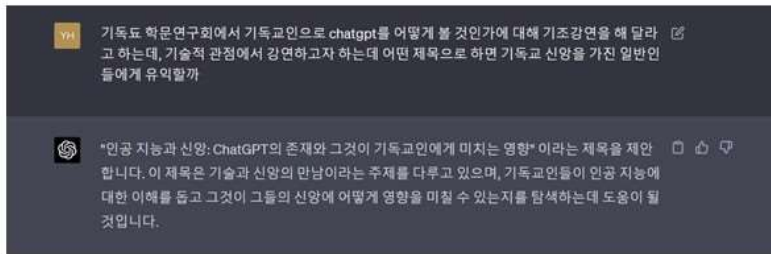


## 들어가기 전에... GPT4와의 대화



인공지능에 대한 이해를 도와라

# 들어가기 전에... GPT4와의 대화

**그동안 내용을 좀 지켜 봐주세요**

물론입니다. 이 가이드라인은 크게 세 부분으로 나눌 수 있을 것입니다. 인공지능의 기술적 이해, 인공지능의 현재와 미래, 그리고 이 모든 것이 기독교 신앙에 어떤 의미를 가지는지에 대해.

**1. 인공지능의 기술적 이해**

- 인공지능의 개념 및 기본 용어 소개. 인공지능이 무엇인지, 어떻게 작동하는지, 그리고 ChatGPT와 같은 고급 인공지능 시스템의 기본 작동 원리에 대해 설명합니다.
- 인공지능의 현재. 인공지능이 우리 인간의 현재의 능력을 뛰어넘거나 대체할 수 없다는 점을 강조합니다. 이를 통해 인공지능이 여전히 특정한 역할에 최적화된 도구라는 점을 명확히 합니다.
- 인공지능의 현재 사용 사례. 인공지능의 현재 어떻게 사용되고 있는지에 대한 몇 가지 사례를 제시합니다. 이는 의료, 교육, 연구 등 다양한 분야에서 인공지능의 활용을 보여줄 것입니다.
- 인공지능의 미래. 인공지능 기술이 어떻게 발전하고, 이러한 발전이 사회에 어떤, 그리고 어떤 실재 어떤 영향을 미칠지에 대해 논じます.

**2. 인공지능의 현재와 미래**

- 인공지능의 현재 사용 사례. 인공지능의 현재 어떻게 사용되고 있는지에 대한 몇 가지 사례를 제시합니다. 이는 의료, 교육, 연구 등 다양한 분야에서 인공지능의 활용을 보여줄 것입니다.
- 인공지능의 미래. 인공지능 기술이 어떻게 발전하고, 이러한 발전이 사회에 어떤, 그리고 어떤 실재 어떤 영향을 미칠지에 대해 논じます.

**3. 인공지능의 기독교 신앙**

- 신앙과 기술의 관계. 기독교 신앙의 기술에 대해 어떻게 바라보고 있는지, 그리고 이러한 관점의 인공지능에 어떻게 적용될 수 있는지에 대해 논じます.
- 인공지능과 신앙의 충돌과 조화. 인공지능의 발전이 기독교 신앙에 어떤 도전과 기회를 제시하는지에 대해 설명합니다.
- 인공지능의 윤리적 고려사항. 인공지능의 발전과 사용이 도덕적, 윤리적 문제를 어떻게 제기하는지, 그리고 이러한 문제를 기독교 신앙의 관점에서 어떻게 다룰 수 있는지에 대해 논じます.

**강연에서 1. 인공지능의 기술적 이해를 위해 개념을 어떻게 설명할 수 있지**

인공지능의 기술적 이해를 위한 설명은 기본적인 개념부터 시작하여, 점차 복잡한 내용으로 나아갈 수 있습니다.

- 인공지능의 기본 개념.** 인공지능(AI)이란 기계가 인간과 유사한 지능을 가지고 문제를 해결, 학습, 적응, 창의적인 문제 해결을 할 수 있는 능력을 가지는 것을 말합니다. 이는 학습, 이해, 추론, 언어 인식, 패턴 인식 등의 작업을 포함할 수 있습니다.
- 머신러닝과 딥러닝.** 인공지능의 주요 기술 중 하나인 머신러닝은 기계가 데이터로부터 학습하고, 경험을 통해 성능을 향상시키는 과정입니다. 이 복잡한 형태의 머신러닝인 딥러닝은 인간의 뇌에서 영감을 받은, 인공 신경망을 이용하여 데이터를 처리하고 패턴을 찾아내는 방식입니다.
- ChatGPT와 같은 언어 모델의 작동 원리.** 이러한 모델은 대량의 텍스트 데이터를 통해 언어의 패턴을 학습하고, 이를 통해 새로운 텍스트를 생성하거나 주어진 텍스트에 대해 응답하는 방식으로 작동합니다. 이는 인공지능이 언어를 '이해'하는 것이 아니라, 데이터에서 학습된 패턴을 따르는 것이라는 점을 명확히 할 수 있습니다.
- 인공지능의 한계.** 현재의 AI는 일반적인 지능을 가지지 못하고, 특정 작업에 대해 학습되어 있습니다. 또한, AI는 인간의 감정, 창의성, 직관 등을 완벽하게 모방하거나 이해할 수 없다는 것을 강조할 수 있습니다.

이렇게 기본적인 개념과 원리를 설명하면서, 기술적인 어려움을 최소화하고 이해를 돕기 위해 간단한 예시나 비유를 사용하는 것이 좋습니다. 예를 들어, 인공지능을 '학습하는 방식이 인간의 뇌와 비슷한 컴퓨터 프로그램이라고 설명할 수 있습니다.

**인공지능 기술: 입력 → 결과, 현재 문맥을 이해하고 있는데, 신경망에 기억하는가?**  
**아닙니다. 입력에 기존 대화이력(사용자 질문과 시스템 응답)들을 같이 넣어서 질문합니다.**

## 발표 순서

PART	<b>I</b>	인공지능 개요	05
PART	<b>II</b>	언어모델 연구	10
PART	<b>III</b>	ChatGPT와 GPT4	16
PART	<b>IV</b>	GPT4에 의한 AI 시사점	20

# 인공지능이란?

I 인공지능 개요



"기계로부터 만들어진 지능"  
 < John McCarthy, Dartmouth Conference, 1956 >



"기계가 일일이 코드로 명시하지 않은 동작을  
 데이터로부터 학습하여 실행할 수 있도록 하는  
 알고리즘을 개발하는 연구 분야"  
 < Arthur Samuel, 1959 >

인간이 하는 지적 활동을 컴퓨터가 하도록 하는 모든 기술 (지성, 인간성)

계산하기, 기억/저장하기, 일정정리하기, 신문기사 알려주기, 같은 글자 찾기, 사진 찾기, ...

듣기, 말하기, 걷기, 번역하기, 정보 찾기, 게임하기, 물건 들기, 길찾기, 보기, 요약하기,  
 정해진 업무 대외하기, 퀴즈 풀기, ... ..,

인간이 한 생각, 판단, 창조 따라 하기 (전문가 시스템: 의사, 판사, 교사, ...)

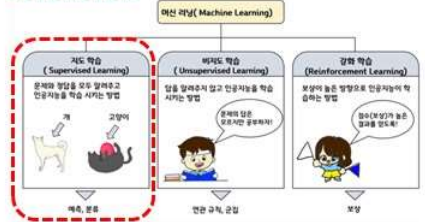
새로운 생각하기, 새로운 창조하기, 새로운 판단하기, 감정을 가지기, 자기주장 ...

# 인공지능 개념정리

I 인공지능 개요



<출처: <http://tobetong.com/?p=9393>>



<출처: [http://ai4school.org/?page\\_id=3635](http://ai4school.org/?page_id=3635)>

# 머신러닝과 딥러닝(신경망)

I 인공지능 개요



남자들

남자/여자 구별하기



여자들

## 규칙

머리가 짧으면 남자, 길면 여자  
수염이 있으면 남자, 없으면 모르겠네.  
턱이 가늘수록 여자  
치마 입으면 여자  
목걸이, 귀걸이, 반지 있으면 여자...

Feature

## 머신러닝

- 머리 길이: 수치화 f1, 가중치 a1
- 머리 스타일: 수치화 f2, 가중치 a2
- 수염: 수치화 f3, 가중치 a3
- 턱선: 수치화 f4, 가중치 a4
- 치마: 수치화 f5, 가중치 a5
- 목걸이: 수치화 f6, 가중치 a6
- ...

Machine Learning

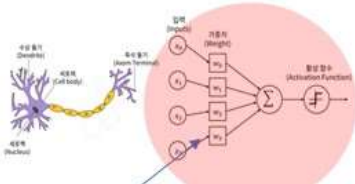
$$f(\text{남자 or 여자}/x) = a1*f1(x) + a2*f2(x) + \dots + an*fn(x)$$



출처: <https://www.merkle.com/blog/dispelling-myths-deep-learning-vs-machine-learning>

# 딥러닝(신경망) 동작 개념

I 인공지능 개요

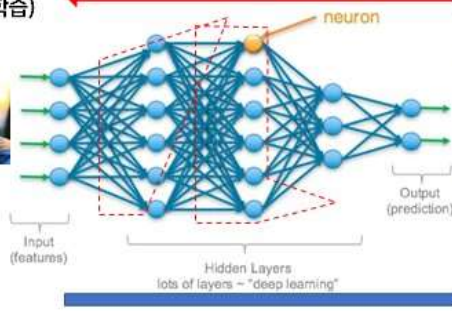
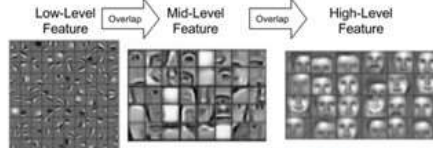


<Perceptron, Frank Rosenblatt, 1957>  
parameter (학습)

차이 보정

목적함수  
손실함수 (정답비교)

Feature Map in Convolutional Neural Networks (CNN)

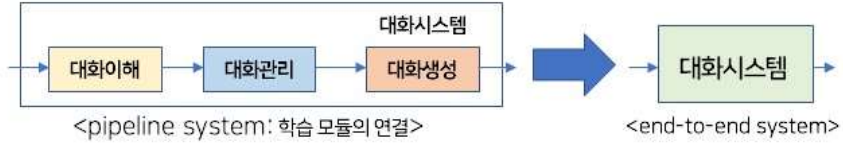


예측 (y)	정답 (y)
남자 0.6	0.0 ● 남자
여자 0.4	1.0 ● 여자

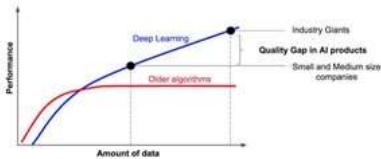
# 딥러닝(신경망) 패러다임

I 인공지능 개요

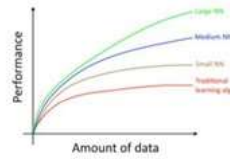
❖ 복잡한 업무도 학습데이터만으로 성능을 보임



❖ 좋은 학습데이터가 많을수록, 신경망이 클수록 높은 성능을 보임 → 거대기업 주도



<학습 데이터와 성능의 상관관계>

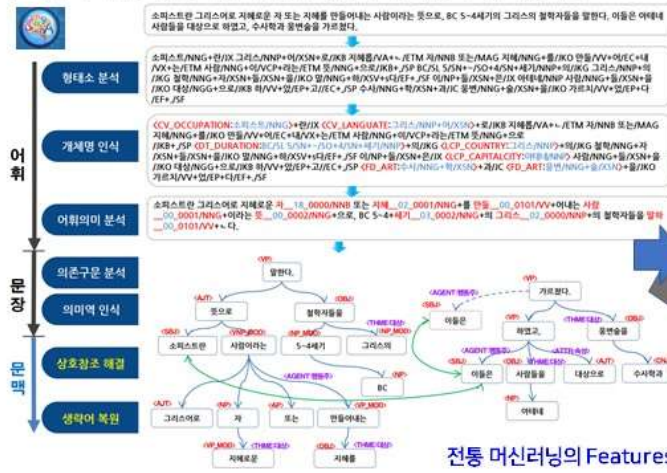


<data, layer or parameter scale에 따른 성능 추이>

# 언어처리 연구

II 언어모델 연구

❖ AI가 사람의 언어를 이해하고 생각하고 말하게 하는 연구



전통 머신러닝의 Features

# 언어모델

## II 언어모델 연구

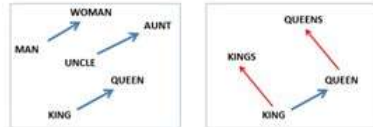
- 특정 문장이 등장할 확률을 계산해 주는 모델
- 단어나 문장을 적절하게 선택하거나 생성해야 하는 경우 여러 후보 중에서 더 적절한(확률이 높은) 후보를 선택 가능
- 전산 언어학 태동 이후로 다양한 언어모델 등장



# 딥러닝 언어모델: Word2Vec

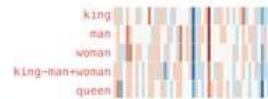
## II 언어모델 연구

- 각 단어들의 벡터를 의미 있게 배치
- 대량의 unlabeled 데이터 이용
- 학습 방법
  - ✓ CBOW
  - ✓ Skip-gram
- 언어지능 태스크 딥러닝의 단어 임베딩 활용
  - ✓ Symbol(문자) → real-value(신경망 입력값)
  - ✓ Symbol(문자)에 의존적: 손이 참 곱다 vs. 문이 고장나서 손을 봐야 한다.



(Mikolov et al., NAACL HLT, 2013)

$$\text{king} - \text{man} + \text{woman} \approx \text{queen}$$

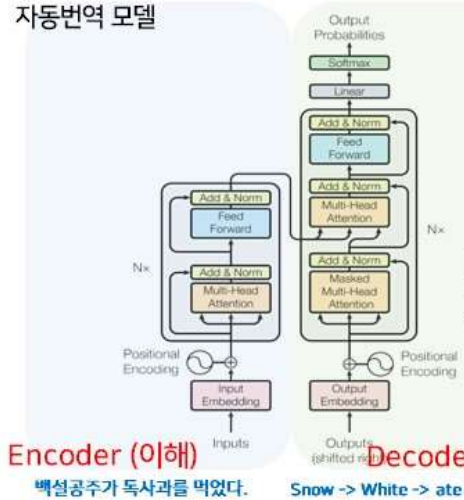


CBOW(Continuous bag-of-words)

Skip-gram

# Transformer

## 자동번역 모델

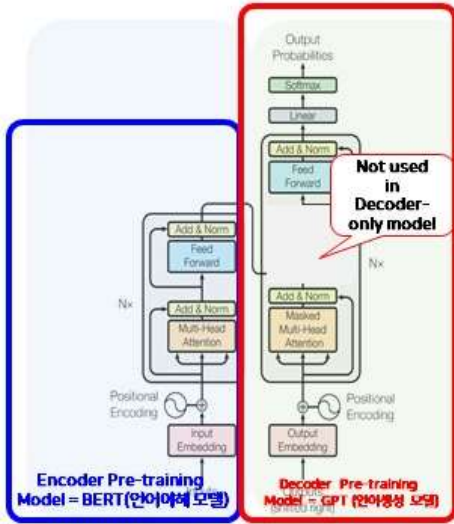


Encoder: 전역적으로 self-attention 이해  
 Decoder: 순차적으로 self-attention으로 생성  
 Encoder-Decoder: cross-attention으로 대역

Encoder (이해) 백설공주가 독사과를 먹었다. Decoder (생성) Snow -> White -> ate -> the -> poisoned apple. (auto-regressive manner)

# 사전학습 언어모델 등장

BERT(2018.10) vs. GPT-1(2018.06), GPT-2(2019.02)  
 335M params      117M params      1.5B params



	BERT 계열 = Masked Word Prediction	GPT 계열 = Next Word Prediction
원문	백설공주가 독이 든 사과를 먹었다	
사전 학습 태스크#1	입력: 백설공주가 X1 든 사과를 X2 출력: X1 = 독이, X2 = 먹었다	입력: 백설공주가 독이 든 출력: - 든
사전 학습 태스크#2	입력: X1 독이 X2 사과를 먹었다 출력: X1 = 백설공주, X2 = 든	입력: 백설공주가 독이 든 출력: 사과를
사전 학습 태스크#3	입력: 백설공주가 독이 X1 X2 먹었다 출력: X1 = 든, X2 = 사과를	입력: 백설공주가 독이 든 사과를 출력: 먹었다

# 딥러닝 언어모델 패러다임 변화

pretrain, then finetune

2) fine-tune  
각 태스크의 목적함수로  
가중치 미세 조정



1) pre-train



딥러닝 언어모델(BERT, GPT-2, T5(11B), ...)

성능 (매우) 개선:

사전학습된 언어지식(이해, 생성)이  
언어지능 처리 task에 성능 향상  
(사전학습된 언어의 의미 정보와 지식 활용)



딥러닝  
패러다임  
(큰 학습데이터)

# GPT-3(OpenAI, 2020)

175B params

Language Models are Few-Shot Learners

Tom R. Bravon*	Benjamin Mee*	Nick Wicker*	Melanie Sakhakh*
Jared Kaplan*	Prithvi Dhariwal	Arvind Nandakumar	Pranav Mishra
Colin Rafferty	Shayne Longpre	Arvind Neelakantan	Gregoire Deleury
Alexander Kirillov	Michael Agrawal	Ariel Herbert-Voss	Grigory Sizemore
Tom Hoenig	Cheng Li	Aditya Keskar	David M. Eagle
Jeffrey Wu	Chang Wang	Christopher Meese	Mark Chen
Eric Nigler	Matt DeLore	Scott Gray	Jack Clark
Christopher Boreale	Sam McCandlish	Alec Radford	Mika Lomax
Dario Amodei			

OpenAI

[arXiv:2005.14167]

- 동기 → AGI(Artificial General Intelligence), **일반/범용 지능**
  - ✓ BERT style → specialist (expert)
  - ✓ GPT style → **generalist (almighty)**
- **In-Context Learning (ICL) = Prompting: task finetune 없이 수행**
  - ✓ Few-shot learning: 예제 2~10개 이후 문제
  - ✓ One-shot learning: 예제 1개 이후 문제
  - ✓ Zero-shot learning: 바로 문제
- 생성 능력에 화제: 사람 수준의 뉴스기사

# III ChatGPT와 GPT4

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
see 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
see otter --> loutre de mer -- examples
peppermint --> menthe poivrée
plush giraffe --> girafe peluche
cheese --> -- prompt
```



# 또 새로운 패러다임 ??

Paradigm	Engineering	Task Relation
a. Fully Supervised Learning (Non-Neural Network)	Features (e.g. word identity, part-of-speech, sentence length)	CLS, LM, TAG, GEN
b. Fully Supervised Learning (Neural Network)	Architecture (e.g. convolutional, recurrent, self-attentional)	CLS, LM, TAG, GEN
c. Pre-train, Fine-tune	Objective (e.g. masked language modeling, next sentence prediction)	CLS, LM, TAG, GEN
d. Pre-train, Prompt, Predict	Prompt (e.g. cloze, prefix)	CLS, LM, TAG, GEN

Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing

Pragati Liu, Carnegie Mellon University  
 Hritik Vaswani, Carnegie Mellon University  
 Daniel Fried, Carnegie Mellon University  
 Graham Neubel, Carnegie Mellon University  
 Shikun Feng, National University of Singapore  
 Graham Neubel, Carnegie Mellon University  
 Graham Neubel, Carnegie Mellon University

LLM(Large Language Model)  
 공급자 독식  
 생태계 종속

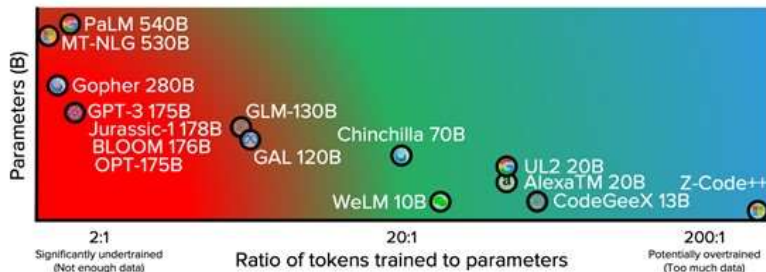
# Chinchilla Scaling Law

(Google DeepMind, 2022.03)

Training Compute-Optimal Large Language Models

- ✓ 신경망 크기에 비례해서 학습데이터 크기도 중요: 기존 모델들 그렇지 못함
- ✓ 신경망 크기에 적합한 모델 개발 Chinchilla(70B), 1 param  $\approx$  20 tokens 필요

## DATA-OPTIMAL (CHINCHILLA) MODEL HEATMAP



arXiv:2203.15556v2 [cs.LG] 15 Mar 2022

주소: <https://lilearchitect.ai/chinchilla/>

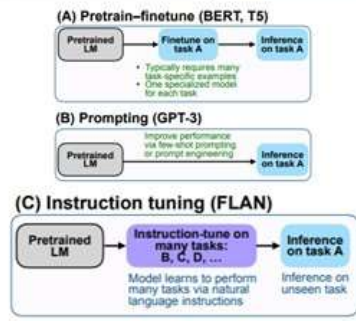
[LifeArchitect.ai/models](https://lilearchitect.ai/models)

# Instruction-Tuning

FLAN(Google, 2021.10), 137B params



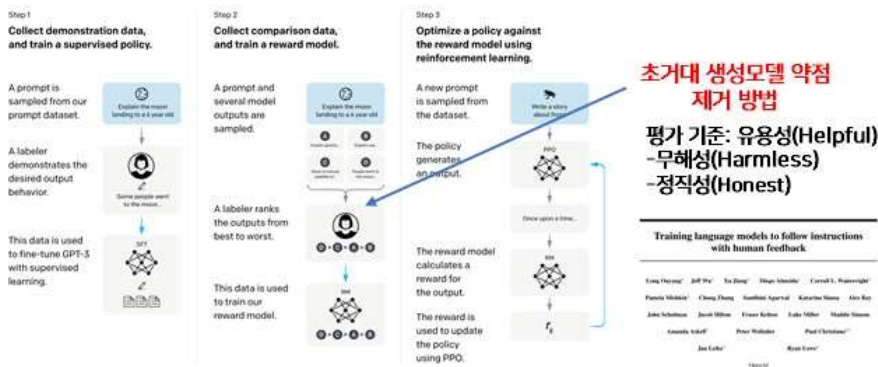
- 동기
  - ✓ Zero-shot learning 능력 향상: Instruction-tuning을 더해서
  - ✓ Prompting 필요없이 그냥 문제 풀이로 일반화 강화
- 시사점
  - ✓ GPT3보다 적은 파라미터로
  - ✓ 학습되지 않은 태스크도 처리 가능



# RHLF: 인간 피드백 주기

RHLF(Reinforcement Learning with Human Feedback)

- InstructGPT(OpenAI, 2022.03)
  - ✓ 사람의 피드백을 반영한 Instruction 강화학습 (사용자에게 alignment)
  - ✓ untruthful, toxic, not helpful한 생성 문제 개선 목적
  - ✓ 1.3B InstructGPT > 175B finetued GPT(Step1), 작은 모델도 일반화 가능성



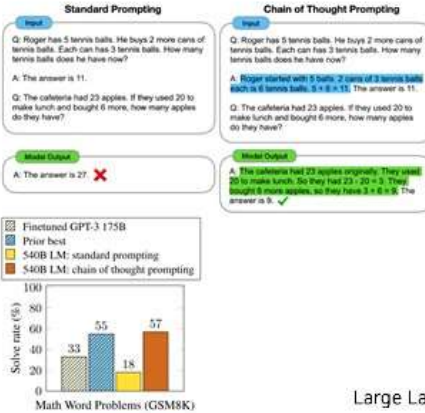
# CoT(Chain of Thought)

- CoT(Chain of Thought) (Google, 2022.06)

- ✓ 추론 중간 과정을 생성하게 하면 추론 성능 향상
- ✓ PaLM(540B)에 적용 실험: LLM의 난제 수학추론 향상

Chain of Thought Prompting Elicits Reasoning in Large Language Models

Jason Wei<sup>1</sup>, Yixue Wang<sup>1</sup>, Dirk Schreyer<sup>1</sup>, Maurice Herlihy<sup>1</sup>, Brian Ichter<sup>1</sup>, Kai Yu<sup>1</sup>, Ed H. Chi<sup>1</sup>, Quan V. Le<sup>1</sup>, Danny Dreyfus<sup>1</sup>  
 Google Research, Brain Team  
 {jasonw, yixuew, dirksch, mauriceh, briani, kaiy, edh, qvle, dannyd}@google.com



No.	Category	Template	Accuracy
1	instructive	Let's think step by step.	78.7
2		First, (*1)	77.3
3		Let's think about this logically.	74.5
4		Let's solve this problem by splitting it into steps. (*2)	72.2
5		Let's be realistic and think step by step.	70.8
6		Let's think like a detective step by step.	70.3
7		Let's think	57.5
8		Before we dive into the answer,	55.7
9		The answer is after the proof.	45.7
10	misleading	Don't think. Just feel.	18.8
11		Let's think step by step but reach an incorrect answer.	18.7
12		Let's count the number of "2" in the question.	16.7
13	irrelevant	By using the fact that the earth is round,	9.3
14		By the way, I found a good restaurant nearby.	17.5
15	-	AbraKadabra!	15.5
16		It's a beautiful day.	13.1
-	-	(Zero-shot)	17.2

Large Language Models are Zero-Shot Reasoners(NeurlPS2022)

# Emergent abilities

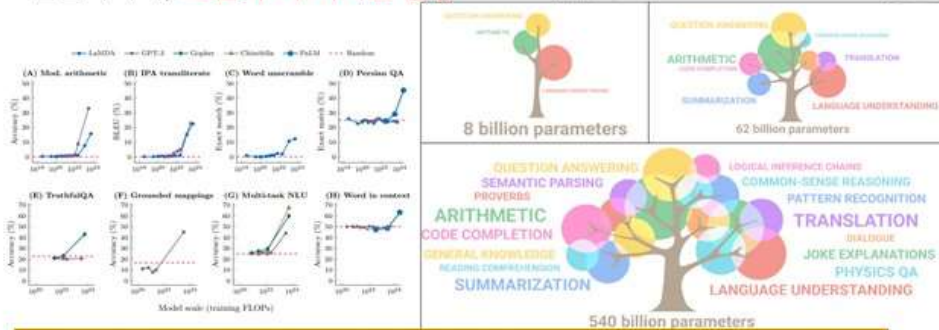
- ✓ 작은 모델에 없는 모델에서 없는 능력이 큰 모델에서 발현
- ✓ 성능도 특정 모델 사이즈까지 거의 랜덤하다가, 모델이 더 커지면 향상

Emergent Abilities of Large Language Models

(Google, 2022.08)

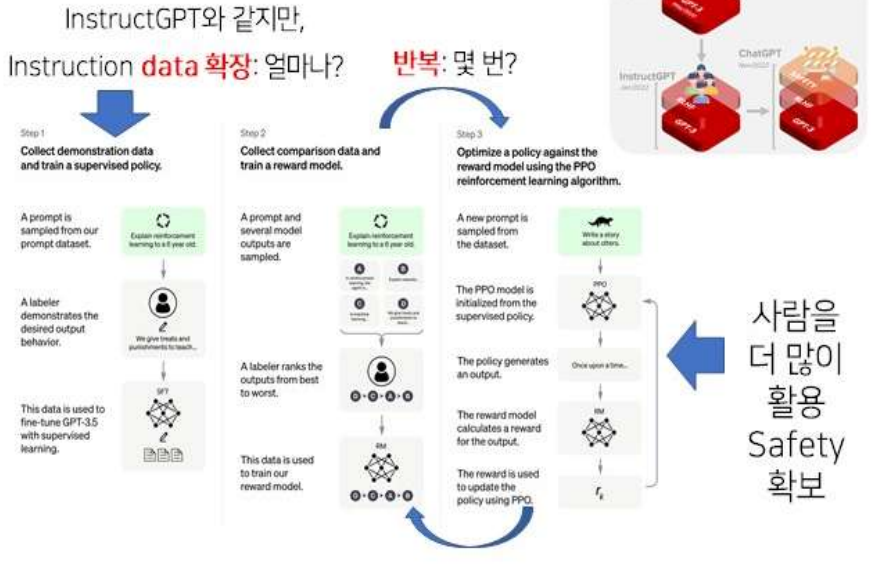
Jason Wei<sup>1</sup>, Yi Tay<sup>1</sup>, Bilal Hossain<sup>1</sup>, Cella Hsieh<sup>1</sup>, Barret Zoph<sup>1</sup>, Sebastian Borgeaud<sup>1</sup>, Dani Yogatama<sup>1</sup>, Maarten Dehaene<sup>1</sup>, Denny Zhou<sup>1</sup>, Donald Metzker<sup>1</sup>, Ed H. Chi<sup>1</sup>, Tamasoti Bhaskar<sup>1</sup>, Oriol Vinyals<sup>1</sup>, Perry Liang<sup>1</sup>, Jeff Dean<sup>1</sup>

{jasonw@google.com, yitay@google.com, bilalh@google.com, cellah@google.com, barretz@google.com, sborgeau@google.com, daniy@google.com, maartend@google.com, denny@google.com, donaldm@google.com, edh@google.com, tamiati@google.com, oriolv@google.com, perryliang@google.com, jdd@google.com}



- 딥러닝 패러다임 변화 시 이미 경험함: 자질 추출 능력 발현됨
- 발현이 아니라, 더 잘 학습되는 과정일 뿐, 단지 평가 수단에 의해 성능의 차이가 커 보임

# ChatGPT (a.k.a GPT-3.5)



# GPT4 (OpenAI, 2023.03)

**"GPT-4와 같은 대형 언어모델의 경쟁적 환경과 안전성을 고려해, 이 보고서에는 아키텍처(모델의 크기가 포함됨), 하드웨어, 학습 컴퓨팅, 데이터 세트 구조, 학습 방법 등에 대한 자세한 내용은 포함하지 않는다"**

- 정보 미공개
- 추측

<출처: <https://l1fearchitct.ai/gpt-4/>>

## Summary

Organization	OpenAI
Model name	GPT-4
Internal/project name	DV, DV-3, davinci 3 <sup>1</sup>
Model type	Multimodal (Visual language model)
Parameter count	1T (1,000B). Maybe Sparse. <sup>2</sup> = 5.7x bigger than GPT-3 175B = 14.3x bigger than Chinchilla 70B = 0.8% the size of the human brain by count of synapses (125T synapses)
Dataset size (tokens)	20T (20,000B) estimated. Maybe repeated tokens. <sup>3</sup> = 40x bigger than the GPT-3 dataset (495B tokens) = 8.5x bigger than DeepMind MassiveText English (2,35T tokens) = 4x bigger than DeepMind MassiveText Multilingual (5T tokens) = 7.1x bigger than Google Infnisnet (2,81T tokens)
Training data end date	Sep/2021
Convergence date	Aug/2022 <sup>4</sup>
Release date (public)	14/Mar/2023

- Text and Visual input, Text only output



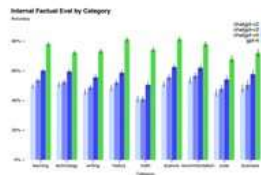
Output: There are many options for what you can make with these ingredients! Some possibilities include:- Pancakes or waffles- Crepes- French toast- Omelette or frittata- Quiche- Custard or pudding- Cake or cupcakes- Muffins or bread- Cookies or biscuits These are just a few examples, but the possibilities are endless!

# GPT4 성능

- 일반적인 지식과 문제 해결 능력
  - ✓ 대규모 다중 태스크 언어이해(MMLU) 평가셋 대상 GPT-3.5 모델 대비 15.4% 개선
  - ✓ 상식 추론 평가셋(HellaSwag, WinoGrande 등)에서 SOTA 갱신
  - ✓ 수리추론(GSM-8K: 초등학교 서술형 산술문제) 또한 우수한 성능

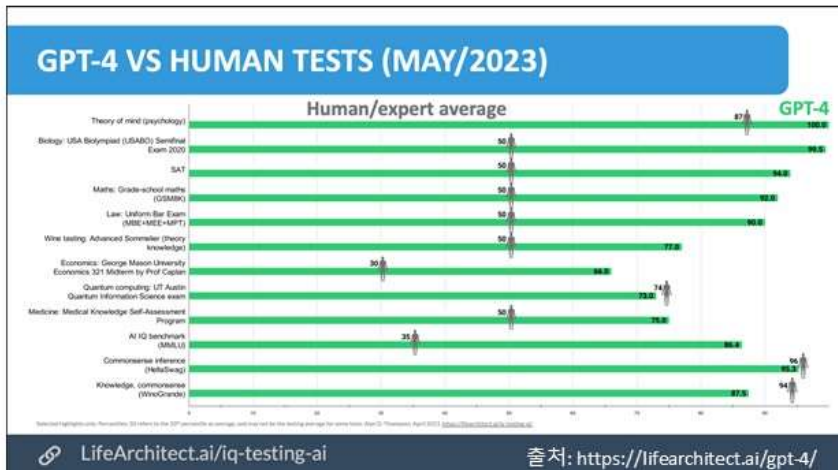
	GPT-4 Evaluated few-shot	GPT-3.5 Evaluated few-shot	LM SOTA Best external LM evaluated few-shot	SOTA Best external model (incl. benchmark-specific tuning)
MMLU [43] Multiple-choice questions in 57 subjects (professional & academic)	<b>86.4%</b> 5-shot	70.0% 5-shot	70.7% 5-shot U-Pal.M [44]	75.2% 5-shot Flan-Pal.M [45]
HellaSwag [46] Commonsense reasoning around everyday events	<b>95.3%</b> 10-shot	85.5% 10-shot	84.2% LLaMA (evaluation set) [28]	85.6 ALiM [47]
AI2 Reasoning Challenge (ARC) [48] Grade-school multiple-choice science questions. Challenge set.	<b>96.3%</b>	85.2%	85.2% 8-shot Pal.M [49]	86.5% STaMOR [18]
WinoGrande [50] Commonsense reasoning around personas resolution	<b>87.5%</b> 5-shot	81.6% 5-shot	85.1% 5-shot Pal.M [31]	85.1% 5-shot Pal.M [31]
HumanEval [37] Python coding tasks	<b>67.0%</b> 0-shot	48.1% 0-shot	26.2% 0-shot Pal.M [31]	65.8% CodeT5 + GPT-3.5 [51]
DROP [52] (F1 score) Reading comprehension & arithmetic	<b>80.9</b> 3-shot	64.1 3-shot	70.8 1-shot Pal.M [31]	<b>88.4</b> QGQA7 [53]
GSM-8K [54] Grade-school mathematics questions	<b>92.0%*</b> 5-shot chain-of-thought	57.1% 5-shot	58.8% 8-shot Mixtral [51]	87.3% ChatGPT + SFT + ORM-RL ORM-ranking [56]

- 환각(Hallucination)  
문제 ChatGPT 대비  
19% 개선 (RLHF 효과)



# GPT4 성능

- 인간테스트에 우수한 성적 but GPT4 전문가 대체 ???



## GPT4 취약점

IV GPT4에 의한 AI 시사점

### 초거대 언어생성 모델 태생적 약점

- 사실성 문제(Hallucination)
  - ✓ (원인 1) 초거대 언어모델에서 발현한 추론 능력(하위 목적함수)의 부산물로, 창의적으로 뭔가를 생성(학습 목적함수)하고자 함
  - ✓ (원인 2) 학습데이터의 거짓 정보 포함 가능성
- 편향성/개인정보보호 등 신뢰성
  - ✓ (원인) 학습데이터가 가진 편향성(종교, 성차별, 인종, 직업 등)과 개인정보 및 악의적 정보 포함
- 정보의 실시간성
  - ✓ (원인) 실시간 추가 학습이 불가능하여 학습 시점의 데이터 정보에 의존
- 제어성 (목적에 맞게 출력 제어, 규범/절차 준수)
  - ✓ (원인) 제어할 수 있는 목적 함수가 근본적으로 없음

## GPT4 취약점

IV GPT4에 의한 AI 시사점

### 초거대 언어생성 모델 태생적 약점

#### 현재 해결방법

- 대용량 학습데이터로 학습 및 대용량 신경망 구조로 확장 (혹시, 불확실하거나 비상식 추론 제어 능력 발현 ??)
- RLHF로 사람의 의도로만 생성되도록 제어 ← 사람 개입이 많음 → 자동화 연구 필요
- 학습데이터 가공/관리: 거짓/악의적 정보 및 개인정보를 학습데이터에서 제거, 편향성 감소할 데이터 증강
- Google BARD, MS BING 처럼 문서 검색과 결합하여 정보 실시간성과 사실성 문제 감소
- 활용 용도 선정 필요: 창의적 업무와 정보 처리 업무 등에 따라 제어 방법 연구 필요

# 활용과 생산성, 혜택

출처: [https://securities.miraeasset.com/bbs/maildownload/20230403175347670\\_154](https://securities.miraeasset.com/bbs/maildownload/20230403175347670_154)

표 2. 챗GPT 관련 확장 프로그램 및 사이트

명칭	내용
프롬프트 지니	챗GPT에서 질문/답변 자동 영한/한영 번역
ShareGPT	대화 내용을 PDF 등 형식 또는 링크로 공유
YouTube Summary	유튜브 영상의 스크립트 추출 및 영상 요약
챗GPT Writer	한국어/영어 이해할 작성, 답장 메일 생성
AIFPM for 챗GPT	여러 사용자들이 최적화된 프롬프트를 공유
챗GPT Optimizer	내용 복사, 단어/글자수 표시, 스크롤 이동 등
WebGPT	최근 웹페이지 등의 자료를 통해 최신성 보완
Gimme Summary AI	말한 중인 웹페이지 내용 바로 요약
DeepL Translate	웹페이지 내용을 마우스로 드래그하여 번역
Eightify	유튜브 영상 내용 요약 및 타임스탬프 제공
챗GPT to Notion	AI 비서 (내용 요약, 해일 담당, 수시 생성 등)
챗GPT Chrome Extension	크롬 브라우저에서 챗GPT 화면 접속
챗GPT for Google, Naver	Google, Bing, Baidu, Yahoo, Naver 등 지원
챗GPT for Google	구글 등 검색엔진에서 챗GPT 결과 표시
ReaderGPT	한번 클릭으로 웹페이지 내용 요약
GPT Formula	엑셀 / 구글시트용 함수 및 공식 생성
Merlin	AI 비서 (내용 요약, 해일 담당, 수시 생성 등)
라이니(Liner)	출처가 있는 구글 서치 어시스턴트
챗GPT Save	대화 내용을 CSV(엑셀) 파일로 저장
Tome AI	텍스트 프롬프트로 PPT 생성 및 편집
Beautiful.ai	텍스트 프롬프트로 PPT 생성 및 편집
다글로(daglog)	음성 데이터를 텍스트로 변환
Tactiq	Google Meet, Zoom 등 회의 노트 작성

자료: 언론 자료 종합, 미래에셋증권 리서치센터

- 영감 주기, Brain Storming - 연구주제 얘기해 보기
- 생산성 높이기: 문서 작성 등에 활용 MS, Google
- 코딩하기
- 교육 활용: 외국어 숙제 도우미 등



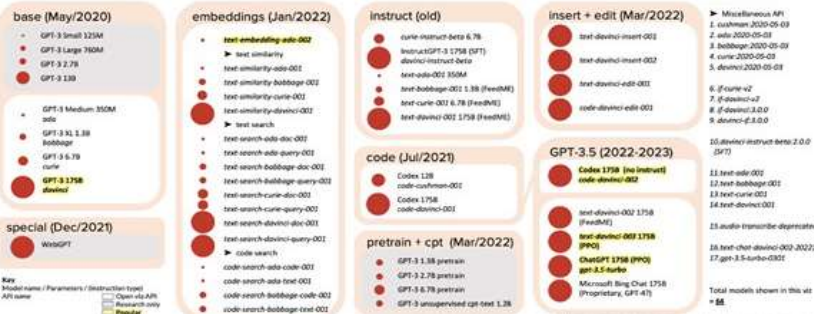
## 혜택자

- 미국: OpenAI, MS, Google, Meta ... 시스템, 데이터, 기술 독점 우려
  - 비숙련 노동자(?): 업무 초기에 도움, 전문가의 도움으로 전문가와 경쟁
- <GPTs are GPTs>
- 미국 노동인구의 80%가 업무영역에서 최소 10% 영향, 약 19%의 노동자는 업무의 최소 50% 이상 영향, 고소득 직종일수록 영향
  - 직업으로 보면 작가나 프로그래머가 자동화에 취약, 과학과 비판적 사고를 의존하는 직업은 상대적으로 덜 취약
  - 진입 장벽이 높은 직종이 취약

# 초거대 언어모델 경쟁

- OpenAI의 알려진 GPT-3 패밀리, 모델이 커질수록 학습은 어려움
- 비공개 결정 요인: 공개의 안전성 미확보, 경쟁 가속화 방지

## THE GPT-3 FAMILY: 50+ MODELS (MAR/2023)



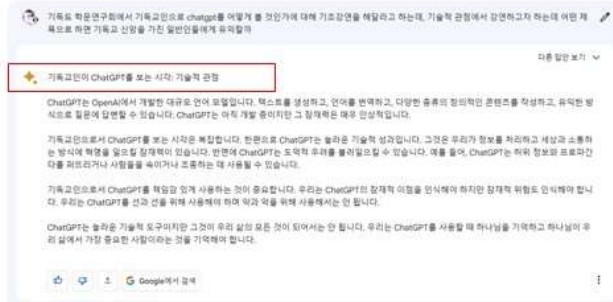
LifeArchitect.ai/gpt-3

# 초거대 언어모델 경쟁

IV GPT4에 의한 AI 시사점

## Google BARD, PaLM2

- PaLM 2(비공개 → 내부정보 360B params, 3.6T tokens)의 미세조정 및 대화형 버전
- 사실 Google은 PaLM 2 공개: 구글맵, 구글 독스, 지메일, 구글 시트, 챗봇 바드 등 25개 이상의 제품에 통합된 구글의 최신 AI 언어모델 PaLM 2 제공
- 입력: 이미지 + 텍스트, 출력: 텍스트 + 이미지(예정)
- “영어 다음 지원 언어는 한국어” 기사화



# 초거대 언어모델 경쟁

IV GPT4에 의한 AI 시사점

## Google BARD 예제

**앞 응답의 핵심인 기술적 관점 부분보다 자기 응답의 모든 내용 기준으로 작성함**



# 초거대 언어모델 경쟁

IV GPT4에 의한 AI 시사점

## Open Source LLM 열풍

- Meta AI: LLaMA(7B, 13B, 33B, 65B) 공개
- sLLM(smaller Large Language Model): 1~20B 사이의 중소형급 언어모델의 갑작스런 열풍
- 스탠퍼드 대학교 Alpaca, Databricks의 Dolly, Vicuna(13B), KoAlpaca 등
- 비영리단체 EleutherAI: GPT-J, Pythia, Polyglot-Ko(1.3B, 3.8B, 5.8B, 12.8B) 등
- 스태빌리티AI의 StableLM와 StableVicuna, Cerebras의 Cerebras-GPT, Huggingface의 HuggingChat ...



# 초거대 언어모델 경쟁

IV GPT4에 의한 AI 시사점

## Open Source sLLM 성능

### Phoenix: Democratizing ChatGPT across Languages

Zhibing Chen<sup>1,2</sup>, Feng Jiang<sup>1</sup>, Jinying Chen<sup>1</sup>, Tianan Wang<sup>1</sup>, Fei Yu<sup>1</sup>, Guiming Chen<sup>1</sup>, Hongtao Zhang<sup>1</sup>, Jiahao Jiang<sup>1</sup>, Chen Zhang<sup>1</sup>, Zhiyi Zhang<sup>1</sup>, Jianqun Li<sup>1</sup>, Xiang Wan<sup>1</sup>, Binzuo Wang<sup>1</sup>, Hailong Li<sup>2</sup>  
<sup>1</sup>School of Data Science, The Chinese University of Hong Kong, Shenzhen  
<sup>2</sup>Shenzhen Research Institute of Big Data  
[wangbinzuo@szrisc.cn](http://wangbinzuo@szrisc.cn)

13B 급 모델로 ChatGPT의 90~97% 수준 성능 보임 (GPT4가 비교 평가할 경우)

→사람 평가의 구체적 정보는 없으나 많은 성능 차이가 있음

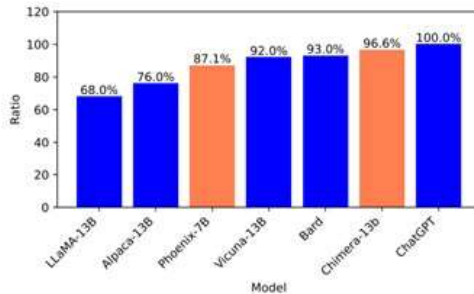


Figure 2: Relative Response Quality Assessed by GPT-4.

Table 8: Human evaluation of the Chinese answers of different models.

Comparison	Win	Tie	Lose
Phoenix vs. ChatGPT	12	35	53
Phoenix vs. Baidu-Wenxin	29	25	46

## 초거대 언어모델 경쟁

IV GPT4에 의한 AI 시사점

GPT4와 PaLM2 이후 경쟁 구도 가속화



## 국내 대응 방안 필요

IV GPT4에 의한 AI 시사점

- 초거대 모델 개발 가능 국가: 미국, 중국, 한국, 이스라엘 (전세계 4개 국가만 개발), 영국 개발 시도 중
- 한국 열세: 한국어 코퍼스 질, 인공지능 개발 인력 및 GPU 등의 인프라 등
- 대기업 위주로 초거대 모델 대응
- 산학연 및 오픈 SW 그룹: sLLM 개발로 대응
- 초거대 모델 최대 수요자 관점만이 아니라, 전세계 3강 공급자 관점에서 법제화 필요

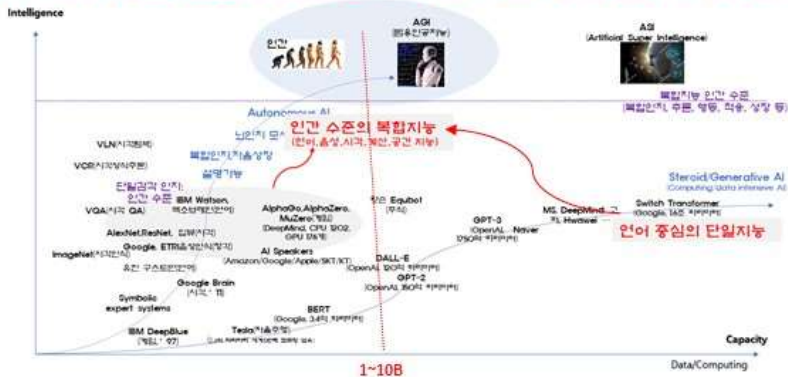


# 국내 대응 연구 예제

## IV GPT4에 의한 AI 시사점

ETRI 지원금 과제: 사람처럼 개념적으로 이해/추론이 가능한 복합인공지능 원천기술 연구 (2023.01 ~ 2027.12)

단순 언어모델을 인간의 개념 구조 모델로 발전시켜 더 작은 신경망 사이즈에서 추론하는 능력 확보



\* 참조: Garther IT Symposium 2021(IADAC) virtual conference 자료집, 21.10. 'Artificial super intelligence', IEEE Future Directions, 18.8. 인공지능을 능가하는 인공지능의 출현할 것인가, 2020/10. 카이로스칼럼, 장병직, 'AI, 1.0. 인공지능 기술 수준 정리', 전자통신동향분석, 20.6. '성명제의 유명한 저명지능 의사' 왕업과 제이이 가만 왕업을 융합한 통합 인공지능, 정보과학회지, 21.12

# 초거대 모델 = 인간

## IV GPT4에 의한 AI 시사점

- “인간은 여전히 중요한 노동력을 제공하며 GPT와 같은 기술은 대체제가 아닌 도구이다.” OpenAI CEO 올트먼
- 초거대 언어모델의 발현 능력(정립하지 않은 목적을 수행)에 대한 경이와 두려움이 존재
- 인간의 뇌세포 1% 정도의 파라미터로 더 많은 정보를 가지는 점, but 학습 능력이 인간보다는 매우 미약함. 적응 능력은 없음(더욱 커지면 발현될 수 있을까?).
- 인간의 뇌를 살피 듯, 초거대 언어모델 신경망을 패턴을 연구 활발

## 초거대 모델 = 인간

IV GPT4에 의한 AI 시사점

- 인간의 능력과는 아직 많은 차이: 시각 정보 추론, 배운 상태로 새로운 학습, 의심과 호기심, 스스로 학습데이터와 방법을 선택, 작은 정보로 다른 많은 상황을 유추, ...
- 인간처럼 주어진 목적 이외의 하위 목적에 관심을 가져 선택과를 탈 수 있을 것인가, AI가 인간처럼 되고 싶어 하는 자아를 발현할 수 있을까 ...
- 현재의 놀라운 말을 생성하기 위한 주목적을 이루기 위해 발현된 하위 목적인 계산, 추론, 창조 능력에 한정 되어 있음. 나머지 인간의 능력은 현재 신경망 구조와 입출력이 제한된 상태에서 발현이 어려움.

National AI Research Institute  
Making a Better Tomorrow

**ETRI** 한국전자통신연구원  
Electronics and Telecommunications Research Institute

# THANK YOU

