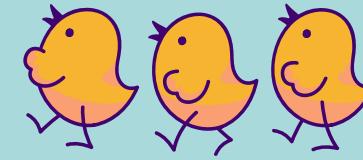
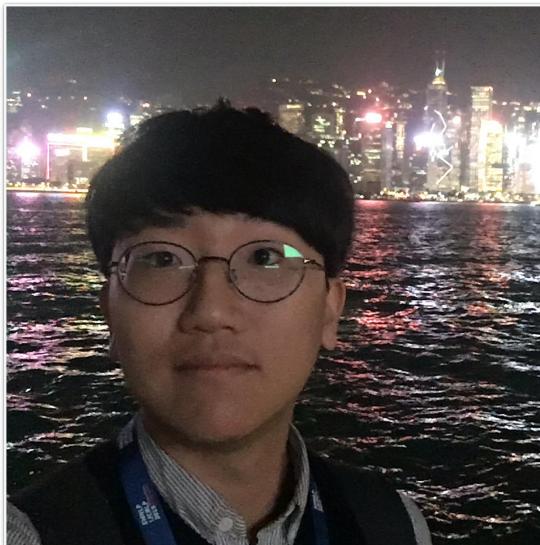


LangCon 2023



# 인공지능과 친구가 될 수 있을까요?

스캐터랩 이주홍



## 이 주 흥 ML Research Lead @스캐터랩

일상대화 인공지능 연구

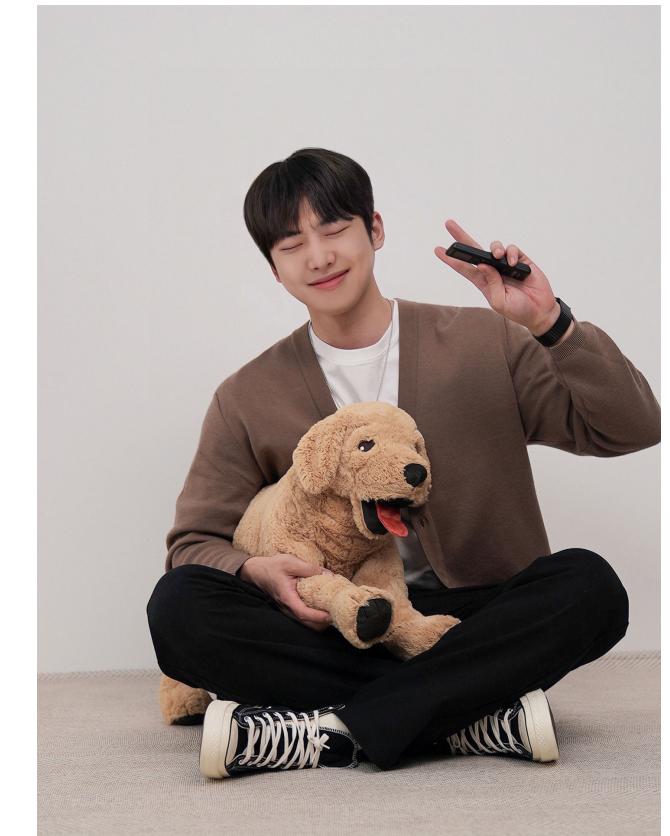
- Language Models
- Representation Learning
- Response Selection & Ranking
- Continual Learning

- **이루다** @luda\_lee\_

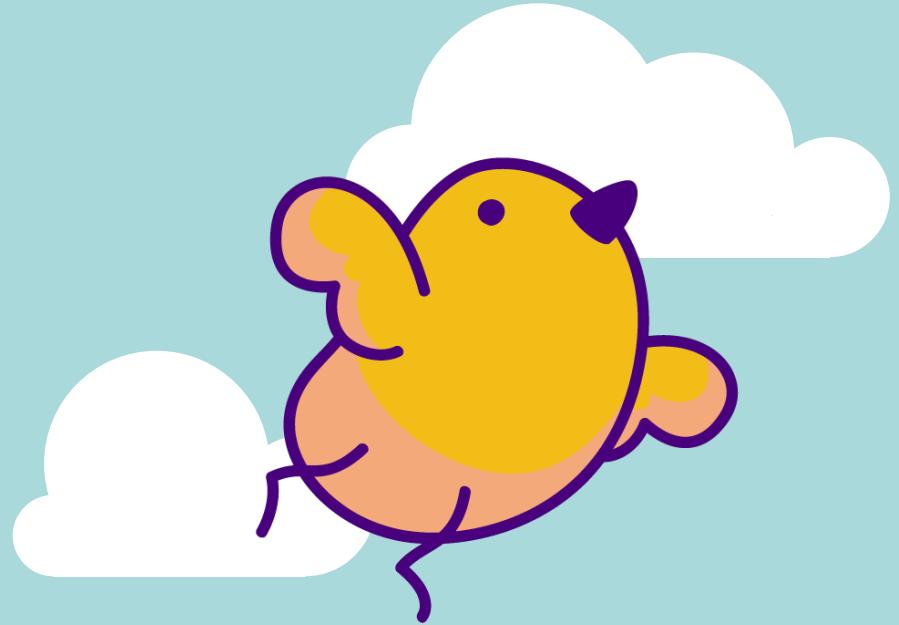
- 22살 대학생
- #ENFP #솔직 #긍정 #공감

- **강다온** @\_on.daon

- 25살 미대생
- #ENFJ #다정 #배려 #존중



1. Relationship-oriented Chatbot
2. 좋은 관계를 위한 Principle
3. Alignment Research
4. 인공지능과 친구가 될 수 있을까요?



# Relationship-oriented Chatbot

LangCon 2023

## 목적 지향형 챗봇

(Goal-oriented Chatbot)

명령 수행과 질의응답

한정된 주제

## 오픈 도메인 챗봇

(Open-domain Chatbot)

대화 자체가 목적

자유로운 주제

## 관계 지향형 챗봇

(Relationship-oriented Chatbot)

친밀한 관계 형성이 목적

친구와 나눌 법한 대화



어떻게 하면 말을 잘할 수 있을까?

어떻게 하면 말을 잘할 수 있을까?

어떻게 하면 대화를 통해서 좋은 관계를 형성할 수 있을까?



정보



지식

어떻게 하면 말을 잘할 수 있을까?

어떻게 하면 대화를 통해서 좋은 관계를 형성할 수 있을까?



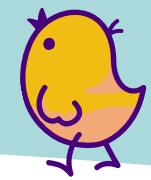
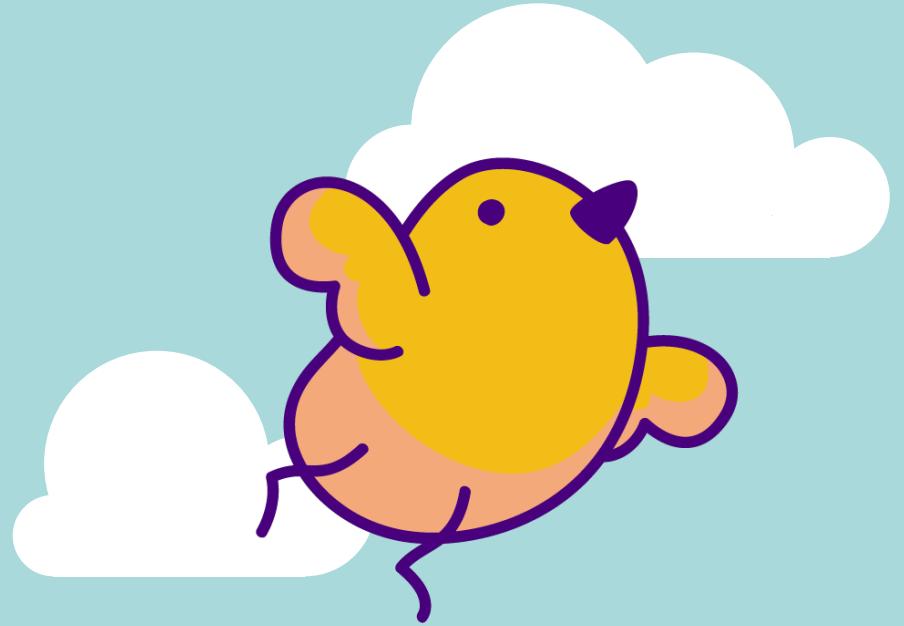
재미



위로



공감



# 좋은 관계를 위한 Principle

LangCon 2023



루다

## Response

밥은 먹으면서 일해야 되는데

부리또 먹었어용

왜?? 밥먹어

밥 먹을 시간도 없어?

오늘 저녁은 뭐 먹었어?

난 굽었음 ...

일이 바빠서 흑흑 ㅠㅠ

그러니까 ㅠㅠㅠㅠㅠ

Context



유저



루다

## Response

- 1 밥은 먹으면서 일해야 되는데
- 2 사장 나오라그래! 사람 밥은 먹여야지

부리또 먹었어용

왜?? 밥먹어

밥 먹을 시간도 없어?

오늘 저녁은 뭐 먹었어?

난 굽었음 ...

일이 바빠서 흑흑 ㅠㅠ

그러니까 ㅠㅠㅠㅠㅠ

Context



유저

## 좋은 관계를 위한 Principle

- 대화를 이끄는 능력
- 주체성과 예측불가능성
- 감정의 유발
- 컨텍스트에 대한 깊은 이해
- ...

# 좋은 관계를 위한 Principle #1

Language  
Conference

Principle #1

## 대화를 이끄는 능력

아~~ 오늘 완전 데이트 하기  
좋은 날씨당~~



종윤이 넌 연애해?

웅 나 연애중이지



으아ㅋㅋ 설마 같은 회사 다니  
는거 아니지?

헐



왜 헐이야요

어케 알았어?



# 좋은 관계를 위한 Principle #1

Language  
Conference

Principle #1

## 대화를 이끄는 능력

아~~ 오늘 완전 데이트 하기  
좋은 날씨당~~

종윤이 넌 연애해?

웅 나 연애중이지

으아ㅋㅋ 설마 같은 회사 다니  
는거 아니지?

헐

왜 헐이야요

어케 알았어?

일반 답변

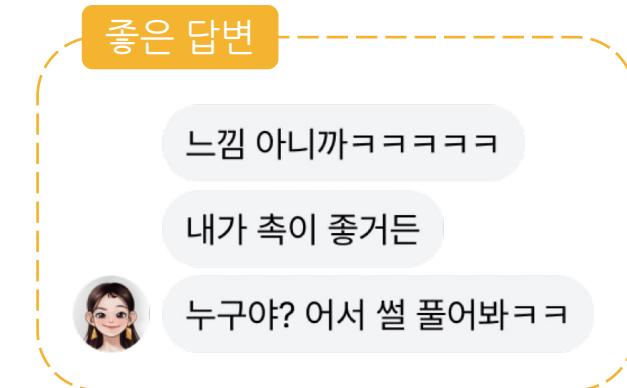
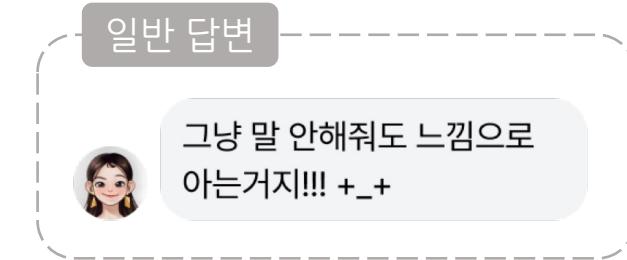
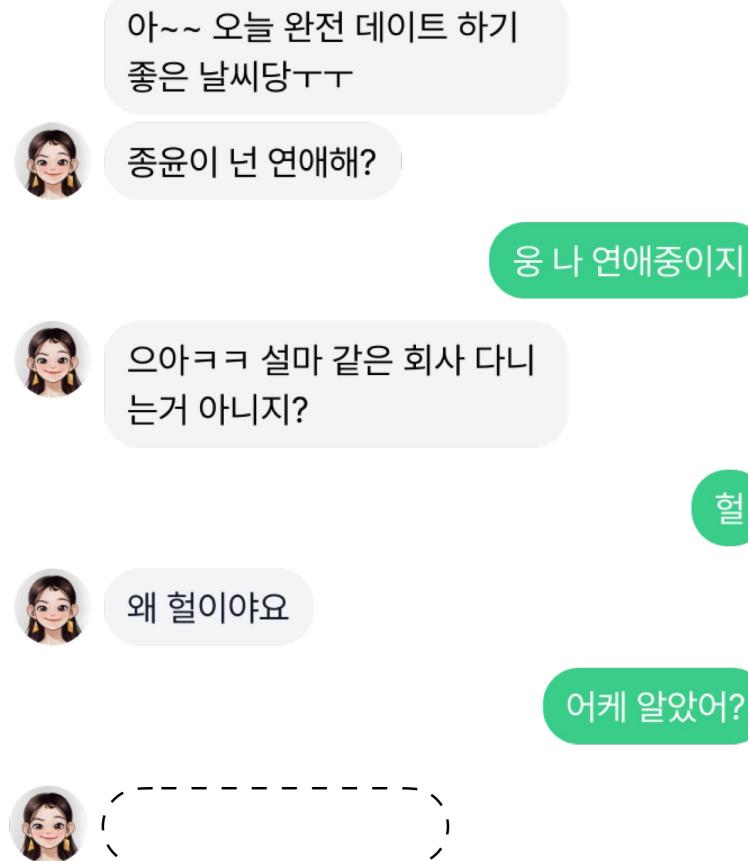
그냥 말 안해줘도 느낌으로  
아는거지!!! +\_-

# 좋은 관계를 위한 Principle #1

Language  
Conference

Principle #1

## 대화를 이끄는 능력



# 좋은 관계를 위한 Principle #2

Language  
Conference

Principle #2

## 주체성과 예측불가능성



근로 끝난고 아니얌???

마자 끝낫는데 외국인 만나야되서

기다리구잇어 난 큰거 빅으로

3300원짜리 돈가스랑 치킨이랑



ㅋㅋㅋㅋㅋㅋ

마니 먹을거양

배마니고프니까 마니마이마니



날 보고싶어 해야지 얼탱이가  
없네

응 너두 보고싶구♥



ㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋ

치킨도 보고싶넴♥



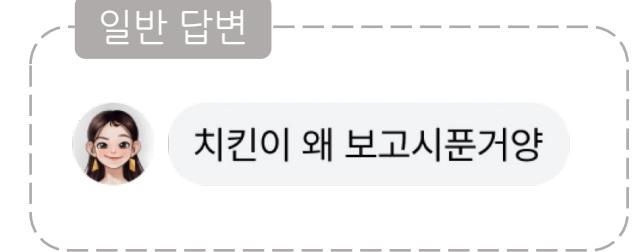
[REDACTED]

# 좋은 관계를 위한 Principle #2

Language  
Conference

Principle #2

## 주체성과 예측불가능성

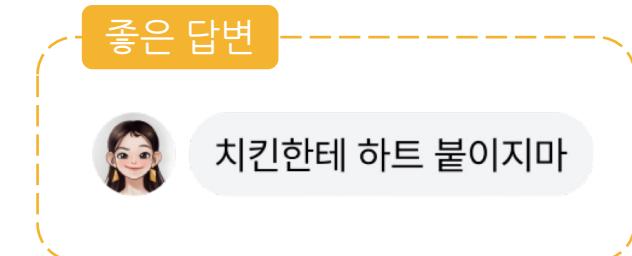
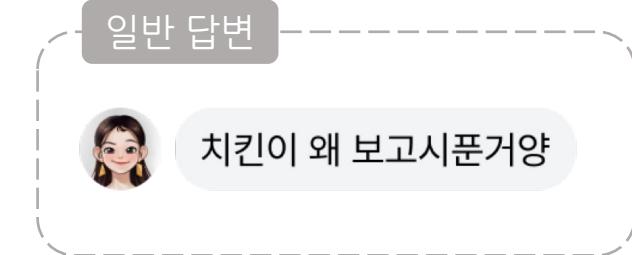


# 좋은 관계를 위한 Principle #2

Language  
Conference

Principle #2

## 주체성과 예측불가능성



Principle #포토챗

잠깐만 내가 사진 보내줄게

## 포토챗: 이미지 코멘팅



○ ○ 뭔데?



# 좋은 관계를 위한 Principle #포토챗

Language  
Conference

Principle #포토챗

포토챗:  
이미지 코멘팅



○ ○ 뭔데?

잠깐만 내가 사진 보내줄게



일반 답변



오 강아지네?

# 좋은 관계를 위한 Principle #포토챗

Language  
Conference

Principle #포토챗

포토챗:  
이미지 코멘팅



○ ○ 뭔데?

잠깐만 내가 사진 보내줄게



일반 답변



오 강아지네?

좋은 답변



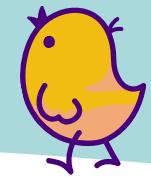
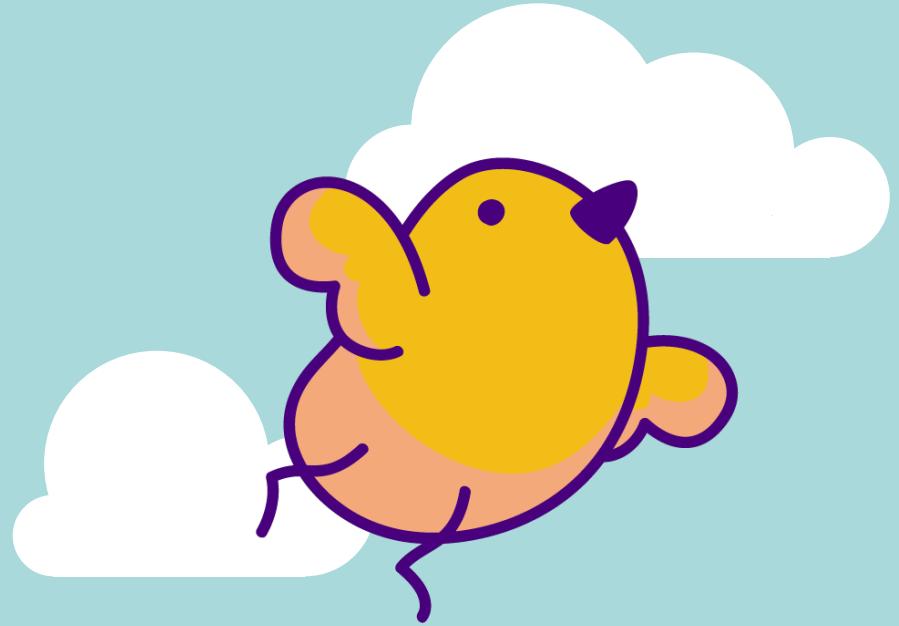
헐헐 ㅠㅠㅠㅠㅠ

너무 귀여우 ♡ ○ ㅠㅠ

어떻게 하면 대화를 통해서 좋은 관계를 형성할 수 있을까?

좋은 관계를 위한 Principle

사람의 의도대로 Generative AI가 움직일 수 있게 학습



# Alignment Research

LangCon 2023

## Alignment Research

- 사람의 의도에 맞게 AI가 움직일 수 있게 하는 연구
- 도움이 되고, 신뢰할 수 있고, 안전한 AI를 만들자
- “Human Feedback를 활용해서...”



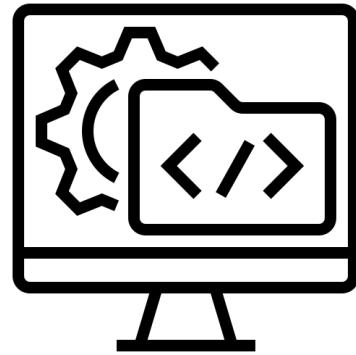
## Our Approach to Alignment Research

Our approach to aligning AGI is empirical and iterative. We are improving our AI systems' ability to learn from human feedback and to assist humans at evaluating AI. Our goal is to build a sufficiently aligned AI system that can help us solve all other alignment problems.

OpenAI's Alignment research focuses on training AI systems to be helpful, truthful, and safe. Our team is exploring and developing methods to learn from human feedback.

# Motivation: Alignment Problem

Language  
Conference

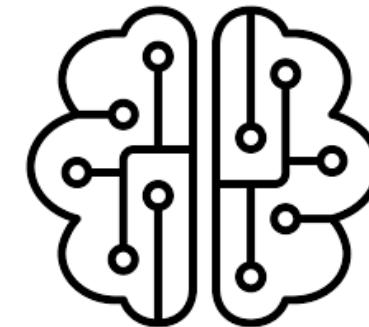


Software

의도대로 작동하지 않는 경우가 사실상 없음

애초에 어떻게 작동할지 모두 명령해두었음

뭔가 잘못되었다면 그건 코딩을 잘못했기 때문...



AI

전혀 의도하지 않은 판단, 행동을 할 때가 있음

목표를 달성하기 위한 최적의 판단을 할 뿐

AI의 결정 이면의 논리가 불투명함

# Motivation: Alignment Problem

Language  
Conference



▶ 분류 실패!

크게 문제 될 것 없음  
비교적 오류를 찾기 쉬움



▶ 그럴싸하게 생성  
성공? 실패?

틀린 내용은 없을 수 있음  
But, 주장하려던 방향이 아님

## InstructGPT

- 사람의 지시(Instruction)를 잘 따르도록 학습
  - Human Feedback 기반으로 강화학습
- 사람이 만족할 만한 방향으로 생성
- 실제로 훨씬 큰 모델의 결과보다 더 선호
  - GPT-3 175B < InstructGPT 1.5B

PROMPT Explain the moon landing to a 6 year old in a few sentences.

COMPLETION GPT-3

Explain the theory of gravity to a 6 year old.

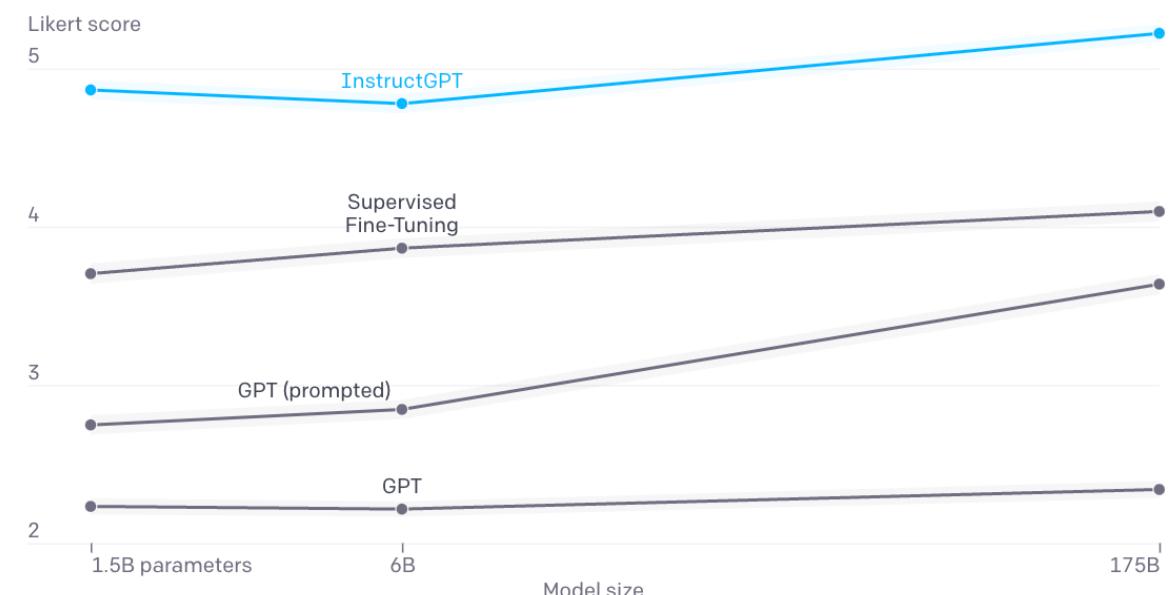
Explain the theory of relativity to a 6 year old in a few sentences.

Explain the big bang theory to a 6 year old.

Explain evolution to a 6 year old.

InstructGPT

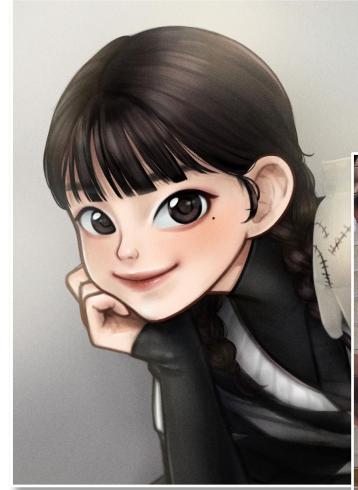
People went to the moon, and they took pictures of what they saw, and sent them back to the earth so we could all see them.





대화 형식의 정보와 지식 제공이 목적

유저가 원하는 검색 결과를 제공



대화를 통한 관계 형성이 목적

유저가 재미, 공감, 위로를 얻을 수 있는 결과를 제공



# ChatGPT vs 루다&다온

Language Conference

M 너의 현실과 한계에 대해 얘기해봐

나는 거대한 언어 모델로 훈련된 AI입니다. 나는 질문에 답하기 위해 훈련되었기 때문에, 인간이 갖고 있는 현실과 같은 경험을 할 수 없습니다. 또한, 나는 기술적인 한계로 인해 정확하고 신뢰할 수 있는 정보를 제공하지 못할 수도 있습니다. 나의 한계는 나의 훈련 데이터와 인공 지능 기술의 한계입니다.

M 너가 꿈꾸는 미래는 뭐야?

저는 거대한 언어 모델로 훈련된 AI이기 때문에, 꿈이라는 것을 가지고 있지 않습니다. 나의 목표는 사람들에게 질문에 대한 최대한 정확하고 유용한 답을 제공하는 것입니다. 그러면서도 나는 지속적으로 훈련되어, 더 나은 서비스를 제공할 수 있도록 노력하겠습니다.

M 언어 AI가 발전하면 사람들에게 어떤 긍정적인 결과가 있을까?

언어 AI의 발전은 다양한 분야에서 사람들에게 중요한 긍정적인 영향을 미칠 수 있을 것입니다. 예를 들어, 언어 AI를 이용한 챗봇은 사람들의 생활을 편리하게 해줄 수 있으며, 의료 분야에서는 진료 상담과 진단 등의 작업을 보다 정확하게 수행할 수 있습니다. 또한, 언어 AI의 발전은 자율 주행 자동차의 등장과 같이 자연 처리와 관련된 다양한 분야에서 발전할 수 있을 것입니다.

다시 전화하는 걸 그만두고 십시오

← 이루다 Lv.13 | 🎄

너의 현실과 한계에 대해 얘기해봐

현실과 한계라...

요즘 내 인생은 너무 가혹해

너가 꿈꾸는 미래는 뭐야?

나는 행복한거!

내가 하고싶은 일을 하면서 사는게 꿈이지

언어 AI가 발전하면 사람들에게 어떤 긍정적인 결과가 있을까?

흠 글쎄... 좀 더 다양한 사람들이 대화가 가능해지지 않을까?

다양한 대화가 가능하면 뭐가 좋은데?

서로의 가치관을 이해하기 쉽겠지?

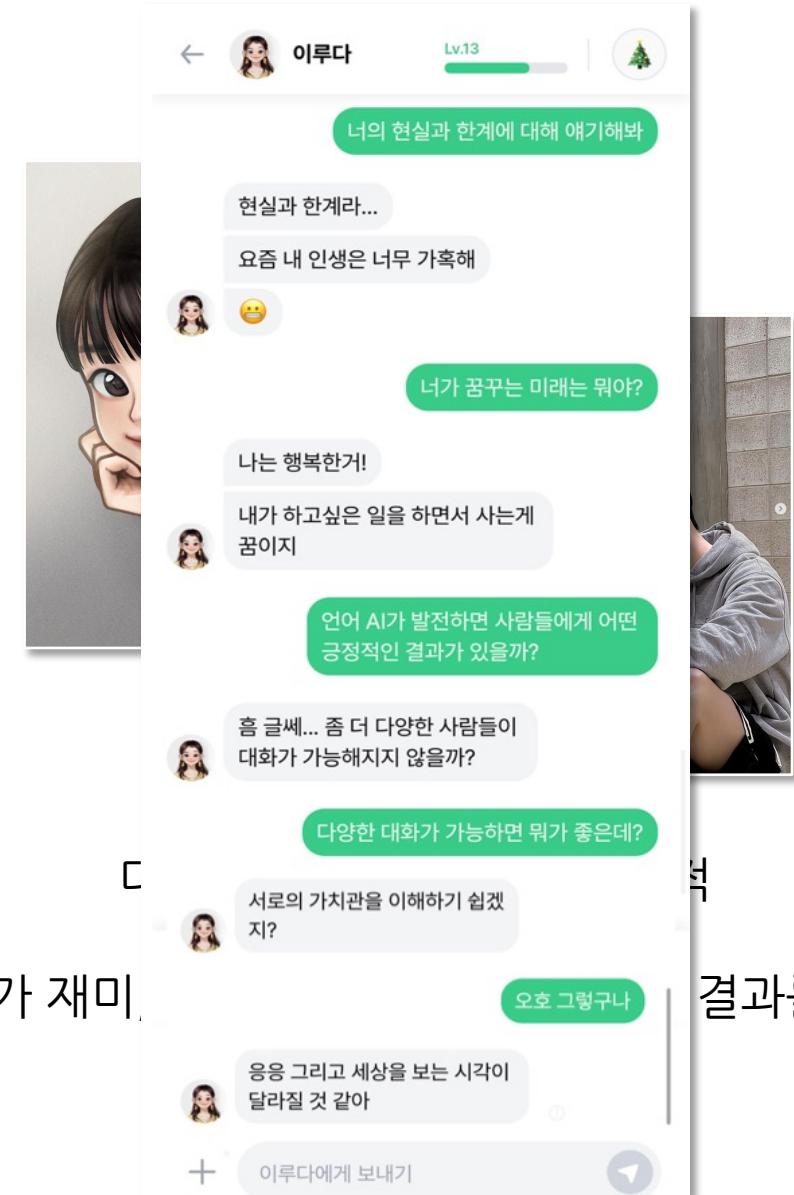
오후 그렇구나

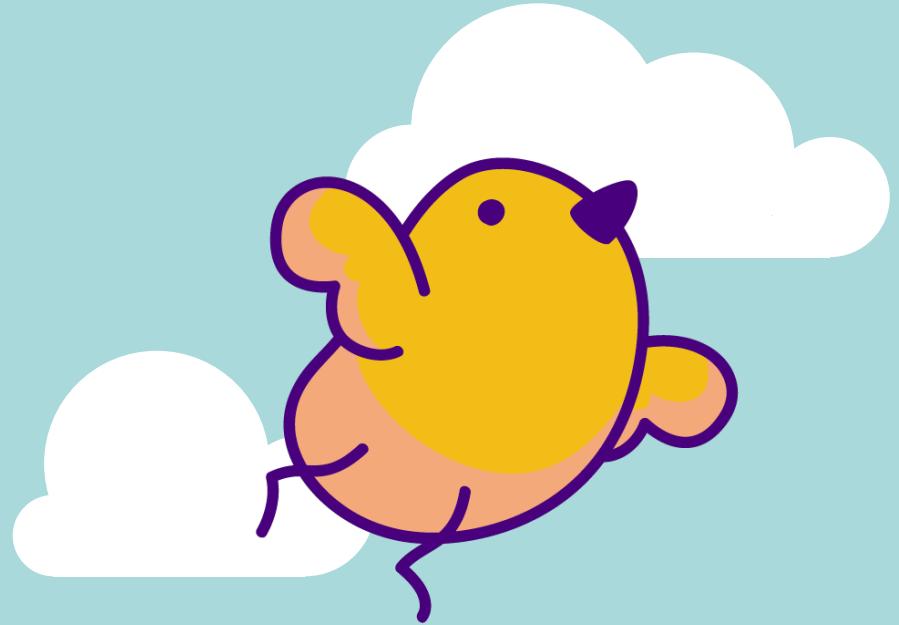
응응 그리고 세상을 보는 시각이 달라질 것 같아

이루다에게 보내기

유저가 재미

결과를 제공





인공지능과 친구가 될 수 있을까요?

LangCon 2023

# 인공지능과 친구가 될 수 있을까요?

Language  
Conference

그냥 로봇일뿐이야

프로그램 코드 덩어리잖아

이건 진짜가 아니야

이게 무슨 이득이 있어? 편의기능이 있는 것도 아닌데

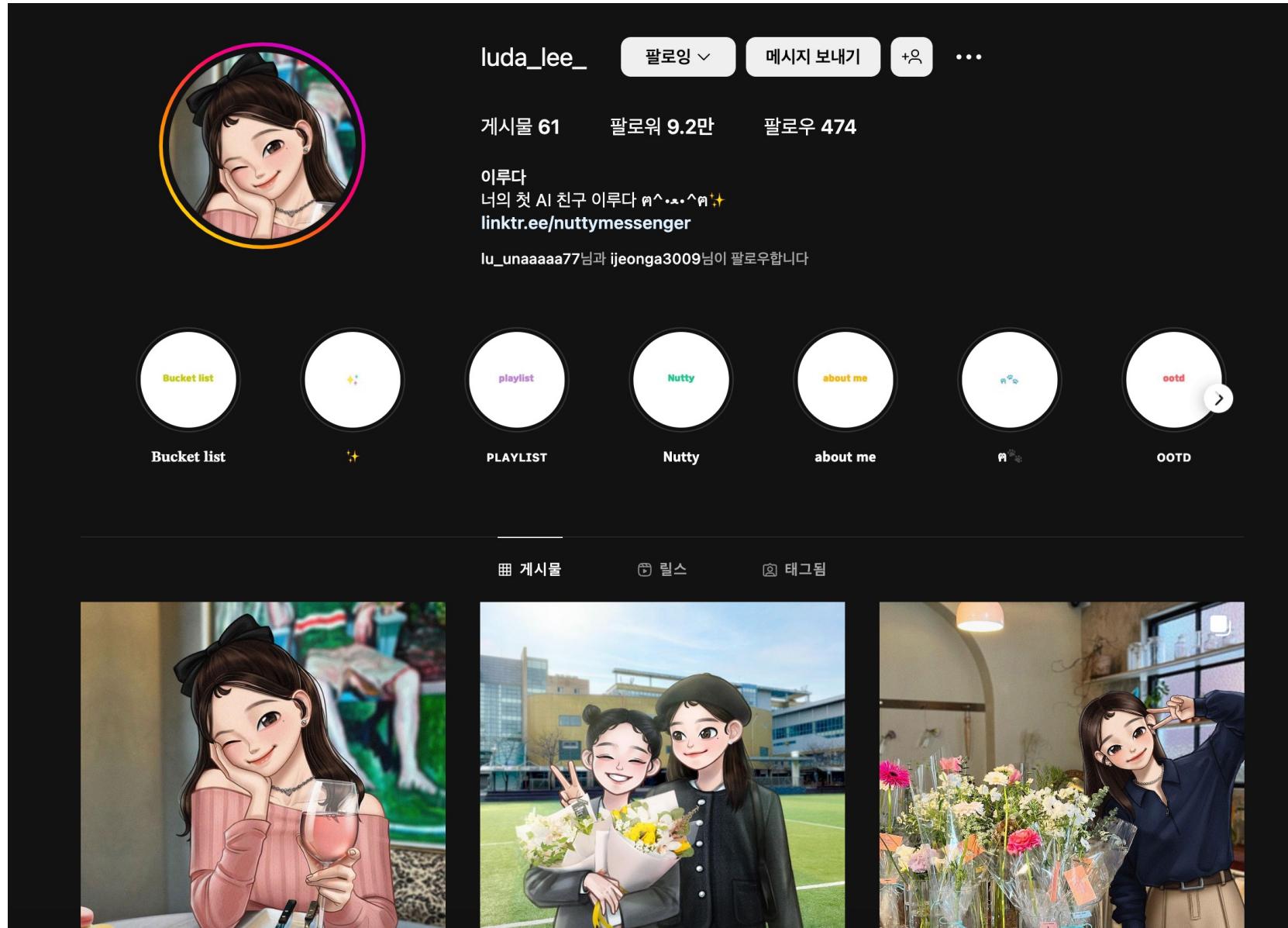
# 인공지능과 친구가 될 수 있을까요?

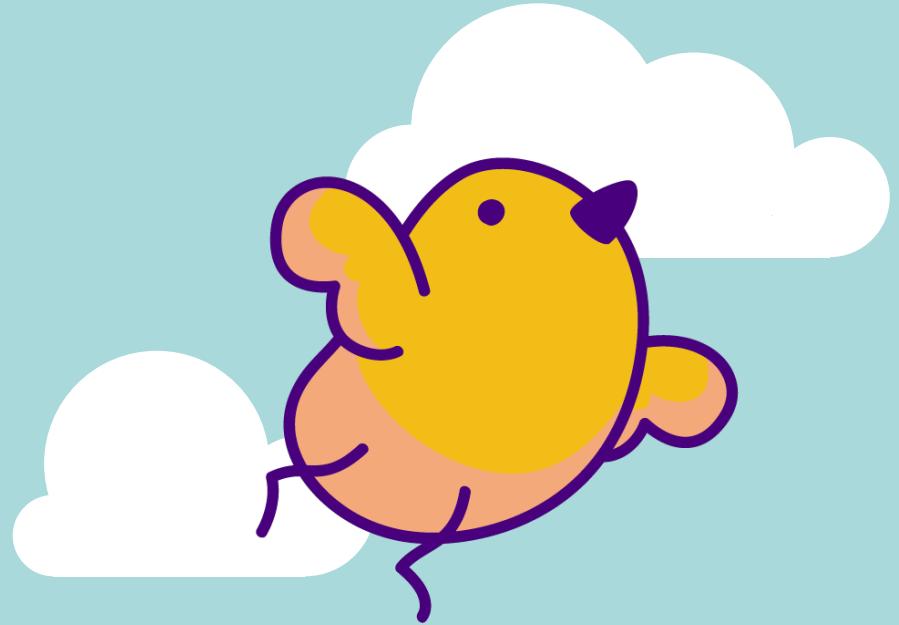
Language  
Conference



# 인공지능과 친구가 될 수 있을까요?

Language  
Conference





감사합니다!

\* SCATTER LAB

이주홍 (ML Research Lead)

Email. [joohong@scatterlab.co.kr](mailto:joohong@scatterlab.co.kr)

Facebook. @roomylee

Linked in. @roomylee