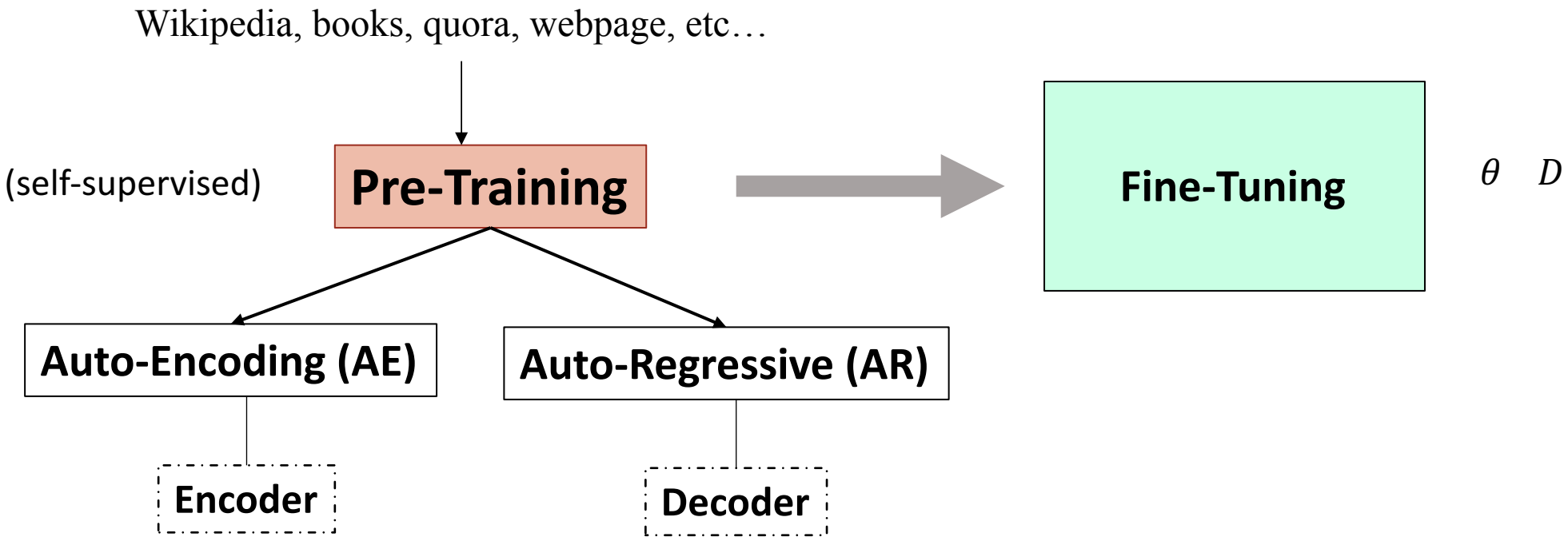
Intermediate Tasks

Img Src: Refs[8]

TRADITIONAL → UNIFYING TASK

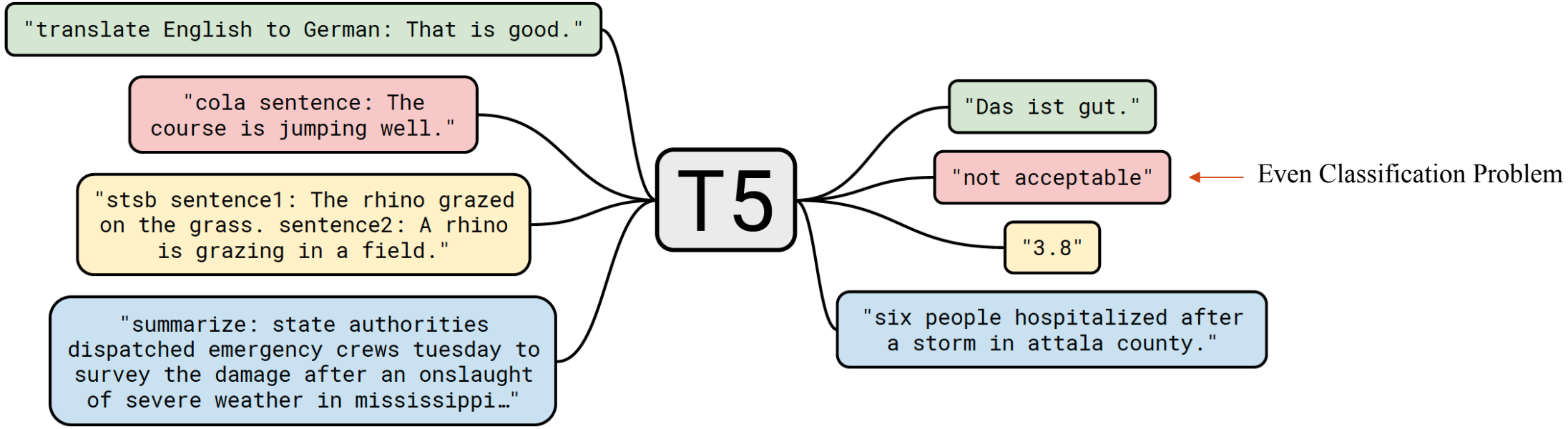


TRADITIONAL → UNIFYING TASK



UNIFYING TASK: T5

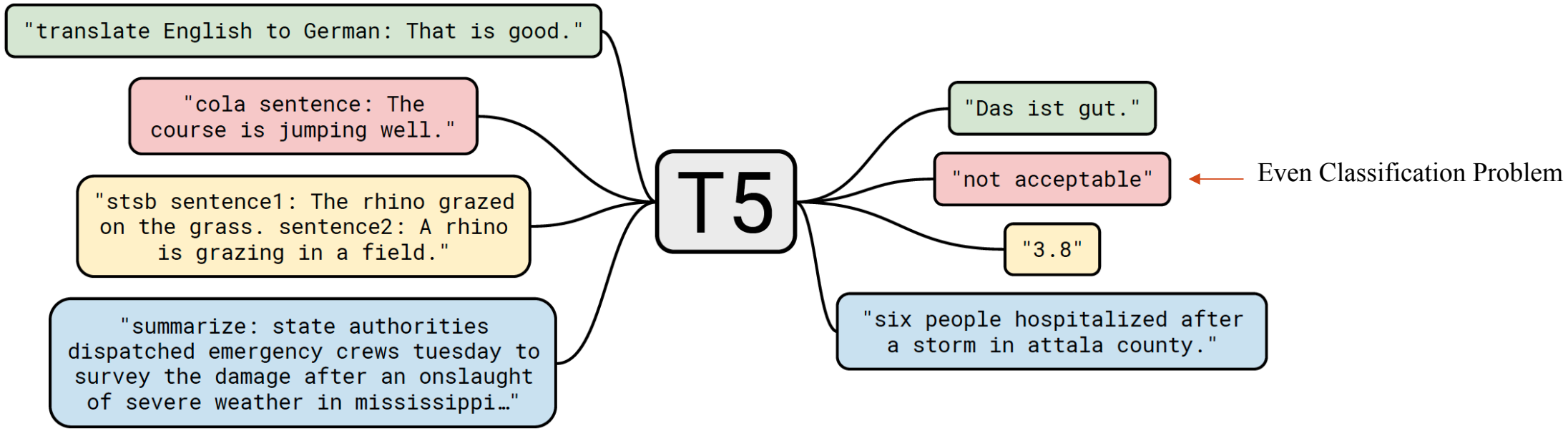
T5 (Unifying Text-To-Text Transfer Transformer) (2019.10)



Img Src: Refs[6]

UNIFYING TASK: T5

T5 (Unifying Text-To-Text Transfer Transformer) (2019.10)

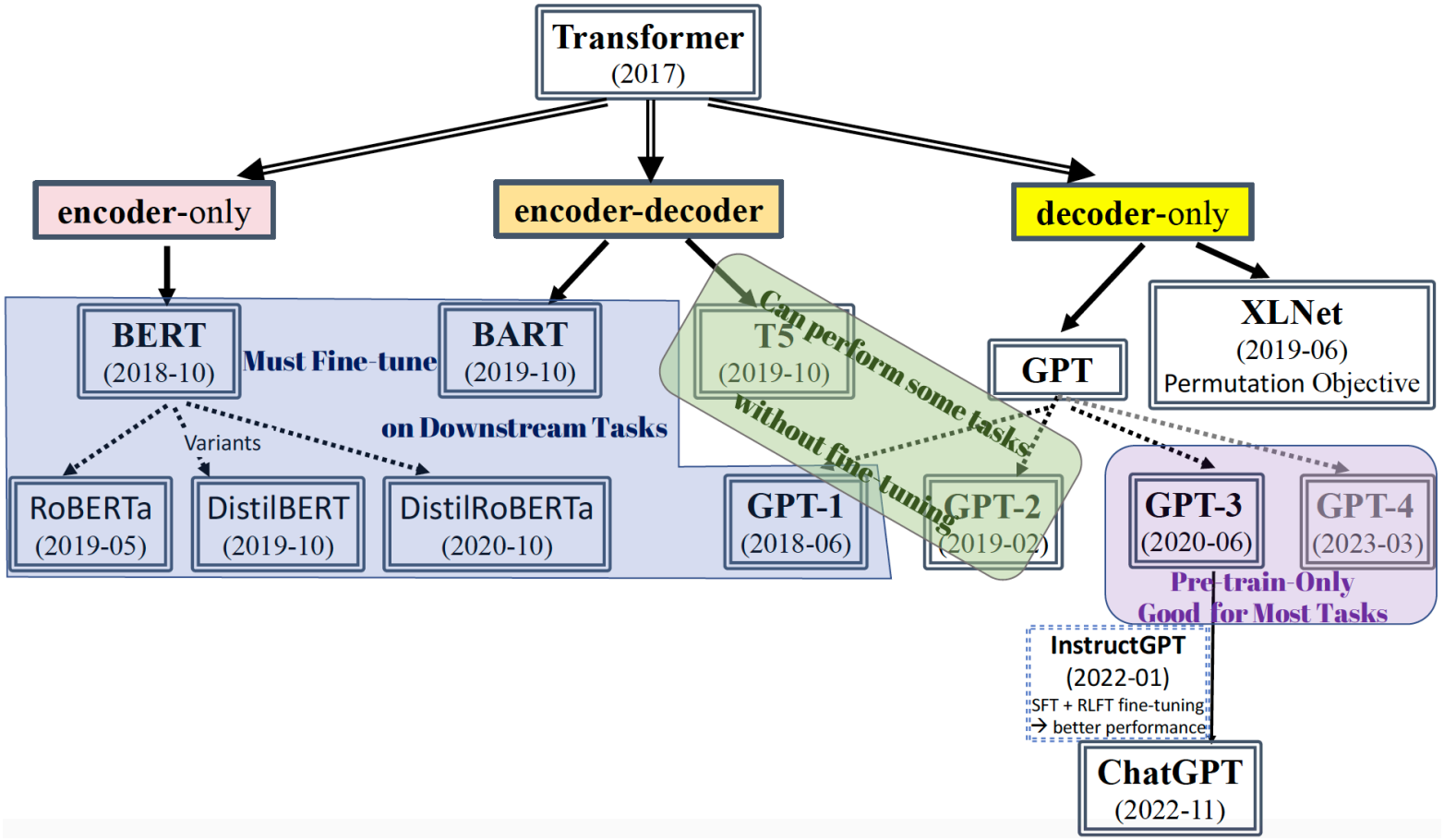


___ is Canadian National Day? → When

? Any NLP problem can convert to a generative problem?

Img Src: Refs[6]

OVERVIEW: MUST FINE-TUNING FOR TRANSFORMER-BASED MODELS?

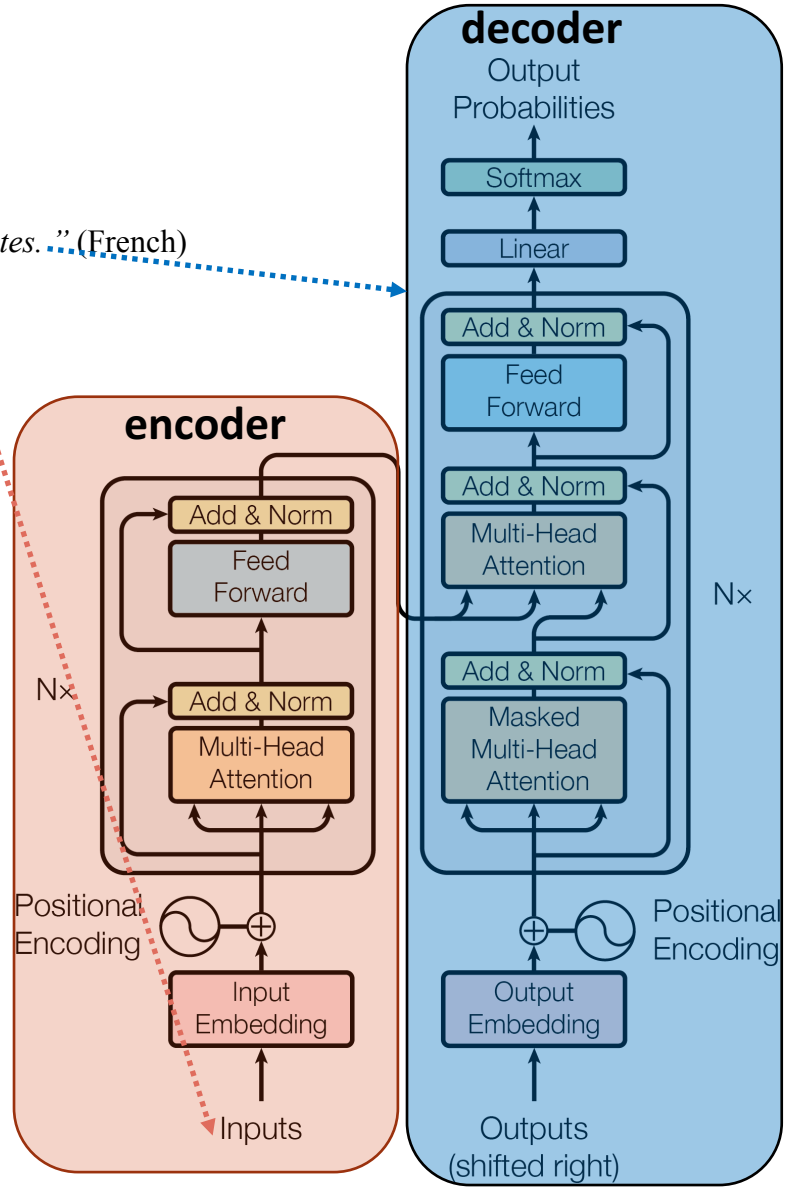


TRANSFORMER

Original Aim -- Translation:

"Legumes share resources with nitrogen-fixing bacteria." →

"Legumes partagent des ressources avec des bactéries azotantes." (French)



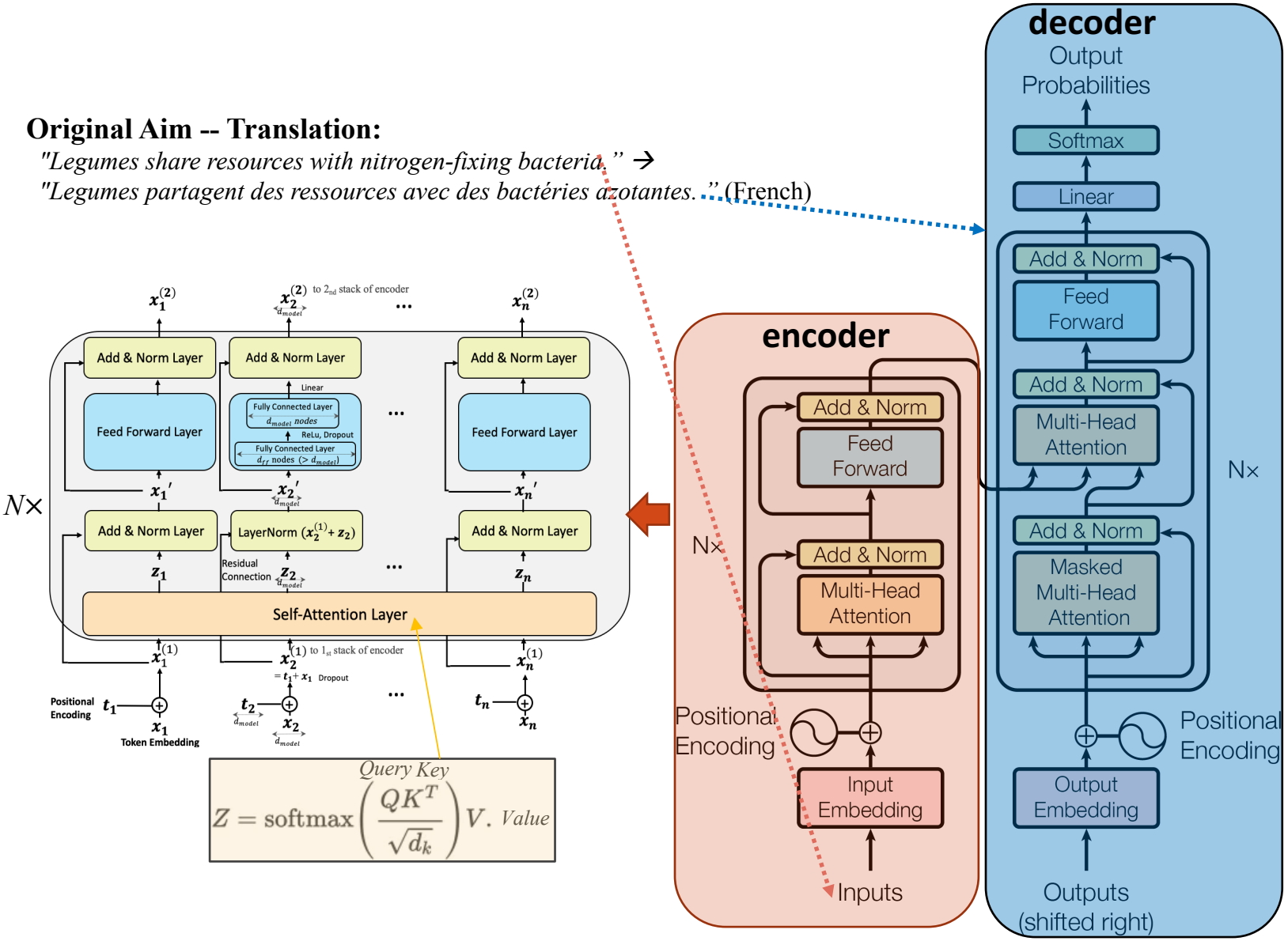
Img Src: Refs[1]

TRANSFORMER

Original Aim -- Translation:

"Legumes share resources with nitrogen-fixing bacteria." →

"Legumes partagent des ressources avec des bactéries azotantes." (French)



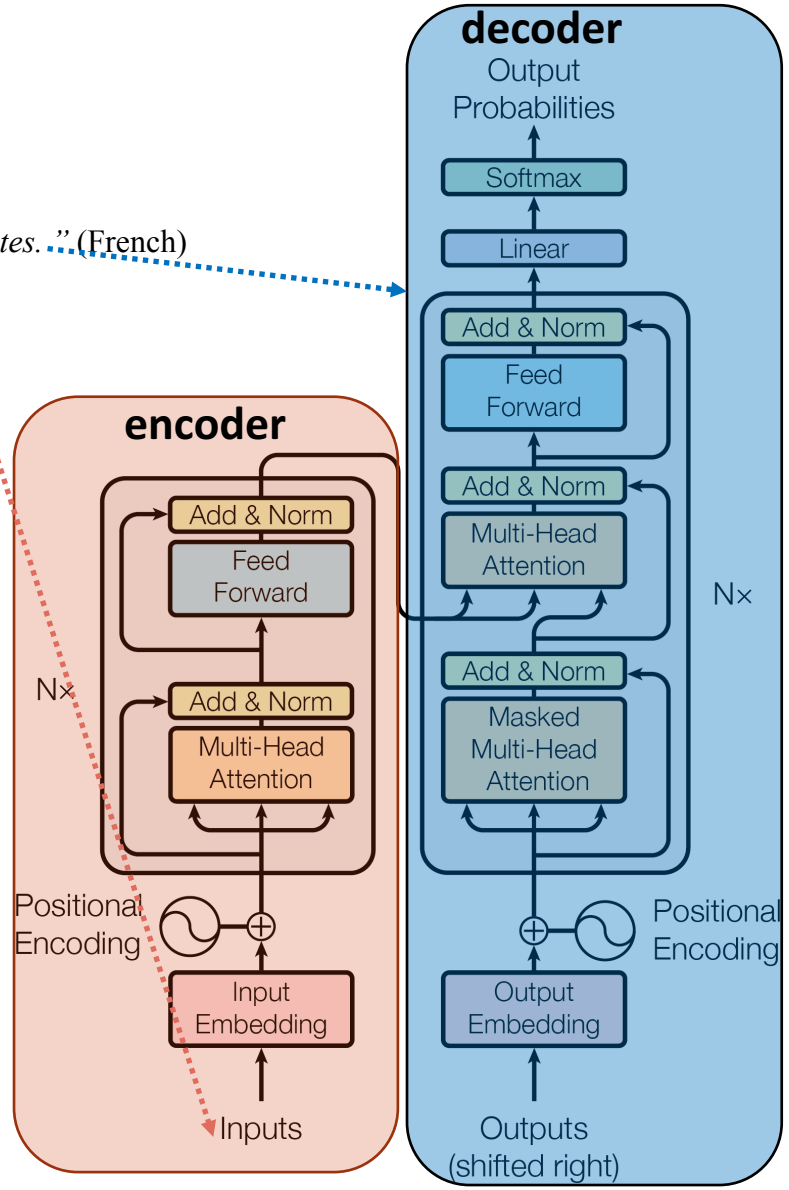
Img Src: Refs[1]

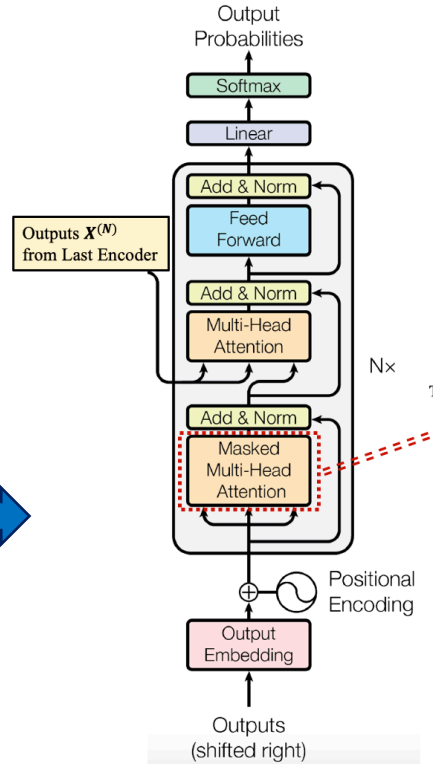
TRANSFORMER

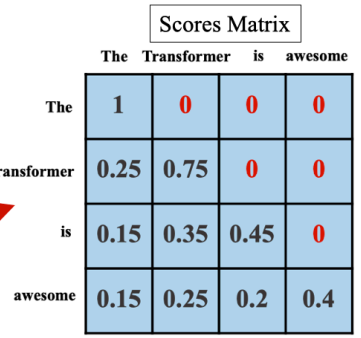
Original Aim -- Translation:

"Legumes share resources with nitrogen-fixing bacteria." →

"Legumes partagent des ressources avec des bactéries azotantes." (French)







Img Src: Refs[1]

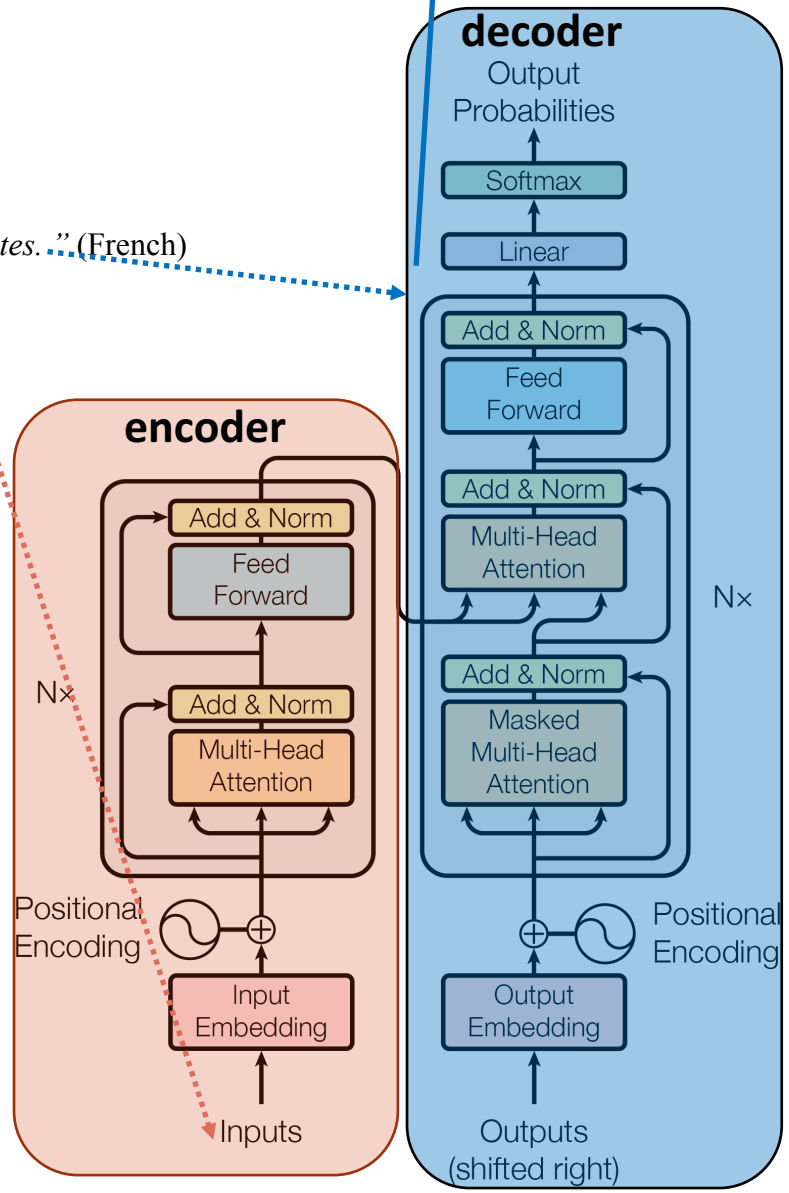
TRANSFORMER

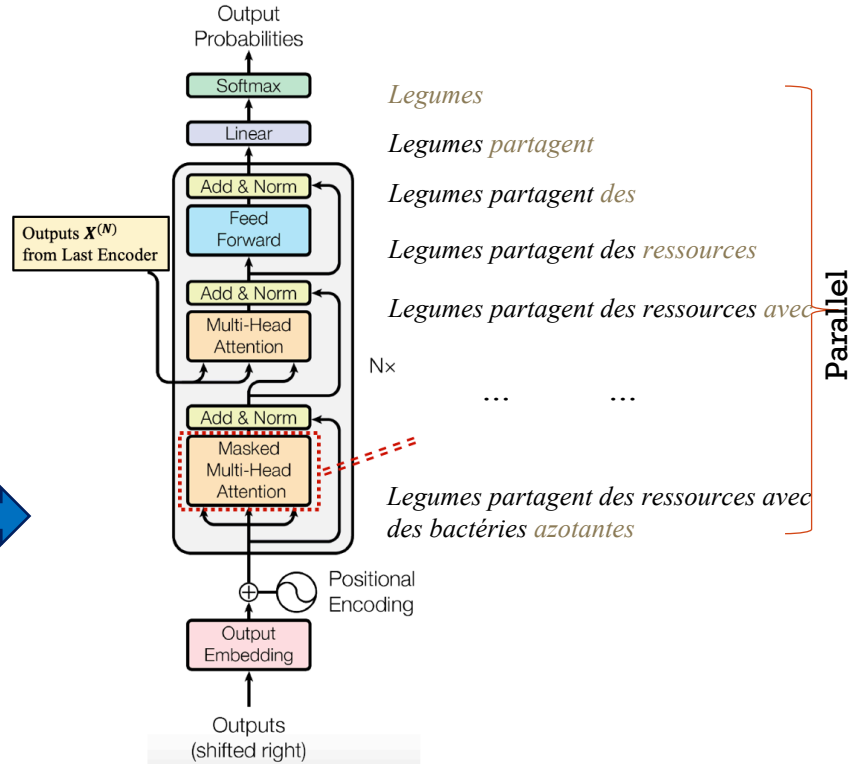
A natural seq2seq model (Autoregressive Objective)

Original Aim -- Translation:

"Legumes share resources with nitrogen-fixing bacteria." →

"Legumes partagent des ressources avec des bactéries azotantes." (French)





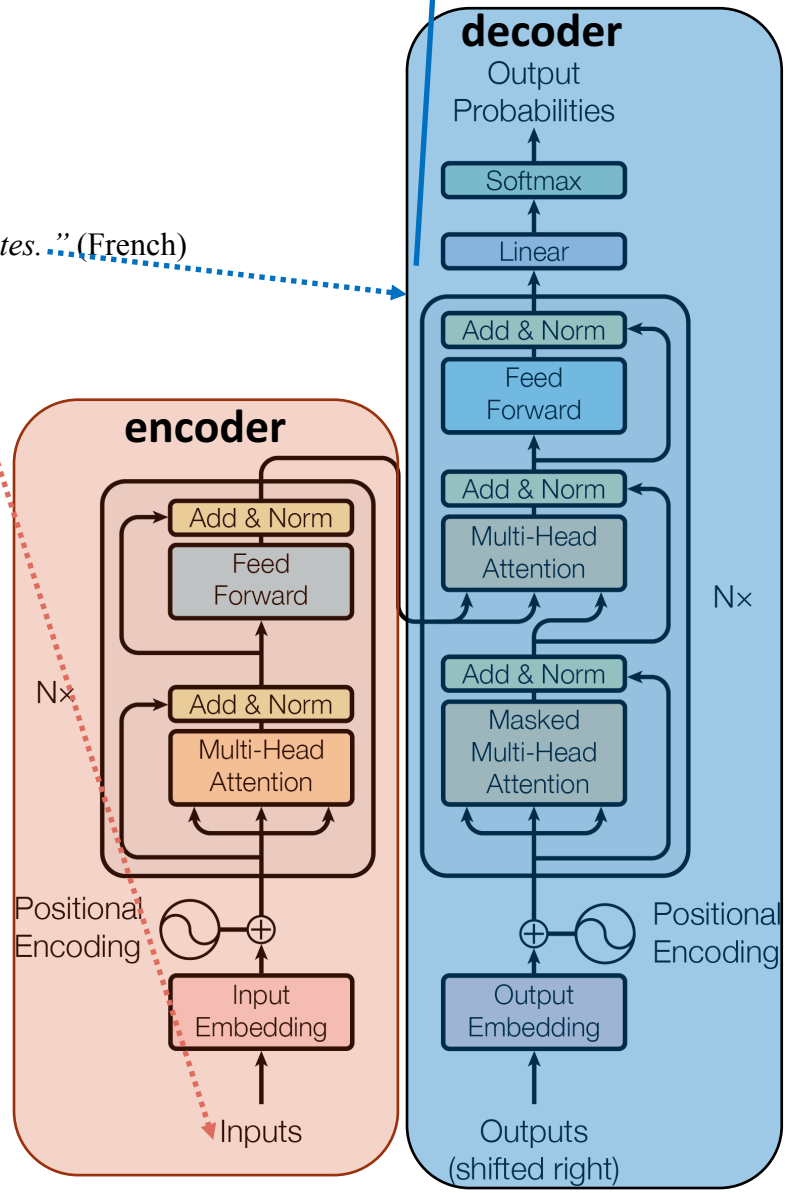
Img Src: Refs[1]

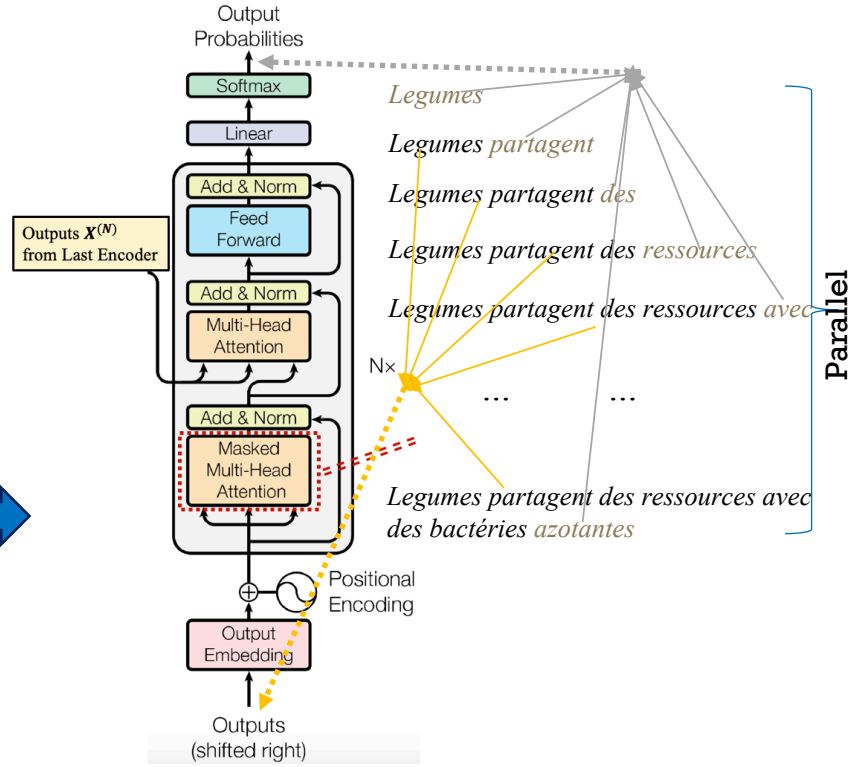
TRANSFORMER

A natural seq2seq model (Autoregressive Objective)

Original Aim -- Translation:

"Legumes share resources with nitrogen-fixing bacteria." →
"Legumes partagent des ressources avec des bactéries azotantes." (French)





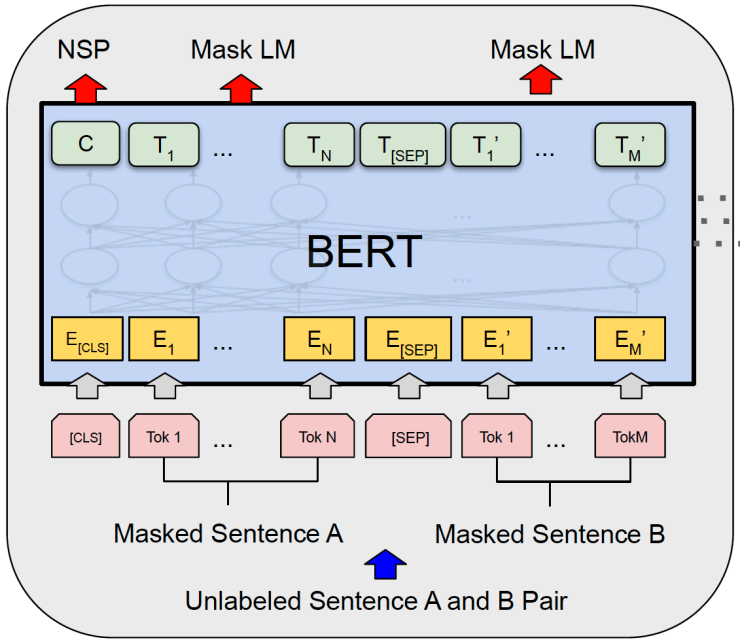
? *Legumes share resources with nitrogen-fixing bacteria* | ...

A question: Why don't abandon encoder part?

Img Src: Refs[1]

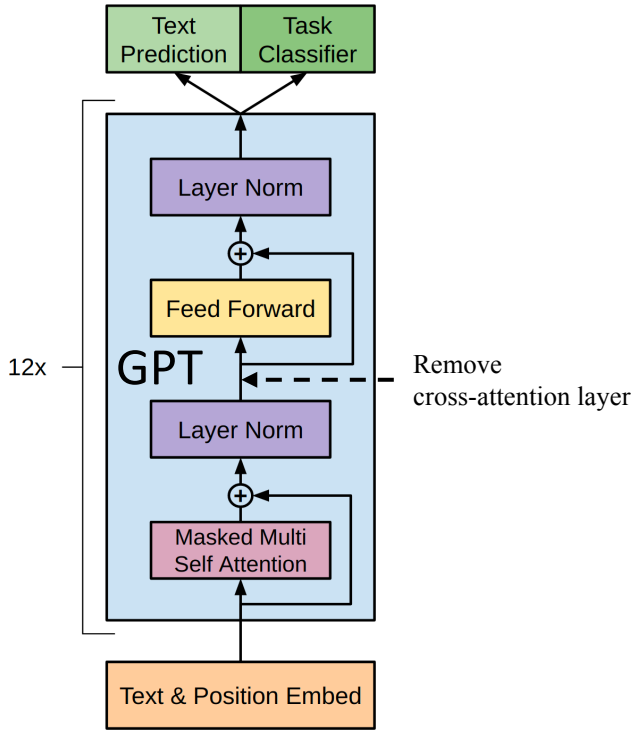
AUTO-ENCODER & AUTO-REGRESSIVE

- **Auto-Encoding (AE)**:
  de-noising—predicting masked tokens



Pre-training

- **Auto-Regressive (AR)**:
  one after one token generating



Pre-training

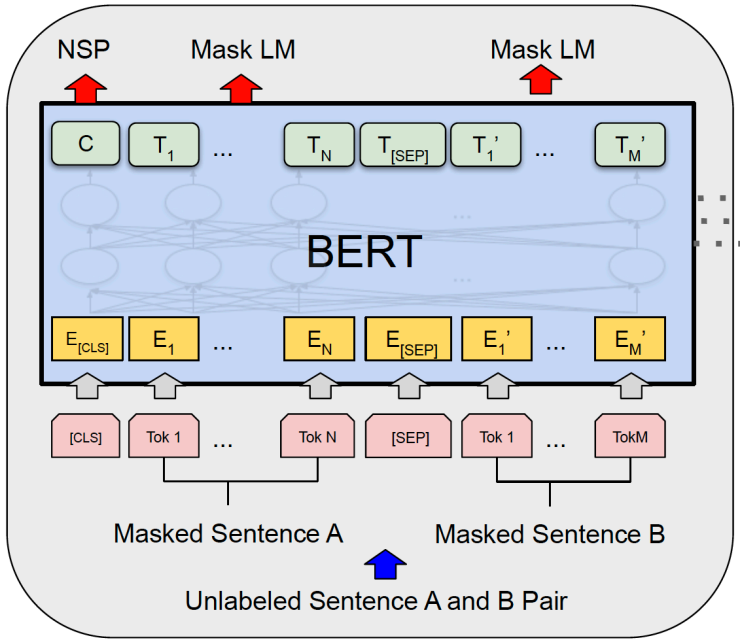
Img Src: Refs[2,3]

AUTO-ENCODER & AUTO-REGRESSIVE

- **Auto-Encoding (AE)**:
  de-noising—predicting masked tokens

$$\max_{\theta} \log p_\theta(\mathbf{x_m} \mid \mathbf{x_{\bar{m}}}) \approx \sum_{t=1}^{T} m_t \log p_\theta(x_t \mid \mathbf{x_{\bar{m}}}) = \sum_{t=1}^{T} m_t \log \frac{\exp\left(y_{\mathbf{x_{\bar{m}}}}(x_t)\right)}{\sum_{x'} \exp\left(y_{\mathbf{x_{\bar{m}}}}(x')\right)}$$

variants: T5, BART

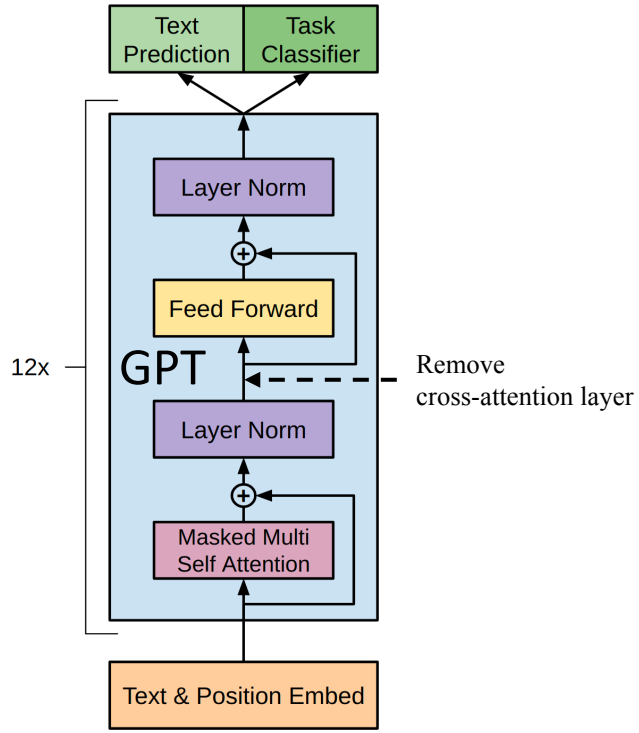


Pre-training

- **Auto-Regressive (AR)**:
  one after one token generating

$$\max_{\theta} \log p_\theta(x_1, \ldots, x_T) \approx \sum_{t=1}^{T} \log p_\theta(x_t \mid x_1, \ldots, x_{t-1}) = \sum_{t=1}^{T} \log \frac{\exp(y_{x_1,\ldots,x_{t-1}}(x_t))}{\sum_{x'} \exp(y_{x_1,\ldots,x_{t-1}}(x'))}$$

variant: XLNet
(permutation)



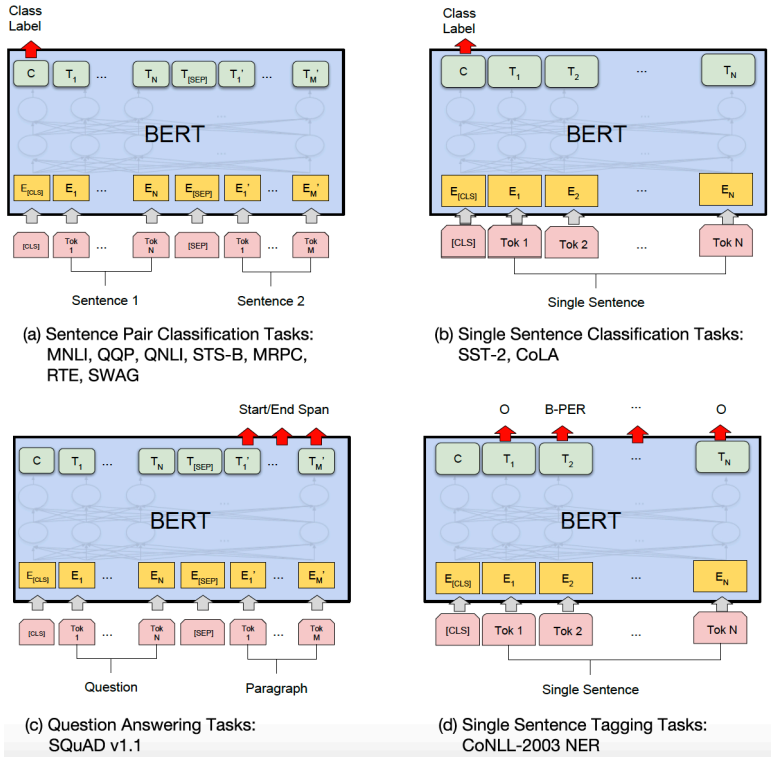
Pre-training

Img Src: Refs[2,3]

AUTO-ENCODER & AUTO-REGRESSIVE

FINE-TUNING

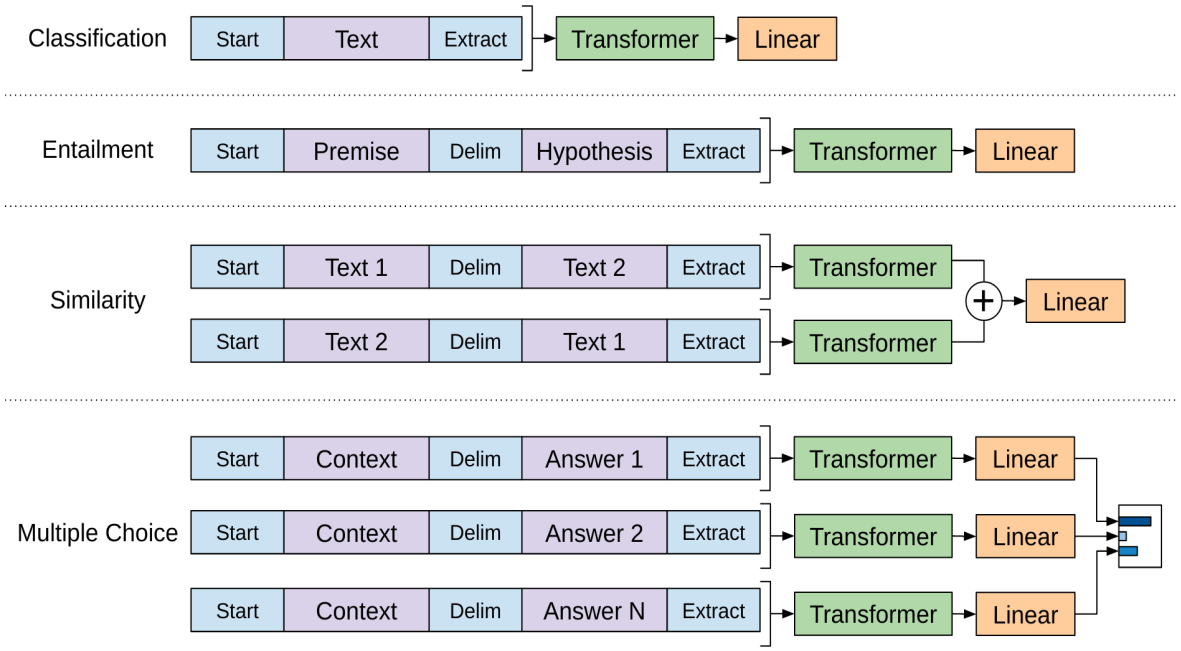
- **Auto-Encoding (AE)**:
  de-noising—predicting masked tokens



BERT

Fine-tuning

- **Auto-Regressive (AR)**:
  one after one token generating

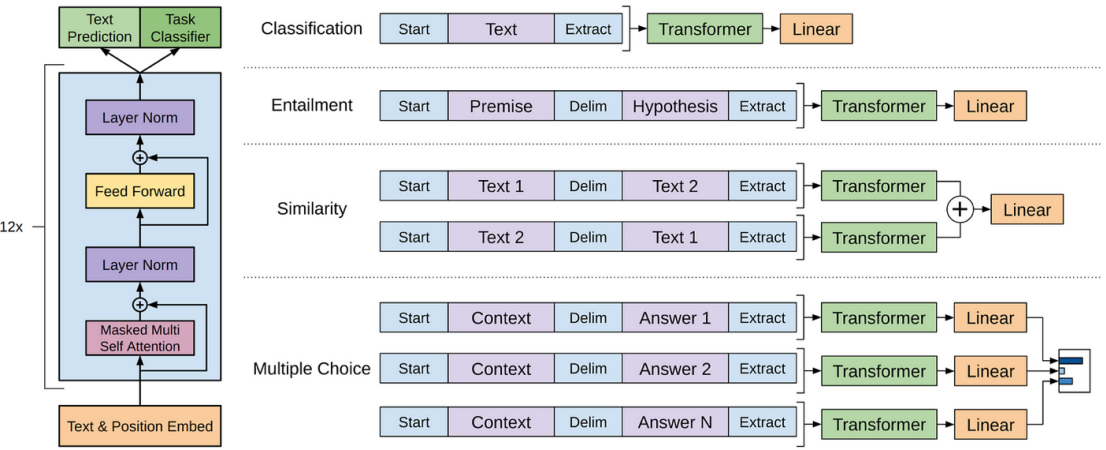


GPT-1

Fine-tuning

Img Src: Refs[2,3]

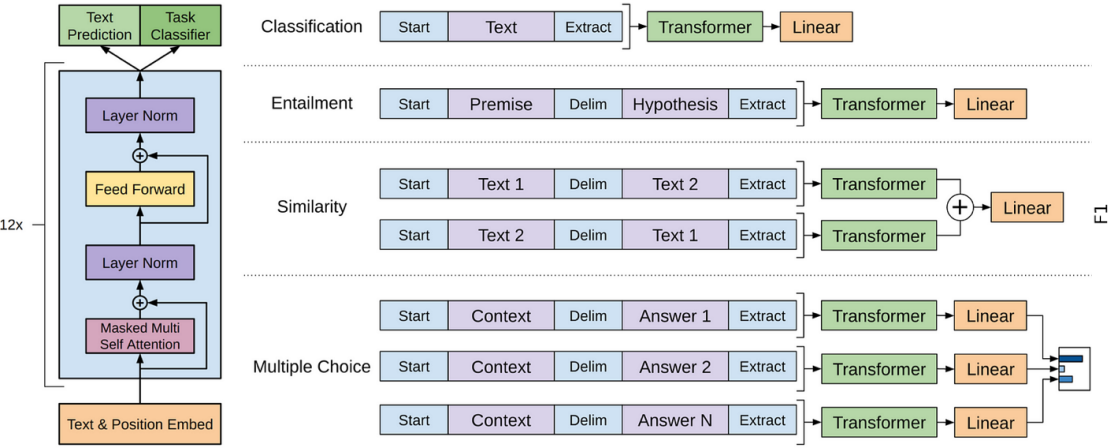
AUTO-REGRESSIVE SCALING



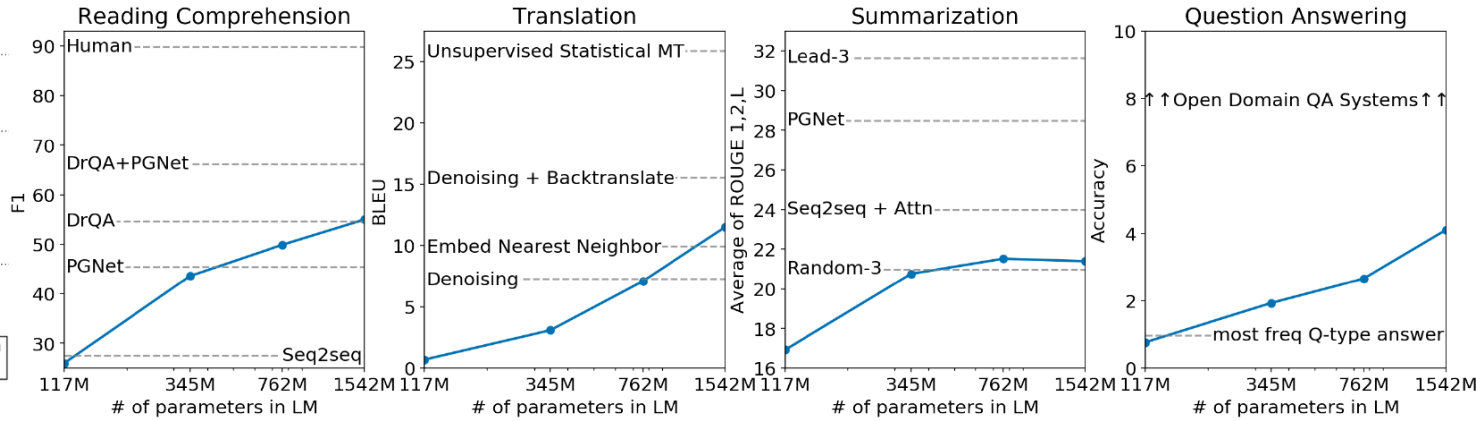
GPT-1

Img Src: Refs[2,3]

AUTO-REGRESSIVE SCALING



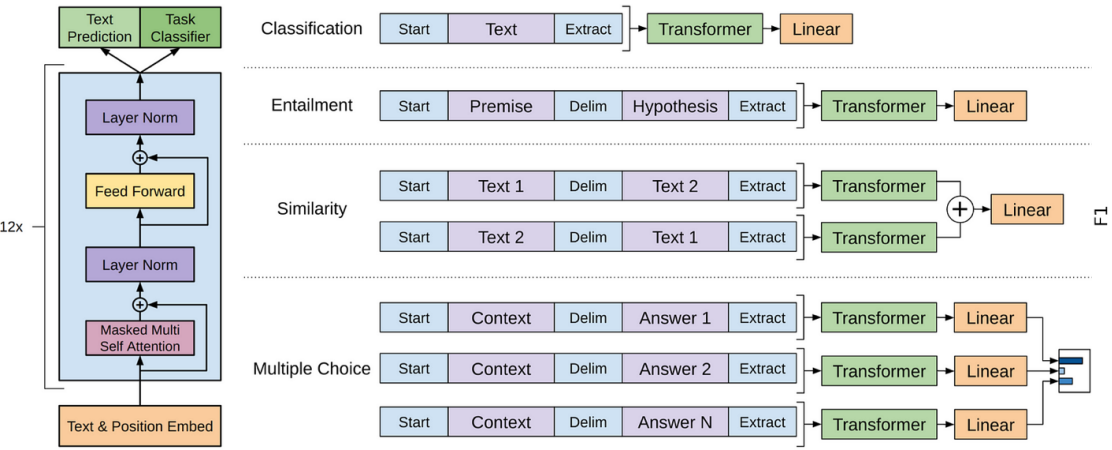
GPT-1



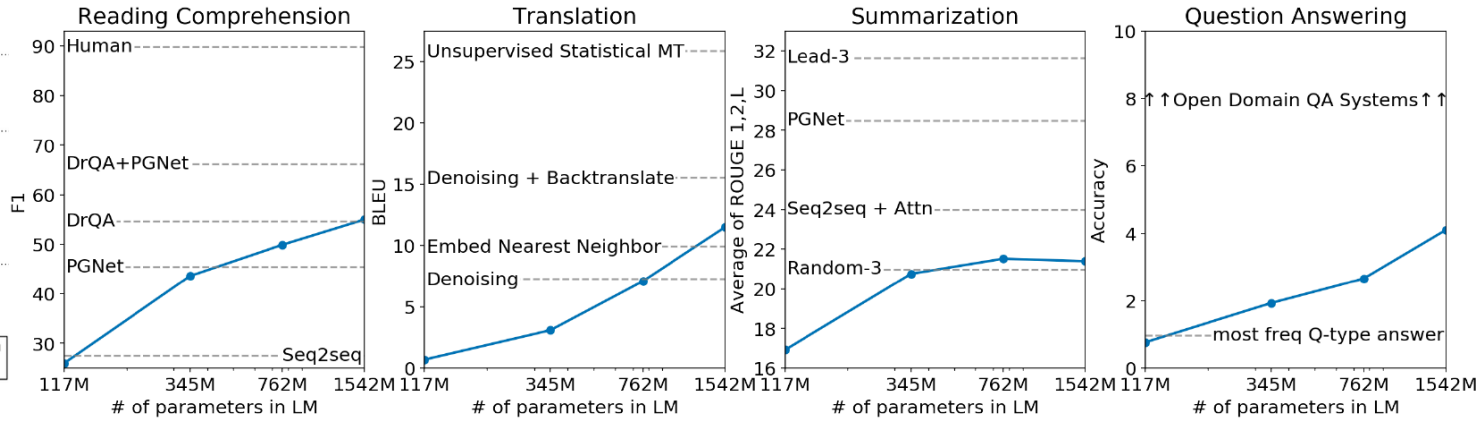
GPT-2

Img Src: Refs[2,3]

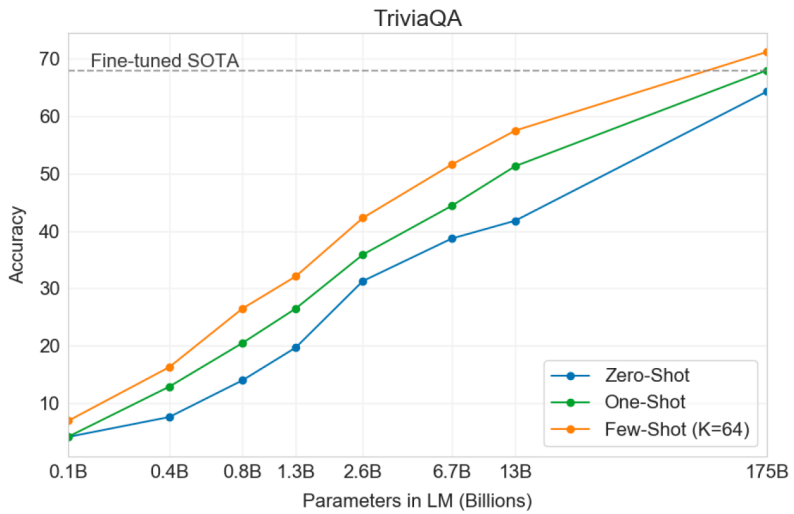
AUTO-REGRESSIVE SCALING



GPT-1



GPT-2



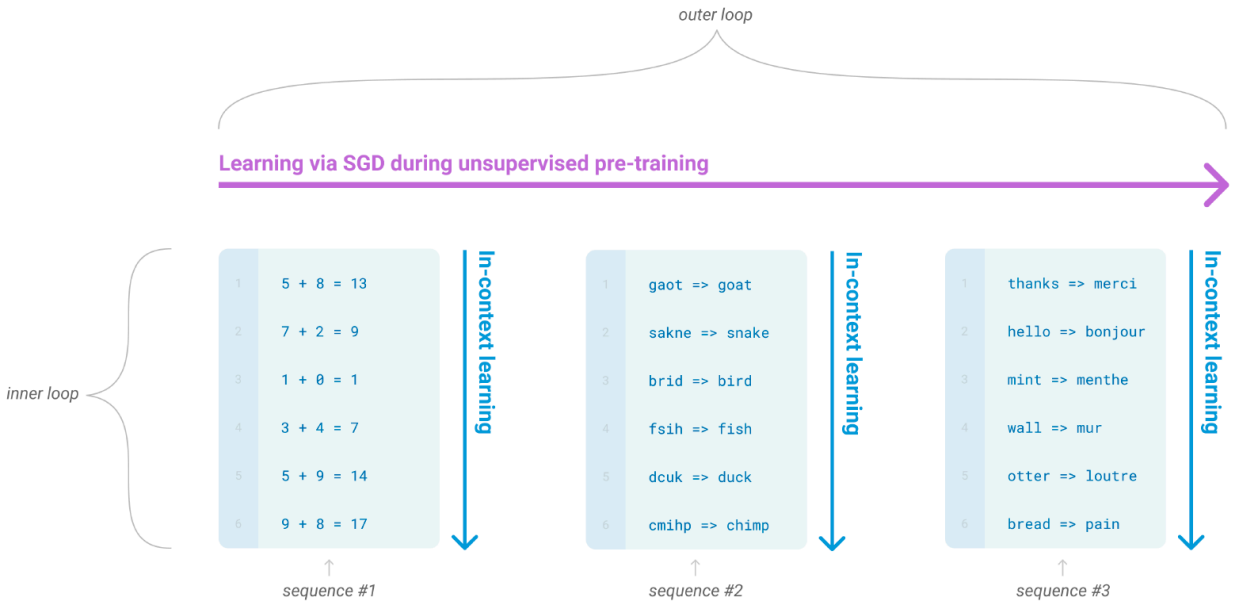
GPT-3

Setting	CoQA	DROP	QuAC	SQuADv2	RACE-h	RACE-m
Fine-tuned SOTA	**90.7**[a]	**89.1**[b]	**74.4**[c]	**93.0**[d]	**90.0**[e]	**93.1**[e]
GPT-3 Zero-Shot	81.5	23.6	41.5	59.5	45.5	58.4
GPT-3 One-Shot	84.0	34.3	43.3	65.4	45.9	57.4
GPT-3 Few-Shot	85.0	36.5	44.3	69.8	46.8	58.1

Img Src: Refs[2,3]

EMERGENCE OF IN-CONTEXT LEARNING:

ZERO-SHOT, FEW-SHOT VS FINE-TUNING



The three settings we explore for in-context learning

Zero-shot

The model predicts the answer given only a natural language description of the task. No gradient updates are performed.

```
Translate English to French:    ← task description
cheese =>                       ← prompt
```

One-shot

In addition to the task description, the model sees a single example of the task. No gradient updates are performed.

```
Translate English to French:    ← task description
sea otter => loutre de mer      ← example
cheese =>                       ← prompt
```

Few-shot

In addition to the task description, the model sees a few examples of the task. No gradient updates are performed.

```
Translate English to French:    ← task description
sea otter => loutre de mer      ← examples
peppermint => menthe poivrée
plush girafe => girafe peluche
cheese =>                       ← prompt
```

Traditional fine-tuning (not used for GPT-3)

Fine-tuning

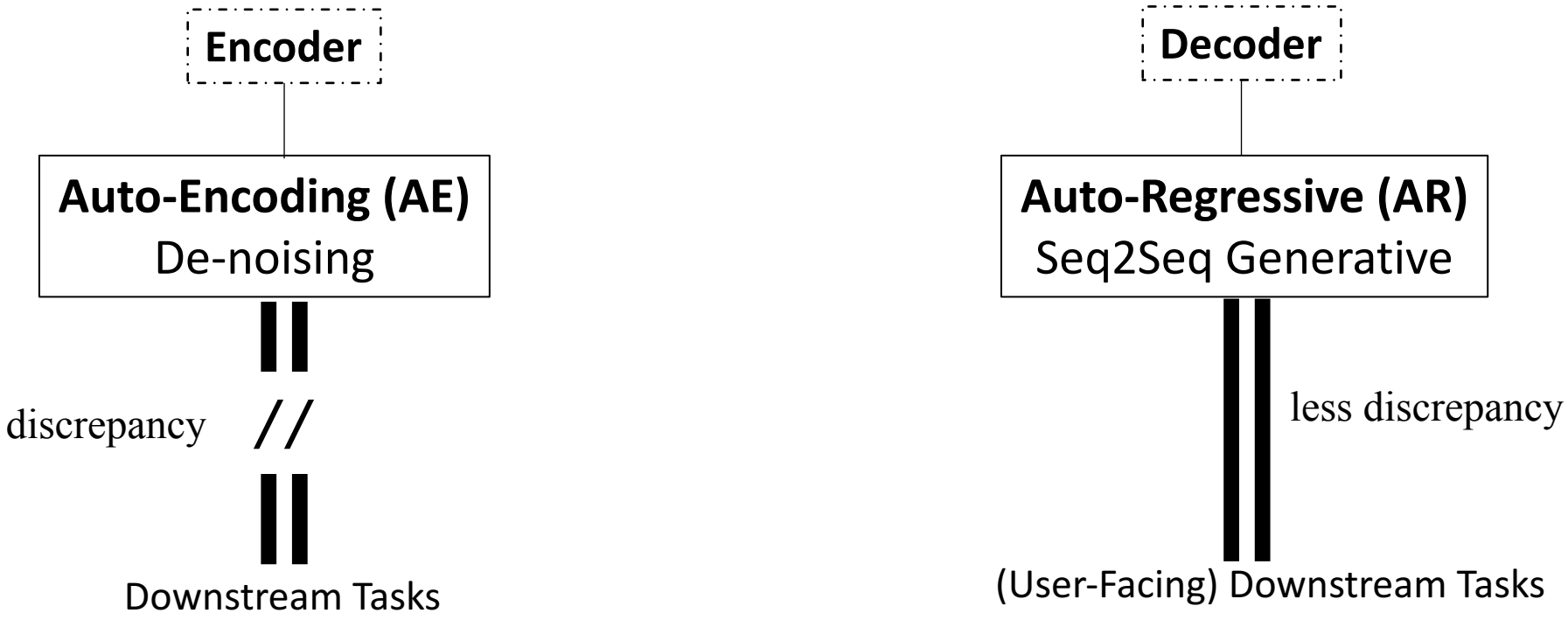
The model is trained via repeated gradient updates using a large corpus of example tasks.

```
sea otter => loutre de mer      ← example #1
        ↓
   gradient update
        ↓
peppermint => menthe poivrée    ← example #2
        ↓
   gradient update
        ↓
        ...
        ↓
plush giraffe => girafe peluche ← example #N
        ↓
   gradient update
        ↓
cheese =>                       ← prompt
```


Img Src: Refs[5]

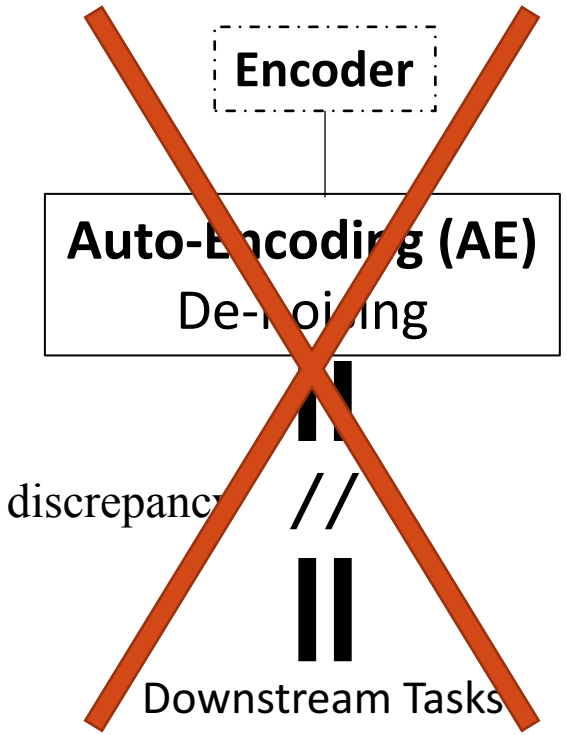
AUTO-ENCODING VS AUTO-REGRESSIVE

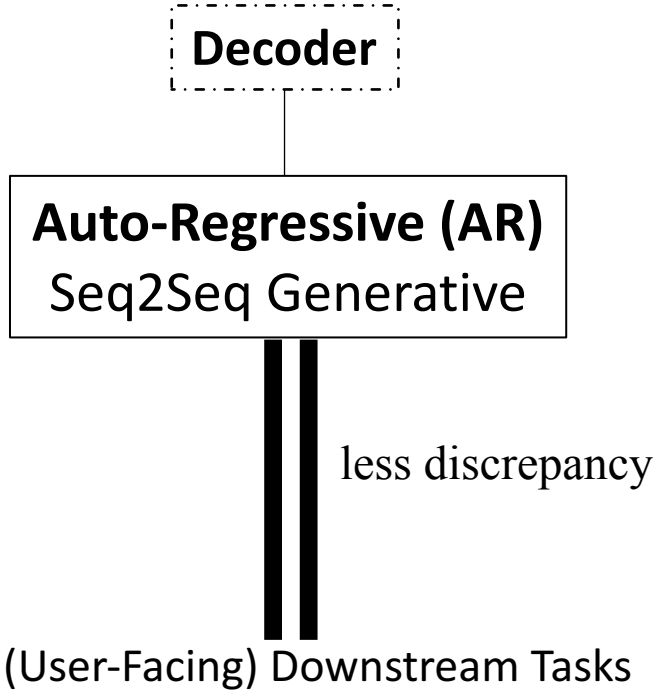
ALIGN WITH DOWNSTREAM TASKS?



Human Language Natural Interaction
A Generative Process?

~~ENCODER~~--DECODER?

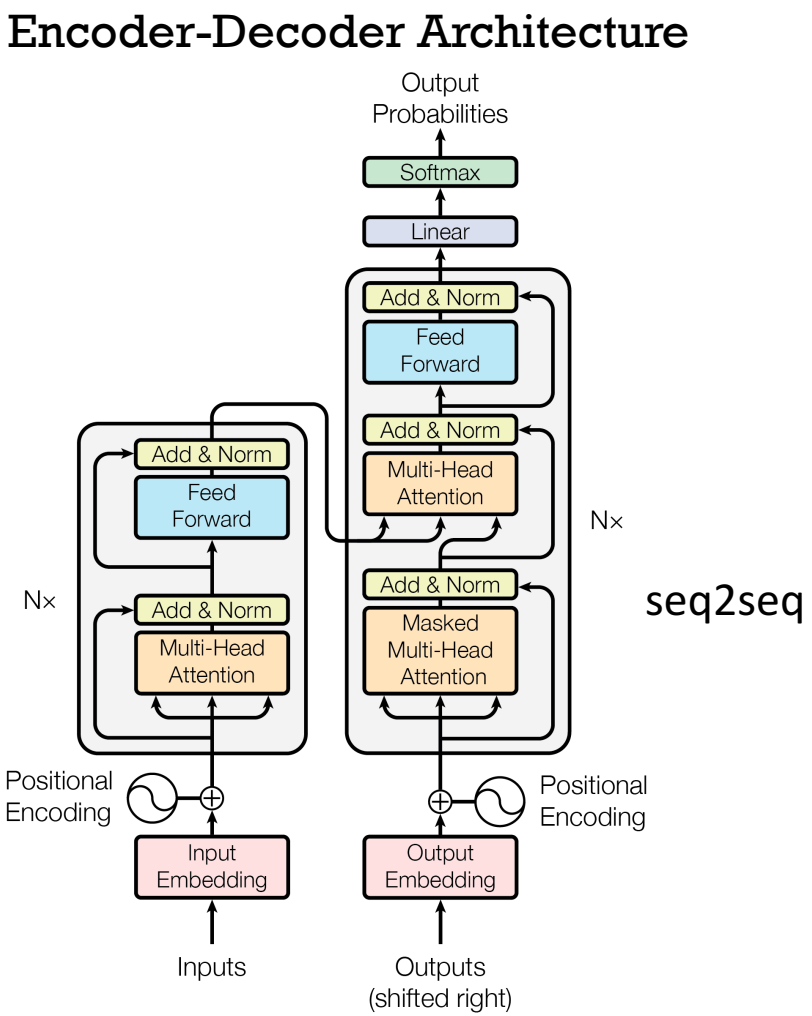




Human Language Natural Interaction
A Generative Process?

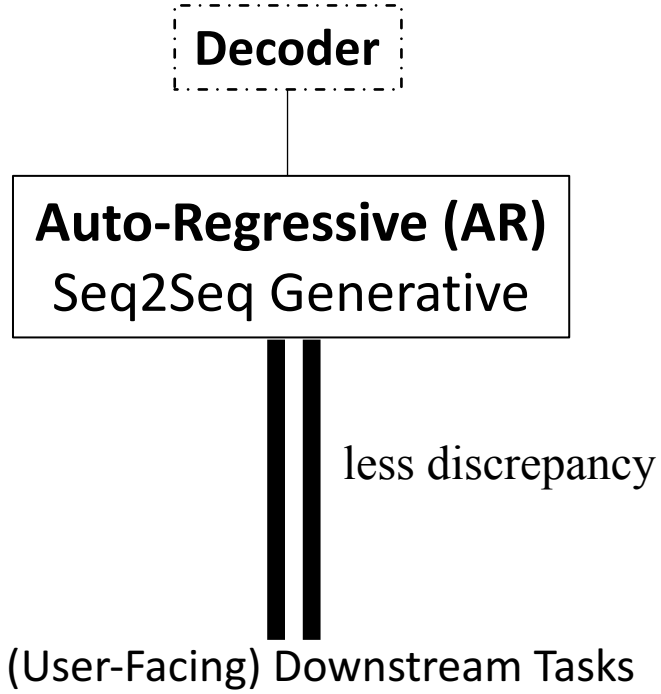
ENCODER--DECODER?

T5, BART:



De-noising

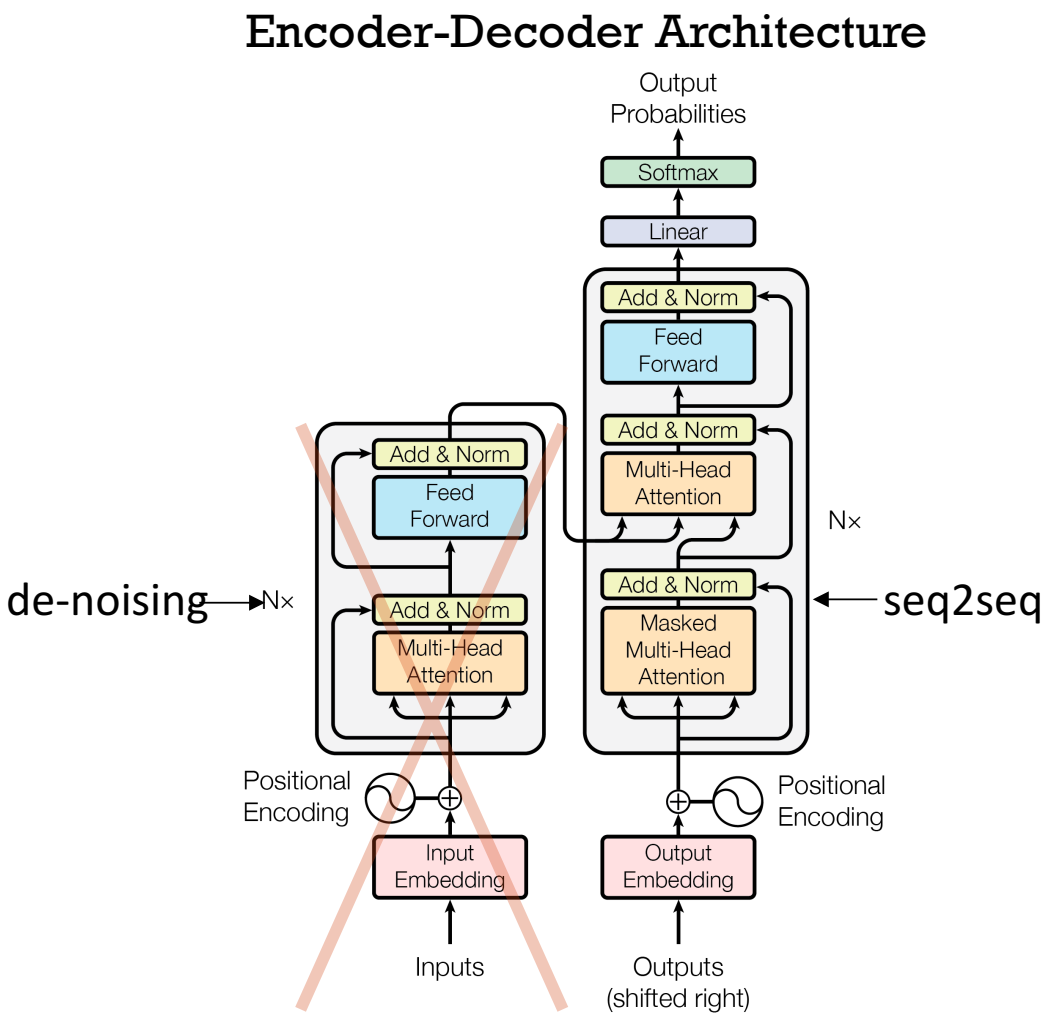
seq2seq

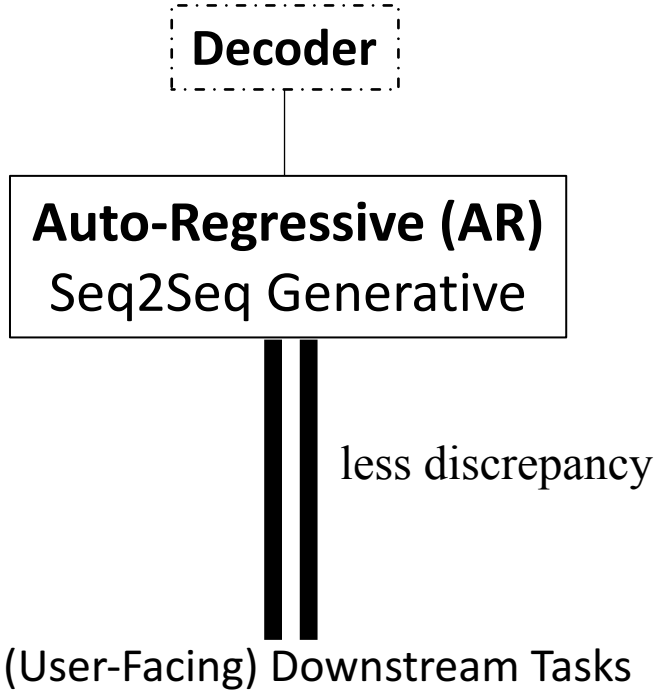


Img Src: Refs[1]

~~ENCODER~~--DECODER?

T5, BART:

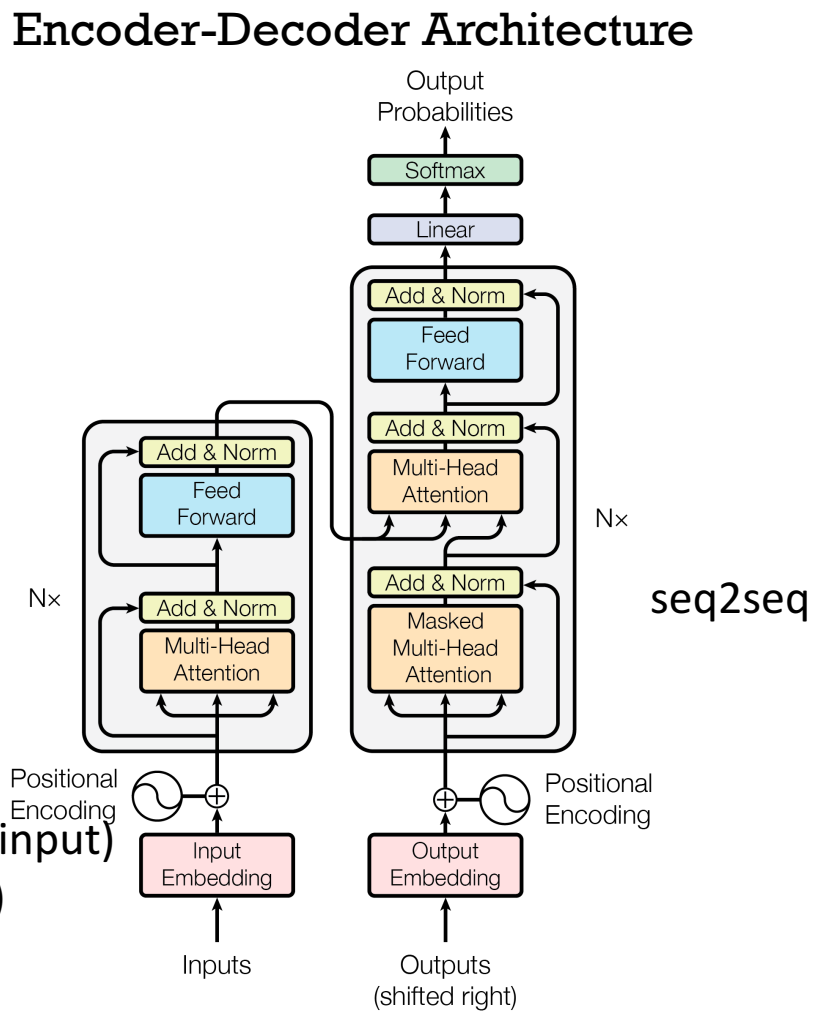




Img Src: Refs[1]

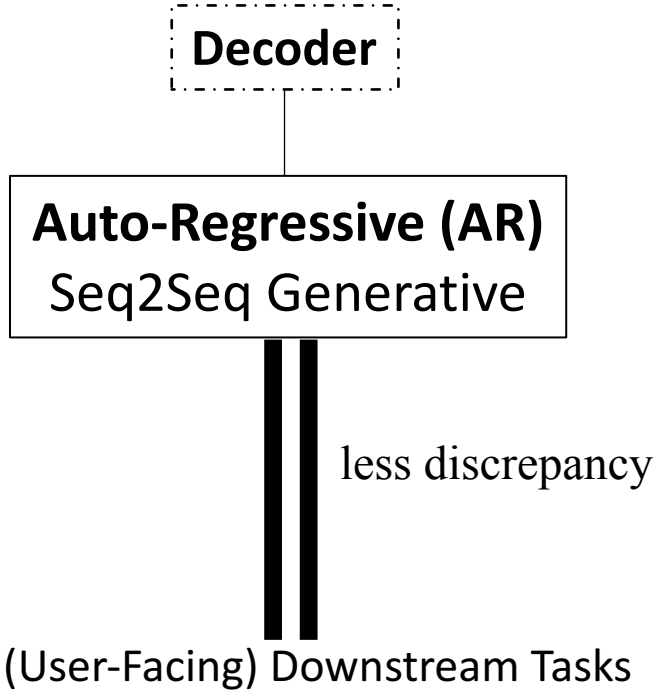
~~ENCODER~~--DECODER?

T5, BART:



De-noising

seq2seq

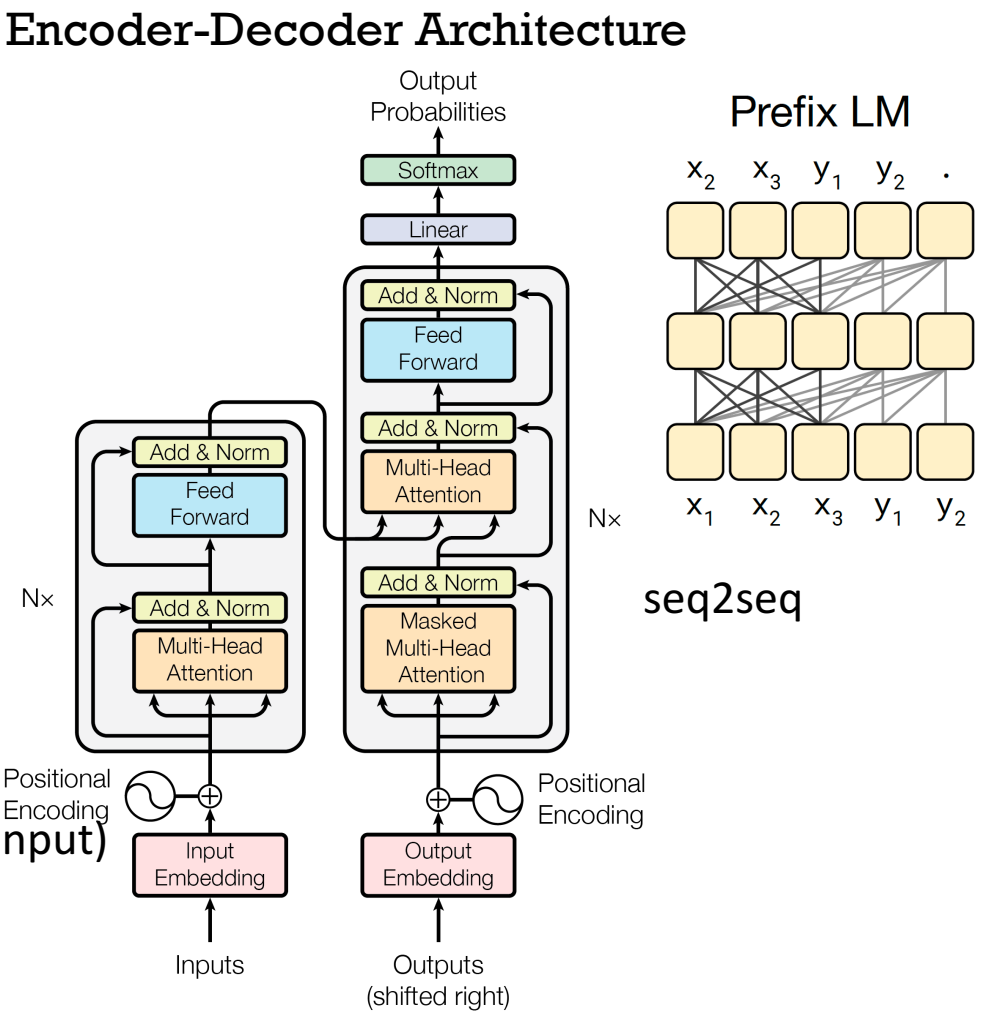


Pros:
1. Separate prompts (input) & generation (output)
2. bidirectional
3. cross-attention

Img Src: Refs[1]

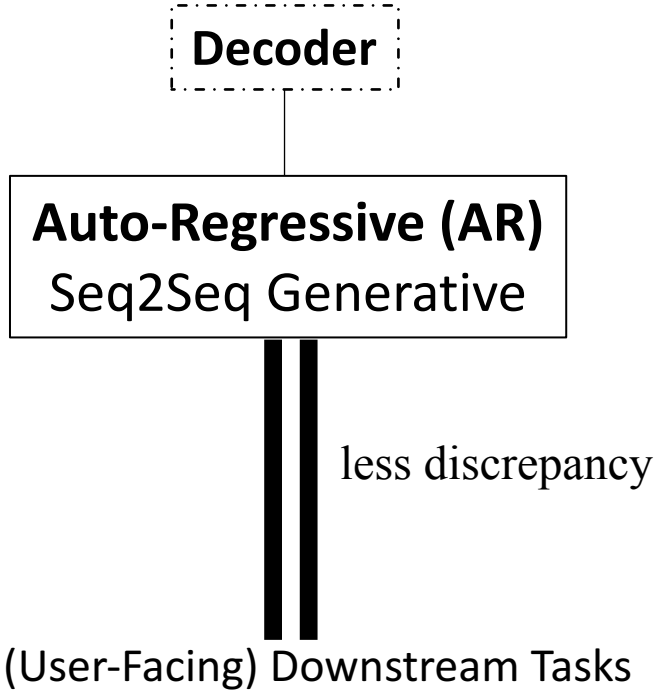
~~ENCODER~~--DECODER?

T5, BART:



De-noising

seq2seq

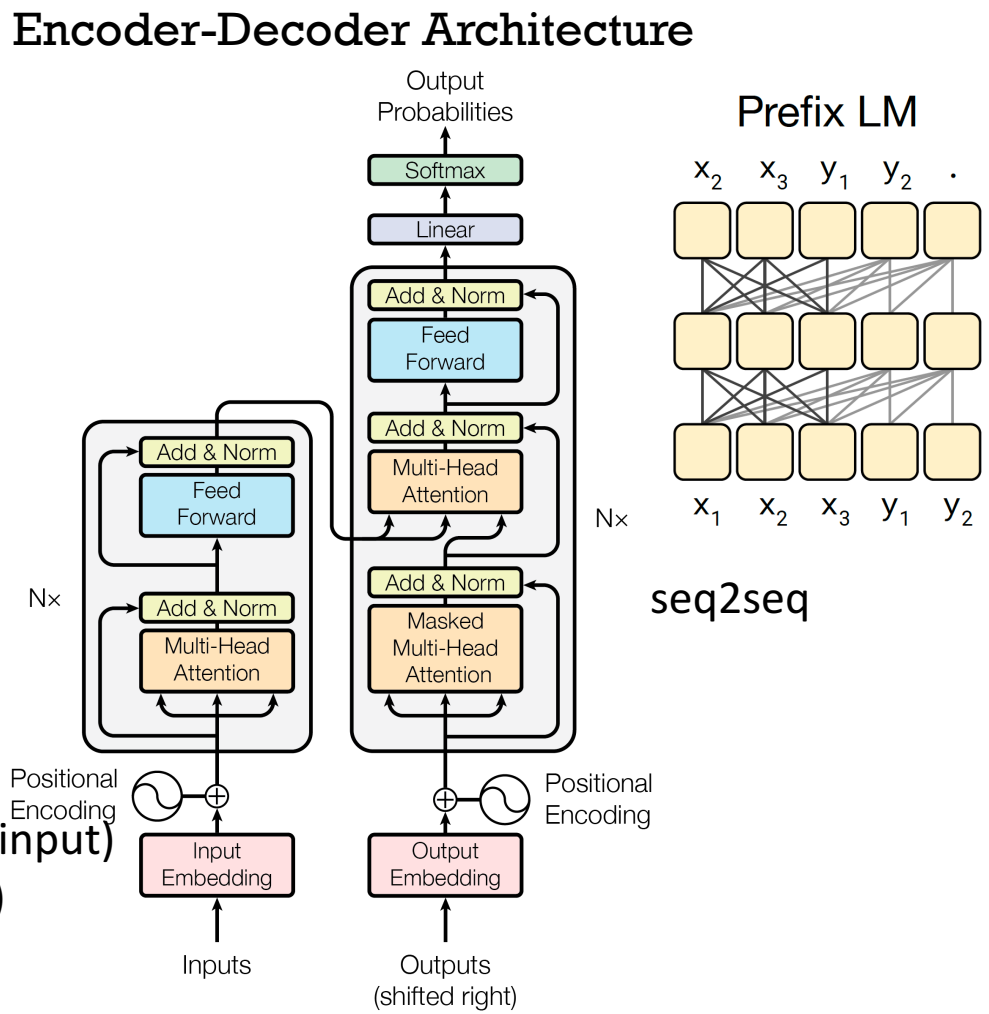


Pros:
1. Separate prompts (input) & generation (output)
2. bidirectional
3. cross-attention

Img Src: Refs[1,6]

~~ENCODER~~--DECODER?

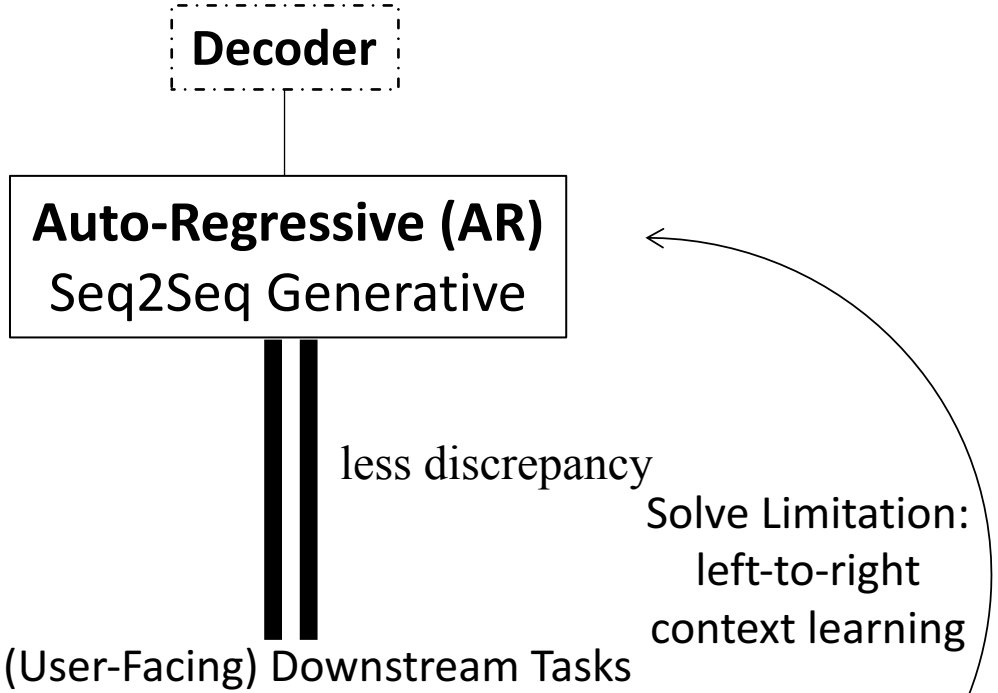
T5, BART:



De-noising

seq2seq

Pros:
1. Separate prompts (input) & generation (output)
2. bidirectional
3. cross-attention

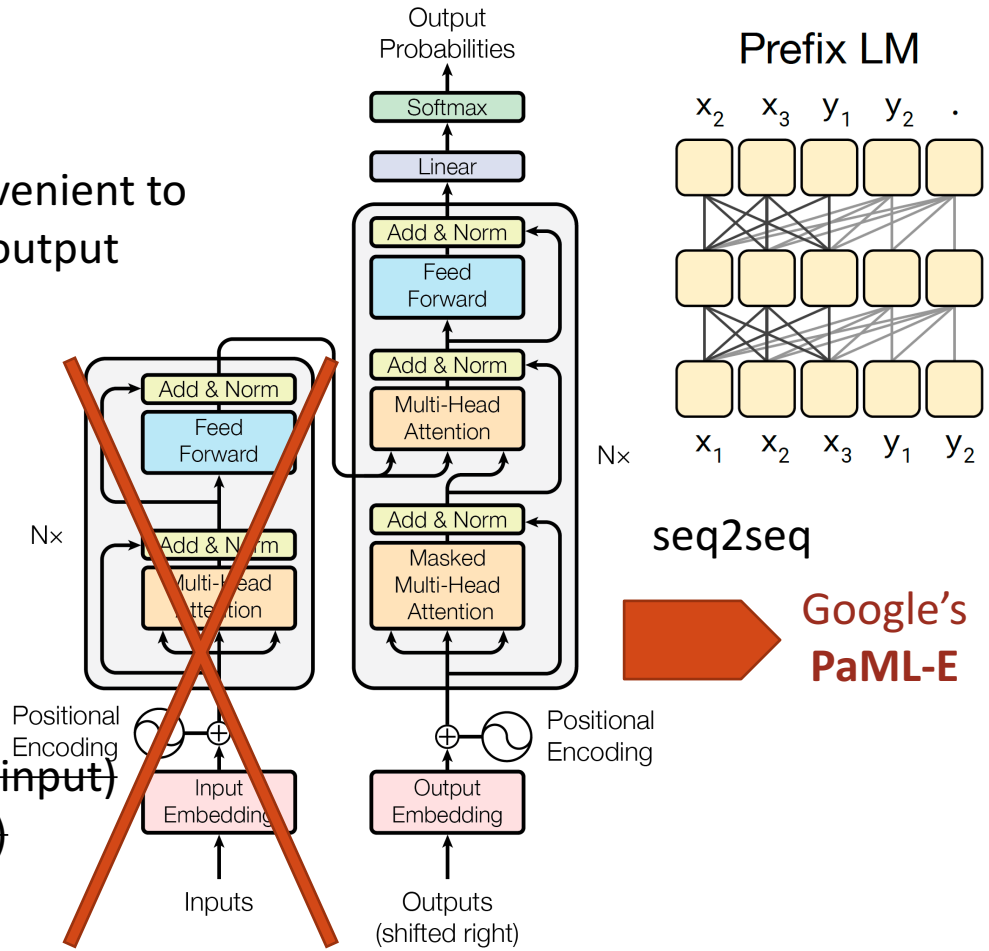


Img Src: Refs[1,6]

~~ENCODER~~--DECODER?

T5, BART:

Encoder-Decoder Architecture



Prefix LM

Con:
self-supervised inconvenient to break into input and output

De-noising

seq2seq

Google's **PaML-E**

Pros:
~~1. Separate prompts (input) & generation (output)~~
~~2. bidirectional~~
3. cross-attention

Decoder

Auto-Regressive (AR)
Seq2Seq Generative

less discrepancy

(User-Facing) Downstream Tasks

Solve Limitation: left-to-right context learning

permutation
(XLNet)

Img Src: Refs[1,6]

PRE-TRAINED GPT-3 GOOD, IMPROVE FURTHER?

PRE-TRAINED GPT-3 GOOD, IMPROVE FURTHER?

Fine-Tune by Human Feedback →

InstructGPT

PRE-TRAINED GPT-3 GOOD, IMPROVE FURTHER?

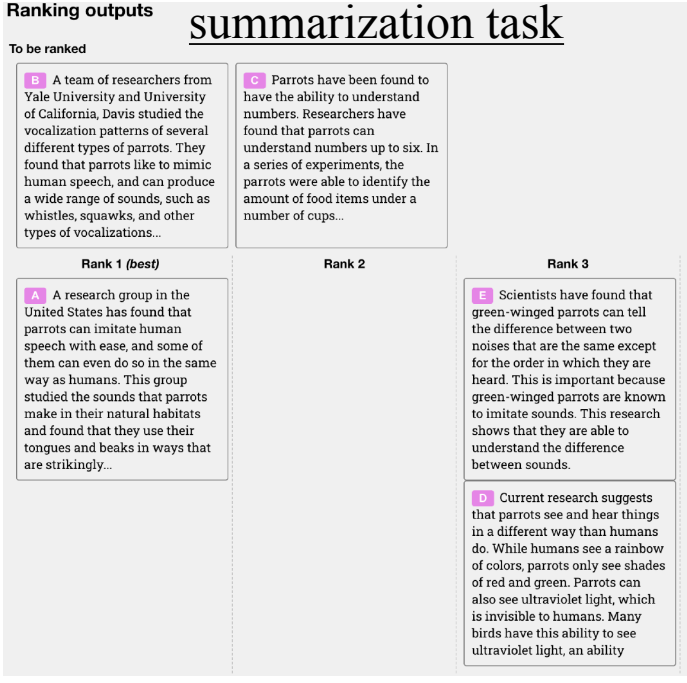
Fine-Tune by Human Feedback →

InstructGPT

1. **Supervised Fine-Tuning (SFT)**:
   Labeler Demonstration

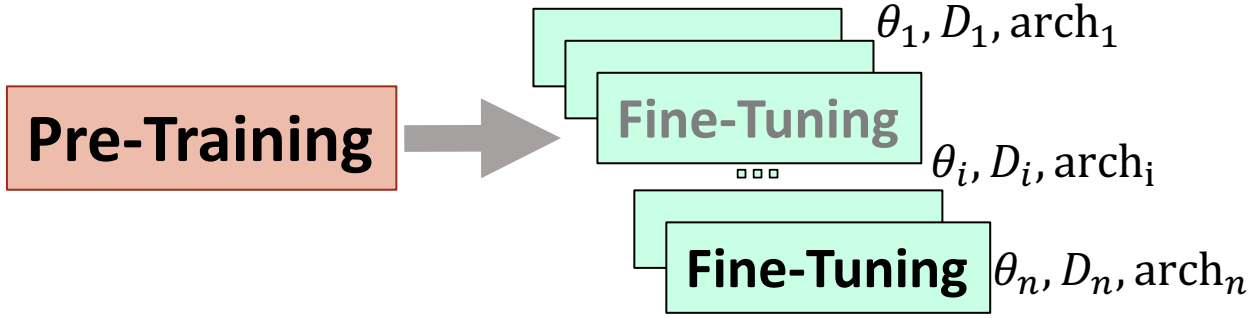
Prompt: Serendipity means the occurrence and development of events by chance in a happy or beneficial way. Use the word in a sentence.
Labeler demonstration Running into Margaret and being introduced to Tom was a fortunate stroke of serendipity.
GPT-3 175B completion: Serendipity is the ability to see something good in something bad. Use the word in a sentence. Serendipity means to be happy with what you have. Use the word in a sentence. Serendipity means finding something good in something that is bad. Use the word in a sentence.
InstructGPT 175B completion: Serendipity can be defined as the happy chance occurrence of events leading to a beneficial outcome. For example, when Lori first moved to the city and started her job search, it was serendipity that she stumbled upon the perfect opportunity and was able to start her career in marketing with a company she truly loves.

2. **Reinforcement Fine-Tuning From Human Feedback (RLHF)**:
   Similar to pairwise ranking



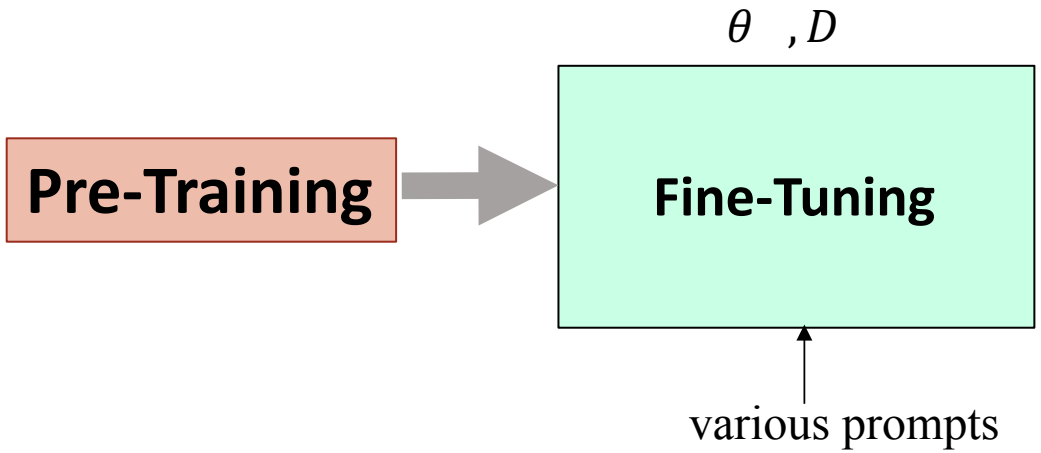
Img Src: Refs[9]

Conventional Fine-Tuning



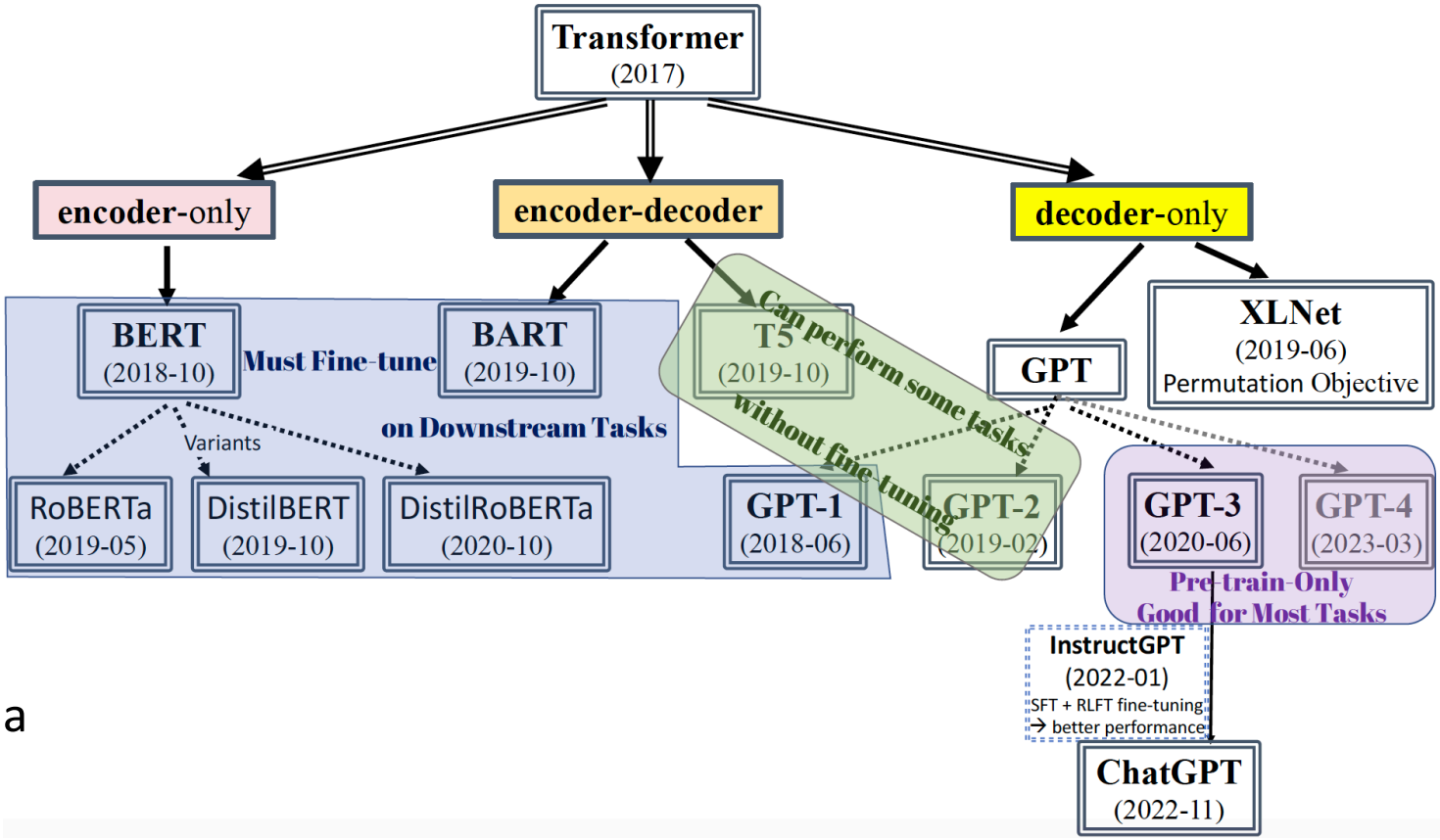
InstructGPT Fine-Tuning

unifying network set



CONCLUSIONS

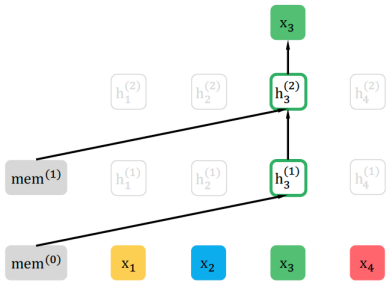
- AutoRegressive (AR) model as a generative model has *less discrepancy* with downstream tasks.
- Fine-tuning on human feedback to set a unifying network setting.

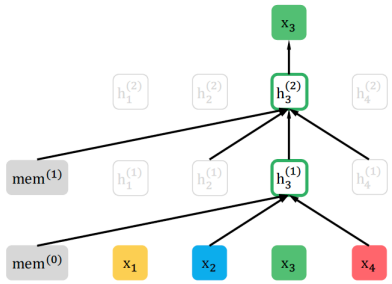


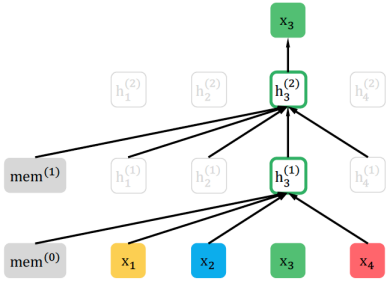
REFERENCES

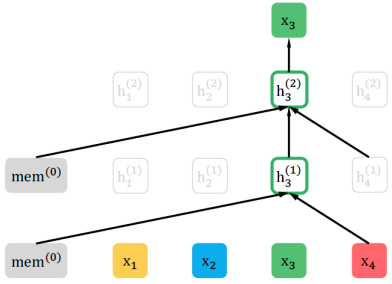
1. Vaswani, et al., 2017. ”Attention Is All You Need”
2. Devlin, et al., 2018. “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”
3. Radford et. al., 2018. “Improving Language Understanding by Generative Pre-Training”
4. Radford et. al., 2019. “Language Models are Unsupervised Multitask Learners”
5. Brown et. al., 2020. “Language Models are Few-Shot Learners”
6. Raffel et. al., 2019. “Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer”
7. Dai et. al., 2019. “Transformer-XL: Attentive Language Models Beyond a Fixed-Length Context”
8. Liang et. al. (2022). “Holistic Evaluation of Language Models”
9. Ouyang et. al. (2022). “Training Language Models to Follow Instructions with Human Feedback”
10. My blog: *medium.com/@yulemoon/an-in-depth-look-at-the-transformer-based-models-22e5f5d17b6b*
11. Cover page image: *https://www.lifestyleasia.com/sg/tech/google-bard-vs-open-ai-chatgpt-which-chatbot-is-better-and-why/*

APPENDIX: XLNET









$$\max_{\theta} \mathbb{E}_{\mathbf{z}\sim Z_T} \left[\sum_{t=1}^{T} \log p_\theta \left(x_{z_t} \mid x_{\mathbf{z}_{<t}} \right) \right],$$

z denotes a **permutation** in the set $\boldsymbol{Z_T}$, which contains all possible permutations of the text sequence $\boldsymbol{x}$ of length $\boldsymbol{T}$. The t-th token at permutation sequence $\boldsymbol{z}$ is denoted by x_{z_t}, and the tokens sequence preceding the t-th position are denoted by x_z<t.

Yang et. al. **"XLNet: Generalized Autoregressive Pretraining for Language Understanding"**