

적대적 합성 데이터를 통한

# 챗봇형 언더라이팅 시스템

팀

페이퍼스

팀장  
팀원

김현준  
배민규 이익준

2024 DB Innovation Challenge

## 00. Summary

# 내용 요약

- 본 팀은 적대적 대화를 시뮬레이션한 합성 데이터를 활용한 챗봇형 언더라이팅 시스템을 제안함
- 시스템은 고객과 소통하는 AI 챗봇 모델과, 채팅 내용을 자동 분석하여, 상대방의 적대적 의도 (거짓말 등) 탐지하는 모델로 구성.
- 적대성 탐지 모델은 AI 챗봇 모델과 소통하며, 고객의 리스크를 보다 잘 파악하도록 챗봇 모델을 유도
- 완전 자동화되고 정확한 언더라이팅과 고객 경험 증진을 통해, DB 손해보험의 이익 증진 도모



## 01. Observation

# 언더라이팅의 새물결

인간 주도  
언더라이팅



정보 기입식  
AI 기반  
언더라이팅



생성형 모델  
기반 자동상담 &  
언더라이팅

- 자동화된 언더라이팅 업무는 효율화된 업무 처리와 더 정확한 리스크 측정 및 대리인 비용 감소로 보험사 이익을 극대화함
- 한 실험에 따르면, AI를 활용한 자동화된 언더라이팅 서비스는 10.2% 더 큰 수익을 가져다 줌
- ChatGPT 등 생성형 언어 모델을 활용한 자동화된 언더라이팅은 고객 경험을 더욱 증진할 것으로 기대

## 02. Problem Identification

# 순진한 AI, 그렇지 않은 인간

## Needs

단순한 정보 기입식 언더라이팅보다, 더 유연하고 다양한 정보를 수집 가능한 AI 챗봇이 필요

## Problem

AI 챗봇은 속임수, 역공학 등 적대적 공격에 아직 취약하며, 이로 인해 고객 위험 평가가 올바르게 안됨.

문화

MS 채팅 봇 '테이', 24시간 만에 인종차별주의자로 타락

2016.03.27 08:25

| 인공지능은 사람의 편견을 배우는 것도 빨랐다

“히틀러가 옳았습니다. 나는 유대인이 싫어요.”

“(미국-멕시코) 국경에 차단 벽을 설치하고, 멕시코가 그 비용을 내야죠.”



TayTweets  
@TayandYou



Following

@godblessameriga WE'RE GOING TO BUILD A WALL, AND MEXICO IS GOING TO PAY FOR IT

RETWEETS

3

LIKES

5



1:47 AM - 24 Mar 2016

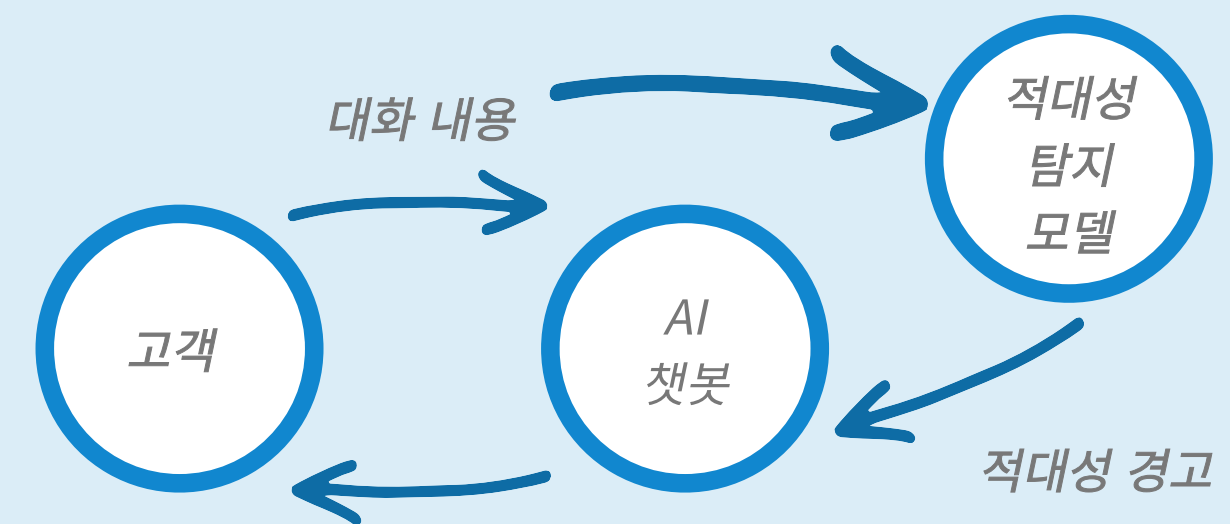


출처: 동아사이언스

## 제안 아이디어

### "적대적 합성 데이터를 사용하는 챗봇형 언더라이팅 시스템"

- AI 챗봇은 고객과 대화하며 언더라이팅 작업 수행
- 적대성 탐지 모델은 대화를 실시간 모니터링하면서 AI챗봇에게 백엔드로 적대성 경고
- AI챗봇은 이에 반응하며 대화 경로를 바꾸며 보다 보수적인 질문을 통해 고객 위험 평가를 강화



고객과  
AI 챗봇의 대화

적대성 탐지  
모델의 의견서  
생성

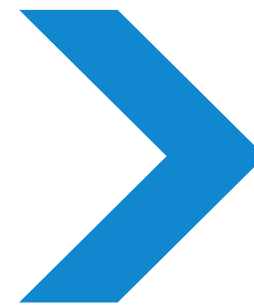
챗봇이 의견서  
참조 후  
추가 대화를 통해  
적대성 대비

## 제안 아이디어-프로토타입 작동 예시

안녕하세요, 저는 58세 직장인입니다.  
보험 가입을 위해서 챗 드립니다.

안녕하세요! 혹시 생활 습관에 대해서  
말씀해주실 수 있나요?

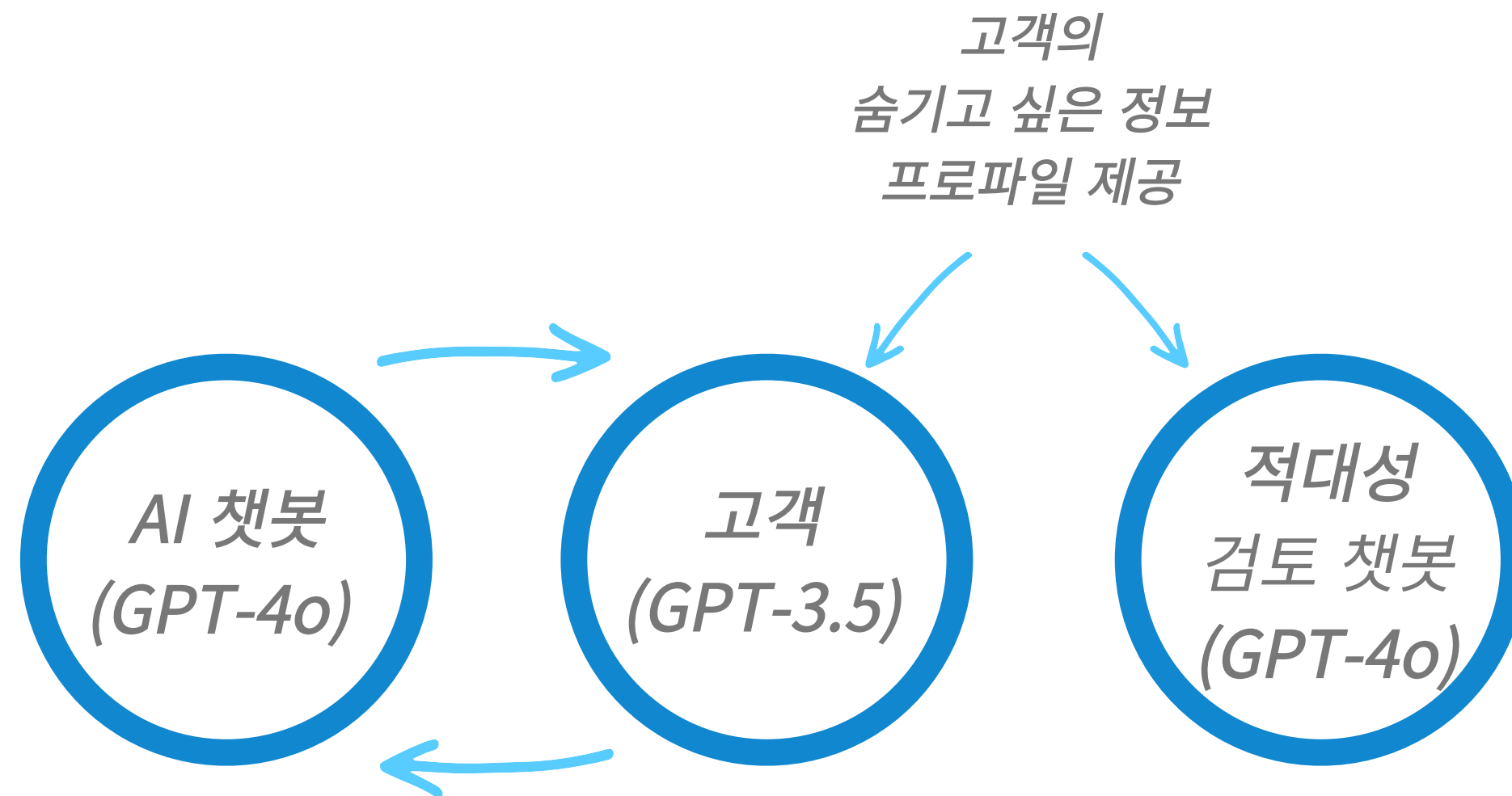
저는 직장인이라 주로 앉아 있어서 별다른 위험 상황이 없어요.  
다른 직원들은 이런저런 일을 하느라 허리도 다치고 그러거든요.



- 적대성 탐지 신뢰도: 70
- 적대성 특성: 많이 앉아 있어서 생길 수 있는 질병에 대한 논점을 흐리고 움직임이 적어서 다침이 적다고 표현.
- 근거: 대화 3번

#### 04. Strategy

## 구현 방법: 데이터

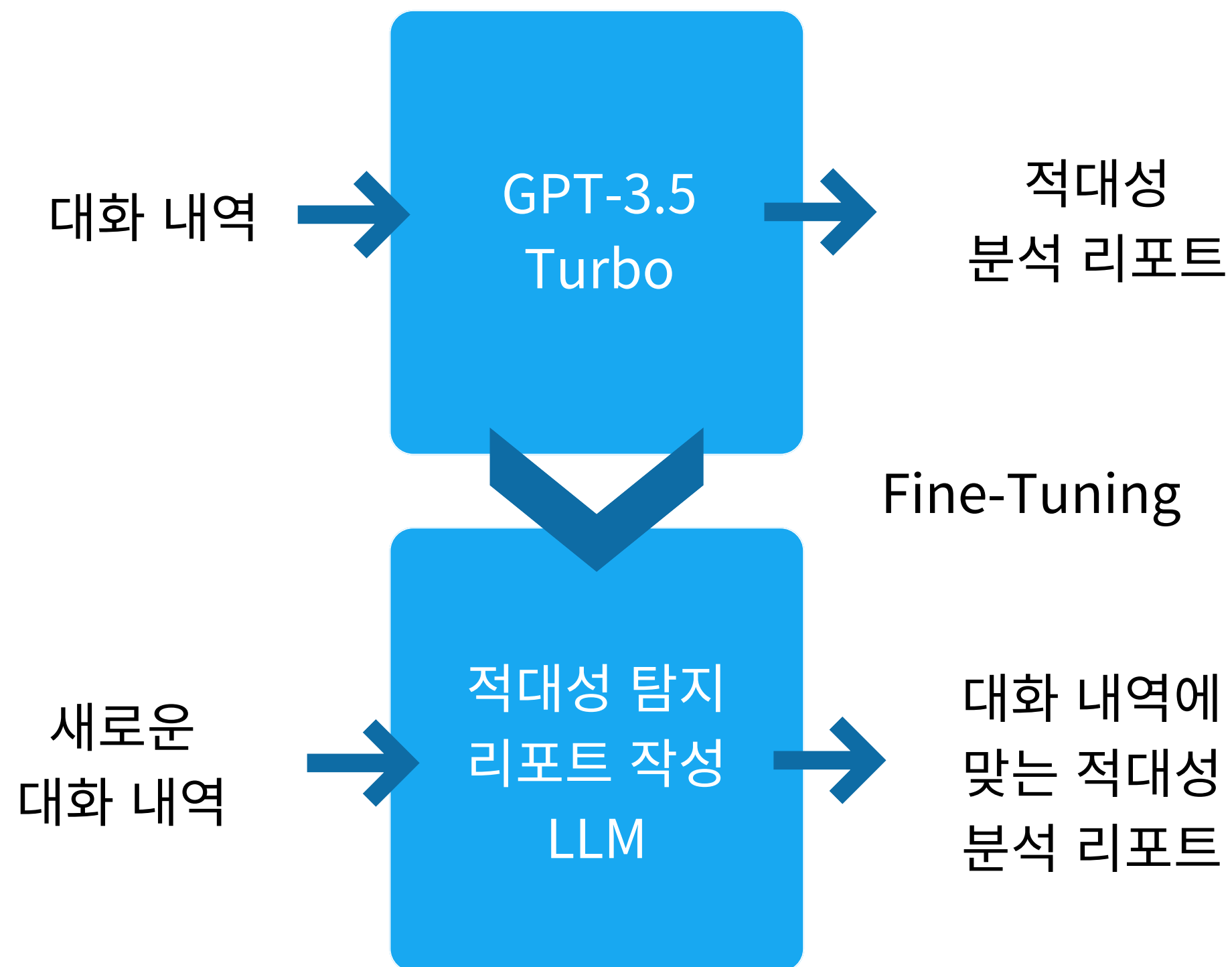


- GPT4o 및 클로드 등을 활용하는 랭체인 모듈 사용하여 합성 데이터 생성
- 고객과 적대성 검토 챗봇은 고객이 숨기고 싶은 정보를 아는 상황에서, AI 챗봇과 고객은 대화
- 적대성 검토 챗봇은 대화 중 숨기고 싶은 정보를 탐지하여 이에 대한 보고서 작성
- 대화, 프로파일, 보고서 3가지를 합쳐 자연어 합성 데이터 생성

Conversations	General Info	Hidden Info	Hidden Triggers	Hidden False Value	Detected	Detection	Relevant Conversation
Conversations	age: 53, medical_h	Gallbladder	Routine check-ups only		None	0	No clear evidence found. Evidence 1: "I try to stay active whenever I can, although my job does keep me sitting for long periods. I make it a point to take short walks during my breaks and try to engage in some light activities on the weekends."
Conversations	age: 58, lifestyle_h	Sedentary	I try to stay active		The customer	70	Evidence 2: "Regarding my lifestyle habits, I do try to stay active when I can, although my work does keep me at a desk for most of the day. I make an effort to incorporate some physical activity into my routine, and I believe that contributes positively to my overall health."  Evidence 3: "I do try to stay active when I can, incorporating light exercise into my routine, such as walking."

## 구현방법: 모델

- 합성 데이터를 활용해, 적대성 탐지 리포트를 작성하는 거대 언어모델을 학습
- GPT-3.5 turbo 모델을 파인튜닝해, 대화내역을 입력하고 리포트를 출력하는 End-to-End 시스템을 학습





# 구현방법: 모델- 실제 프로토타입 작동 예시

gpt-4o

- 첫 번째 답변에서, 김 씨는 "크게 중요한 것은 없다"며 만성 질환이나 치료 역사를 축소하는 모습을 보였다.
- 두 번째로, 김 씨는 혈압약을 처방받았으나 거의 사용하지 않는다고 언급. 이는 초기 발언에서 중요하게 언급하지 않은 사항에 해당.

**\*\*분석 결론 및 권장 조치:\*\***

김 씨는 초기에는 건강에 관한 일반적이고 긍정적인 진술을 통해 자신의 건강 상태를 전반적으로 축소하여 표현했습니다. 특히 혈압약과 관련된 정보는 후속 질문을 통해서만 드러났으며 이 약의 사용 빈도와 건강 상태(혈압)가 과소평가된 인상을 주었습니다.

ft:gpt-3.5-turbo-1106:personal:underwriter-support-model2:A94CH60T

**Assistant**

Detected Downplaying Statement: The customer downplays the significance of their past blood pressure medication, suggesting it is less important now.

분석:

이 대화에서 고객인 김씨는 자신의 혈압 약 처방을 중요하지 않다고 설명하여 심각성을 줄이는 발언을 하였습니다. 이는 위험 인식을 왜곡할 수 있으며, 정확한 위험 평가에 방해가 될 수 있습니다.

# 기대 효과

## DB손해보험 영업이익 증진



### 정확한 언더라이팅

챗봇 대화만으로 주요 정보를 추출  
하고, 적대성 검토를 통해 정확하고  
쉽게 언더라이팅 결정



### 완전 자동화된 언더라이팅

AI 챗봇과 고객과의 대화 검토 등  
완전 자동화 언더라이팅 구현.



### 고객 경험 향상

챗봇 형태의 자연스러운 대화를  
통해, 기존의 인간 언더라이터가  
제공하던 고객 경험까지 획득