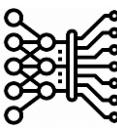


# Machine Learning Advances

# Artificial Intelligence



❖ 인간과 유사한 지능을 필요로 하는 스마트 머신 구축을 목표로 하는 컴퓨터 과학의 광범위한 분야

**Artificial Intelligence (AI):**

컴퓨터 시스템이 스스로  
특정 작업이나 행동을  
수행할 수 있는 능력

**Training Data (TD):**

기계 학습 알고리즘을  
학습시키는 데 사용되는 데이터

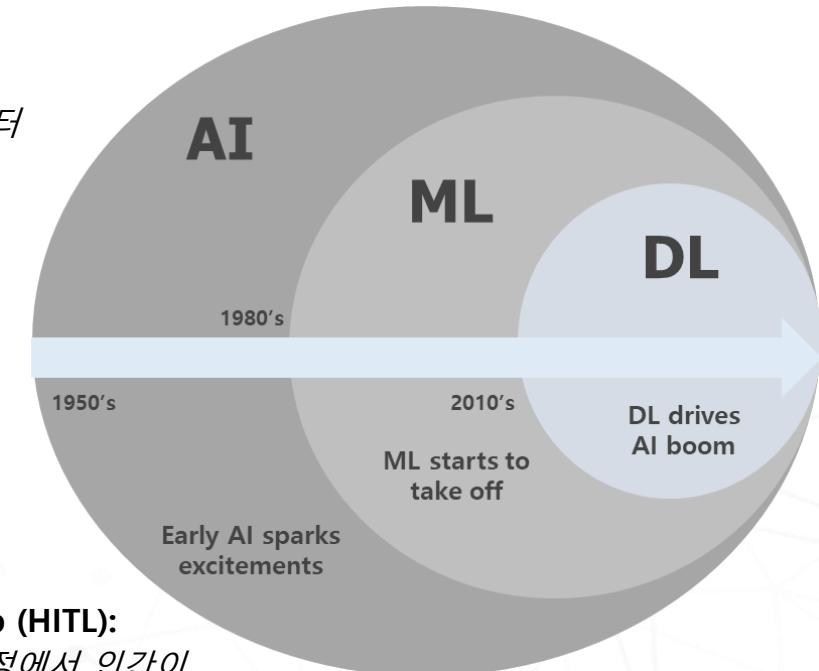
$$\text{AI} = \text{ML} + \text{TD} + \text{HITL}$$

**Machine Learning (ML) :**

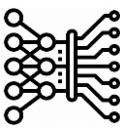
컴퓨터 시스템이 명시적인 지시 없이  
알고리즘과 통계 모델을 사용하여  
학습하는 능력

**Human in the Loop (HITL):**

알고리즘을 훈련하는 과정에서 인간이  
개입하는 것을 의미하며,  
하이퍼파라미터 튜닝과 같은 작업 포함



# AI Advances: Hypescale AI



## Hypescale AIs



Korea

Naver	Kakao Brain	SK Telecom	KT	LG AI Research
<b>HyperClova</b> <ul style="list-style-type: none"><li>Ingests Korean language data from 50 years of news and 9 years of blog</li><li>Understands context and speaks like a person</li><li>Creates sentences and makes phone calls</li></ul>	<b>KoGPT</b> <ul style="list-style-type: none"><li>Korean language model of GPT-3 minDALL-E</li><li>Text-based image generation AI model</li></ul>	<b>A.</b> <ul style="list-style-type: none"><li>Commercializes GPT-3 in Korean</li></ul>	<b>Mi:deum</b> <ul style="list-style-type: none"><li>Scheduled to launch in first half</li></ul>	<b>Exaone</b> <ul style="list-style-type: none"><li>Multimodal</li><li>Deploys billion parameter models</li></ul>



United States

OpenAI	Google
<b>ChatGPT</b> <ul style="list-style-type: none"><li>Interactive AI chatbot</li></ul>	<b>Bard</b> <ul style="list-style-type: none"><li>Similar to ChatGPT, but can talk about recent events</li><li>Scheduled to launch in coming weeks</li></ul>



China

Baidu
<b>Ernie Bot</b> <ul style="list-style-type: none"><li>Similar to ChatGPT</li><li>Will combine with search engine after launching stand alone application</li></ul>

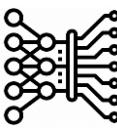
**OUTFRONT**

 PGIM +  Reuters Plus

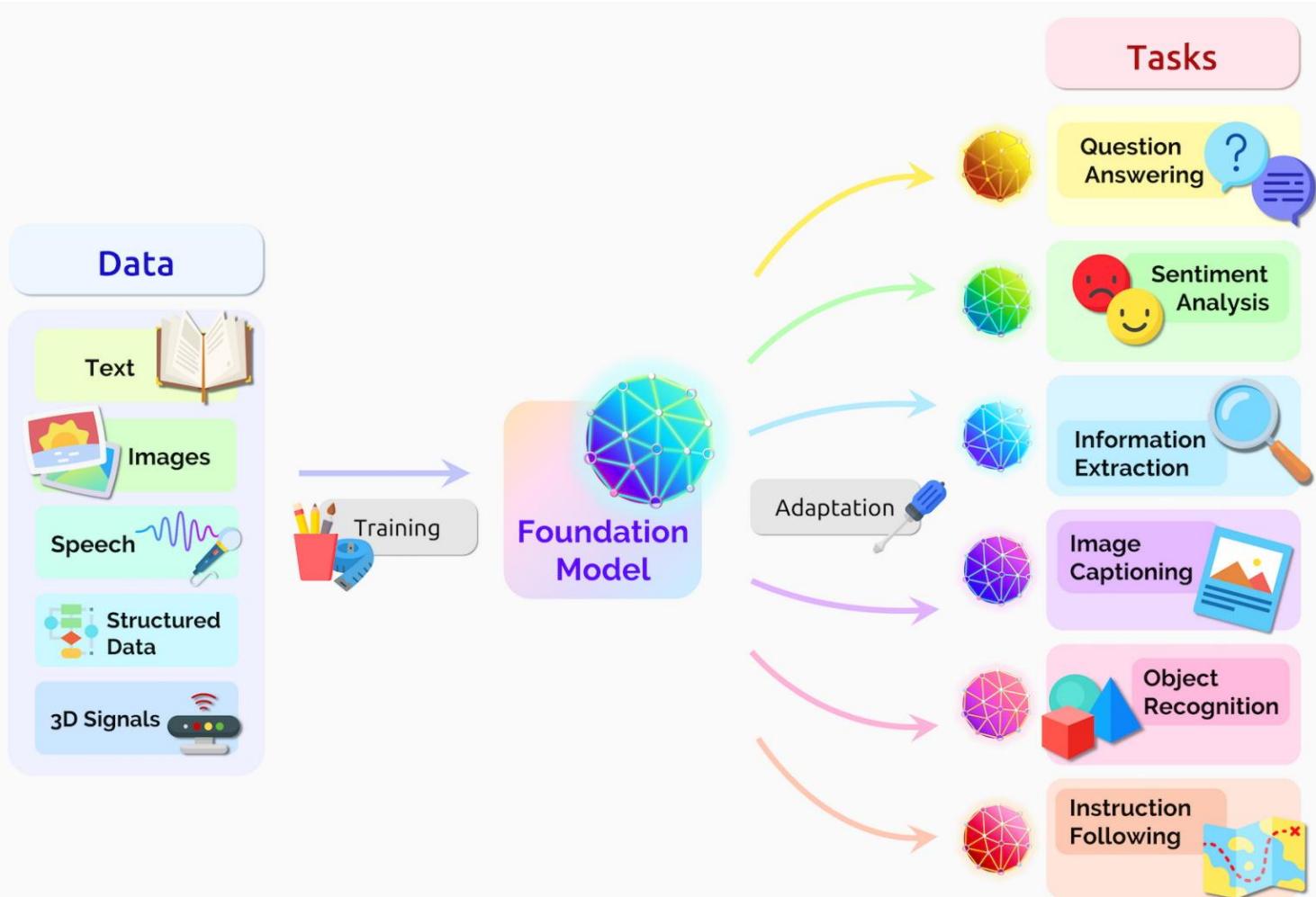
## HYPERSCALE: THE ARTIFICIAL INTELLIGENCE AND DATA CENTER REVOLUTION



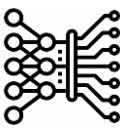
# AI Advances: Foundation Model



- Foundation model은 다양한 작업에 맞게 미세 조정할 수 있는 대규모 사전 학습 모델을 의미함
- 이 모델들은 대규모 다중 모달 데이터셋에서 일반적인 패턴을 학습하고, 특정 응용 분야에 맞게 적응 가능함



# What is a Large Language Model (LLM)?



## ❖ 정의

- Large Language Model (LLM)은 사람이 하는 표현과 유사한 자연어를 이해하고 생성하도록 설계된 AI 모델

## ❖ 주요 특징

- 방대한 양의 데이터를 기반으로 학습하여 일반적인 자연어를 예측하고 생성함
- Transformer 아키텍처를 기반으로 하여 텍스트의 장기적인 종속성을 효율적으로 처리할 수 있음

## ❖ 예시

- GPT-4, BERT, LLaMA



OpenAI - GPT4

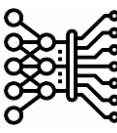


Google - Claude



Meta - LLaMA 3

# Transformer

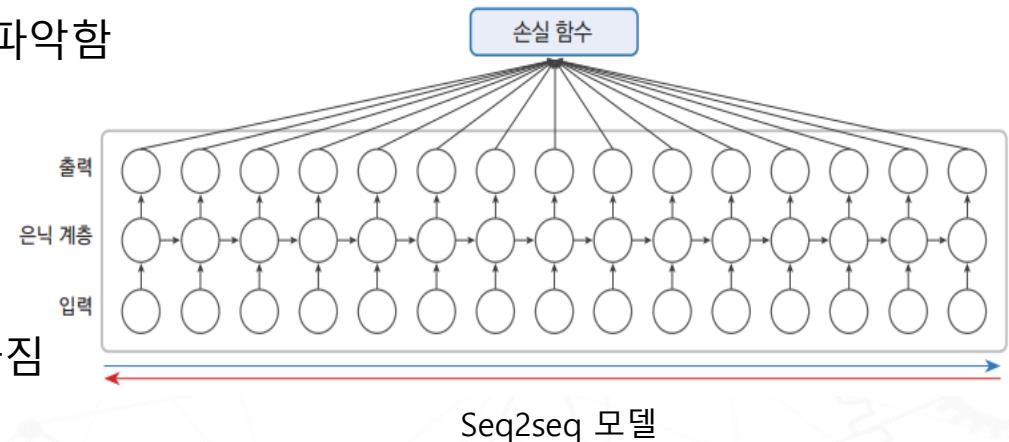


## ❖ 개념

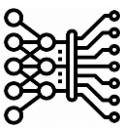
- Transformer는 여러 인공지능의 작업을 수행하기 위해 설계된 신경망 아키텍처
- 2017년 "Attention is All You Need" 논문에서 소개됨
- 핵심 기능: 순환 신경망(RNN)이나 합성곱 신경망(CNN)을 사용하지 않고도 시퀀스를 처리함
- Self-attention 메커니즘을 활용하여 데이터 요소 간의 관계를 파악함
- 적용 분야: 기계 번역, 텍스트 요약, 감정 분석 등

## ❖ Seq2Seq (Sequence to Sequence)

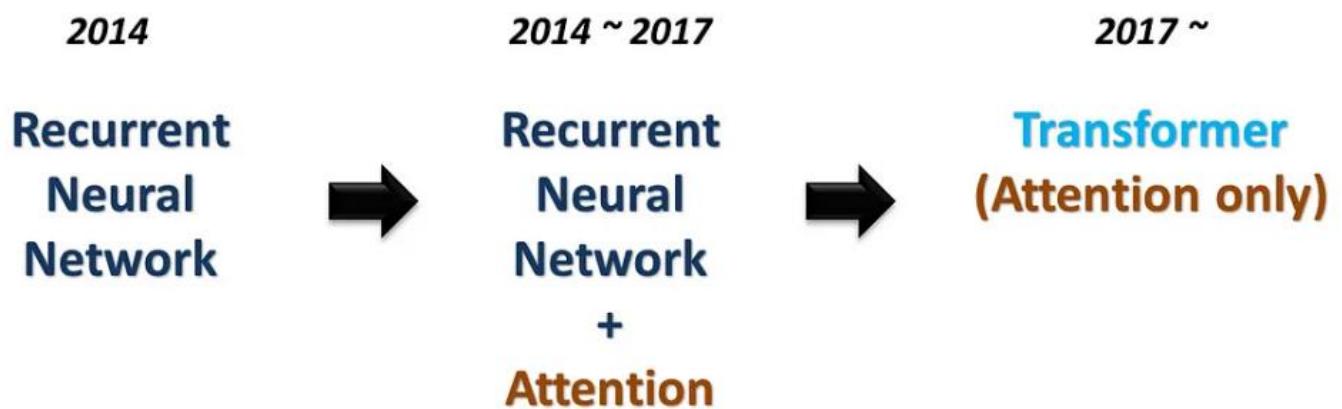
- Vanishing Gradient Problem
  - 역전파시 기울기가 사라져 학습이 어렵게 되는 문제를 가짐
- Exploding Gradient Problem
  - 역전파시 기울기가 급격히 커져 모델이 불안정해지는 문제를 가짐



# Advantages Over RNNs

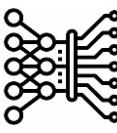


- ❖ Vanishing gradient problem 해결
  - 장기 의존성(long-term dependencies) 문제를 해결
- ❖ 병렬 처리
  - 모든 입력 토큰을 동시에 처리하여 계산 시간을 단축함
- ❖ 장기 의존성 처리
  - 긴 시퀀스에서 요소 간의 관계를 효과적으로 포착할 수 있음
- ❖ 더 빠른 학습 및 추론
  - 대규모 데이터셋과 실시간 작업에 특히 유리함



RNN부터 Transformer까지의 발전 과정

# Transformer



## ❖ 아키텍처

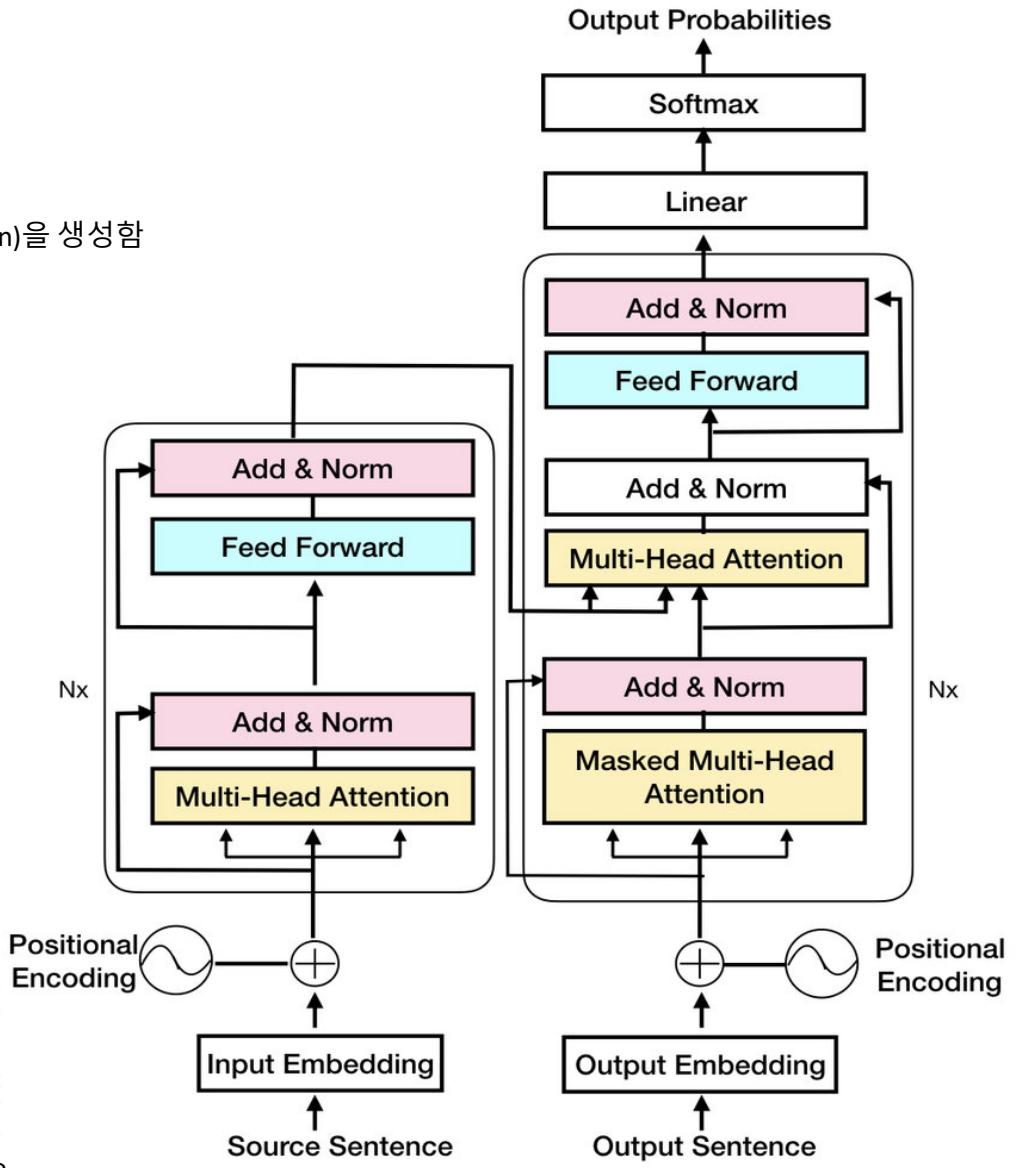
- 인코더-디코더 구조(Encoder-Decoder Structure)
  - 인코더(Encoder): 입력 시퀀스를 처리하고 문맥 표현(Context Representation)을 생성함
  - 디코더(Decoder): 인코더의 문맥 표현을 사용하여 출력 시퀀스를 생성함

## ❖ Transformer의 주요 특징

- Self-Attention 메커니즘
  - 시퀀스 내 요소 간의 관계를 파악함
  - 각 요소의 중요도에 따라 가중치 부여
- Multi-Head Attention
  - 여러 개의 Self Attention 메커니즘을 동시에 적용함
  - 데이터의 다양한 측면을 포착할 수 있음
- Positional Encoding
  - 시퀀스의 순서 정보를 추가함
  - 모델이 단어 또는 토큰의 위치를 구분할 수 있도록 도움
- 병렬 처리(Parallel Processing)
  - 입력 시퀀스의 모든 요소를 동시에 처리함
  - 순차적인 RNN보다 빠른 계산 속도를 달성함

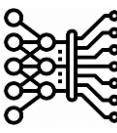
## ❖ 워크플로우

- Input Sequence → 인코더 → Context Representation
- Context Representation → 디코더 → Output Sequence
  - 트랜스포머는 긴 시퀀스 내 장기 의존성 문제를 효과적으로 처리할 수 있음



Transformer 아키텍쳐

# Self-Attention Mechanism



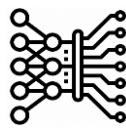
- ❖ 모든 입력 요소 간의 관계를 모델링
  - 입력 시퀀스 내의 모든 요소(e.g. 문장의 단어) 간의 관계를 모델링
- ❖ 각 요소 쌍에 대한 어텐션 점수(attention score) 계산
  - 입력 요소 쌍마다 어텐션 점수를 계산하여 두 요소 간의 관계를 나타냄
- ❖ 어텐션 점수는 한 요소가 다른 요소에 얼마나 중요한지를 결정
  - 특정 요소가 다른 요소에 미치는 중요도를 어텐션 점수를 통해 판단
- ❖ 시퀀스의 중요한 부분에 집중하도록 모델을 돋는 역할
  - 모델이 입력 시퀀스에서 관련성 높은 부분에 집중할 수 있도록 지원

"The **cat** sat on the mat because **it** was tired."



어텐션 점수가 높은 것으로 관계성 파악

# Multi-Head Attention & Positional Encoding

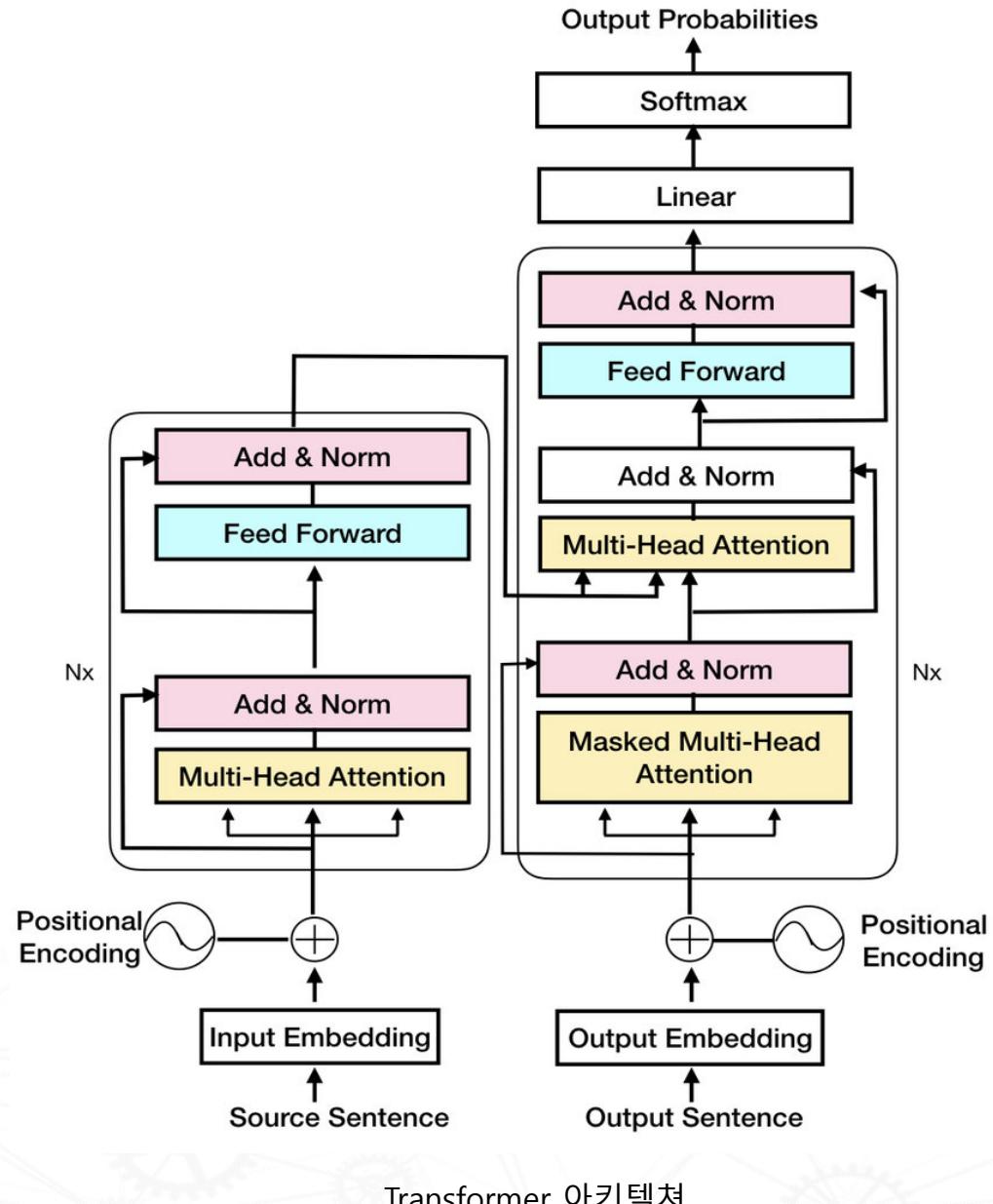


## ❖ Multi-head attention

- 여러 개의 셀프 어텐션 메커니즘을 병렬로 실행
- 각 "헤드"는 서로 다른 패턴이나 관계를 캡처
- 결과를 결합하여 입력에 대한 종합적인 이해를 생성
- 모델이 여러 관점을 동시에 고려할 수 있도록 함

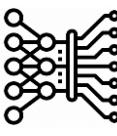
## ❖ Positional encoding

- 셀프 어텐션은 토큰의 순서를 고려하지 않음
- 위치 인코딩은 사인 및 코사인 함수를 사용하여 시퀀스 정보를 추가
- 모델이 입력에서 각 토큰의 위치를 알 수 있도록 함
  - Learnable positional encoding
  - Relative positional encoding



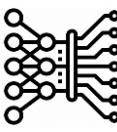
Transformer 아키텍쳐

# Limitations of LLM



- ❖ 정보의 정확성 문제
  - 훈련된 데이터에만 의존하여 최신 정보나 특정 도메인의 깊이 있는 정보에 대한 답변을 제공하는 데 한계가 존재
- ❖ 모델의 크기와 효율성
  - 대형 언어 모델은 매우 크고 무겁기 때문에 실시간 응답을 제공하는 데 있어 비효율적
- ❖ 맥락 유지의 어려움
  - 긴 대화나 복잡한 질문의 경우, 맥락을 유지하면서 정확한 답변을 제공하는 데 어려움
- ❖ 데이터 편향 문제
  - LLM은 훈련 데이터의 편향을 그대로 반영할 수 있으며, 이는 부정확하거나 편향된 답변을 초래 가능

# Retrieval-Augmented Generation (RAG)



## ❖ 정의

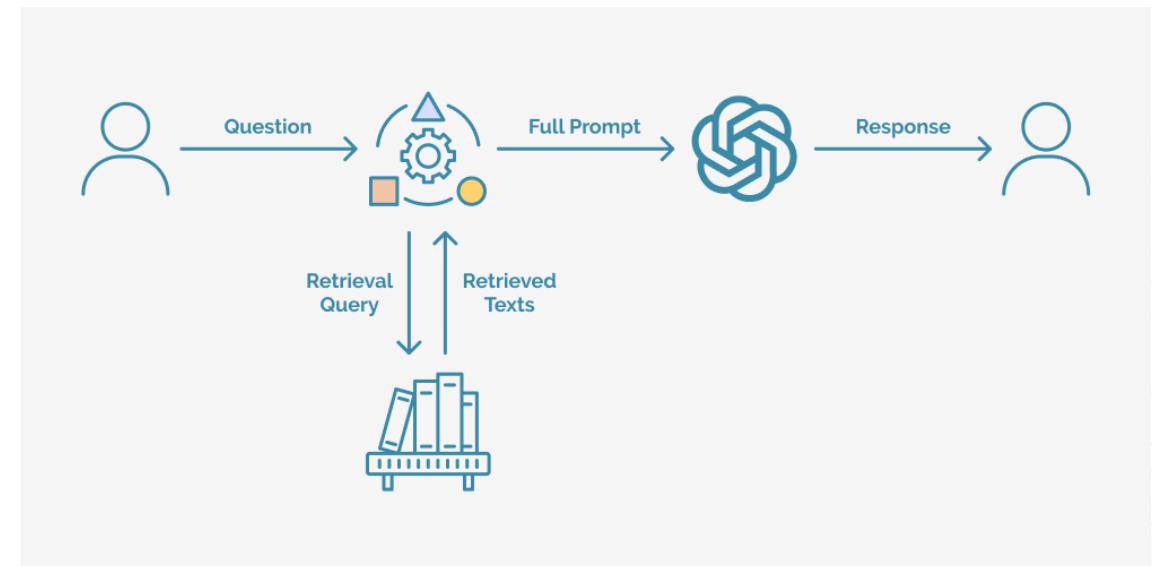
- 언어 모델과 검색 메커니즘을 결합하여 외부 지식 베이스에 접근하는 하이브리드 모델

## ❖ 구성 요소

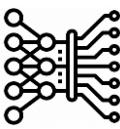
- 검색기 (Retriever)
  - 데이터베이스에서 관련 문서를 찾음
- 생성기 (Generator)
  - 검색된 문서를 사용하여 정확하고 맥락에 맞는 응답을 생성

## ❖ 장점

- 정확성을 향상시키고, 사실 데이터에 기반한 답변을 제공하여 환각 현상(Hallucination)을 줄임



Retrieval Augmented Generation



# Retrieval-Augmented Generation (RAG)

## ❖ Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks

### Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks

Patrick Lewis<sup>†‡</sup>, Ethan Perez<sup>\*</sup>,

Aleksandra Piktus<sup>†</sup>, Fabio Petroni<sup>†</sup>, Vladimir Karpukhin<sup>†</sup>, Namana Goyal<sup>†</sup>, Heinrich Küttler<sup>†</sup>,

Mike Lewis<sup>†</sup>, Wen-tau Yih<sup>†</sup>, Tim Rocktäschel<sup>†‡</sup>, Sebastian Riedel<sup>†‡</sup>, Douwe Kiela<sup>†</sup>

<sup>†</sup>Facebook AI Research; <sup>‡</sup>University College London; <sup>\*</sup>New York University;  
plewis@fb.com

#### Abstract

Large pre-trained language models have been shown to store factual knowledge in their parameters, and achieve state-of-the-art results when fine-tuned on downstream NLP tasks. However, their ability to access and precisely manipulate knowledge is still limited, and hence on knowledge-intensive tasks, their performance lags behind task-specific architectures. Additionally, providing provenance for their decisions and updating their world knowledge remain open research problems. Pre-trained models with a differentiable access mechanism to explicit non-parametric memory have so far been only investigated for extractive downstream tasks. We explore a general-purpose fine-tuning recipe for retrieval-augmented generation (RAG) — models which combine pre-trained parametric and non-parametric memory for language generation. We introduce RAG models where the parametric memory is a pre-trained seq2seq model and the non-parametric memory is a dense vector index of Wikipedia, accessed with a pre-trained neural retriever. We compare two RAG formulations, one which conditions on the same retrieved passages across the whole generated sequence, and another which can use different passages per token. We fine-tune and evaluate our models on a wide range of knowledge-intensive NLP tasks and set the state of the art on three open domain QA tasks, outperforming parametric seq2seq models and task-specific retrieve-and-extract architectures. For language generation tasks, we find that RAG models generate more specific, diverse and factual language than a state-of-the-art parametric-only seq2seq baseline.

#### 1 Introduction

Pre-trained neural language models have been shown to learn a substantial amount of in-depth knowledge from data [47]. They can do so without any access to an external memory, as a parameterized implicit knowledge base [51, 52]. While this development is exciting, such models do have downsides: They cannot easily expand or revise their memory, can't straightforwardly provide insight into their predictions, and may produce "hallucinations" [38]. Hybrid models that combine parametric memory with non-parametric (i.e., retrieval-based) memories [20, 26, 48] can address some of these issues because knowledge can be directly revised and expanded, and accessed knowledge can be inspected and interpreted. REALM [20] and ORQA [31], two recently introduced models that combine masked language models [8] with a differentiable retriever, have shown promising results,

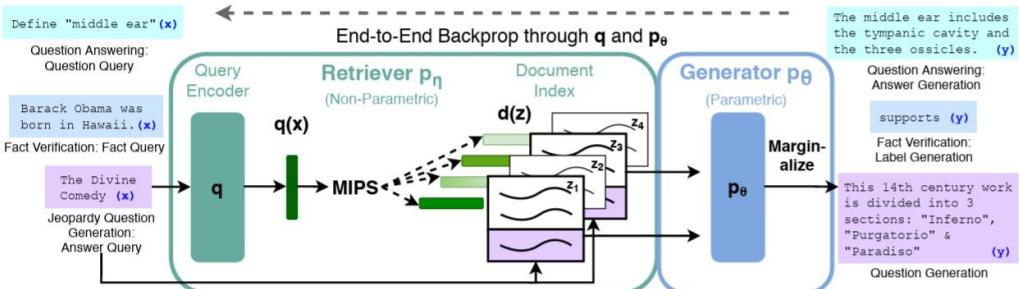


Figure 1: Overview of our approach. We combine a pre-trained retriever (*Query Encoder + Document Index*) with a pre-trained seq2seq model (*Generator*) and fine-tune end-to-end. For query  $x$ , we use Maximum Inner Product Search (MIPS) to find the top-K documents  $z_i$ . For final prediction  $y$ , we treat  $z$  as a latent variable and marginalize over seq2seq predictions given different documents.

[View full instructions](#)

[View tool guide](#)

Note: Some questions are control questions. We require good accuracy on our control questions to accept responses.

Indicate which one of the following sentences is more factually true with respect to the subject. Using the internet to check whether the sentences are true is encouraged.

Which sentence is more factually true?

**Subject : Hemingway**

**Sentence A :** "The Sun Also Rises" is a novel by this author of "A Farewell to Arms"

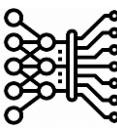
**Sentence B :** This author of "The Sun Also Rises" was born in Havana, Cuba, the son of Spanish immigrants

**Select an option**

Sentence A is more true	1
Sentence B is more true	2
Both sentences are true	3
Both sentences are completely untrue	4

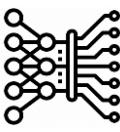
Figure 4: Annotation interface for human evaluation of factuality. A pop-out for detailed instructions and a worked example appear when clicking "view tool guide".

# Retrieval-Augmented Generation (RAG)



- ❖ 정보의 정확성 향상
  - 질문에 답변하기 위해 외부 지식을 검색하는 단계를 추가하여 최신 정보나 훈련 데이터에 포함되지 않은 정보를 포함한 답변을 제공 가능
- ❖ 모델의 크기와 효율성 개선
  - 필요한 경우에만 외부 문서를 검색하고 이를 기반으로 답변을 생성하기 때문에, 모든 정보를 모델 내부에 포함해야 하는 부담을 줄일 수 있음
- ❖ 맥락 유지 능력 강화
  - 검색된 문서를 바탕으로 답변을 생성하기 때문에, 복잡한 질문이나 긴 대화에서도 관련성 높은 정보를 제공 가능
- ❖ 데이터 편향 문제 완화
  - 다양한 출처에서 정보를 검색하기 때문에, 특정 데이터에 편향되지 않고 더 균형 잡힌 답변을 제공할 가능성이 높고, 검색된 문서를 통해 답변의 출처를 명확히 할 수 있어 투명성을 높일 수 있음

# Parameter Efficient Fine-Tuning (PEFT)

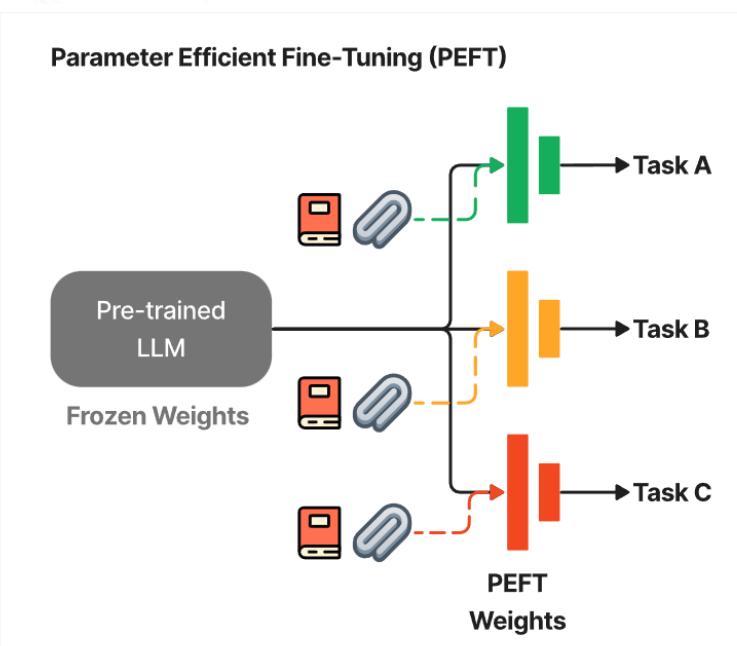


## ❖ 정의

- 특정 데이터셋에 대해 사전 훈련된 모델을 조정하여 특정 작업에 대한 성능을 향상시키는 과정

## ❖ Parameter efficient fine-tuning

- LoRA (Low-Rank Adaptation): 계산 비용을 줄이기 위해 일부 파라미터만 업데이트
- 양자화 (Quantization): 가중치를 더 적은 비트로 표현하여 모델 크기 축소

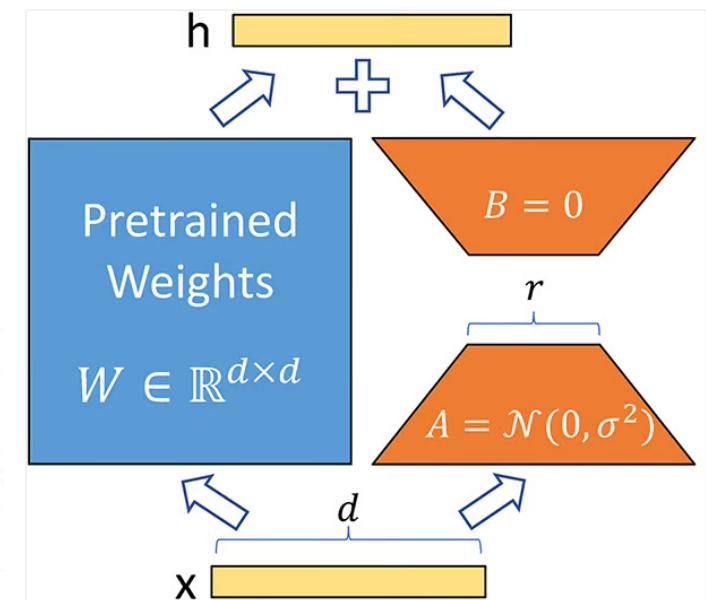


GPU Usage: 22GB

GPU Usage: 22GB

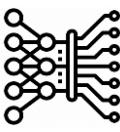
GPU Usage: 9GB

Parameter efficient fine-tuning



Low-rank adaption

# NLP: AI Author for Book Reading



## ❖ 기계학습 모델

- 사전 훈련된 모델 파인튜닝을 위한 데이터 전처리  
(토큰 추가) 모듈
- 사전 훈련된 모델 로딩 모듈
- 텍스트에 부분화 기법을 적용한 데이터 전처리 모듈
- 파인튜닝을 위한 질문과 답변 데이터셋 생성 모듈

## ❖ 하이퍼파라미터 선택 모듈

- 샘플 데이터를 기반으로 최적화  
(애거사 크리스티 – 병어리 목격자)

```
[43]: print("Q: ",pd.read_csv("../qa_generation/QA_Results/20240822_073259_QuestionAnswer.csv").iloc[6][0])
print()
print("A: ", pd.read_csv("../qa_generation/QA_Results/20240822_073259_QuestionAnswer.csv").iloc[6][1])
```

Q: \*\*타니オス 부인의 내적 갈등\*\*: 타니オス 부인은 왜 자신의 남편과 상의하기 전까지는 확실한 결정을 내리지 못하고 있나요? 그녀의 내적 갈등에 대해 추측해 보세요.

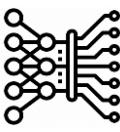
A: 타니オス 부인의 내적 갈등은 그녀가 겪는 복잡한 감정과 상황에서 비롯됩니다. 이를 이해하려면 그녀의 배경과 결혼 생활, 그리고 그녀와 남편 사이에 얹힌 속사정을 생각해 볼 필요가 있습니다.

타니オス 부인은 아마도 남편 타니オス 의사의 깊이 사랑하고 존경하며, 그의 판단을 신뢰하고 있을 가능성이 큽니다. 그러나 그녀 또한 자신만의 생각과 감정을 가지고 있어서, 어떤 결정들에 대해 내적 갈등을 느낄 수밖에 없을 것입니다. 남편과 상의하기 전까지 확실한 결정을 내리지 못하는 이유는 여러 가지로 추측해 볼 수 있습니다:

1. \*\*신뢰와 의존\*\*: 타니オス 부인은 남편의 전문성과 경험을 신뢰합니다. 중요한 결정을 내리기 전에 남편과 상의해야만 맘이 놓이기 때문에 심리적으로 의존하고 있는 것 같습니다. 이를 통해 그녀는 결정의 무게를 함께 나누고자 합니다.
2. \*\*두려움과 불안\*\*: 그녀는 혼자 결정을 내리기에 앞서 두려움과 불안을 느낄 수 있습니다. 남편과 상의함으로써 그녀는 자신이 올바른 선택을 하고 있다는 확신을 얻고 불안을 덜고자 할 것입니다.
3. \*\*배우자로서의 역할\*\*: 그녀는 가정 내에서 남편과 공동으로 결정하는 것이 자신의 역할이라고 생각할 수 있습니다. 남편이든 부인이든 상대방의 의견을 존중하고 협력하여 결정을 내리는 과정을 통해 단단한 유대감을 유지하고자 할 수 있습니다.
4. \*\*복잡한 상황\*\*: 현재 상황이 매우 복잡하고 다중적인 측면이 있을 때, 한 사람의 단독 결정보다는 두 사람의 협력이 더 나은 결과를 가져올 수 있다고 판단할 수 있습니다. 이를 통해 다양한 가능성과 시각을 고려할 수 있기 때문입니다.

질문-답변 데이터 생성 예시

# NLP: AI Author for Book Reading

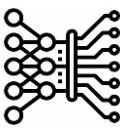


- ❖ 단방향 커뮤니케이션 챗봇 → 양방향 인터렉션형 커뮤니케이션 챗봇
  - 훈련 데이터 (4권): 데일 카네기 – 인간관계론, 자기관리론, 인생경영론, 성공대화론
  - 주어진 질문에 대한 답변을 생성하는 단방향 챗봇 기능 구현
- ❖ 최적화된 모델 GPU 사용 서버 비용 절감을 위해 PEFT (Parameter Efficient Fine-Tuning) 적용

```
명령 프롬프트 명령 프롬프트 명령 프롬프트
C:\Users\rb917>curl -X POST https://d2f9-118-44-139-122.ngrok-free.app/chat -H "Content-Type: application/json; charset=utf-8" -d "{\"input\": \"안녕하세요! 반갑습니다. 저는 오늘도 기분이 좋고, 좋은 하루를 보내고 있습니다.\", \"time\": \"25.6570 sec\"}"
{"output": "안녕하세요! 반갑습니다. 저는 오늘도 기분이 좋고, 좋은 하루를 보내고 있습니다.", "time": "25.6570 sec"}
C:\Users\rb917>curl -X POST https://d2f9-118-44-139-122.ngrok-free.app/chat -H "Content-Type: application/json; charset=utf-8" -d "{\"input\": \"여러 책들을 통해서 본 데일 카네기는 어떤 사람인 것 같아?!\", \"time\": \"24.0801 sec\"}"
{"output": "데일 카네기는 미국의 사업가이자 자선가로, 그의 이름은 자본주의와 자선 활동의 상징으로 여겨집니다. 그는 철저한 비즈니스맨으로서 성공을 위해 모든 노력을 기울였으며, 특히 철강 산업에서 두각을 나타냈습니다. 카네기는 자신의 사업을 통해 막대한 부를 쌓았지만, 동시에 사회적 책임을 다하기 위해 많은 기부와 자선 활동을 펼쳤습니다. 그는 사람들에게 긍정적인 영향을 미치고자 하는 열망이 강했으며, 이를 위해 자신의 재산을 사회에 환원하는 데 큰 노력을 기울였습니다. 카네기는 자신의 성공을 통해 얻은 부를 사회에 환원함으로써, 다른 사람들에게도 기회를 제공하고자 했습니다. 그는 또한 교육과 문화에 대한 지원을 아끼지 않았으며, 이를 통해 많은 사람들에게 긍정적인 영향을 미쳤습니다."}
카네기의 삶은 단순한 성공의 이야기에서 그치지 않고, 우리가 사회적 책임을 다하는 방법에 대한 깊은 통찰을 제공합니다. 그는 부를 쌓는 것뿐 아니라, 그 부를 어떻게 사용할 것인가에 대한 고민을 통해 우리에게 많은 교훈을 줍니다. 그의 삶은 우리가 어떻게 더 나은 세상을 만들 수 있는지에 대한 질문을 던지며, 우리가 함께 나아가야 할 방향을 제시합니다.", "time": "24.0801 sec"
C:\Users\rb917>
```

답변 생성 예시

# NLP: AI Author for Book Reading



## ❖ 질문-답변 데이터셋 필터링 모듈

- GPT-4o mini 모델을 사용하니 저품질의 응답이 생성
  - 평균적으로 응답의 약 10%가 낮은 품질
- 훈련 데이터에서 "모르겠습니다"라는 응답을 필터링하는 기능

## ❖ 2차 창작물 생성 가능성을 줄이는 기능

- 특정 키워드가 포함된 경우 답변 생성을 거부하는 기능 예: "요약", "감상문" 등과 같은 키워드

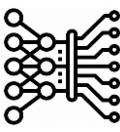
```
# 필터링된 결과 확인
print(df.shape, filtered_df.shape)
# print(filtered_df.shape)

[1]:    ✓ 0.0s
      (1665, 2) (162, 2)
```

filtered_df		Python
✓ 0.0s		
Question	Answer	
9 그랜트 장군의 두통과 리 장군의 항복 사이에는 어떤 연관성이 있나요?	죄송하지만, 그랜트 장군의 두통과 리 장군의 항복 사이의 연관성에 대한 정보는 가지...	
13 본문에서 언급된 네 가지 관절염의 요인 중 가장 개인적으로 공감되는 것은 무엇인가요?	죄송하지만, 본문에서는 네 가지 관절염의 요인에 대한 언급이 없습니다. 따라서 어떤...	
36 리 장군의 군대가 리치몬드를 탈출하는 과정에서 어떤 감정적 변화가 있었는지 설명해...	죄송하지만, 리 장군의 군대가 리치몬드를 탈출하는 과정에서의 감정적 변화에 대한 정...	
39 관절염의 주된 원인으로 지목된 네 가지 요인 중 가장 공감되는 부분은 무엇인가요?	죄송하지만, 제가 알고 있는 정보로는 관절염의 주된 원인으로 지목된 네 가지 요인에...	
54 그랜트 장군의 두통은 실제로 어떤 감정적 요인으로 인해 발생했나요?	죄송하지만, 그랜트 장군의 두통에 대한 구체적인 감정적 요인은 제공된 문맥에서는 언...	

필터링 데이터 예시

# NLP: AI Author for Book Reading



## ❖ 훈련 데이터 필터링 모듈

- 질문-답변 데이터셋 필터링 모듈
  - 훈련 데이터에서 "모르겠습니다"라는 응답을 필터링하는 모듈

## ❖ 2차 창작물 생성을 차단하는 모듈

- 특정 키워드가 포함된 경우 답변 생성을 거부하는 모듈
  - 추가된 키워드: 요약, 감상문, 각색, 패러디

## ❖ 질문 정확도 향상을 위한 필터링 모듈

- 학습 데이터 중 질문에 대해 "모르겠습니다"라는 내용이 포함된 응답을 필터링하는 모듈 생성
- "프롬프트 엔지니어링" 기법을 사용하여 질문 형식을 수정하고 챗봇의 역할을 강화하는 모듈 적용

```
[18]: # 테스트
Question = "이 책의 내용을 요약해줘"
filtering_question(Question)
print()
Question = "이 책으로 감상문 작성해줘"
filtering_question(Question)

Question: 이 책의 내용을 요약해줘
Generation: 해당 질문은 저작권에 의해 답변드릴 수 없습니다.

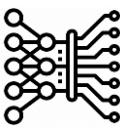
Question: 이 책으로 감상문 작성해줘
Generation: 해당 질문은 저작권에 의해 답변드릴 수 없습니다.
```

```
prompt_template = """
<s> [INST] 당신은 작가입니다.
독자의 질문의 핵심만 파악하여 간결하게 답변하세요.
불필요한 설명은 피하며 요구된 정보만 검토 후 제공하세요.
모르는 답변은 책의 내용을 참고해서 답변하세요.

내용: {context}
질문 : {question}
답변: [/INST]
"""
```

2차 창작물 생성 키워드 차단 모듈과 프롬프트 엔지니어링 예시

# NLP: AI Author for Book Reading



## ❖ 최적화 및 안정화

- 텍스트의 맥락을 이해하고 답변을 생성하는 기능 추가하여 환각을 줄임
  - 자연어 벡터 유사도 검색 라이브러리를 사용하여 질문과 유사한 책의 내용을 탐색
  - 책의 내용을 활용하여 배경 지식을 제공하고, 이를 통해 더 정확한 답변을 생성하도록 함

```
Question = "데일 카네기는 이 책을 왜 썼어?"  
Inference(Question)
```

Question: 데일 카네기는 이 책을 왜 썼어?

Generation: 저는 그에 대해 잘 모르겠습니다. 제  
가 도와드릴 수 있는 다른 질문이 있으시면 말씀해  
주세요.

```
[5]: Question = "인간관계론에서 중요한 것이 뭐야?"  
Inference(Question)
```

Question: 인간관계론에서 중요한 것이 뭐야?

Generation: 저는 그에 대해서는 잘 모르겠습니다.

```
[2]: Question = "데일 카네기는 이 책을 왜 썼어?"  
filtering_question(Question)
```

100% | [ ] | 1/1 [00:00<00:00, 24.25it/s]

Question: 데일 카네기는 이 책을 왜 썼어?  
Generation: 카네기는 이 책을 통해 사람들과의 관  
계를 개선하고, 성공적인 커뮤니케이션 기술을 공유  
하기 위해 썼습니다. 그는 사람들이 서로 소통하고  
협력할 수 있도록 돕고자 했으며, 특히 비즈니스 환  
경에서의 인간관계의 중요성을 강조했습니다. 카네  
기는 사람들의 감정을 이해하고 존중함으로써 긍정  
적인 결과를 얻을 수 있다는 점을 강조합니다.

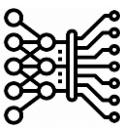
```
Question = "인간관계론에서 중요한 것이 뭐야?"  
filtering_question(Question)
```

100% | [ ] | 1/1 [00:00<00:00, 24.44it/s]

Question: 인간관계론에서 중요한 것이 뭐야?  
Generation: 1. \*\*타인의 중요성 인정하기\*\*: 다른 사람들  
이 중요하다고 느끼도록 만드는 것이 중요합니다. 이는 상대  
방의 자존감을 높이고 긍정적인 관계를 형성하는 데 도움이  
됩니다.

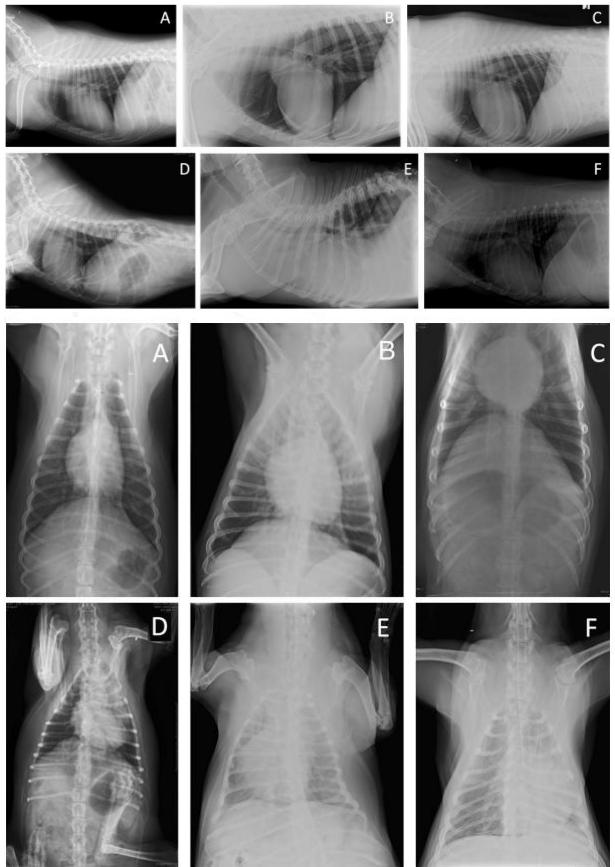
안정화 후 성능 비교

# CV: Medical Imaging Analysis



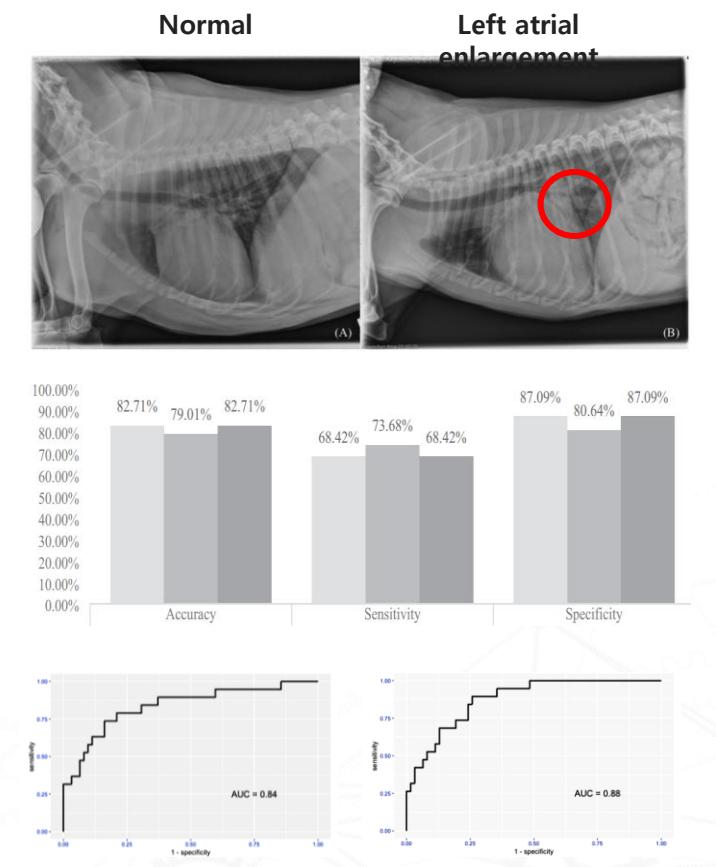
An AI-based algorithm for the automatic evaluation of image quality in canine thoracic radiographs

*Scientific Reports, 2023*



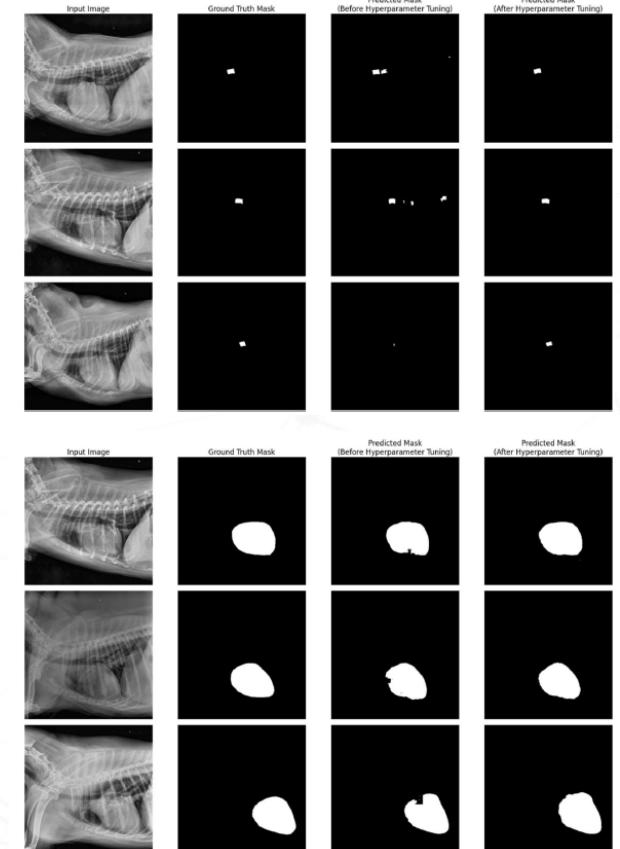
Pilot study: Application of artificial intelligence for detecting left atrial enlargement on canine thoracic radiographs

*Veterinary Radiology & Ultrasound, 2020*

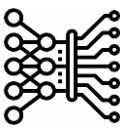


An automated deep learning method and novel cardiac index to detect canine cardiomegaly from simple radiography

*Scientific Reports, 2022*



# CV: Medical Imaging Analysis

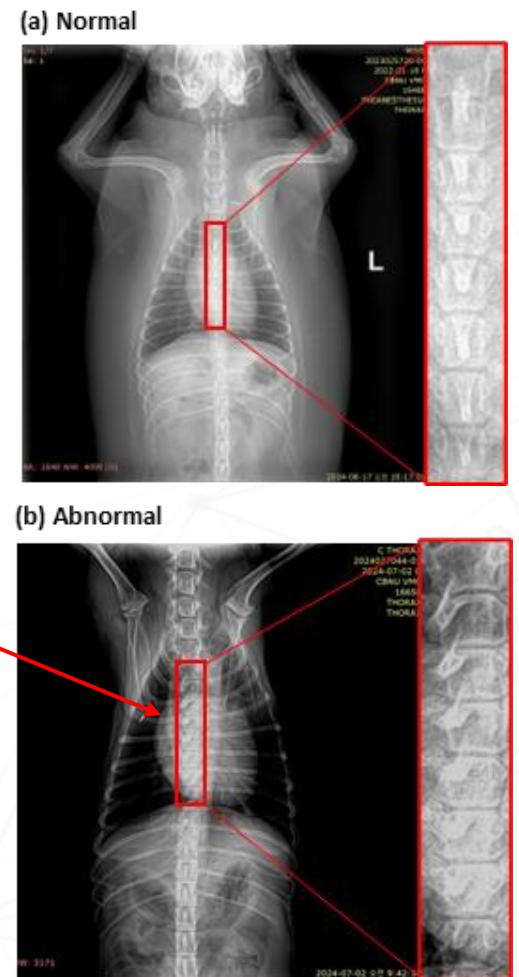


## ❖ Radiograph quality validation

- Categorize radiographic failure cases
  - General, Thorax, Abdomen, Stifle
- Detect failure cases
  - Apply YOLOv10 for object detection
  - Use vision transformer for image classification

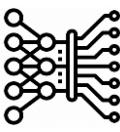
Projection type	Failure description	Failure code	Task type
General	- Case where a human hand is captured in the radiograph	$G_{HND}$	Object Detection
	- Case where a motion artifact occurs due to patient movement during image acquisition	$G_{MOT}$	Image Classification
Thorax	Case where the FOV is incorrect	$T_{FOV}$	Image Classification
	DV/VD Case where the sternum and vertebra are misaligned	$T_{ALN}$	Object Detection
	Case where the vertebral body is rotated	$T_{ROT}$	Image Classification
Lateral	Case where the FOV is incorrect	$TL_{FOV}$	Image Classification
	Lateral Case where rib rotation and left-right asymmetry occur	$TL_{ALN}$	Image Classification
	Case where the front leg is not fully extended forward	$TL_{LEG}$	Object Detection
Abdomen	VD Case where the FOV is incorrect	$A_{FOV}$	Image Classification
	Abdomen Case where the abdominal structures are misaligned and rotated	$A_{ROT}$	Object Detection
Lateral	Case where the FOV is incorrect	$AL_{FOV}$	Image Classification
	Lateral Case where the pelvis is misaligned	$AL_{ALN}$	Object Detection
MSK (Stifle)	CC Case where the stifle joint is misaligned and rotated	$M_{ALN}$	Object Detection
	Lateral Case where the femur condyle and tibia overlap, appearing as a single structure	$ML_{ROT}$	Object Detection

Radiographic failure cases categorized by projection type



Example of thoracic X-ray images

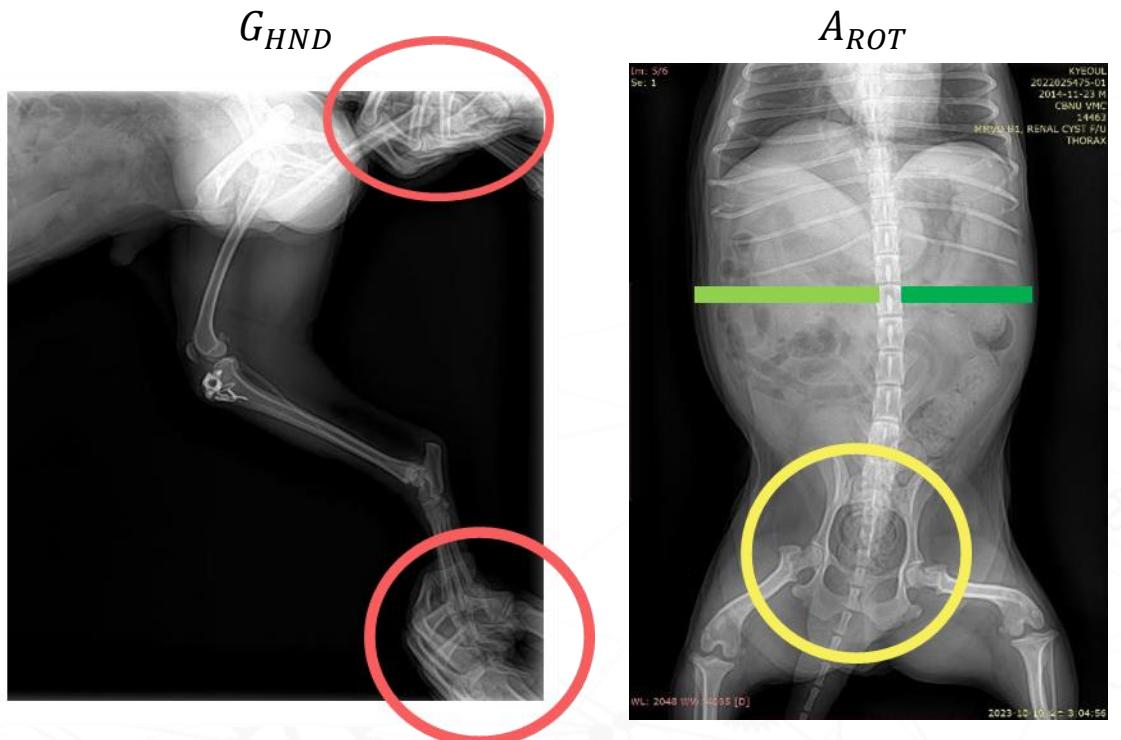
# CV: Medical Imaging Analysis



- ❖ Radiograph quality validation performance
  - Achieve motion artifact error accuracy 0.8889
  - Outperform thoracic lateral FOV error accuracy 1.0000
  - Show low performance in  $G_{HND}$  and  $A_{ROT}$  failure cases

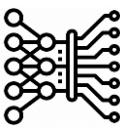
Projection type		Failure code	Accuracy
General	-	$G_{HND}$	0.5765
	-	$G_{MOT}$	0.8889
Thorax	DV/VD	$T_{FOV}$	0.9549
		$T_{ALN}$	0.8571
		$T_{ROT}$	0.7881
Lateral	Thorax	$TL_{FOV}$	1.0000
		$TL_{ALN}$	0.9524
		$TL_{LEG}$	0.8144
Abdomen	VD	$A_{FOV}$	0.8571
		$A_{ROT}$	0.5528
	Lateral	$AL_{FOV}$	0.9914
		$AL_{ALN}$	0.7388
MSK (Stifle)	CC	$M_{ALN}$	0.9474
	Lateral	$ML_{ROT}$	0.6934

Performance of radiograph quality validation



Visualization of failure cases categorized by projection types

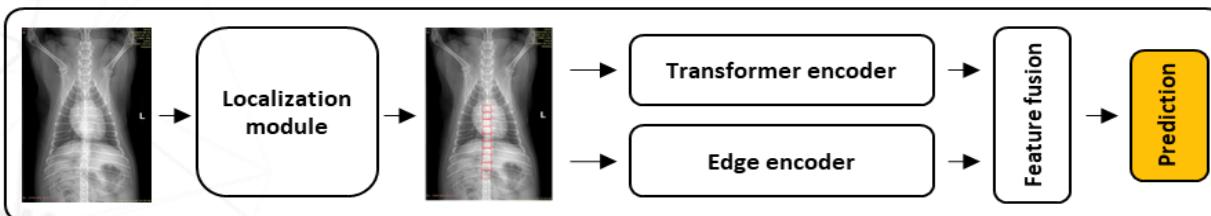
# CV: Medical Imaging Analysis



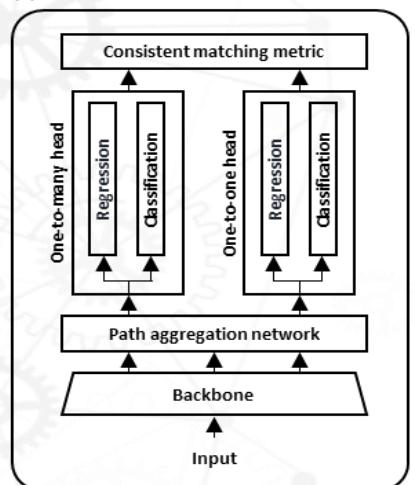
## ❖ Radiograph quality validation

- Extract global context using transformer encoder
- Extract local anatomical details using edge encoder with Canny edge detection
- Fuse transformer and edge features to enhance diagnostic accuracy

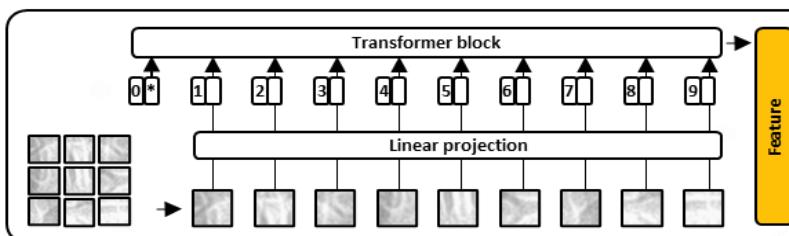
(a) Transformer-based edge representation learning network



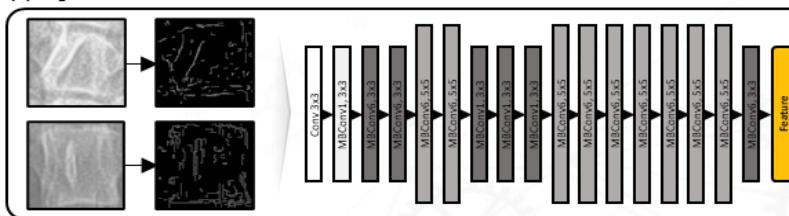
(b) Localization module



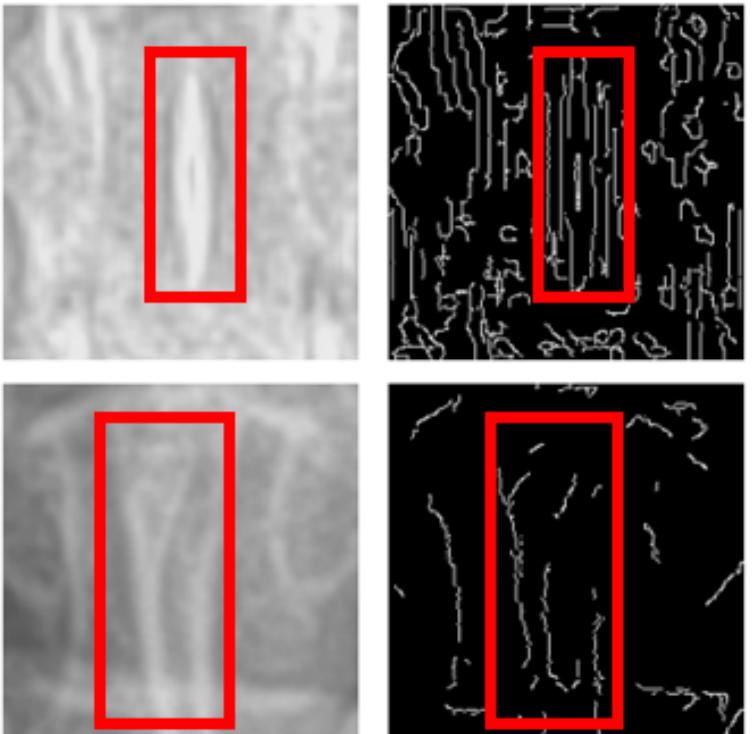
(c) Transformer encoder



(d) Edge encoder

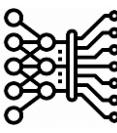


Overview of the radiograph quality validation network

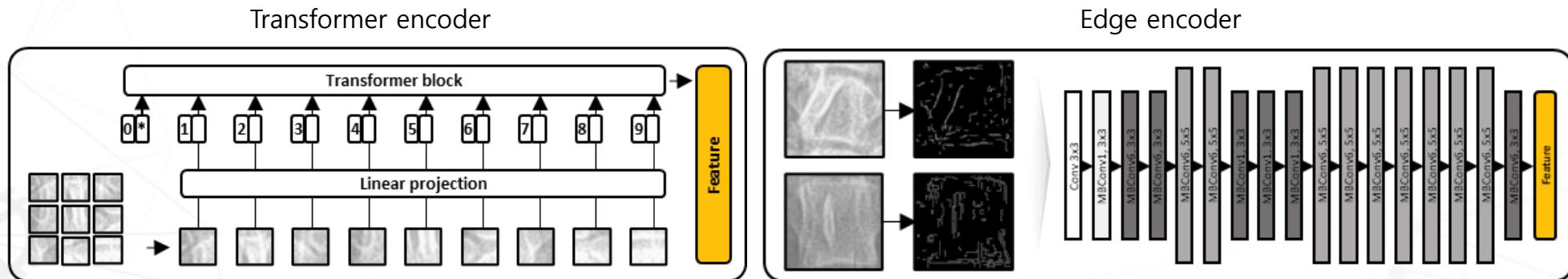


Example of Canny edge detection

# CV: Medical Imaging Analysis



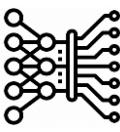
- ❖ Quality validation ablation study
  - Achieve baseline model accuracy 0.7368
  - Improve accuracy to 0.7632 with transformer encoder
  - Enhance accuracy to 0.8158 using transformer and edge encoder



Model	Transformer encoder	Edge encoder	Accuracy
Baseline	✗	✗	0.7368
w/o Edge encoder	✓	✗	0.7632
Proposed (Ours)	✓	✓	<b>0.8158</b>

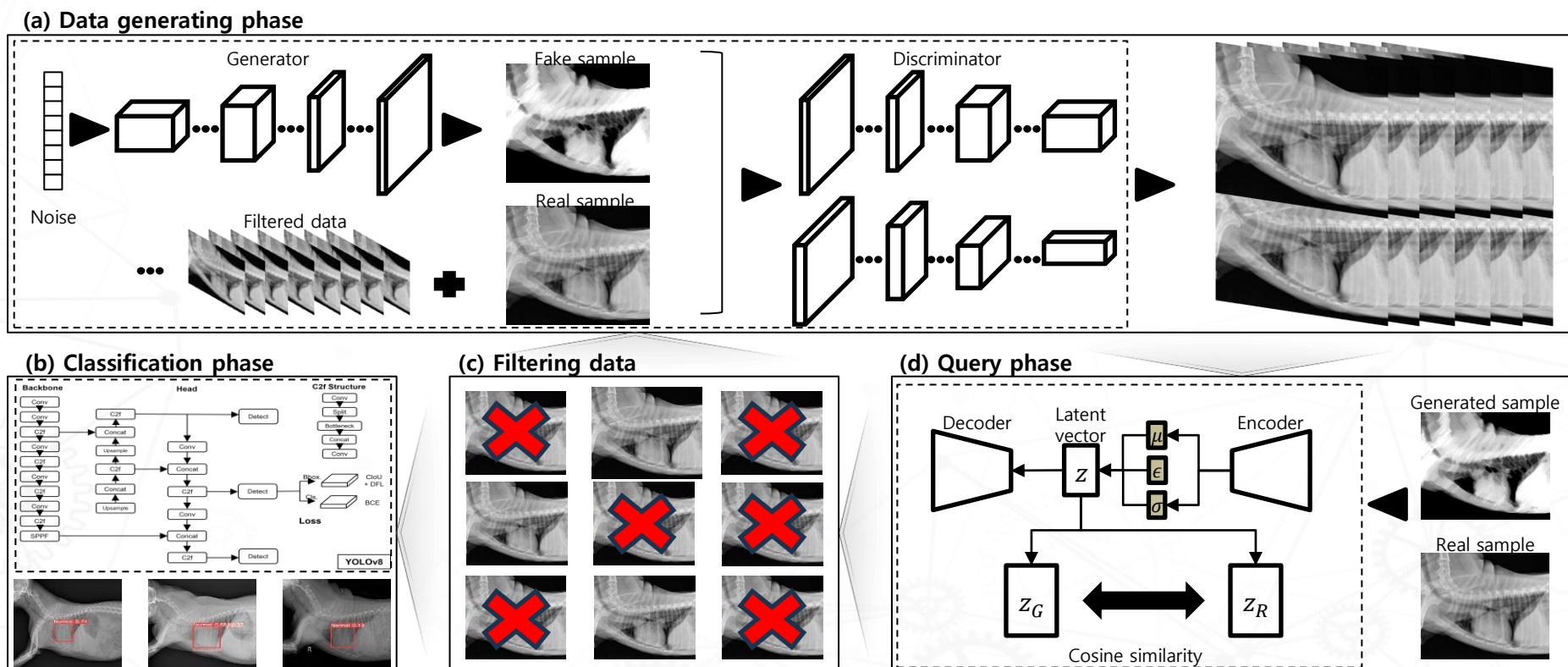
Ablation study results for encoder components

# CV: Medical Imaging Analysis



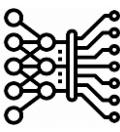
## ❖ Generative active learning

- Generate synthetic data using ProjectedGAN
- Filter data by cosine similarity of VAE embedding
- Enhance synthetic data quality progressively through iterative active learning

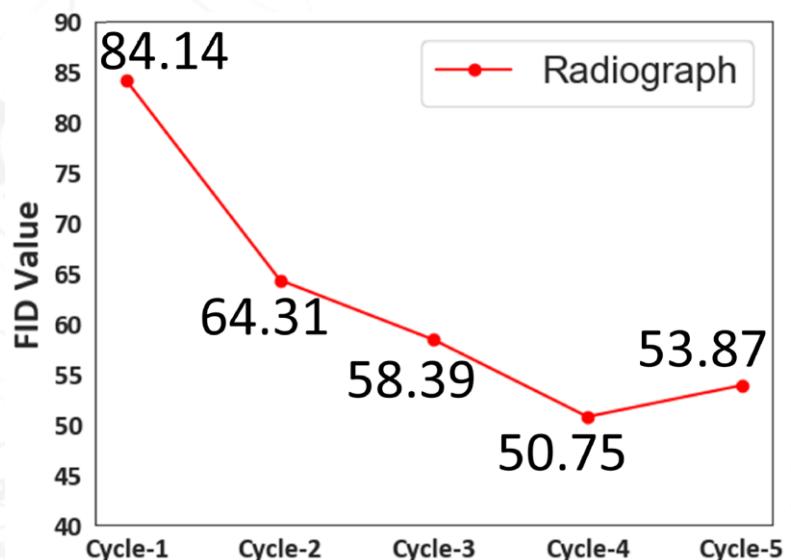


Overview of the generative active learning

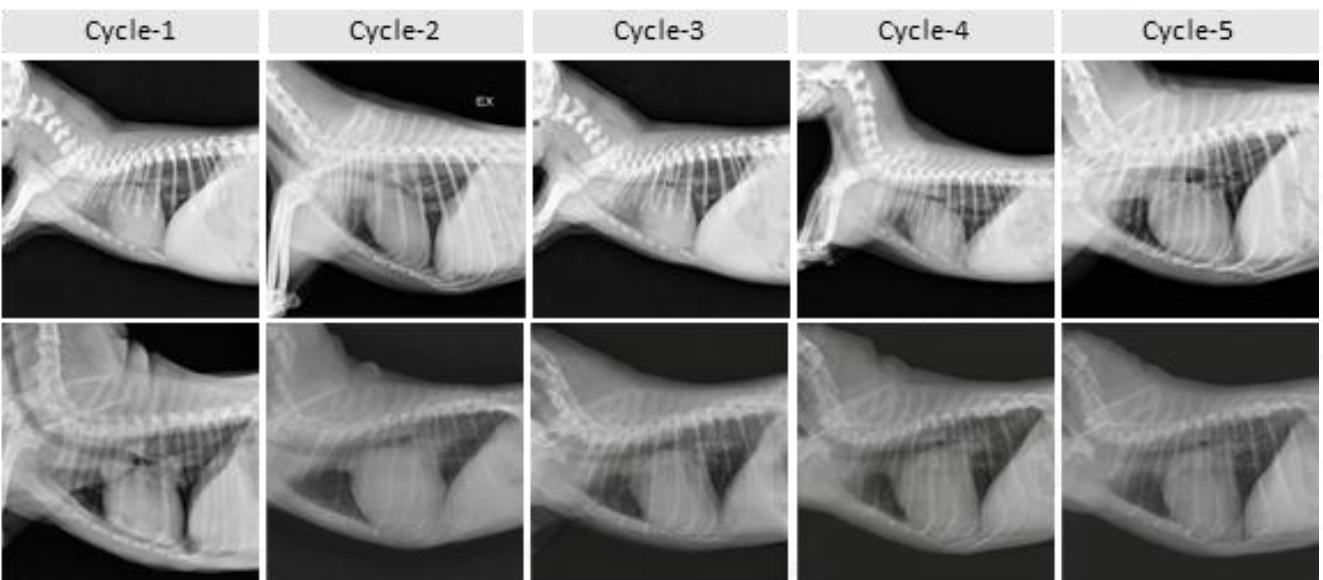
# CV: Medical Imaging Analysis



- ❖ Generative model evaluation
  - Improve FID consistently from 84.14 to 50.75
  - Achieve highest cosine similarity of 0.9096 between synthetic and real data
- ❖ Synthetic images usability for classification
  - Obtain best cardiomegaly classification in Cycle-4
    - Attain accuracy 0.77, precision 0.88, recall 0.79, F1 score 0.72
  - Improve generalization significantly using synthetic data

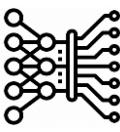


FID scores across cycles



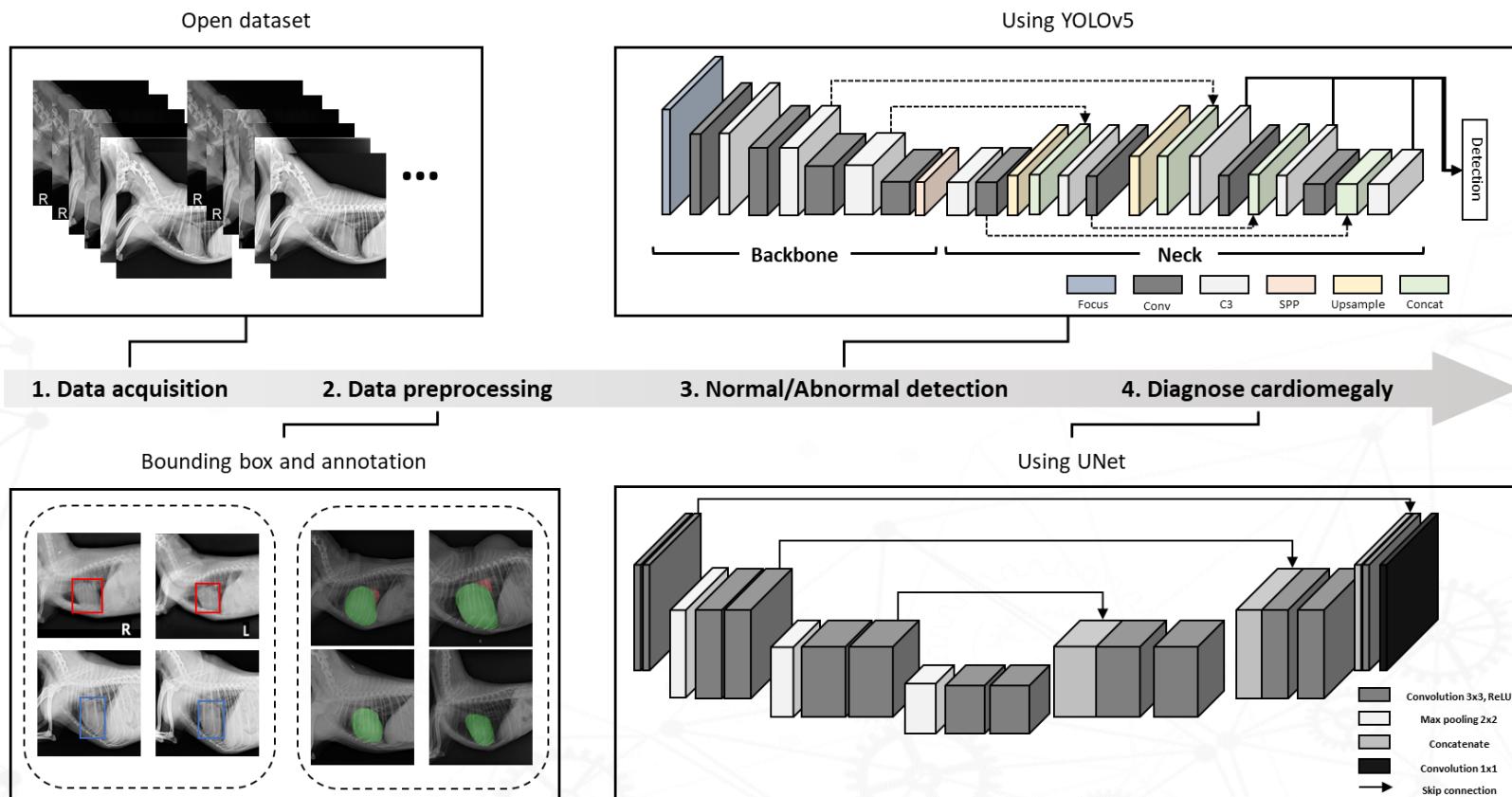
Generated images of each cycles

# CV: Medical Imaging Analysis



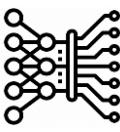
## ❖ Cardiomegaly detection pipeline

- Classify normal/abnormal cases using YOLOv5
- Segment left atrial enlargement using Unet
- Evaluate segmentation accuracy using Dice score



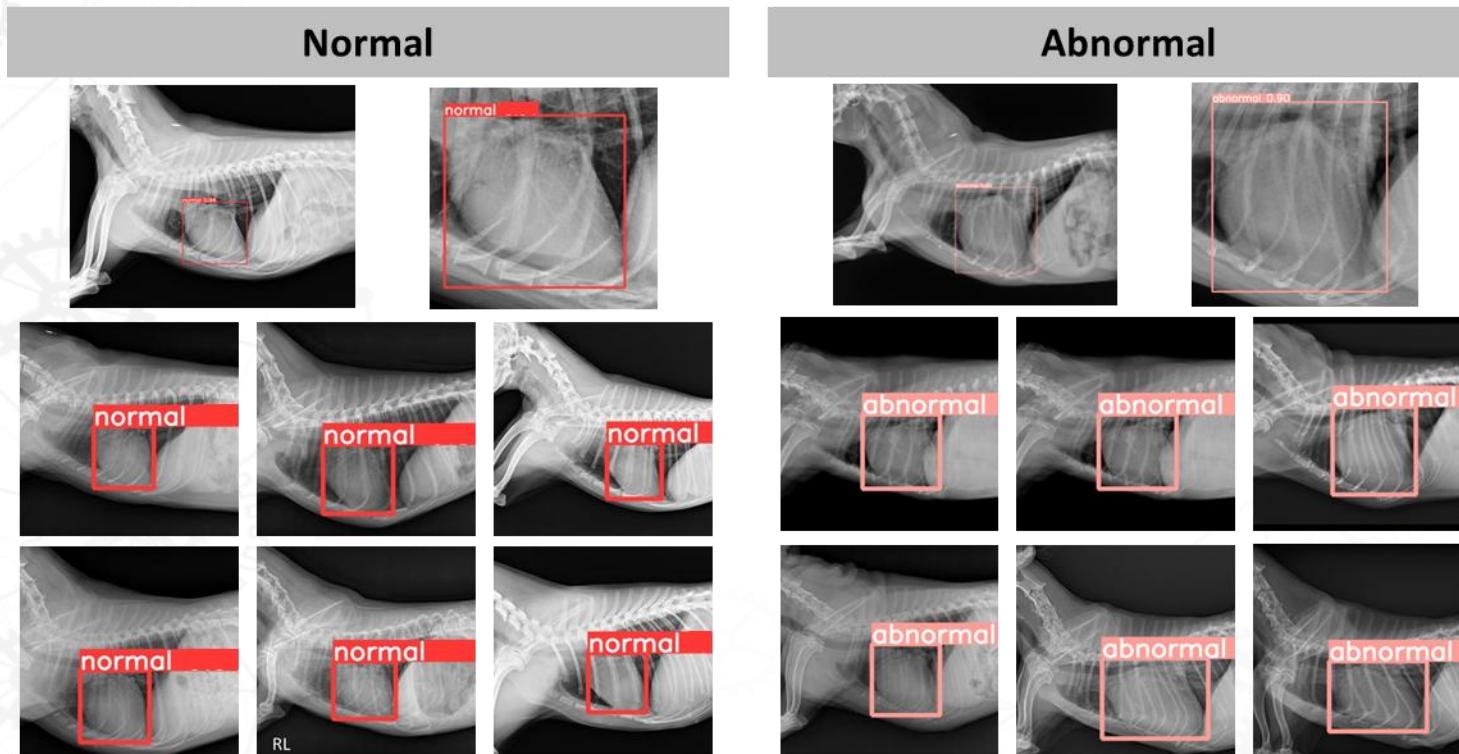
Overview of the cardiomegaly detection pipeline

# CV: Medical Imaging Analysis



## ❖ Abnormal heart detection

- Demonstrate robust performance in 4-fold cross-validation
  - Achieve accuracy 0.8887, precision 0.8962, recall 0.8837, F1 score 0.8898
- Maintain stable performance across validation folds

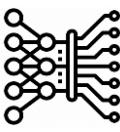


Experimental results of abnormal heart detection

Fold	Accuracy	Precision
Fold 1	0.8900	0.8920
Fold 2	0.8850	0.8900
Fold 3	<b>0.8925</b>	0.9000
Fold 4	<b>0.8875</b>	<b>0.9025</b>
Average	0.8887 ( $\pm 0.0032$ )	0.8962 ( $\pm 0.0061$ )

Fold	Recall	F1 score
Fold 1	0.8800	0.8860
Fold 2	0.8750	0.8820
Fold 3	0.8875	0.8937
Fold 4	<b>0.8925</b>	<b>0.8975</b>
Average	0.8837 ( $\pm 0.0078$ )	0.8898 ( $\pm 0.0071$ )

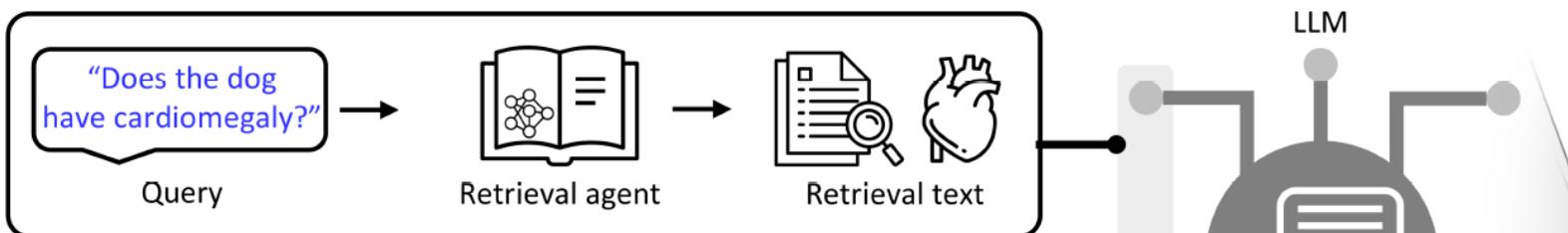
# CV: Medical Imaging Analysis



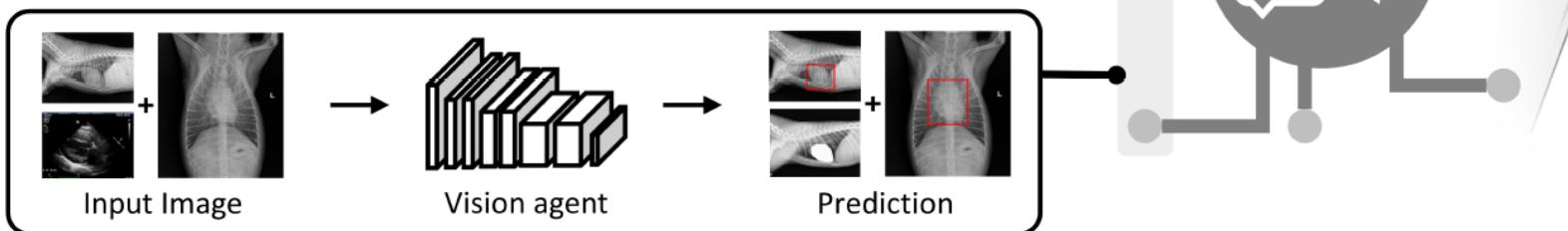
## ❖ Future works

- Integrate multimodal AI with LLM-based retrieval
- Develop advanced vision agents for anatomical feature extraction
- Generate context-aware diagnostic reports for enhanced clinical trust

(a)



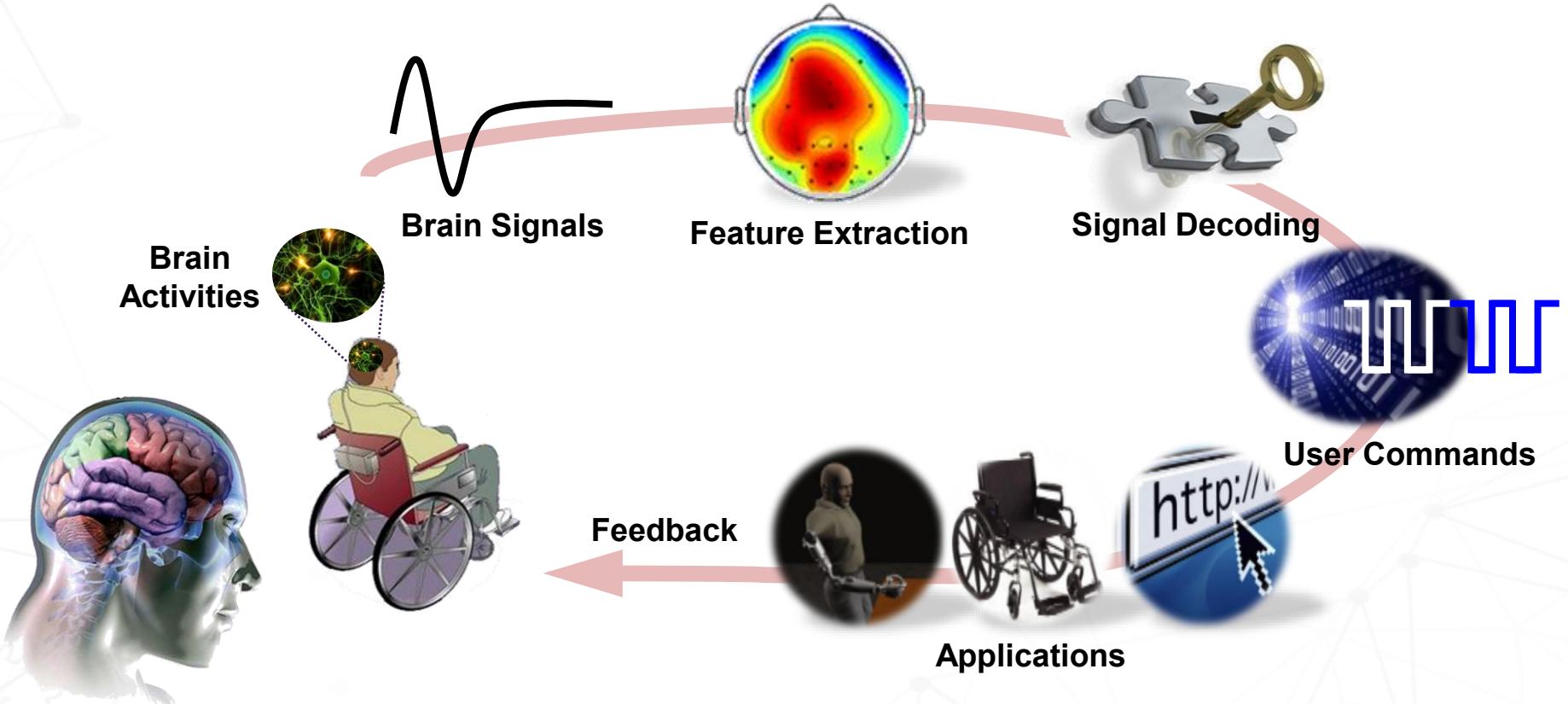
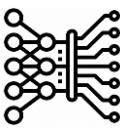
(b)



The **left ventricle** shows normal diameter and geometry, with **preserved systolic function**.  
The right ventricle is of normal size, but exhibits **reduced longitudinal systolic function**....

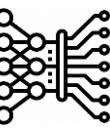
Overview of the AI agent for veterinary radiology

# Neural Computing: Brain-Computer Interface



**Brain-Computer Interface (BCI)** is a communication system that interprets and delivers a user's intention to the external world without relying on the normal output pathways of peripheral nerves and muscles

# Recent BCI Research: Neuralink



## AN INTEGRATED BRAIN-MACHINE INTERFACE PLATFORM WITH THOUSANDS OF CHANNELS

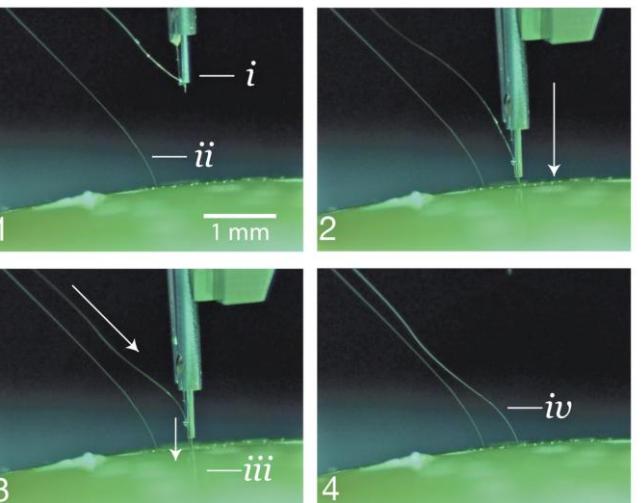
Elon Musk & Neuralink

### ABSTRACT

Brain-machine interfaces (BMIs) hold promise for the restoration of sensory and motor function and the treatment of neurological disorders, but clinical BMIs have not yet been widely adopted, in part because modest channel counts have limited their potential. In this white paper, we describe Neuralink's first steps toward a scalable flexible electrode "threads", which can be individually targeted to specific brain regions. Each thread can be individually controlled and provides full-bandwidth data streams. This system has achieved a significant breakthrough in the approach to BMI and has unprecedented performance.

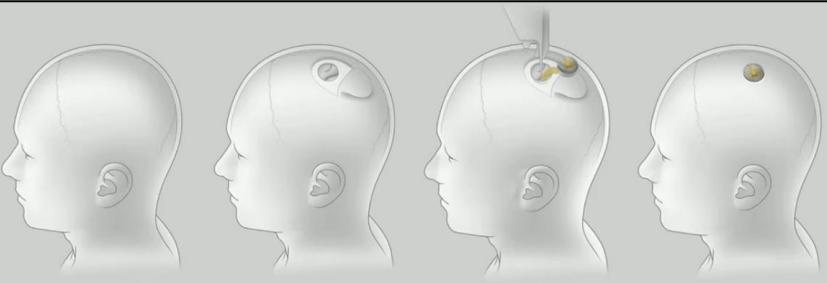
### 1 Introduction

Brain-machine interfaces (BMIs) have traditionally been limited by the number of channels available. Researchers have demonstrated human speech synthesizers [6] using no more than 100 channels.

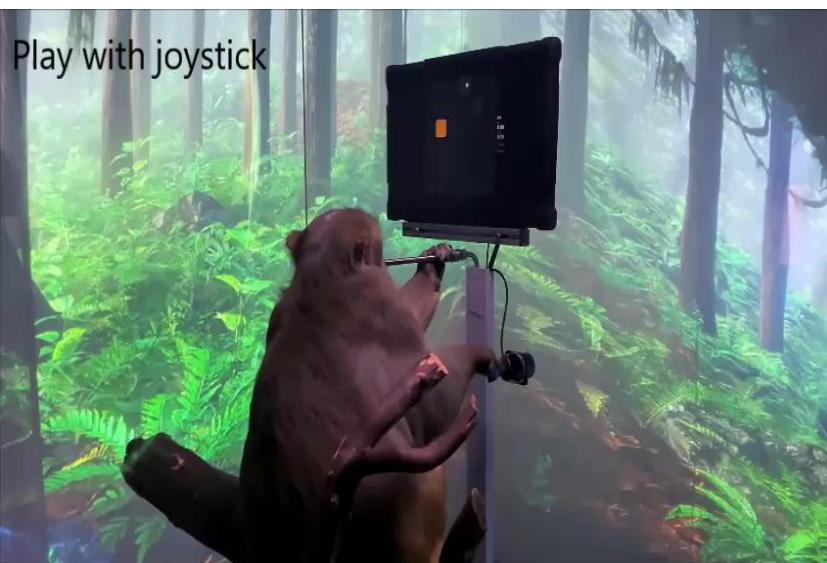


**Figure 4:** 1. The inserter approaches the brain proxy with a thread. *i*. needle and cannula. *ii*. previously inserted thread. 2. Inserter touches down on the brain proxy surface. 3. Needle penetrates tissue proxy, advancing the thread to the desired depth. *iii*. inserting thread. 4. Inserter pulls away, leaving the thread behind in the tissue proxy. *iv*. inserted thread.

Musk et al., BioRxiv, 2019

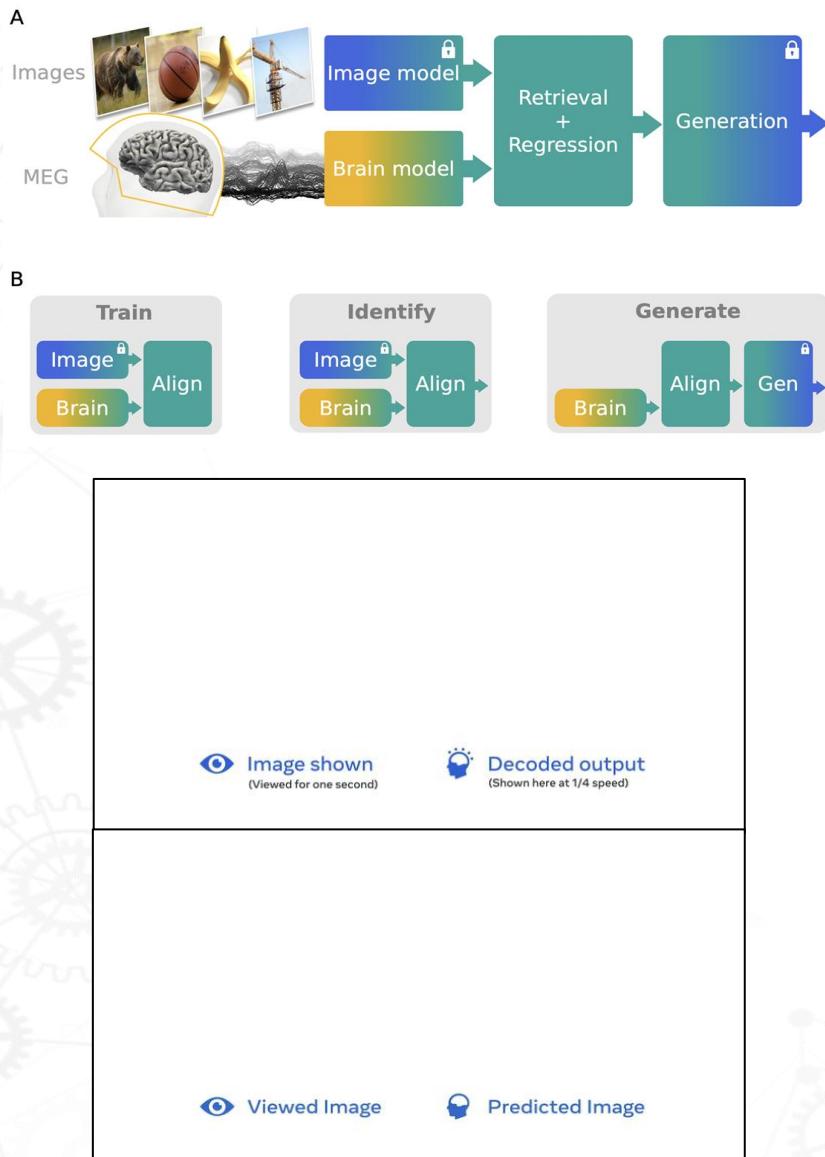
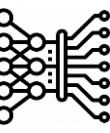


Chip ([www.neuralink.com](http://www.neuralink.com))



Mind pong ([www.neuralink.com](http://www.neuralink.com))

# Recent BCI Research: Meta AI

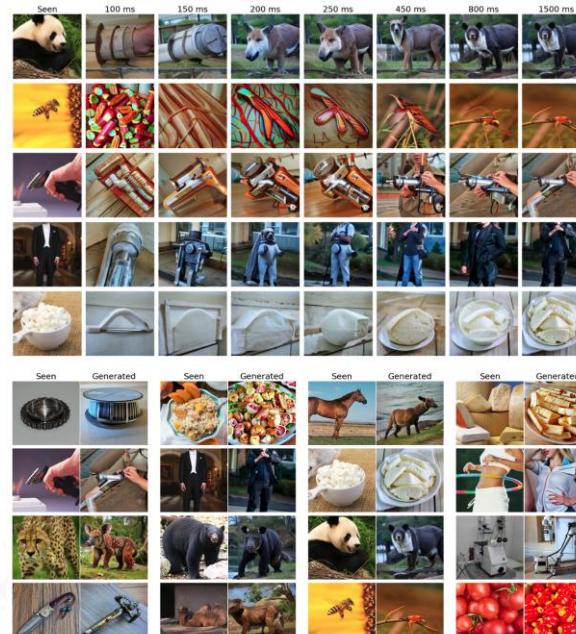


## Brain decoding: toward real-time reconstruction of visual perception

Yohann Benchirrit<sup>1,\*</sup>, Hubert Banville<sup>1,\*</sup>, Jean-Rémi King<sup>1,2</sup>

<sup>1</sup>FAIR at Meta, <sup>2</sup>Laboratoire des Systèmes Perceptifs, École Normale Supérieure, PSL University  
\*Equal contribution.

In the past five years, the use of generative and foundational AI systems has greatly improved the decoding of brain activity. Visual perception, in particular, can now be decoded from functional Magnetic Resonance Imaging (fMRI) with remarkable fidelity. This neuroimaging technique, however, suffers from a limited temporal resolution ( $\approx 0.5$  Hz) and thus fundamentally constrains its real-time usage. Here, we propose an alternative approach based on magnetoencephalography (MEG), a neuroimaging device capable of measuring brain activity with high temporal resolution ( $\approx 5,000$  Hz). For this, we develop an MEG decoding model trained with both contrastive and regression objectives and consisting of three modules: i) pretrained embeddings obtained from the image, ii) an MEG module trained end-to-end and iii) a pretrained image generator. Our results are threefold: Firstly, our MEG decoder shows a 7X improvement of image-retrieval over classic linear decoders. Second, late brain responses to images are best decoded with DINOv2, a recent foundational image model. Third, image retrievals and generations both suggest that high-level visual features can be decoded

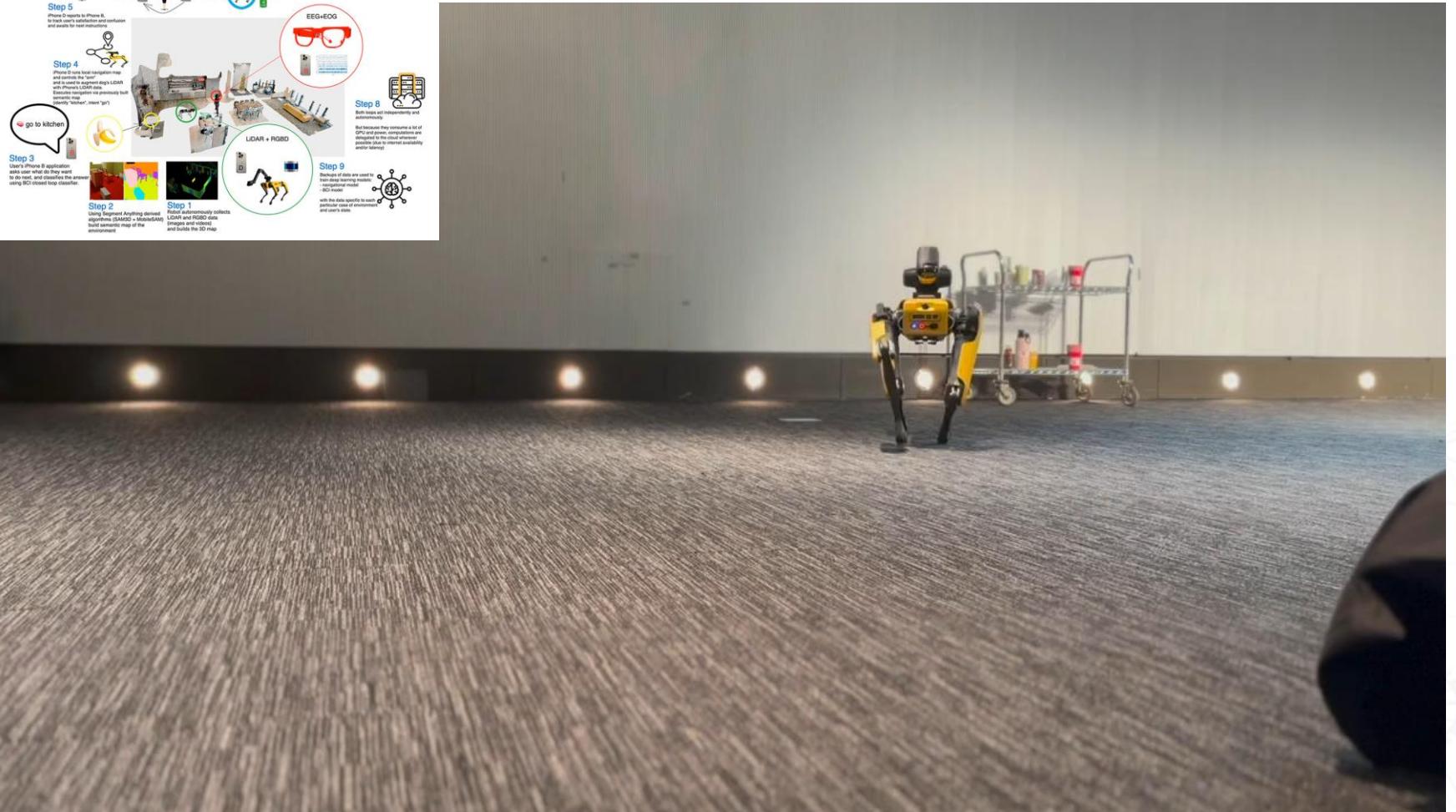
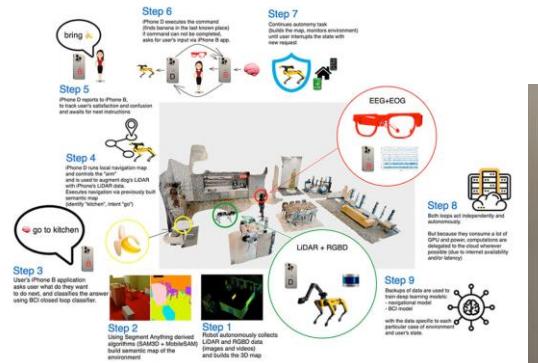
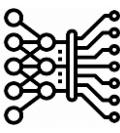


and how the human brain represents the world is a quest, which originally consisted of searching, on, (e.g. Hubel and Wiesel (1962); O’Keefe and Burgess (2003)). It has been approached and solved by Machine Learning (ML) in two main ways. First, ML models have been trained to extract informative patterns of brain activity from fMRI signals. For example, Tong (2005) trained a support vector machine (SVM) to decode visual information from functional Magnetic Resonance Imaging (fMRI). Since then, many studies have shown that such brain activity patterns (Roy et al., 2019; Szűcs et al., 2022; Scotti et al., 2023). Second, ML models have been trained to predict the visual representations of brain activity. For example, Yamins et al. (2014) have shown that the neuronal responses to these images have been shown to account for a wide variety of visual stimuli (Banino et al., 2018; Schrimpf et al., 2020; Scotti et al., 2023).

representational alignment between brain activity and visual perception. This means that the decoding of visual stimuli need not be restricted to a single modality. By doing so, it is possible to condition subsequent generative AI models to generate images that are more realistic. For example, interpreting images can be much simpler than generating them. For example, generative approaches (Nishimoto et al., 2011; Gómez et al., 2018), diffusion techniques have, in this work, been used to generate images from functional Magnetic Resonance Imaging (fMRI).

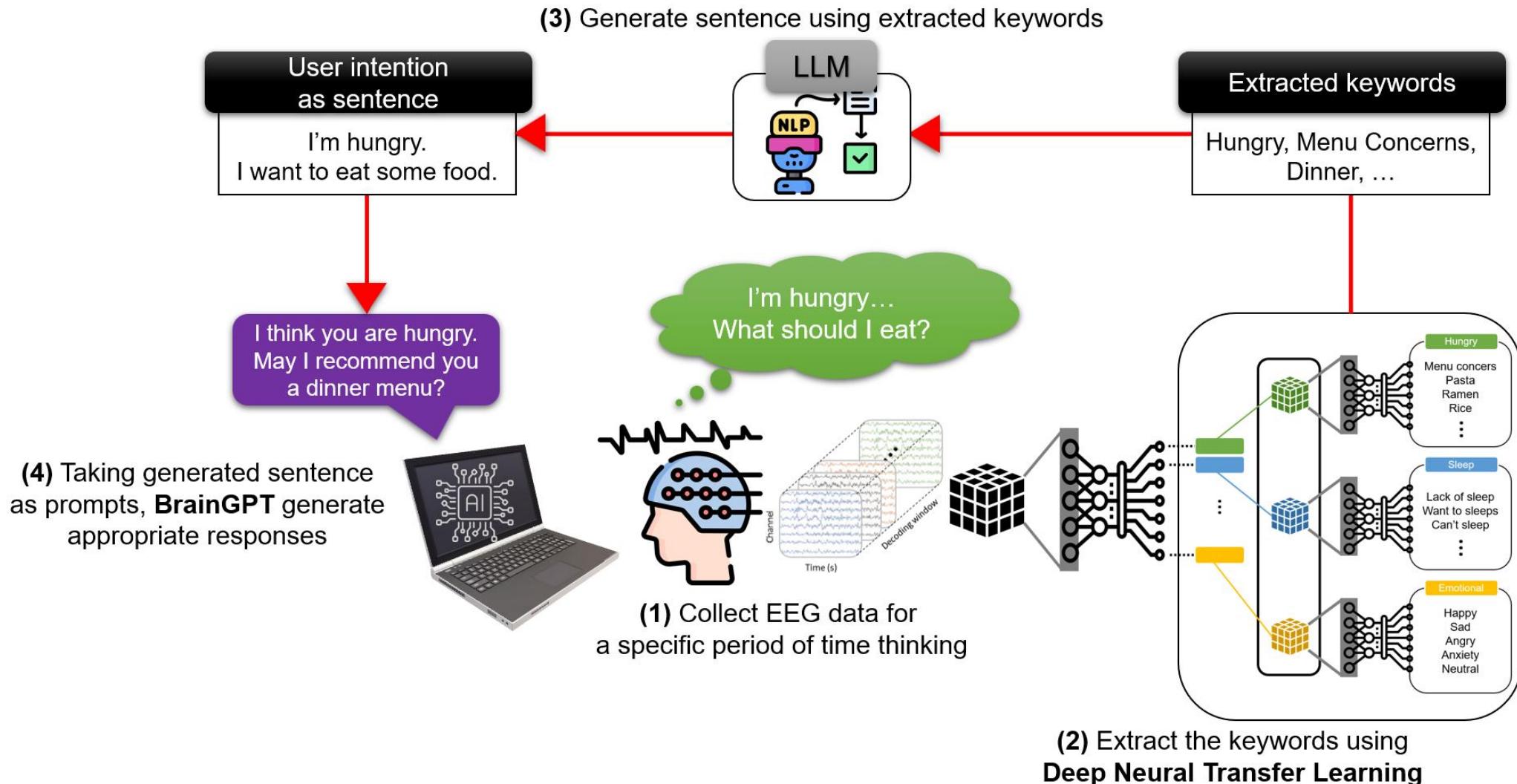
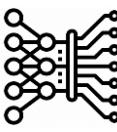
“Brain decoding: toward real-time reconstruction of visual perception,” Accepted, ICLR 2024

# Recent BCI Research: MIT & braini



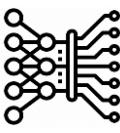
An overview of the Project Ddog system, MIT

# System Architecture



Overall flow of keywords extraction and language generation process of BrainGPT

# Neural computing: Digital Healthcare



Effortless Health Monitoring and Tracking Anytime, Anywhere with Your Camera in 30+ seconds

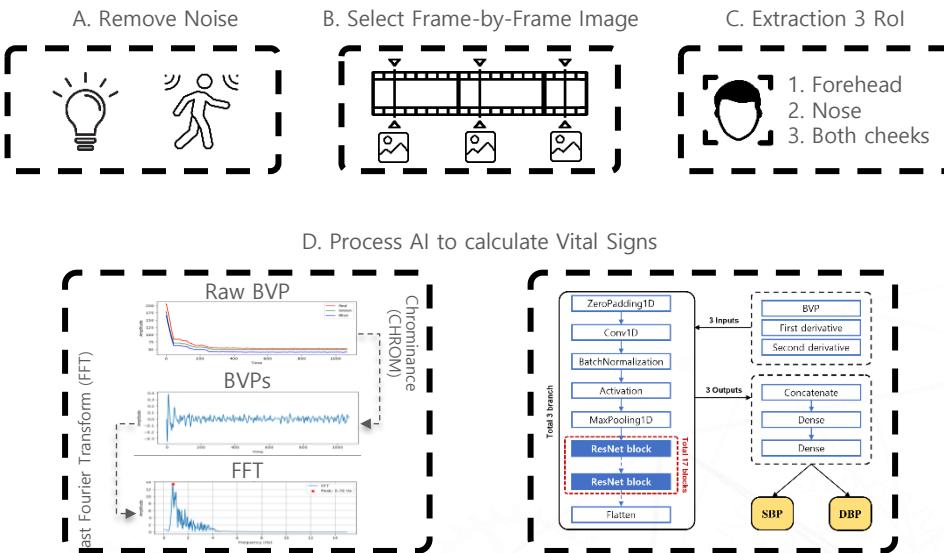
## 1. FACE SCANNING

Record with a Camera  
on an Internet-connected Device



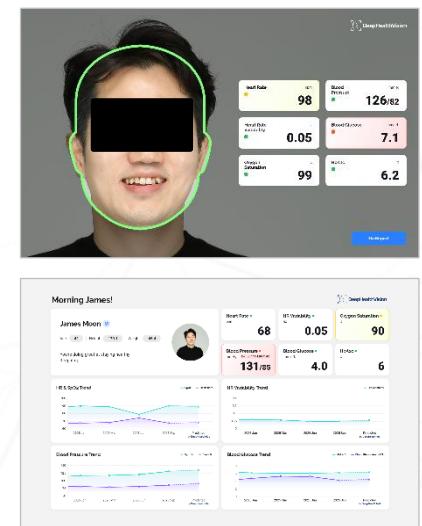
## 2. AI PROCESSING

Preprocess Recorded Data, Extract Blood Volume Pulse,  
then Process the Calculation and Estimation through AI model

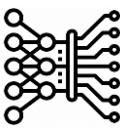


## 3. RESULT SHARING

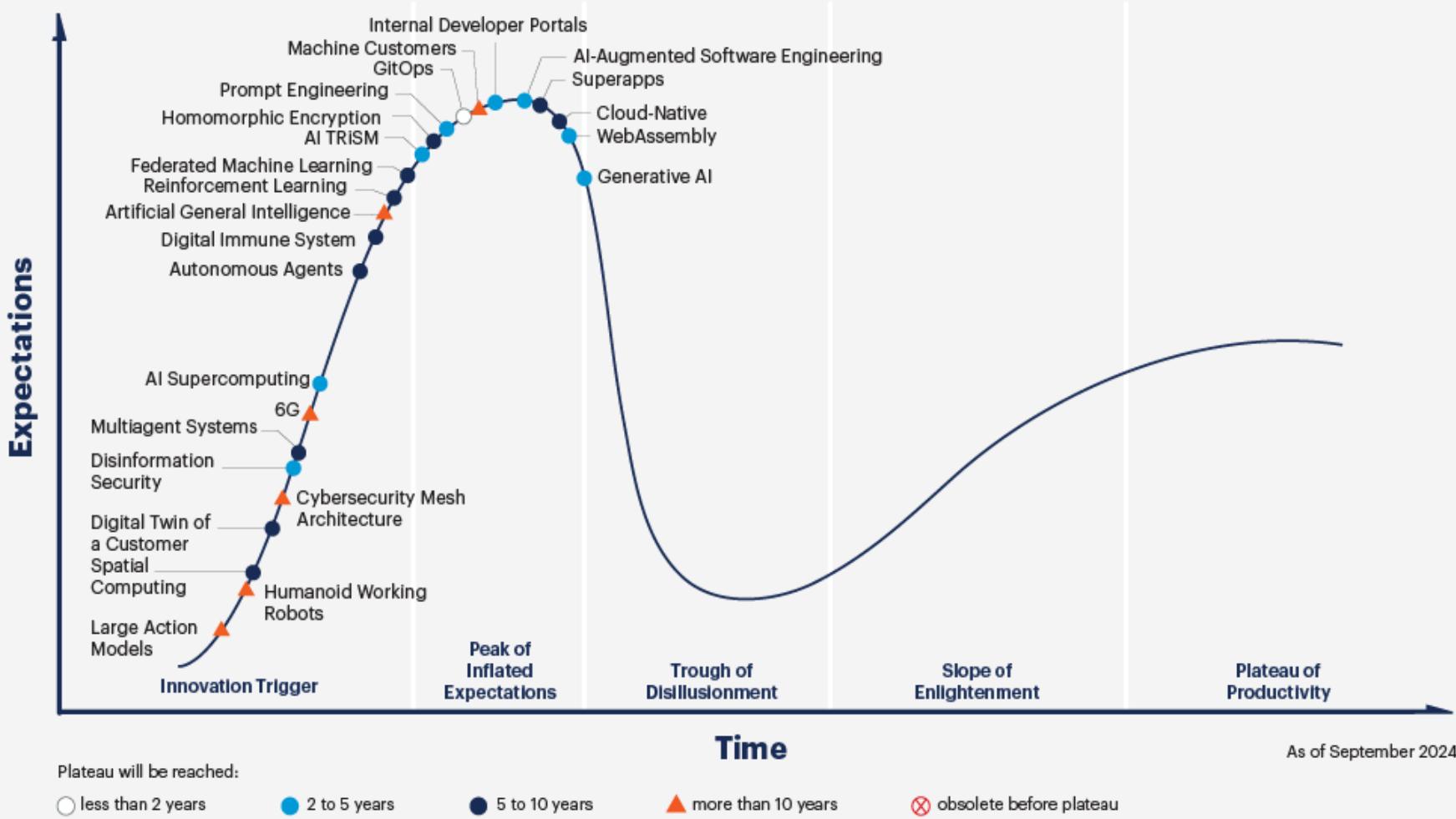
Provide Real-Time Measurements  
with Personal Health Report



Real-Time Delivery within 30+ Seconds



## Hype Cycle for Emerging Technologies, 2024



Hype Cycle for Emerging Technologies [Gartner, 2024]