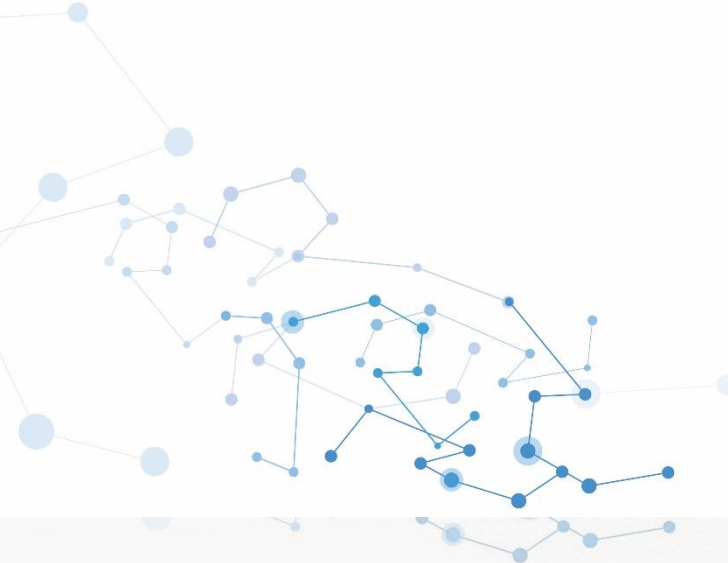


# 세계최고 수준의 차세대 AI 기술 개발을 통한 우수 전문 인력 양성 및 산·학·연 협력을 통한 가치 창출

고려대학교 Human-inspired AI 연구원

박기남



2024.10

# 발표 순서 및 내용

## ● 연구원 소개

1. 연구원 목표 및 비전
2. 연구원 보유기술
3. 연구원 현황

## ● 전략기술

1. LLM
2. 전략기술 실적
3. 전략기술 응용



# “고려대학교 HIAI 연구소는 세계최고 수준의 인공지능 연구소를 지향합니다.”

## 연구 및 기술 개발

“세계 최고 수준의 차세대 기술 연구개발”

## 전문인력 양성

“전문 연구인력 육성을 위한 우수 인재 양성 시스템 구축 및 인력 활용 시스템 운영”

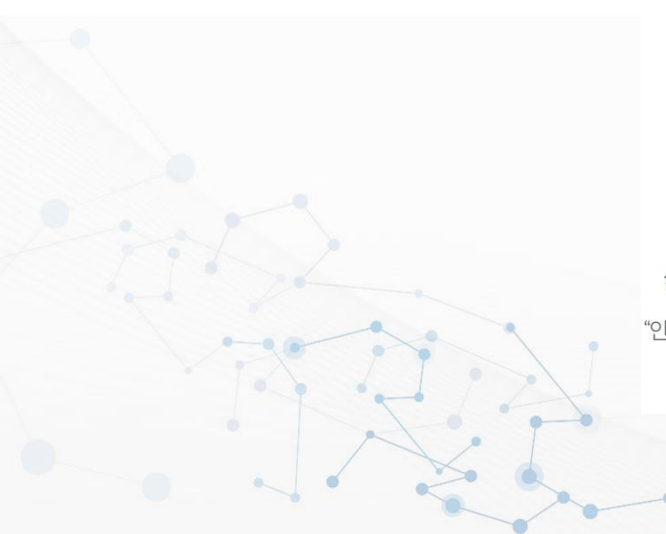


## 학·연·산 협력체 구축

“인공지능 기술 기반 산업체 애로기술 해결 및 맞춤형 인력 양성”

## 자생력 확보

“정부 및 산업체 연구과제 수주 및 중점기술 발굴”



# 다양한 분야에 최적화 된 자연어처리기반 언어지능 기술

## 1. 자연어처리

- 한국어 띄어쓰기 자동 교정기
- 딥러닝을 이용한 영어 문법 오류 교정기
- 통계 및 확률 기반 형태소 분석 기술
- 딥러닝 기반 형태소 분석 기술
- 개체명 인식기 (Named Entity Recognition)
- 문서 자동 분류 기술
- Bag of Characters를 응용한 Character-Level Word Representation 기술
- 병렬 코퍼스를 이용한 이중언어 워드 임베딩
- Stack-Pointer Network를 이용한 한국어 의존 구문 분석
- 의존구문분석 (Dependency Parser)
- Small Data의 한계를 극복하기 위한 전이 학습 모델
- 통계기반 한국어 뉴스 감성분석
- 대화속 화자의 감성 분석 (Emotion Recognition in Conversation)
- 자연어 추론에서의 교차 검증 앙상블 기법
- Denoising Transformer기반 한국어 맞춤법 교정기
- 지식 임베딩 심층학습을 이용한 단어 의미 중의성 해소
- Attentive Aggregation(주의적 종합)기반 크로스 모달 임베딩
- Poly-encoder를 이용한 COVID-19 질의응답시스템
- 외부지식정보를 이용한 상식추론 질의응답시스템
- 사전 학습된 Transformer 언어 모델의 이중 언어 간 전이 학습을 통한 자원 희소성 문제 극복
- 한국어 특성을 반영한 한국어 관계추출 기술
- Whisper 기반 음성인식기 API 개발
- 한국어 상식 추론 모델
- Phone Scam 탐지기 우회 문구 생성 기술

## 2. 대화 시스템

- 대화 시스템에서의 자연스러운 대화를 위한 Memory Attention 기반 Breakdown Detection
- 검색 기반 대화 시스템에서의 정답 예측 기술
- 딥러닝 기반 자동 질의응답 시스템
- 딥러닝 방법을 이용한 발화의 공손함 판단
- 기계 독해(MRC)를 이용한 COVID-19 뉴스 도메인의 한국어 질의응답 챗봇
- 일상대화생성 모델
- 시각 질의응답 시스템
- 화자의 페르소나를 반영한 대화 모델
- 지식 검색 기반 일반 상식 문장 생성기(영어)
- KommonGen: 한국어 일반 상식 추론을 위한 데이터
- 페르소나 및 지식 기반 대화 데이터와 베이스라인 모델 구축
- PEEP-Talk: 상황별 영어 교육을 위한 챗봇
- 유형다양성을 고려한 교육용 질의응답쌍 생성 모델
- 지식과 페르소나를 반영한 RAG기반 대화 모델

## 3. 정보 검색/분류/추출/요약

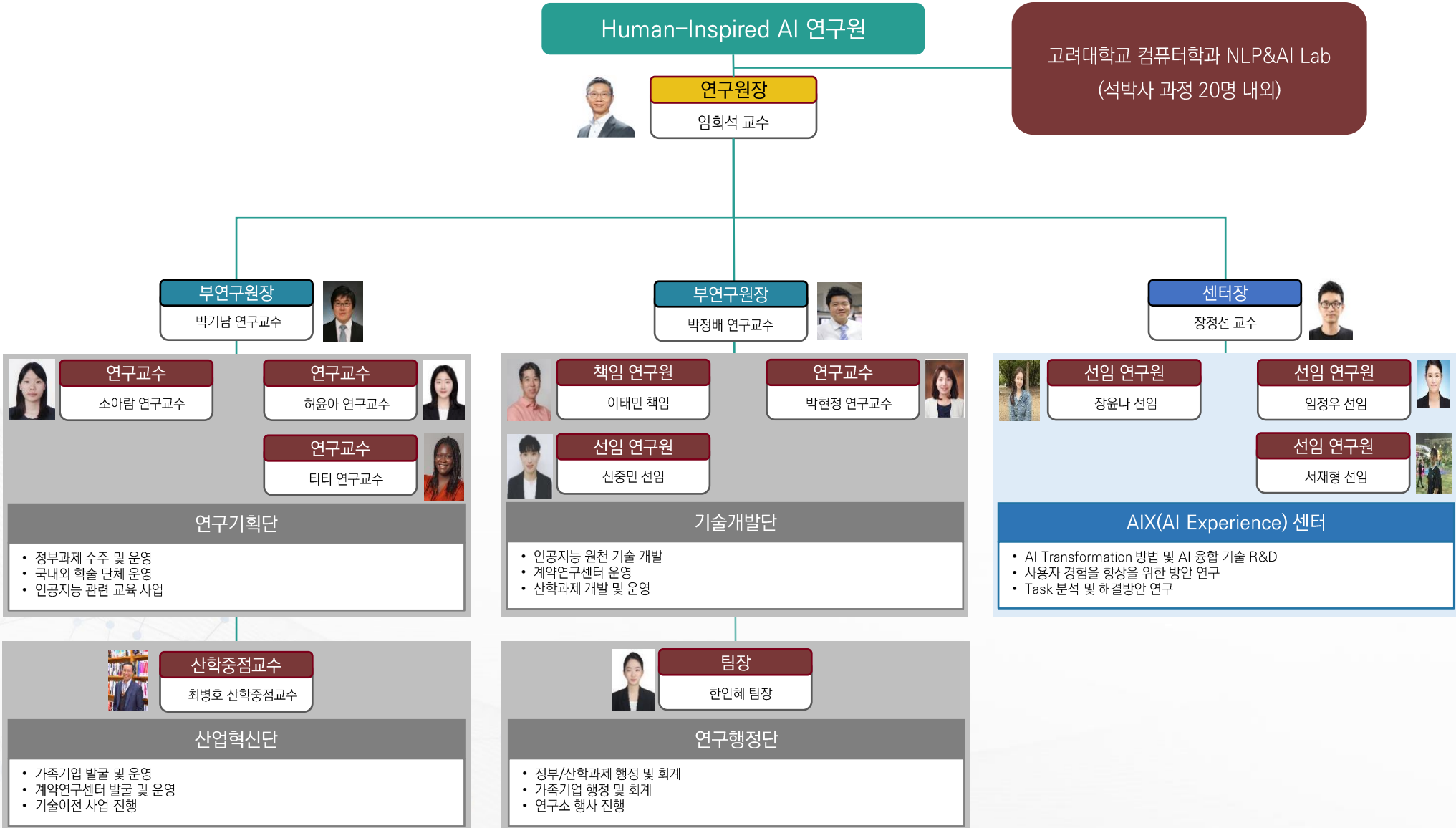
- 머신러닝 기반 보고서 자동 분석 및 키워드 추출 기술
- 메타러닝을 응용한 문서 단위의 관계 추출
- 비정형 위함정보 자동 인식 및 추출
- 머신러닝을 이용한 문서 자동 요약
- 딥러닝을 이용한 유사 문서 검색 및 시각화
- Narrative기반 자동 비디오 분할
- 비지도 학습 알고리즘을 이용한 보고서 자동 분석 및 토픽 자동 추출 기술
- 순차 정보를 이용한 콘텐츠 추천 시스템 개발
- 스케치를 이용한 패션 의류 검색 시스템
- Eye tracking 기반의 휴먼 리딩을 반영한 추출 요약 기법
- Sentence BERT 임베딩을 이용한 과편향 뉴스 판별
- 종교활동을 위한 휴머노이드 질의응답 로봇
- 아이들 교육을 위한 나오 로봇
- GPT2를 활용한 유사 뉴스 기사 추천 시스템
- 나오 로봇을 활용한 이중 언어 교육
- 나오 로봇을 활용한 동화 추천 및 읽기
- Virtual-Try-On Model for Fashion AI
- 사용자 그래프 기반 한국어 가짜뉴스 판별 방법

## 4. 기계번역

- 고려대학교 다국어 신경망 기계번역기
- 딥러닝 기반 한국어 고전번역기
- PicTalky: Text to Pictogram
- COVID19 도메인특화 기계번역기
- 인간의 인지과정을 반영한 도메인 특화 번역기



# 최신 인공지능 원천기술 개발, 정부과제 수행, 산학 및 계약연구 센터 운영 등을 위한 인력 구성



Human-Inspired AI 연구원



**연구원장**  
임희석 교수

고려대학교 컴퓨터학과 NLP&AI Lab  
(석박사 과정 20명 내외)

**부연구원장**  
박기남 연구교수



 <b>연구교수</b> 소아람 연구교수	 <b>연구교수</b> 허윤아 연구교수	 <b>연구교수</b> 티티 연구교수
-----------------------------	-----------------------------	----------------------------

**연구기획단**

- 정부과제 수주 및 운영
- 국내외 학술 단체 운영
- 인공지능 관련 교육 사업

**부연구원장**  
박정배 연구교수



 <b>책임 연구원</b> 이태민 책임	 <b>연구교수</b> 박현정 연구교수	 <b>선임 연구원</b> 신종민 선임
-----------------------------	-----------------------------	-----------------------------

**기술개발단**

- 인공지능 원천 기술 개발
- 계약연구센터 운영
- 산학과제 개발 및 운영

**센터장**  
장정선 교수



 <b>선임 연구원</b> 장윤나 선임	 <b>선임 연구원</b> 임정우 선임	 <b>선임 연구원</b> 서재형 선임
-----------------------------	-----------------------------	-----------------------------

**AIX(AI Experience) 센터**

- AI Transformation 방법 및 AI 융합 기술 R&D
- 사용자 경험을 향상을 위한 방안 연구
- Task 분석 및 해결방안 연구

 <b>산학중점교수</b> 최병호 산학중점교수
---------------------------------

**산업혁신단**

- 가족기업 발굴 및 운영
- 계약연구센터 발굴 및 운영
- 기술이전 사업 진행

 <b>팀장</b> 한인혜 팀장
-------------------------

**연구행정단**

- 정부/산학과제 행정 및 회계
- 가족기업 행정 및 회계
- 연구소 행사 진행

# 세계최고 수준의 차세대 AI 기술 개발을 통한 우수 전문

## 인력 양성 및 학·연·산 협력을 통한 가치 창출

학·연·산 협력을 통한 가치창출



### 차세대 AI 기술 연구개발



Human Inspired AI Research  
KOREA UNIVERSITY



### 우수 전문 인력양성

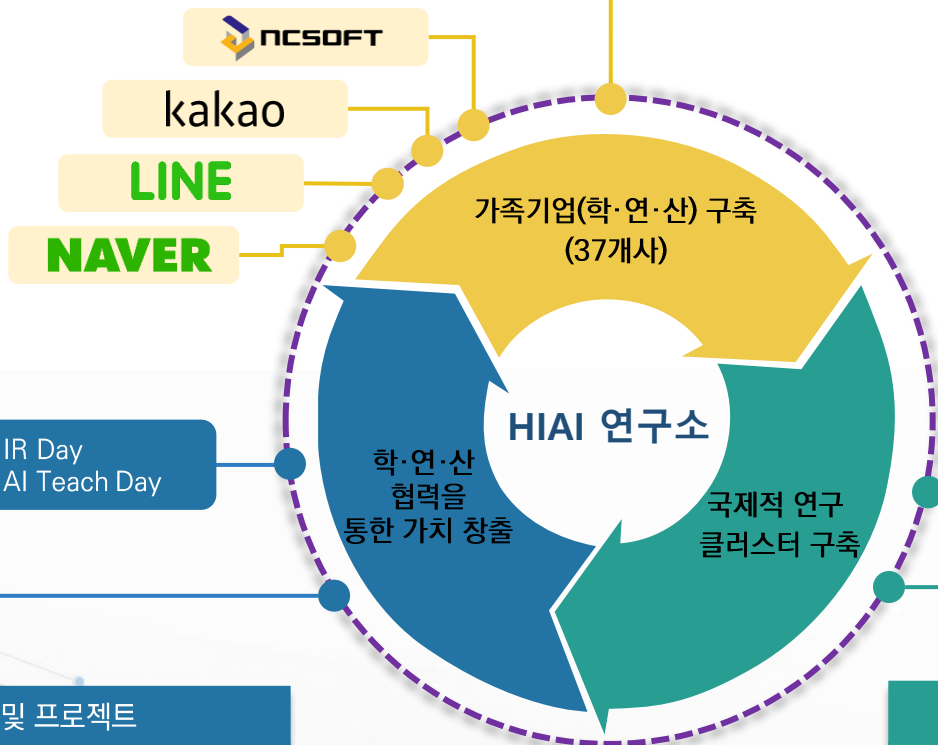
구분	주요 연혁
2023.09	고려대 NLP&AI 연구팀, 'ICT 챌린지 2023' 과학기술정보통신부 장관상 수상
2023.07	Human-inspired AI 연구소 Tech-day 행사개최
2022.12	2022년도 병역지정연구소 신규 선정
2022.12	ICICPE2022 국제학술대회 개최
2022.11	ICT 챌린지 2022 장관상 수상
2022.07	Human-inspired AI 연구소 Tech-day 행사개최
2022.04	계약연구센터(오투오, 플래스, 디엠랩, 케어사피엔스)
2021.09	AI 기술이전 상담회 개최
2021.05	2021년도 산학연 Collabo R&D사업 수주(워드마인드, HIAI 연구소, 약 4.5억)
2021.05	2021년도 ICT R&D 혁신 바우처 사업 수주(오투오, HIAI 연구소, 약 4.2억)
2021.05	2021년도 ICT R&D 혁신 바우처 사업 수주(비즈스프링, HIAI 연구소, 약 4.2억)
2021.05	중점연구소 사업 수주(9년간 총 80억 이상)
2020.12	2020 Okawa Foundation Research Grant Recipients 선정(연구소장)
2020.1	기술이전을 통한 회원사 투자유치 성공(약 20억원 이상)
2020.05	37개 기업과 산학협력체 구성 및 AI Tech Day 개최
2020.02	Human-Inspired AI 연구소로 명칭 변경
2019.12	엔씨소프트 외 30개 회원사와 업무협약 체결(현재 운영 중)
2019.09	글로벌 메이커톤 2019 수상(소속 연구원)
2019.02	인공지능 기계학습 교육아카데미 개최
2018.12	ICICPE2018 국제학술대회 개최
2018.06	Human-Inspired AI 복합지능 연구센터 개소

구분	주요 연혁
2024.11 (현재)	특허 출원 및 등록 50건 이상
	기술이전 20억 돌파
	연구비 수주 370억 돌파(2023년 10억 수주)
	가족기업 30개 회원사와 업무협약(현재 운영중)

# Human-Inspired AI 연구원 최근 5년간 학·연·산 협력을 통한 경제적 가치 추진 실적

**학·연·산 협력체(가족기업) 구축 및 운영**

- Human-Inspired AI 연구소 학·연·산 협력체계 구축을 통한 기술 교류 활성화
- 연구소 주관 IR (Investor Relations) Day 개최 등 가치 창출을 위한 적극적 노력



**학·연·산 협력 및 프로젝트**

**기술 교류회, IR Day, 협력 프로젝트 수행**

- 가족회사 대상 IR Day 개최
- 가족회사 투자 유치 성공(O2O)
- 기술 교류회를 통한 기술이전 및 협력 프로젝트 수행

**국제적 연구 클러스터 구축**

**세계화를 위한 국제연구기관, 대학과의 교류 활성화**

# 세계최고 수준의 차세대 AI 기술 개발을 통한 우수 전문

## 인력 양성 및 학·연·산 협력을 통한 가치 창출

계약연구센터	
기업명	연구내용
(주)가온플랫폼	Private LLM 모델 개발
(주)오투오	Deep Learning 모델 Customizing 기술
케어사피엔스	인공지능 추천 알고리즘
(주)쿠드로보틱스	Qoder-Large Language Model
유라클	Private LLM 모델 개발

가족기업	
기업명	플래티넘 파트너십(7개社)
바이브컴파니	
온더아이티	
에이아이노믹스	
가온플랫폼(주)	
(주)테스트웍스	
(주)심스리얼리티	
(주)파워보이스	
골드 파트너십(12개社)	
CYD	
바이엇	
로마켓	
퀵렌시스	
질경이	
(주)쿠드로보틱스	
크디랩(주)	
유스바이오글로벌	
(주)업템포글로벌	
(주)브이웨이	
생코에듀(주)	

**삼성화재, 고려대와 '디지털 경쟁력 강화' 연구 협약**  
 디지털 서비스 위한 체계적인 협력체계 구축  
 2023.08.03 17:10 | 권민호 기자

**고려대-오투오, 국내 첫 'AI 계약연구센터' 설립**  
 10년 계획 '장기 협력체계' 구축  
 양립 연구성과 공유·역량 내재화  
 기업에 '전담 연구원' 상시 제공  
 지속 가능한 AI 기술 R&D 기대

10년 계획 '장기 협력체계' 구축 양립 연구성과 공유·역량 내재화 기업에 '전담 연구원' 상시 제공 지속 가능한 AI 기술 R&D 기대

10년 계획 '장기 협력체계' 구축 양립 연구성과 공유·역량 내재화 기업에 '전담 연구원' 상시 제공 지속 가능한 AI 기술 R&D 기대

**전자신문 etnews Conference allshowTV English E**  
 경제·금융 전자·모바일러티 통신·미디어·게임 소재·부품 SW·보안 산업·에너지·환경 플랫폼·유통

**AI+바이오, 대학 최초 계약연구센터 만들어진다**  
 “30년지기 친구처럼 대화” 엔씨소프트-고려대, AI 대화 데이터셋 공개

고려대 HI A 공동 R&D 중추로 5개의 연구소가 설립된다

고려대 HI A 연구하는 곳이 원천기술 개발 연구의 '가교'가 된다

고려대 HI A 연구원들의 시너지 연구 성과를 공유하고 있다. 박사과정자

무로영리, 에이라 다 죽어...

물론요딩으로 실천부터 시작하기

# 발표 순서 및 내용

## ● 연구원 소개

1. 연구원 목표 및 비전
2. 연구원 보유기술
3. 연구원 현황

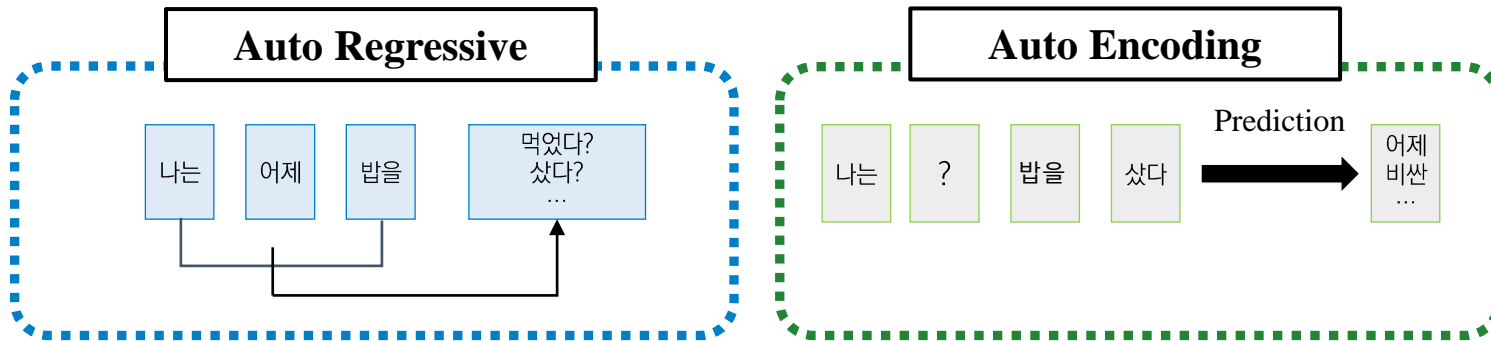
## ● 전략기술

1. LLM
2. 전략기술 실적
3. 전략기술 응용

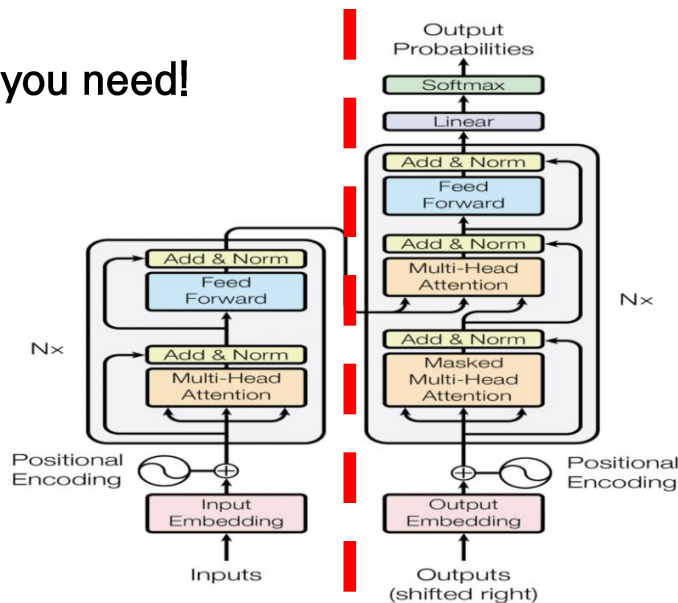


# Large Language Model #1

## 언어모델이란?



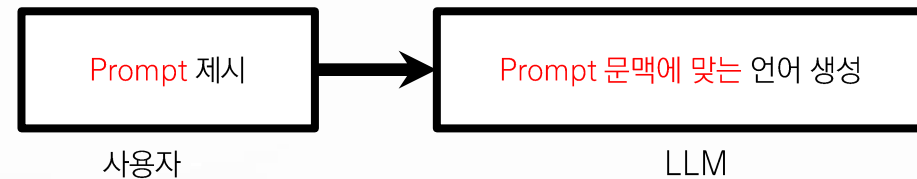
Attention is all you need!



Auto Encoding

Auto Regressive

## Large Language Model



# Large Language Model #2

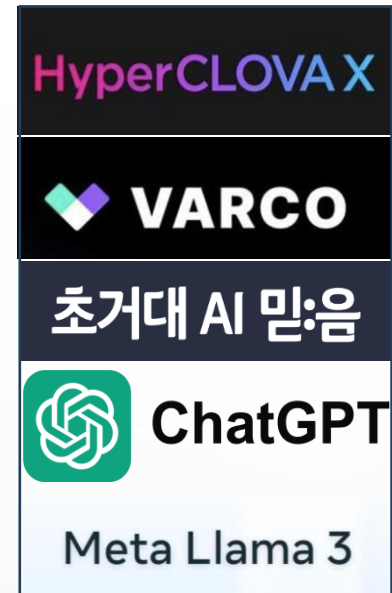
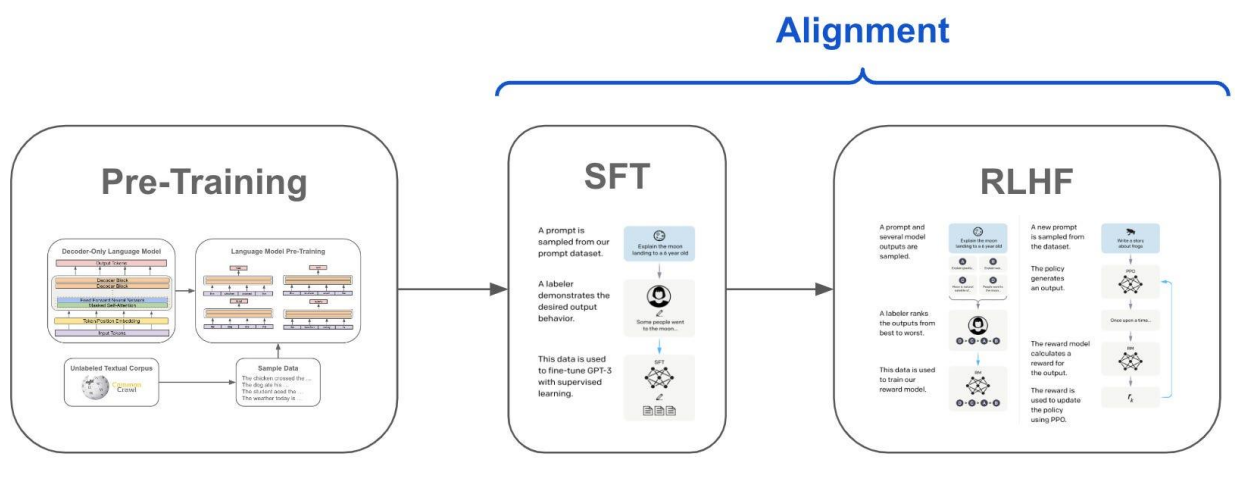
## 일반적인 LLM 개발 절차

- Pre-training (PT):

막대한 양의 코퍼스로 언어능력을 내재화, 가장 비용적으로 비싼 단계, 생성 능력 자체는 뛰어나지 않고 추가 학습이 필요

- Alignment Tuning (AT)

Supervised Finetuning (SFT), Reinforcement Learning from Human Feedback (RLHF) 포함하는 단계  
 사람에게 align되도록 학습하여 사용자가 원하는 방향으로 생성을 하도록 학습, PT에 비해 비용 저렴



대표 모델들

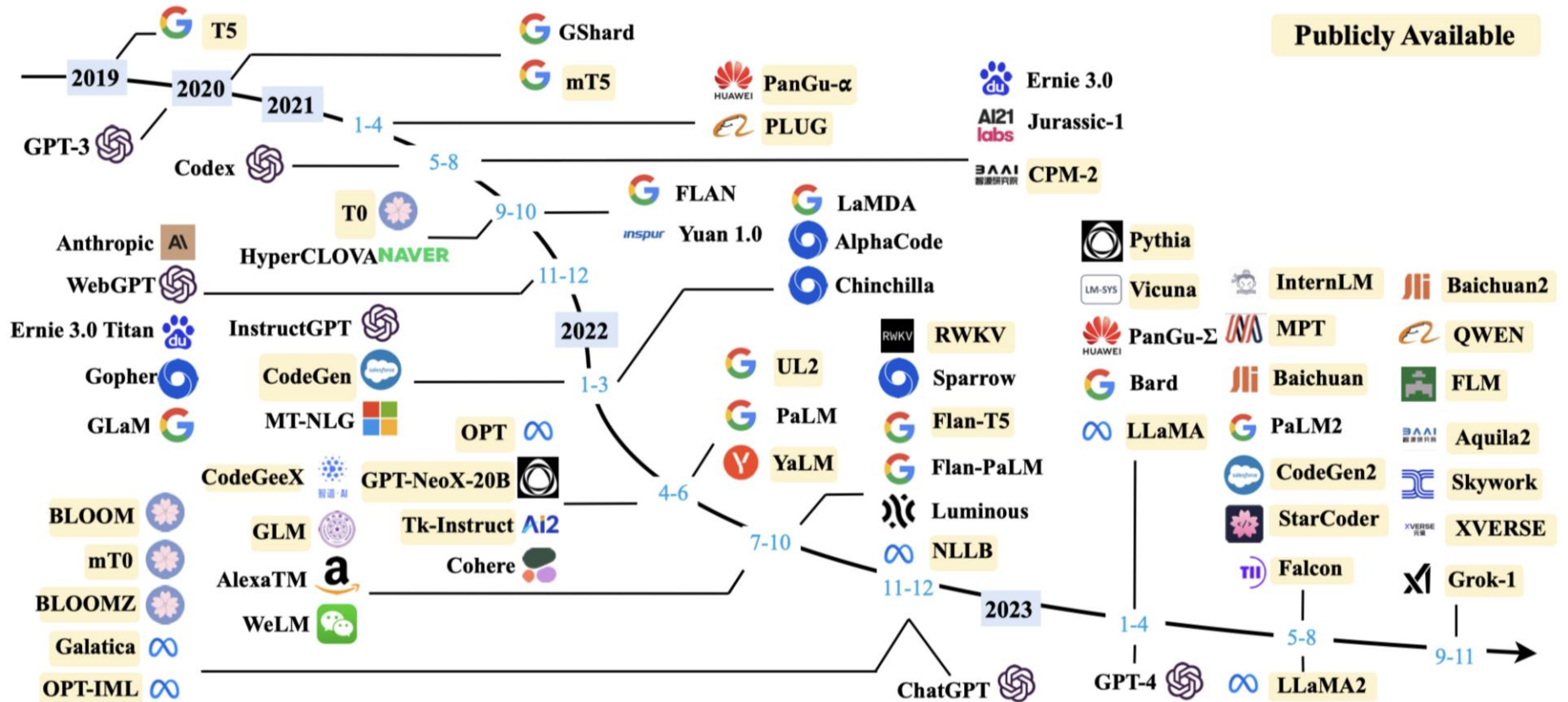


## LLM의 대표적인 응용분야



# Large Language Model #3

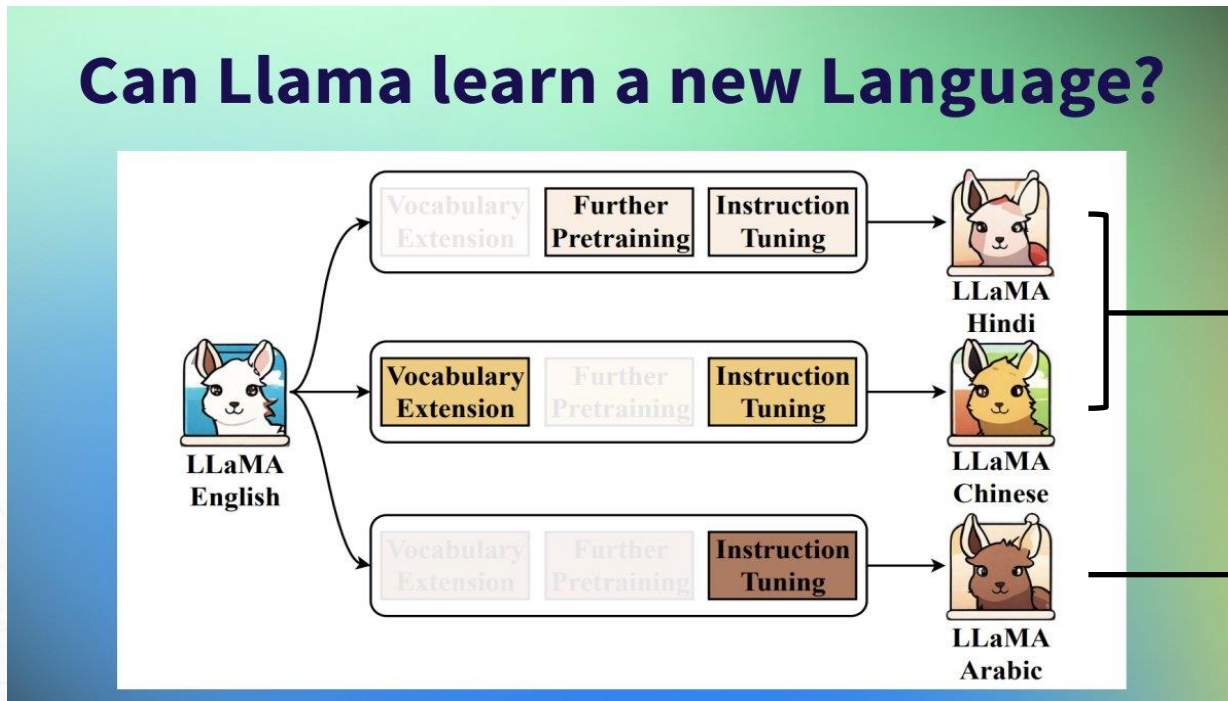
## 다양한 LLM들



# Open AI와 경쟁할 수 있는 LLM을 국내에서 개발가능한가?

## 한국어 특화 sLLM 개발 전략

- PT 모델의 성능을 활용하면서 한국어 능력을 최대화하는 다양한 방법론



한국어 특화 LLM 개발 전략 (2)

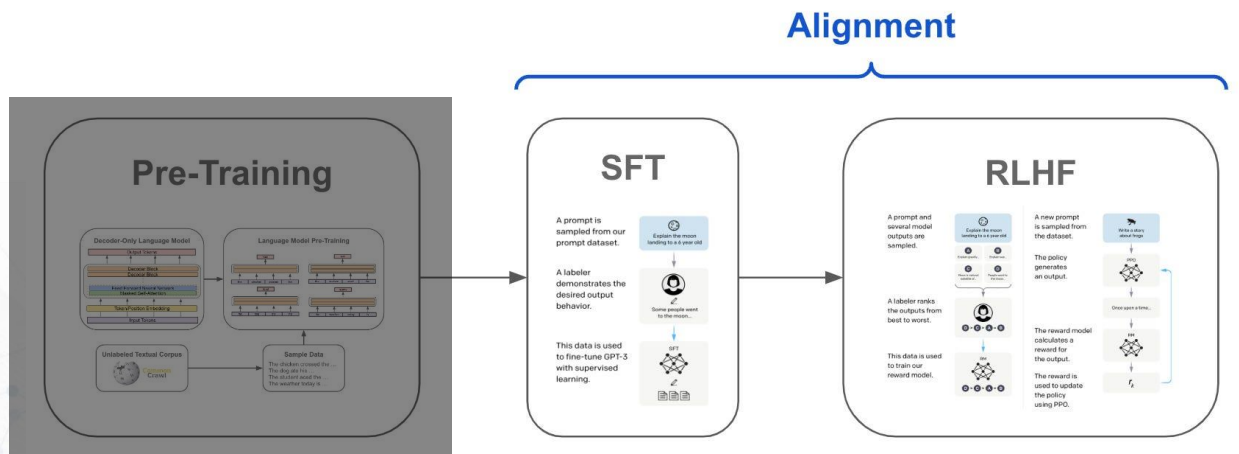
한국어 특화 LLM 개발 전략 (1)

LLaMA Beyond English: An Empirical Study on Language Capability Transfer, Zhao et al., 2024, Arxiv

# Open AI와 경쟁할 수 있는 LLM을 국내에서 개발가능한가?

## 한국어 특화 sLLM 개발 전략

- Multi-lingual 혹은 Korean PT모델을 활용한 Alignment Tuning
  - ✓ 비용적으로 비싼 PT단계 없이 한국어 언어 능력을 가지고 있는 공개된 PT 모델을 활용
  - ✓ Multi-lingual 혹은 Korean PT 모델을 활용해서 한국어 데이터에 Alignment Tuning
  - ✓ High quality의 답변으로 데이터가 구성, 모델이 실제로 사용자에게 주었으면 하는 데이터를 생성하는 것을 학습



## 한국어 특화 LLM 개발 전략



대표 모델

# 01



## KULLM 소개

We provide the best way to your business.

### 개요

- 개발 : 고려대학교 NLP&AI 연구원
- 연구 배경: Leaderboard 순위 및 벤치마크 점수가 '말을 잘 하는' 것을 보장하지 못함
- OpenKo-LLM Leaderboard 순위 != 좋은 모델 순위
- 리더보드용 모델을 제작, 두 평가항목에서 1위 (2024년 2월 19일 기준, 등록된 모델 약 900개)
- 하지만 벤치마크형 리더보드가 모델을 잘 평가하지 못한다는 것이 주류 의견
- 라이선스: 아파치 2.0
- Base Model: upstage/SOLAR-10.7B-v1.0

### Update Logs

- 2024.04.08: [구름3\(KULLM3\) 양자화 모델\(awq-4bit\) 공개](#)
- 2024.04.03: [구름3\(KULLM3\) 공개](#)
- 2023.06.23: [한국어 대화 평가 결과 공개](#)
- 2023.06.08: [Polyglot-ko 5.8B 기반 KULLM-Polyglot-5.8B-v2 fp16 모델 공개](#)
- 2023.06.01: [구름\(KULLM\) 데이터셋 v2 HuggingFace Datasets 공개](#)
- 2023.05.31: [Polyglot-ko 12.8B 기반 KULLM-Polyglot-12.8B-v2 fp16 모델 공개](#)
- 2023.05.30: [Polyglot-ko 12.8B 기반 KULLM-Polyglot-12.8B fp16 모델 공개](#)

### LeaderBoard 1위

The screenshot shows the Open Ko-LLM LeaderBoard interface. At the top, it says "Open Ko-LLM LeaderBoard". Below that, there's a search bar and various filters for model types, precision, and model sizes. A table of models is displayed, with the top two models highlighted in red boxes:

T	Model	Average	Ko-ARC	Ko-HoLaSwag	Ko-MRLU	Ko-TruthfulQA	Ko-CM
1	<a href="#">heorytal/ku-in-solar-5</a>	55.53	70.14	43.54	45.9	71.46	46.64
2	<a href="#">heorytal/ku-in-solar-5</a>	44.31	59.81	32.4	28.49	68.04	48.02

At the bottom, it says "Hosted by NIA, Upstage, Powered by KCloud, Sponsored by..."

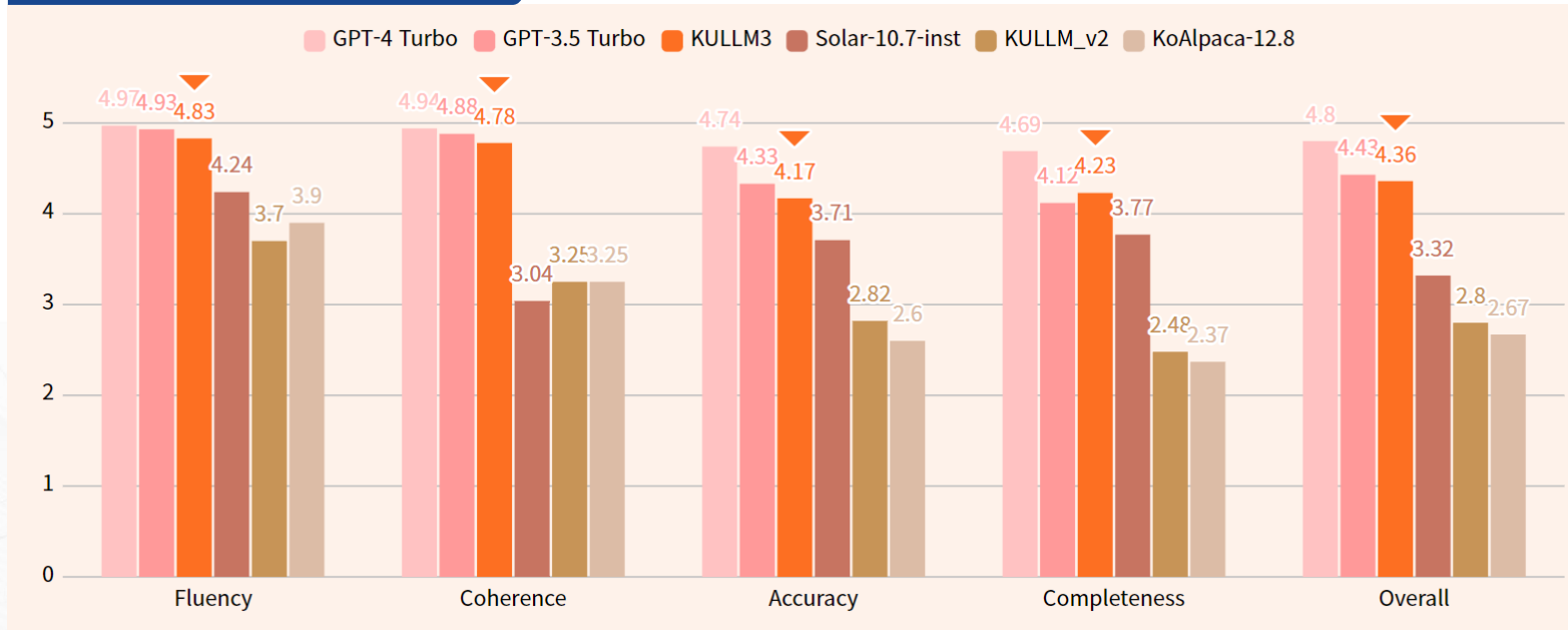
01



# KULLM 소개

We provide the best way to your business.

## KULLM3 대화 성능 평가 결과



- 평가 모델은 GPT-4-Turbo(gpt-4-0125-preview)를 사용하였고, 평가 데이터셋은 [yizhongw/self-instruct](https://github.com/yizhongw/self-instruct)의 휴먼 평가 데이터셋인 user\_oriented\_instructions.jsonl을 deepl로 번역한 데이터셋을 사용함.
- 주어진 prompt 데이터에 대해 모델이 응답을 생성하고, 그 응답을 OpenAI API를 사용하여 평가하는 방식입니다.
- 해당 평가 결과는 [repo](#)에서 재현이 가능합니다.

# 01



## KULLM 특징

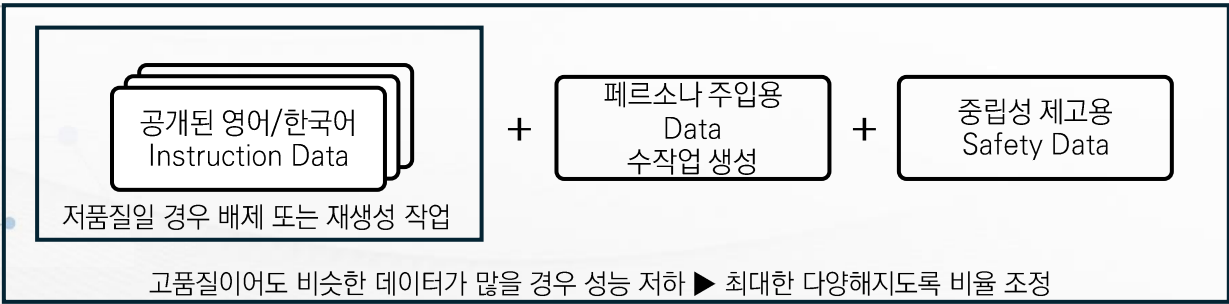
We provide the best way to your business.

### '말'을 잘하는 모델을 만들기 위한 데이터 큐레이션

- 공개된 영어, 한국어 Instruction tuning 데이터셋 활용
- 저품질이라고 판단될 경우 OpenAI API를 활용하여 재생성
- 영어/한국어 데이터셋 병용으로 성능 향상
- 데이터셋의 출처, 범위(domain), 용도에 따라 모델의 품질이 달라짐
- 예를 들어, GPT-4를 이용하여 생성한 고품질 데이터라도 패턴이 비슷한 데이터가 여러 개 들어갈 경우 모델 성능 유의미하게 저하됨
- '나는 고려대학교 NLP&AI 연구실에서 만든 KULLM이다' 라는 페르소나를 주입하기 위한 데이터 수작업으로 생성 및 큐레이션
- 모델 생성 시 중립성을 제고하기 위한 Safety Data 큐레이션
- 하이퍼파라미터 서치를 진행, 동일 데이터에서 더 좋은 모델 훈련

KULLM3 훈련 데이터 =

(더 많이 확보할 경우 성능 더 좋아질 수 있음)







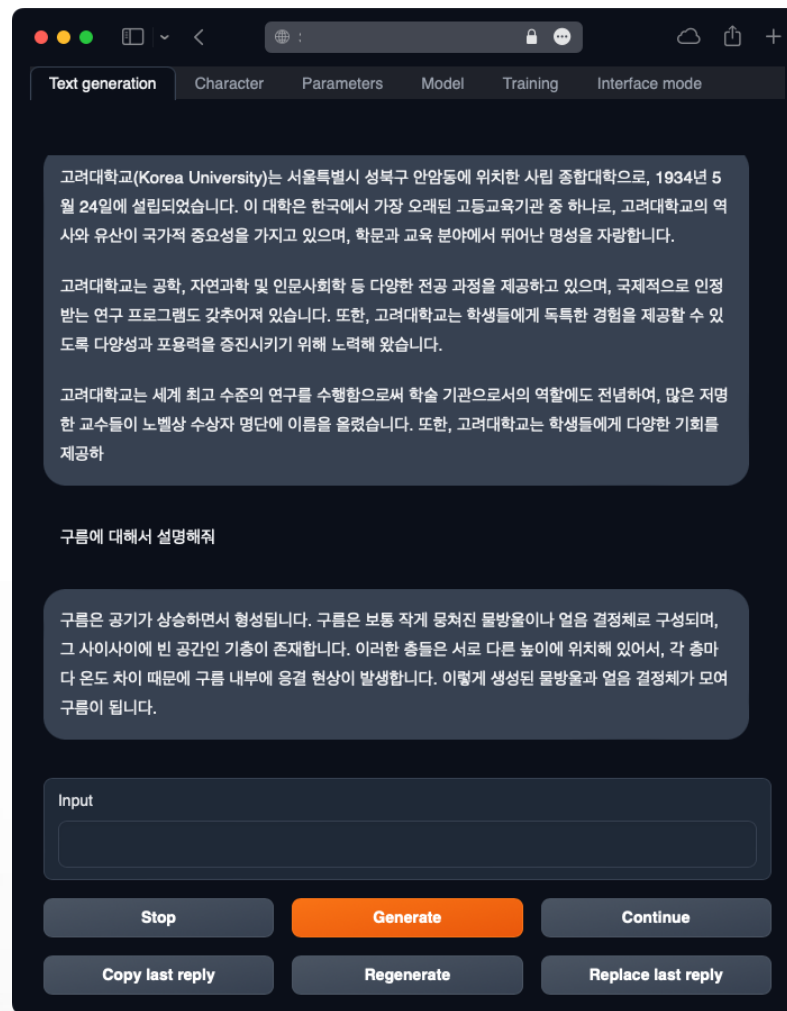
### Update Logs

- 2023.06.23: [한국어 대화 평가 결과 공개](#)
- 2023.06.08: [Polyglot-ko 5.8B 기반 KULLM-Polyglot-5.8B-v2 fp16 모델 공개](#)
- 2023.06.01: [구름\(KULLM\) 데이터셋 v2 HuggingFace Datasets 공개](#)
- 2023.05.31: [Polyglot-ko 12.8B 기반 KULLM-Polyglot-12.8B-v2 fp16 모델 공개](#)
- 2023.05.30: [Polyglot-ko 12.8B 기반 KULLM-Polyglot-12.8B fp16 모델 공개](#)

---

### ☁ KULLM (구름): Korea University Large Language Model

KULLM(구름)은 고려대학교 [NLP & AI 연구실](#)과 [HIAI 연구소](#)가 개발한 한국어 Large Language Model (LLM)입니다. 구름 프로젝트는 한국어 모델 뿐만 아니라, 데이터 셋까지 전면 공개하여 한국어 LLM 생태계에 기여하고자 하였습니다.



HIAI 연구소&NLP&AI 연구실은 2023년 국내 대학 최초 한국어 특화 대규모언어모델 “구름(KuLLM)” 공개

## Human-Inspired AI 연구소 기술이전 상용화 실적

IT·전자

### 엔씨소프트, 고려대 연구팀과 'AI 대화 데이터셋' 글로벌 공개

2022-04-14 11:24:56



가




[빅데이터뉴스 심준보 기자] (주)엔씨소프트(대표 김택진, 이하 엔씨(NC))가 고려대학교 임희석 교수 연구팀과 AI 대화 데이터 'FoCus Dataset(For Customized conversation dataset)'을 공개했다고 14일 밝혔다.

NC-고려대 임희석 교수 연구팀은 세계 최초로 사용자 페르소나(Persona)와 외부 지식을 활용하는 데이터셋 구축

23. 11. 29. 오후 12:34      업스테이지-NIA, 한국어 AI 경쟁력 제고 'Open-Ko LLM 리더보드' 개설 - 전자신문

## 업스테이지-NIA, 한국어 AI 경쟁력 제고 'Open-Ko LLM 리더보드' 개설

발행일 : 2023-09-25 10:41



<사진=업스테이지 제공>

업스테이지가 한국지능정보사회진흥원(NIA)과 함께 한국어 대규모 언어모델(LLM) 리더보드 'Open Ko-LLM 리더보드'를 구축한다.

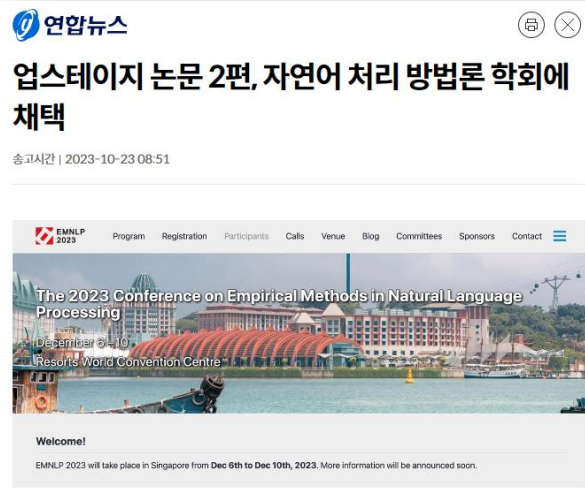
Open Ko-LLM 리더보드는 한국어 LLM 모델을 등록, 경쟁 가능한 공개 플랫폼이다. 허깅페이스가 운영하는 오픈LLM 리더보드의 기존 데이터를 단순 번역한 것이 아니라, 한국어 특성과 문화를 반영한 고품질 데이터를 자체 구축한 것이 특징이다.

또 상식 생성 능력을 살펴보는 '상식생성' 기준을 추가해 한국어 LLM 모델 성능과 다양성을 평가한다. 상식생성은 업스테이지가 임희석 고려대 교수 연구진과 협업해 구축한 데이터셋이다. 역사 왜곡, 환각 오류, 형태소 오류 등 광범위한 유형의 질문지로 구성했다. AI가 주어진 조건에서 만들어 낸 결과물이 일반 상식에 부합하는지 측정한다.

Open Ko-LLM 리더보드는 한국어 LLM 연구 수준과 경쟁력을 높이고, 데이터 양과 질을 개선할 것으로 기대된다. 안정적 운영을 위해 KT클라우드가 인프라를 지원한다.

김성훈 업스테이지 대표는 "NIA와 함께 한국어 LLM 경쟁력을 높일 Open Ko-LLM 리더보드를 개설하게 돼 기쁘다"면서 "고품질 한국어 데이터 공유는 물론, 리더보드를 통한 협업 촉진과 빠른 기술 전파로 한국 AI 생태계 발전에 기여할 것"이라고 말했다.

류태웅 기자 bigheroryu@etnews.com



연합뉴스

### 업스테이지 논문 2편, 자연어 처리 방법론 학회에 채택

송고시간 | 2023-10-23 08:51

The 2023 Conference on Empirical Methods in Natural Language Processing  
December 6-10  
Resorts World Convention Centre

Welcome!

EMNLP 2023 will take place in Singapore from Dec 6th to Dec 10th, 2023. More information will be announced soon.

자연어 처리방법론 학회 EMNLP 홈페이지  
(업스테이지 제공)

(서울=연합뉴스) 홍국기 기자 = 인공지능(AI) 스타트업 업스테이지는 자사가 작성한 논문 2편이 세계 3대 자연어 처리 학회인 '자연어 처리방법론 학회'(EMNLP)에 채택됐다고 23일 밝혔다.

논문을 한국어 관련 자연어 처리(NLP) 연구 성과로, 업스테이지의 박찬준 기술팀장(테크리드) 주도로 고려대 임희석 교수 연구팀과 협업을 통해 진행됐다고 업스테이지는 소개했다.

이 중 한국어 음성 인식 후처리기와 관련한 새 벤치마크 데이터 세트를 구축한 논문(KEBAP: Korean Error Explainable Benchmark Dataset for ASR and Post-processing)은 음성 인식 모델의 약점을 평가·식별하기 위한 새 평가 방법론을 제안한다.

한국어의 특성을 살린 새 데이터 증강 기법을 제안한 다른 논문(CHEF in the Language Kitchen: A Generative Data Augmentation Leveraging Korean Morpheme Ingredients)은 한국어의 특성에 기반한 새 데이터 증강 방법론(CHEF)을 통해 같은 재료로도 자연스러운 문장을 생성하고 생성형 언어 모델이 한국어 형태소 조합을 다양하게 변형해 데이터를 증강하는 방법론을 제안한다고 회사는 설명했다.

고려대 임희석 교수 연구팀은 업스테이지와 공동 논문 집필 그리고 협업을 통해 Open Ko LLM 리더보드 개설 및 운영

## 산학협력 연구과제 참여 협약서

□ 사업명 : 2022년 중소기업 기술애로 컨설팅 지원사업

### □ 과제 개요

- 연구과제명: 금융 빅데이터와 자연어처리 기술을 활용한 금융 도메인 특화 인공 지능 기계 번역 성능 개선 연구
- 연구책임자: 고려 대학교 Human-Inspired AI 연구소
- 연구기간: 2022.10.01. ~ 2022.12.31.
- 총연구비: [REDACTED]

### □ 참여 기업



- 기업명: ㈜웨이커
- 대표자명: 황규종
- 주소: 서울 강남구 선릉로 667, 11층
- 담당자: 황규종
- 연락처: (전화) [REDACTED]

본 연구 과제에 참여하며 아래 내용에 동의함을 확약합니다.

CES 2024 INNOVATION AWARD PRODUCT

# AI-powered Market Data Intelligence Platform

By Waiker, Inc

**Honoree**  
Artificial Intelligence

Waiker is an AI NLP powered stock market data solution that transforms unstructured stock market data into a structured database. Financial services professionals lose valuable investment opportunities for their clients as they spend >80% of their time on processing raw unstructured market data to find valuable investment insights. Waiker's data platform uses AI NLP technology to process and transform raw stock market data into a structured data base in mere seconds. Waiker's Data API solution offers traders and other market participants actionable insights, enabling them to respond swiftly to market dynamics.

본 연구소와 산학협력 연구과제를 수행한 웨이커는 CES 2024 INNOVATION AWARD PRODUCT 에 선정되는 쾌거를 이룸

<https://www.ces.tech/innovation-awards/honorees/2024/honorees/a/ai-powered-market-data-intelligence-platform.aspx>



## 산학협력 연구과제 참여 협약서

□ 사업명 : 2022년 중소기업 기술애로 컨설팅 지원사업

### □ 과제 개요

- 연구과제명: NLP AI를 활용한 콘텐츠 추천 알고리즘 기술 연구 개선
- 연구책임자: 고려대학교 Human-Inspired AI 연구소
- 연구기간: 2022.10 . 01. ~ 2022. 12 . 31 .
- 총연구비: [REDACTED]

### □ 참여 기업

- 기 업 명: ㈜오투오
- 대표자명: 안성민
- 주 소: 강원도 춘천길 강원대학길 1, 1005호
- 담 당 자: 안성민
- 연 락 처: [REDACTED]

본 연구 과제에 참여하며 아래 내용에 동의함을 확약합니다.

O2O만의 차별화된 솔루션

## 자연어처리, 대화시스템, 메타지식 생성 기술

오투오는 자체 AI 솔루션 개발과 거대언어모델(LLM) 연구를 통한 차세대 AI 서비스를 개발하며 고려대학교 Human-Inspired AI 연구소와 공동 AI 계약연구센터를 설립하여 산학기술 협력추진 및 AI 협력연구를 진행하고 있습니다.

<http://o2o.kr/page/solution.php>

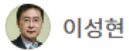


- 자서전적 기억은 i) 시간적 요소, ii) 공간적 요소, 그리고 iii) 개인이 직접 경험한 사건인지의 여부, 세 가지 요소로 이루어짐
- 주어진 주제에 대해 최대한 구체적으로 회상하여 답변하고, 이를 내부적으로 평가 및 피드백을 제공하는 과정을 통해 노년기 어르신들을 대상으로 자서전적 기억훈련을 제공하기 위함. 이를 통하여 노년기 사용자들의 기억의 구체성을 제고하는 데에 도움을 주고자 함
- 주요 적용분야는 치매노인 등 노년기 사용자 대상 케어 서비스

## 충북, 국내 최초 인공지능 육아 상담 프로그램 운영

문화일보 | 입력 2021-03-02 16:28

프린트



이성현

댓글 0 | 폰트 | 공유



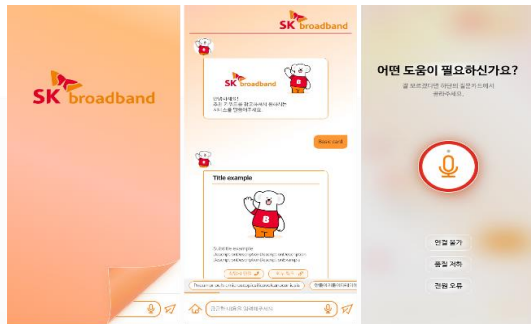
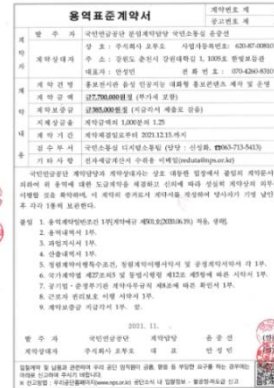
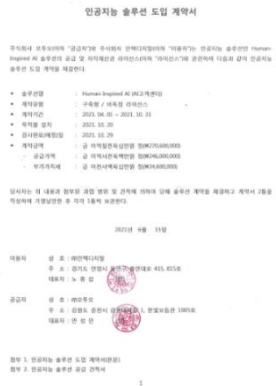
인공지능(AI) 육아상담 챗봇 '우리아이 토닥토닥' 프로그램. 충북육아종합지원센터 제공

충북에서 국내 최초로 육아에 어려움을 겪고 있는 부모를 상대로 인공지능(AI) 육아 상담 프로그램 서비스가 제공된다.

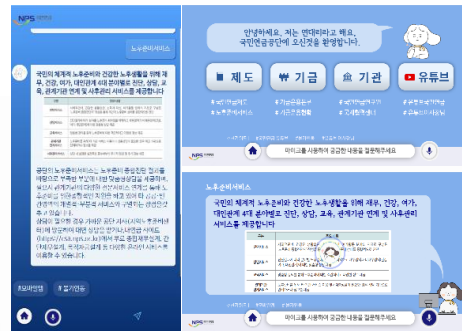
본 연구소과 기술이전을 체결한 협력 업체 (주)유스바이오글로벌은 고려대 Human-inspired AI 연구소로 부터 기술이전 받은 "육아상담 챗봇" 을 충북육아종합지원센터에 제공하여 지역 사회 및 관련업계 종사자와 부모들에게 큰 기여를 수행함



# Human-Inspired AI 연구소 기술이전 상용화 실적



SKBB 향 AI 고객센터  
인택디지털 납품



홍보전시관 음성 인공지능 대화형  
홍보콘텐츠 제작 및 운영

본 연구소과 기술이전을 체결한 협력 업체 (주)오투오는 고려대 Human-inspired AI 연구소로 부터 기술이전 기술을 활용해 SK브로드밴드와 국민연금공단 대화형 홍보콘텐츠 제작 및 운영을 계약함



# Human-Inspired AI 연구소 최근 5년간 차세대 인공지능

## 연구개발 추진 실적

구분		2019~2023년도 (최근 5년)	
연구 실적	SCI급 저널	57	
	KCI 저널	39	
	컨퍼런스 발표논문	국내	29
		국제	100
		컨퍼런스 발표논문 소개	129
논문실적 합계	284		

**SCIENTIFIC DATA**  
OPEN DATA DESCRIPTOR

**Building a PubMed knowledge graph**

Jian Xu<sup>1</sup>, Sunkyu Kim<sup>2</sup>, Min Song<sup>1</sup>, Minbyul Jeong<sup>1</sup>, Donghyeon Kim<sup>2</sup>, Jaewoo Kang<sup>3</sup>, Justin F. Rousseau<sup>4</sup>, Xin Li<sup>5</sup>, Weijia Xu<sup>6</sup>, Vette I. Torvik<sup>1</sup>, Yi Bu<sup>7</sup>, Chongyan Chen<sup>8</sup>, Islam Akel Ebeid<sup>9</sup>, Da Feng Liu<sup>10</sup> & Ying Ding<sup>11</sup>

PubMed<sup>1</sup> is an essential resource for the medical domain, but useful concepts are either difficult to extract or are ambiguous, which has significantly hindered knowledge discovery. To address this issue, we constructed a PubMed knowledge graph (PKG) by extracting bio-entities from 29 million PubMed abstracts, disambiguating author names, integrating funding data through the National Institutes of Health (NIH) EXPORTER, collecting affiliation history and educational background of authors from ORCID<sup>2</sup>, and identifying fine-grained affiliation data from MapAffili. Through the integration of these credible multi-source data, we could create connections among the bio-entities, authors, articles, affiliations, and funding. Data validation revealed that the BioBERT deep learning method of bio-entity extraction significantly outperformed the state-of-the-art models based on the F1 score (by 0.51%), with the author name disambiguation (AND) achieving an F1 score of 98.09%. PKG can trigger broader innovations, not only enabling us to measure scholarly impact, knowledge usage, and knowledge transfer, but also assisting us in profiling authors and organizations based on their connections with bio-entities.

**Expert Systems with Applications**  
Volume 233, 15 December 2023, 120962

**Doubts on the reliability of parallel corpus filtering**

Hyonseok Moon<sup>a</sup>, Chanjun Park<sup>b</sup>, Seonmin Koo<sup>c</sup>, Jungsook Lee<sup>d</sup>, Seungjun Lee<sup>e</sup>, Jaehyung Seo<sup>f</sup>, Suyang Eo<sup>g</sup>, Younna Jang<sup>h</sup>, Hyeonjoong Kim<sup>i</sup>, Hyeon-gu Lee<sup>j</sup>, Heuseok Lim<sup>k</sup>

**Abstract**  
Parallel corpus filtering (PCF) aims to filter out low-quality data residing in parallel corpora. Recently, deep learning-based methods have been employed to assess the quality of sentence pairs in a parallel corpus, along with rule-based filtering that filters out noisy data depending on the pre-defined error types. Despite their utilization, to the best of our knowledge, a comprehensive investigation into the practical applicability and interpretability of PCF techniques remains unexplored. In this study, we raise two doubts on deep learning-based PCF: (i) Can deep learning-based PCF extract high-quality data? and (ii) Are scoring functions of PCF reliable? To answer these questions, we conduct comparative experiments on various PCF techniques with four datasets on two language pairs, English-Korean, and English-Japanese. Through the experiments, we demonstrate that the performance of the deep learning-based PCF highly depends on the targeting parallel corpus, and shows fluctuating adaptability depending on their characteristics. In particular, we figure out that high-scored sentences derived by the PCF technique do not necessarily guarantee high-quality results, rather it shows unintended preference.

**Knowledge-Based Systems**  
Volume 256, 28 November 2022, 109861

**PU-GEN: Enhancing generative commonsense reasoning for language models with human-centered knowledge**

Jaehyung Seo<sup>a</sup>, Dongguk Oh<sup>b</sup>, Suyong Eo<sup>c</sup>, Chanjun Park<sup>d</sup>, Kisu Yang<sup>e</sup>, Hyonseok Moon<sup>f</sup>, Kinam Park<sup>g</sup>, Heuseok Lim<sup>h</sup>

**Highlights**

- Reinterpretation of generative commonsense reasoning via Wittgenstein's picture and use theory.
- We extend the limited knowledge in the CommonGen by leveraging scene and relational knowledge.
- PU-GEN is a knowledge-enhanced language model for developing generative commonsense reasoning.
- PU-GEN can be implemented in another commonsense downstream task: Commonsense Question Answering.

**ACL Anthology**

**Towards Diverse and Effective Question-Answer Pair Generation from Children Storybooks**

Suyong Eo, Hyonseok Moon, Insung Kim, Yuna Hur, Jeongwook Kim, SongJun Lee, Changwoo Chus, Sungsoo Park, Heuseok Lim

**Abstract**  
Recent advances in QA pair generation (QA2) have raised interest in applying this technique to the educational field. However, the diversity of QA types remains a challenge despite its contribution to comprehensive learning and assessment of children. In this paper, we propose a QA2 framework that enhances QA type diversity by producing different interrogative sentences and explicit/implicit answers. Our framework comprises a Q2S-based answer generator, an IntraQA generator, and a relevancy-aware reranker. The two generators aim to expand the number of candidates while covering various types. The reranker trained on the in-content negative samples clarifies the top-N outputs based on the ranking score. Extensive evaluations and detailed analyses demonstrate that our approach outperforms previous state-of-the-art results by significant margins, achieving improved diversity and quality. Our task-oriented processes are consistent with real-world demand, which highlights our system's high applicability.

**ACL Anthology**

**You Truly Understand What I Need: Intellectual and Friendly Dialog Agents grounding Persona and Knowledge**

Jungwoo Lim, Myunghoon Kang, Yuna Hur, Seung Won Jeong, Insung Kim, Younna Jang, Dongbuk Lee, Hyesung Ji, Donghoon Shin, Seungyong Kim, Heuseok Lim

**Abstract**  
To build a conversational agent that interacts fluently with humans, previous studies blend knowledge or personal profile into the pre-trained language model. However, the model that considers knowledge and persona at the same time is still limited, leading to hallucination and a passive way of using persona. We propose an effective dialog agent that grounds external knowledge and persona simultaneously. The agent selects the proper knowledge and persona to use for generating the answer with our candidate scoring implemented with a policy procedure. Then, our model generates the sentence with lesser hallucination and more engagingly utilizing retrieval augmented memory with knowledge persona enhanced query. We conduct experiments to test the persona state-of-the-art performance in grounding and generation tasks on the automatic metrics. Moreover, we validate the answers from the model regarding hallucination and engagingness through human evaluation and qualitative results. We show our retriever's effectiveness in extracting relevant documents compared to the other previous retrieval, along with the comparison of multiple candidate ranking methods. Code is available at <https://github.com/limw/grounding-persona>.

**The Thirty-Sixth AAAI Conference on Artificial Intelligence (AAAI-22)**

**Call for Customized Conversation: Customized Conversation Grounding Persona and Knowledge**

Younna Jang<sup>1</sup>, Jungwoo Lim<sup>1</sup>, Yuna Hur<sup>1</sup>, Dongguk Oh<sup>1</sup>, Sulyone Son<sup>1</sup>, Younna Lee<sup>1</sup>, Donghoo Shin<sup>1</sup>, Seungyong Kim<sup>1</sup>, and Heuseok Lim<sup>1</sup>

**Abstract**  
Humans usually have conversations by making use of prior knowledge about a topic and background information of the person whom they are talking to. However, existing conversational agents lack comprehensive information, and thus they have a limitation in generating the sentences when the knowledge and persona are fused properly. To address this issue, we introduced a call for Customized Conversation (C2C) in the persona state-of-the-art performance evaluation. We evaluate the abilities to make automatic and customized sentences of pre-trained language models, we utilize BART and GPT-2 as well as transformer-based models. We assess their generation abilities with automatic scores and conduct human evaluations for qualitative results. We examine whether the model reflects adequate persona and knowledge with our proposed two tasks, persona grounding (PG) and knowledge grounding (KG). Moreover, we show that the sentences of our data are constructed with the proper knowledge and persona through grounding quality assessment.

**Introduction**  
A person who is asked by a vegetarian to suggest a restaurant in New York City would not usually recommend Wolfgang's Steakhouse. When people give information to others, they consider the background of the person whom they are talking to. Following this manner of human conversation, a conversational agent's ability to have a conversation with customized answers from prior knowledge and user's personal information is crucial for satisfying the users. For example, as exemplified in Figure 1, the answer that considers both the user's persona and knowledge is much more attractive as well as informative.

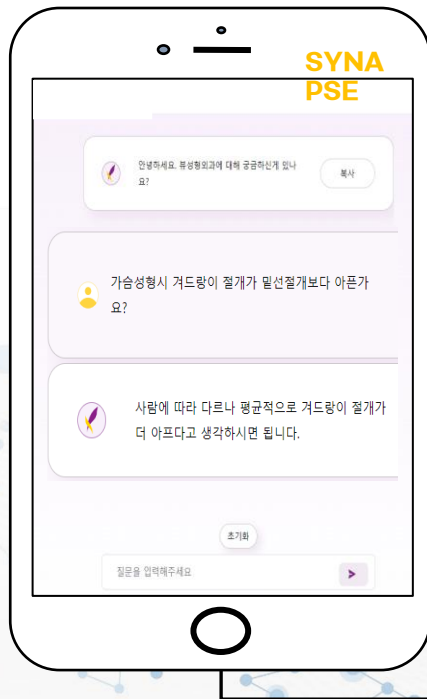
Research for human-machine dialog has achieved significant success recently, owing to the advance of diverse dialog datasets (Adiwilandana et al. 2020; Zhang et al. 2019b; Shuster et al. 2019; Li et al. 2017; Lowe et al. 2015) and pre-trained language models (Radford et al. 2019; Clark et al. 2020; Brown et al. 2020).

Meanwhile, to build a dialog agent that generates more knowledgeable answers, datasets with the informative answers have been released (Drouot et al. 2018; Zhou, Pribbenow, and Black 2018). In these datasets, the dialog agents learn to retrieve the required knowledge from the documents. However, these datasets do not consider the user's persona,

Figure 1: Objective of FoCUS dataset. In contrast to the general answer, which only gives basic information, the machine's answer of FoCUS dataset is more knowledgeable and customized, reflecting both knowledge and persona.

# SYNAPSE 란?

“고객에게 **정확한 정보 전달**과  
 빠르고 **정확한 문제해결**을 자연스러운 대화로 가능한  
 Human-like Bot 서비스”



웹과 카톡을 통한 UI 제공



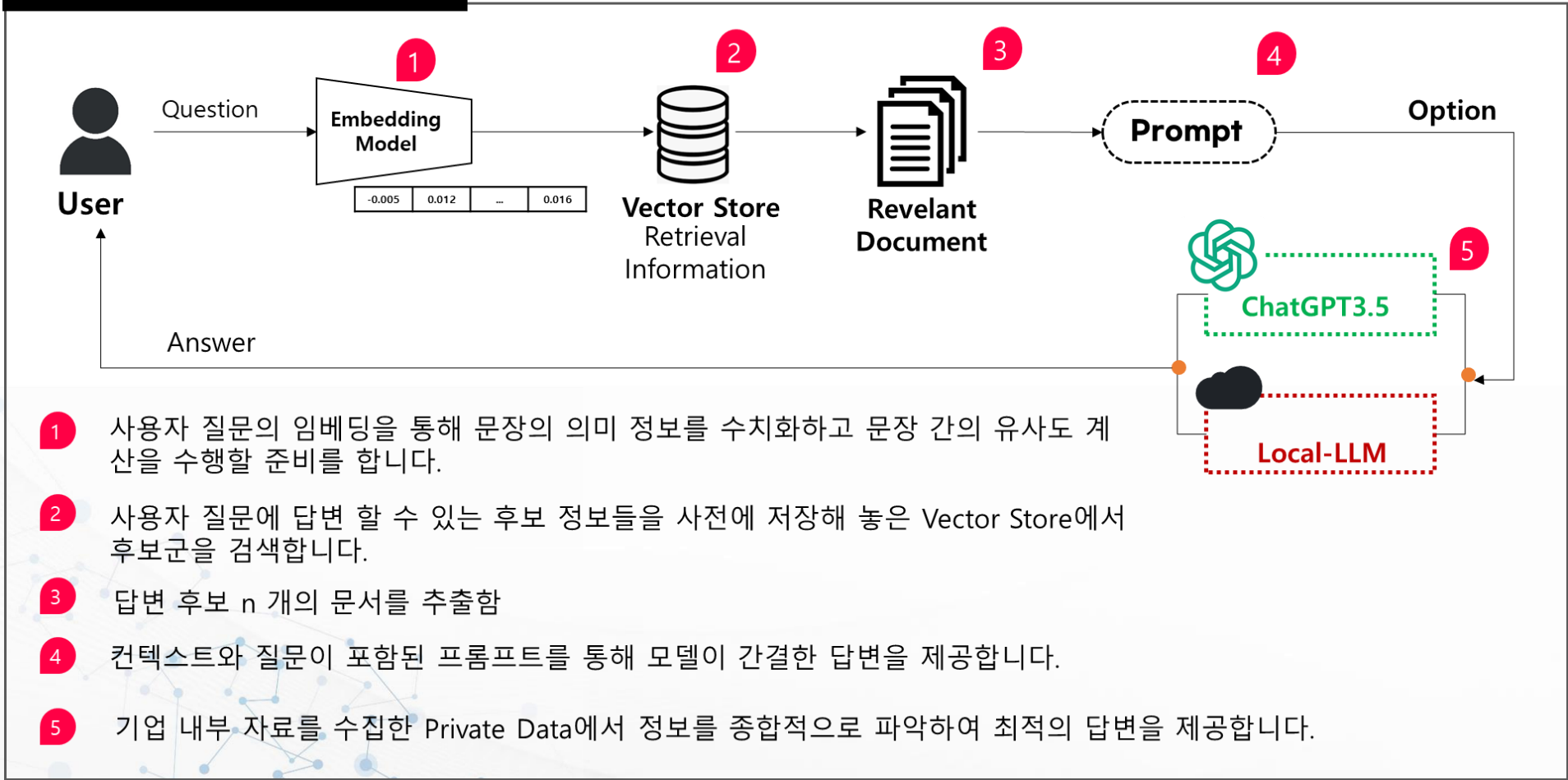
예약시스템 연동



심층 상담신청



## LLM 기반 답변 프로세스



- 1** 사용자 질문의 임베딩을 통해 문장의 의미 정보를 수치화하고 문장 간의 유사도 계산을 수행할 준비를 합니다.
- 2** 사용자 질문에 답변 할 수 있는 후보 정보들을 사전에 저장해 놓은 Vector Store에서 후보군을 검색합니다.
- 3** 답변 후보 n 개의 문서를 추출함
- 4** 컨텍스트와 질문이 포함된 프롬프트를 통해 모델이 간결한 답변을 제공합니다.
- 5** 기업 내부 자료를 수집한 Private Data에서 정보를 종합적으로 파악하여 최적의 답변을 제공합니다.



## 영상 음성 전사

- 다양한 언어의 YouTube 자막을 자동으로 생성하고 해석 가능
- 음성 인식 기술을 이용하여 영상의 음성을 정확하게 전사
- 사용자가 추가적인 편집 없이도 프로페셔널한 수준의 자막 생성 가능

## 영상 자막 번역

- AI 기반의 번역 기술을 사용하여 정확하고 자연스러운 번역 제공
- 자막에 특화된 번역 기능을 제공해 높은 수준의 번역 결과 제시



Turn on the party light  
and play cool music at 8  
pm every other Tuesday

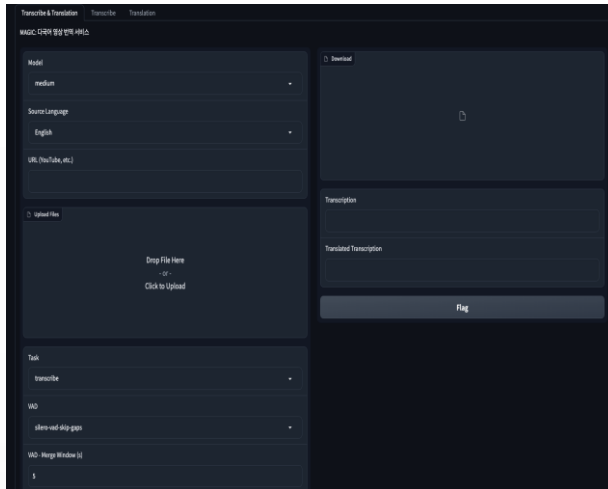
English



Korean

“Turn on the party light  
and play cool music at  
8 pm every other Tuesday”

격주 화요일 오후 8시에 파티 조명을  
켜고 멋진 음악을 재생하세요.



**기존 상용화 번역의  
어려움 극복**  
한국어 전사, 다국어 번역  
특화



Human Inspired AI Research  
KOREA UNIVERSITY