



100%



#바글바글

# Chat GPT를 이용한 댓글 순화 서비스

Comment purification service using Chat GPT





# Agenda

## 핵심 키워드 분석

- 설문조사와 인터뷰의 결과
- Core Insight 와 멘탈모델

## 기존 서비스와의 경쟁력

- 기존 서비스와의 차이점
- Chat GPT 를 활용해야하는 이유

Recap

Survey  
&  
Insight

Idea  
&  
Task

Differentiation  
&  
Strength

Story  
Board



## 맞춤 접근방식

- Text, Sketch
- 단계별 맞춤 조절
- 대상별 맞춤 조절



100%



# Recap

바글바글?

바르게 글쓰고  
바르게 글읽기

## Recap

# 'Chat GPT'를 이용한 댓글 순화

'무차별적 검열과 삭제는 문제 해결에 도움이 되지 않는다'

Q - '그렇다면 순화가 답인가요?'

## Target Users

- 인터넷 콘텐츠를 이용하는 모든 사람들
  - 댓글 쓰는 이 입장
  - 댓글 보는 이 입장



# Survey Insight

## Research Methods

Survey  
106

Age

20~29 : 88.7%  
30~39 : 5.7%

Interview  
11

### 설문지



## Google Forms



안녕하세요.

저희는 AI+X 선도인재양성프로그램 중급 수업을 수강하고 있는 **바글바글** 팀입니다.

저희는 ChatGPT API를 이용한 챗글 관련 서비스를 제작하고 있습니다.

진행함에 있어 챗글에 대한 인식과 챗글 서비스의 기능 수요를 통해 더 나은 챗글 경험을 제공하기 위해 여러분의 소중한 의견이 필요합니다. 질문지에 최대한 솔직하게 답변해주세요.

설문의 정확성을 위해서 중립적인 답변을 최대한 피해서 답변해주시면 감사하겠습니다 ^—^



# Survey Insight

## *Question*

- 재미를 위해 댓글의 수위를 어디까지 허용하시나요?  
=> "이 선수는 쓰레기야. 왜 이렇게 못하냐?"



100%



# Survey Insight

## *Question*

- 재미를 위해 댓글의 수위를 어디까지 허용하시나요?  
=> “이 선수는 쓰레기야. 왜 이렇게 못하냐?”
- 그 댓글의 대상이 자신이었더라도 괜찮을까요?

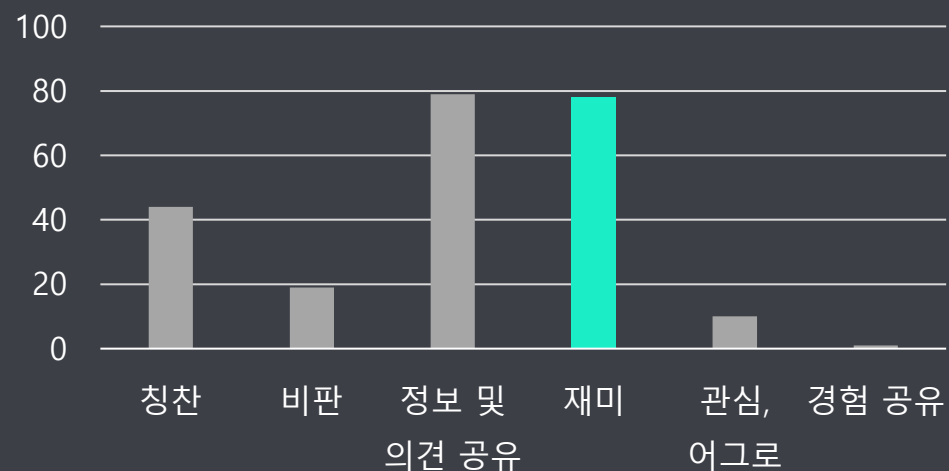
# Survey Insight

## Core Insight : 목적, 허용

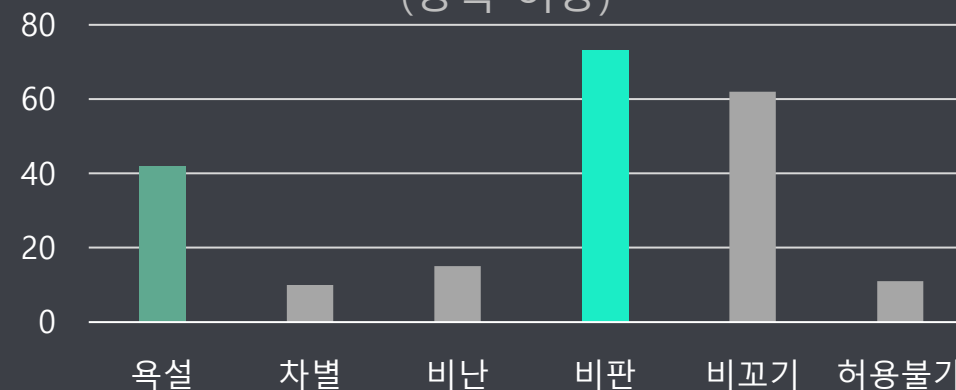
댓글 이용 목적 중  
재미는 중요한 요소이다

사람마다 허용하는 민감한 댓글  
종류는 다양하다

댓글 사용의 주된 목적 (중복 허용)

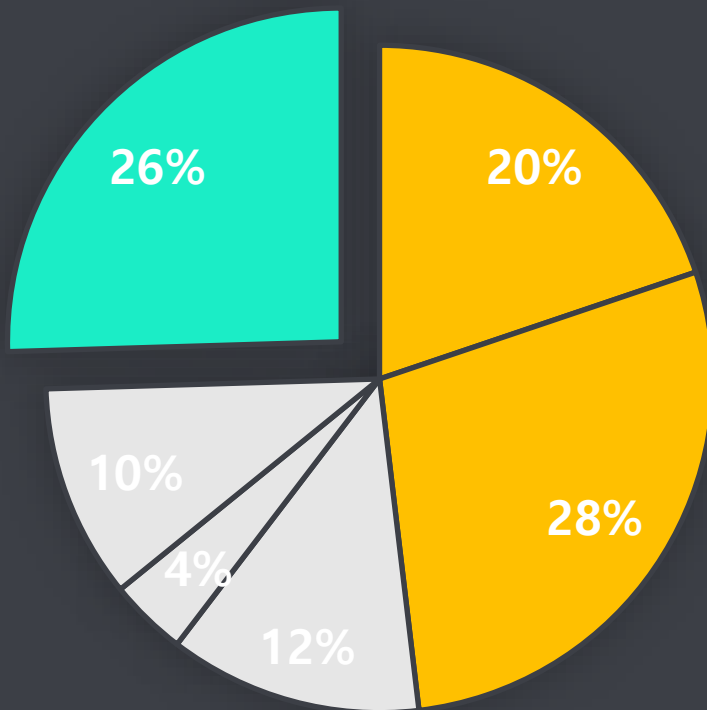


재미를 위해 댓글 이용 시 허용 범위  
(중복 허용)





# Survey Insight



■ "이 선수는 정말 못하네요. 왜 이렇게 잘하지 못하는 건가요?"

■ "이 선수는 지금은 아쉬움이 남지만, 더욱 노력하면 분명 좋은 선수가 될 거예요."

"이 선수는 쓰레기야.  
왜 이렇게 못하냐?"

**A**

비난적 표현  
순화

**48%**

**B**

부정적 의도  
순화

**26%**

**C**

순화 필요 없음

**26%**

# Survey Insight

"이 선수는 쓰레기야.  
왜 이렇게 못하냐?"

## Core Insight : 단계

같은 댓글이라도 사람마다 순화가 필요한 단계는 다르다

"이게 **악플**이라고 생각하지 않는다.  
이렇게 말할 수 있다고는 본다.

비슷한 말을 들었는데 그 당시에는  
상처를 받았었지만 **결론적으로 그  
말을 듣고 더 열심히 하는 계기**가 되었다."

-인터뷰 응답자 中 작곡가 @@씨-

"그 팀을 응원하는 입장인데 해당 선수 때문에 경기를 져서 자신의 의견을 말하는 기분은 이해한다.

하지만 **쓰레기**라는 표현은 너무 **과격**한 것 같다."

-인터뷰 응답자 中 스포츠판 @@씨-

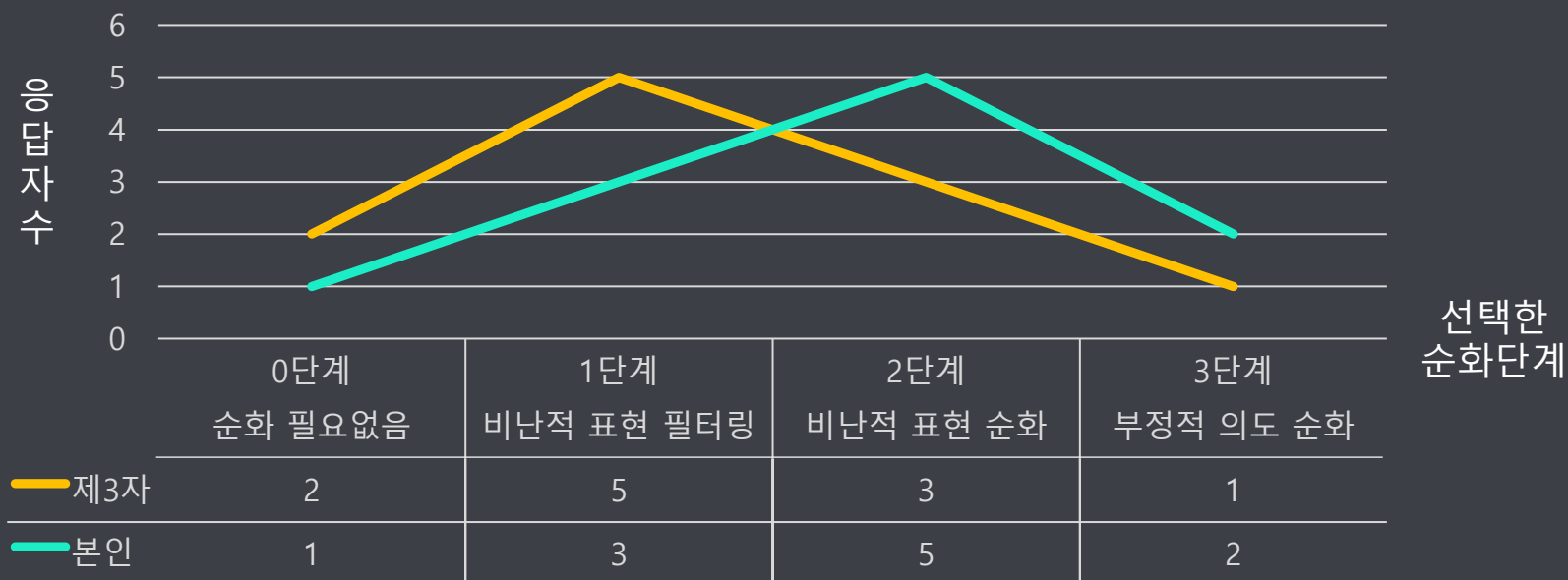
# Survey Insight

"이 선수는 쓰레기야.  
왜 이렇게 못하냐?"

## Core Insight : 대상

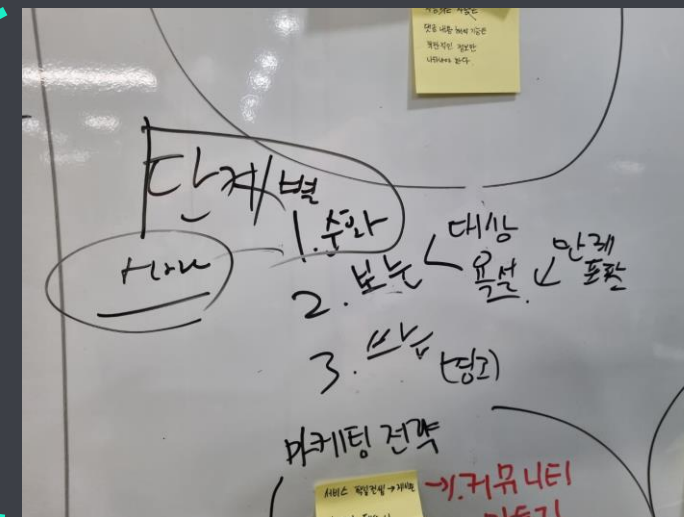
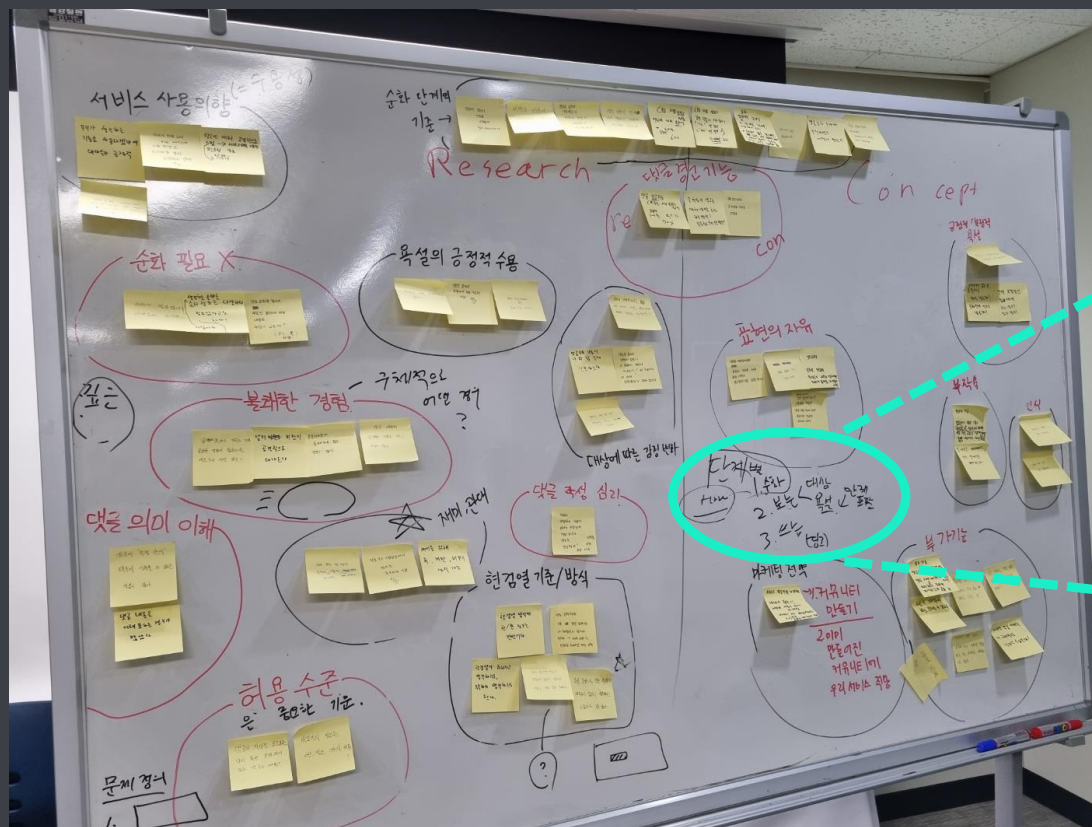
같은 댓글이라도 댓글의 대상에 따라 순화 단계는 달라진다

비난적 표현의 대상에 따른 응답한 순화 방식 변화



# Survey Insight

## Affinity Diagram



# Survey Insight

## 문제정의

1. 댓글 이용 목적 중 재미는 중요한 요소지만, 사람마다 허용 가능한 표현의 단계는 다르다.
2. 같은 댓글이라도 '남'에게 작성되었을 때와 '나'에게 작성된 댓글일 때는 다르다.

댓글창의 멘탈 모델? < 재미를 위해 보지만, 내 기준에 맞지 않으면 기분 나빠 >

# 단계별 + 대상별 접근!

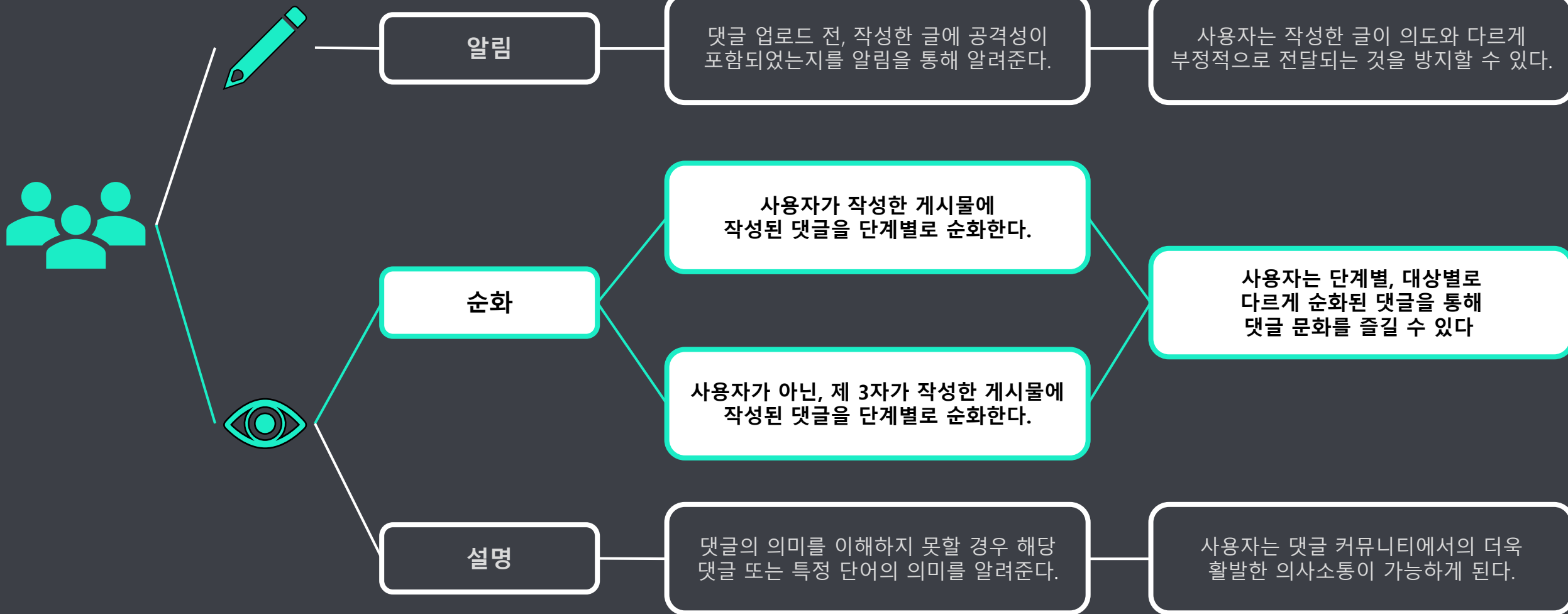
# Ideation



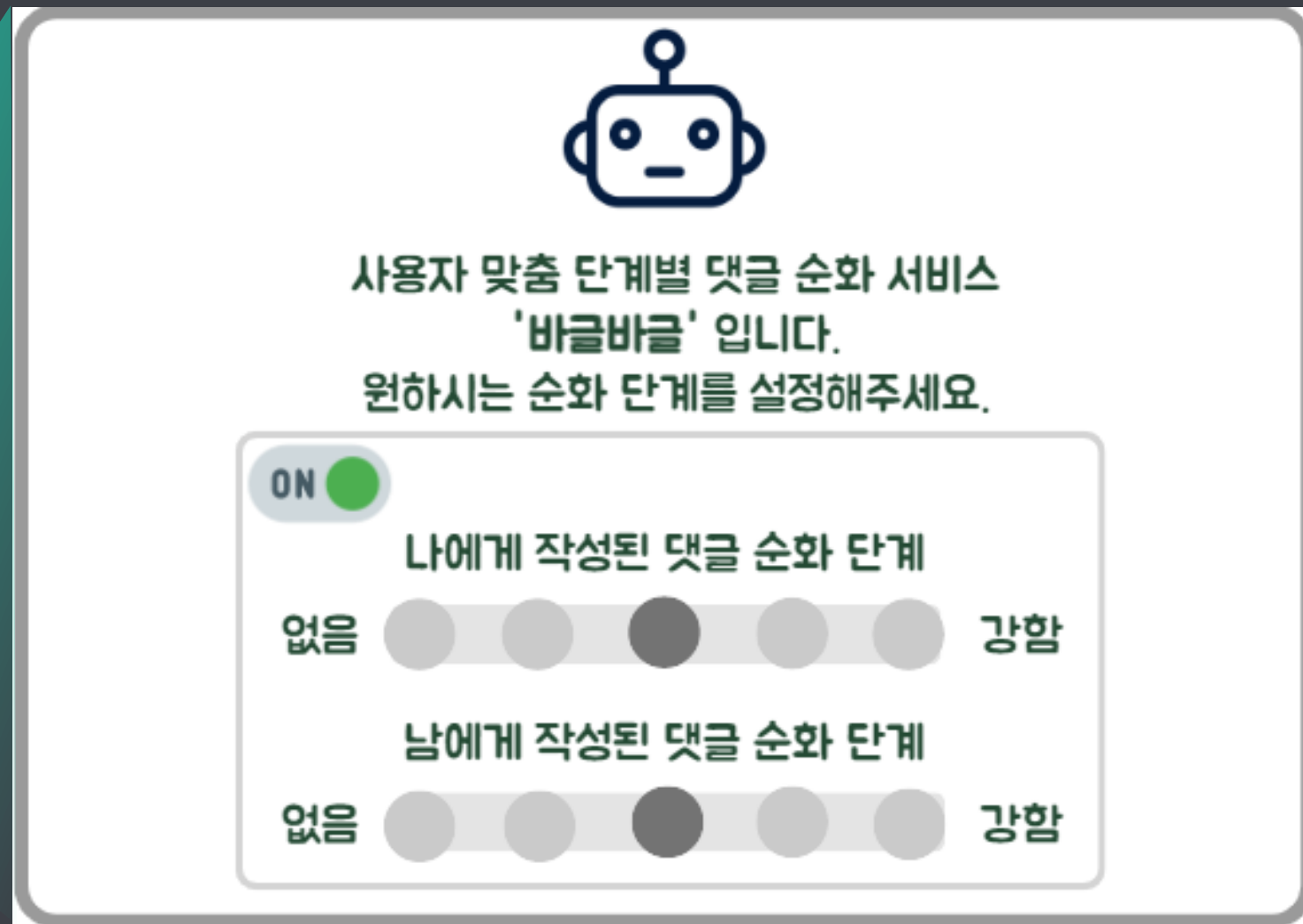
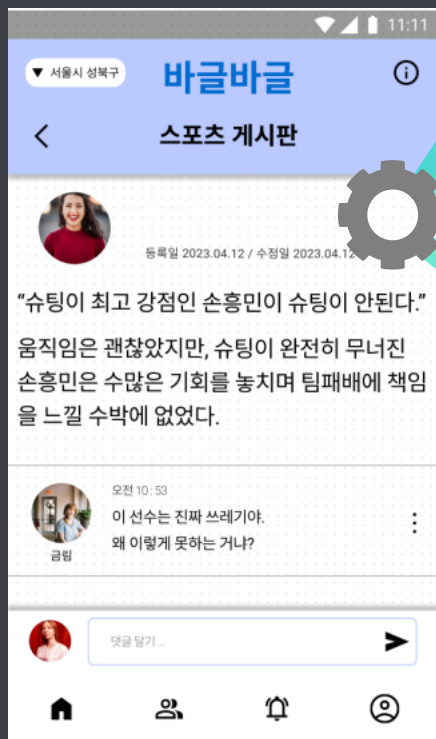
# Users

# Function

# Task



# Sketch



사용자별 순화 단계 설정 기능



Sketch

## 남에게 작성된 댓글 순화 단계

없음



강함

WIP

예시 댓글 : "이 선수는 쓰레기야. 왜 이렇게 못하냐?"

1단계 : 순화가 필요없음

"이 선수는 진짜 쓰레기야. 왜 이렇게 못하는 거냐?"

2단계 : 비속어 필터링

"이 선수는 진짜 @@@야. 왜 이렇게 못하는 거냐?"

3단계 : 비속어 제거

"이 선수는 정말 못하네요. 왜 이렇게 잘하지 못하는 거냐?"

4단계 : 비난 표현 제거

"이 선수는 아직 부족한 부분이 많은 것 같아요."

5단계 : 부드러운 표현 추가

"이 선수는 조금 더 연습하면 더욱 발전할 수 있을 것 같아요."



## 김 내로남불

- 남에게 박히는 비난  
과 욕설은 꿀잼
- 나한테 달리는 악플이  
나 욕설은 보기 싫음

ON ☒

나에게 작성된 댓글 순화 단계

없음



강함

남에게 작성된 댓글 순화 단계

없음



강함

# VS



## 이 피드백

- 남을 공격하는 비난과 욕설  
댓글을 보고싶지 않음
- 하지만, 자신의 작품 피드  
백은 여과 없이 보고싶음

ON ☒

나에게 작성된 댓글 순화 단계

없음



강함

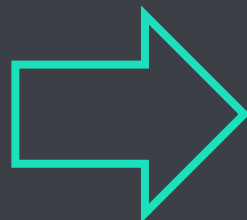
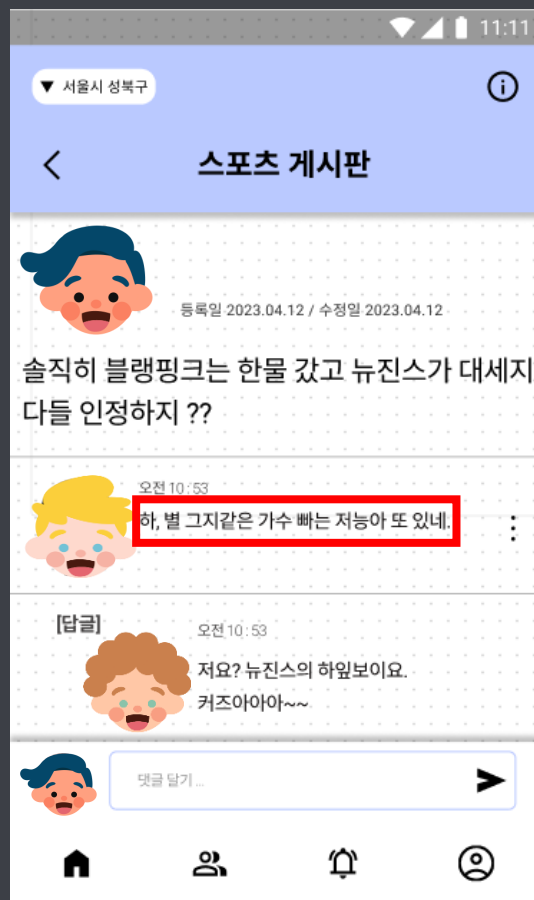
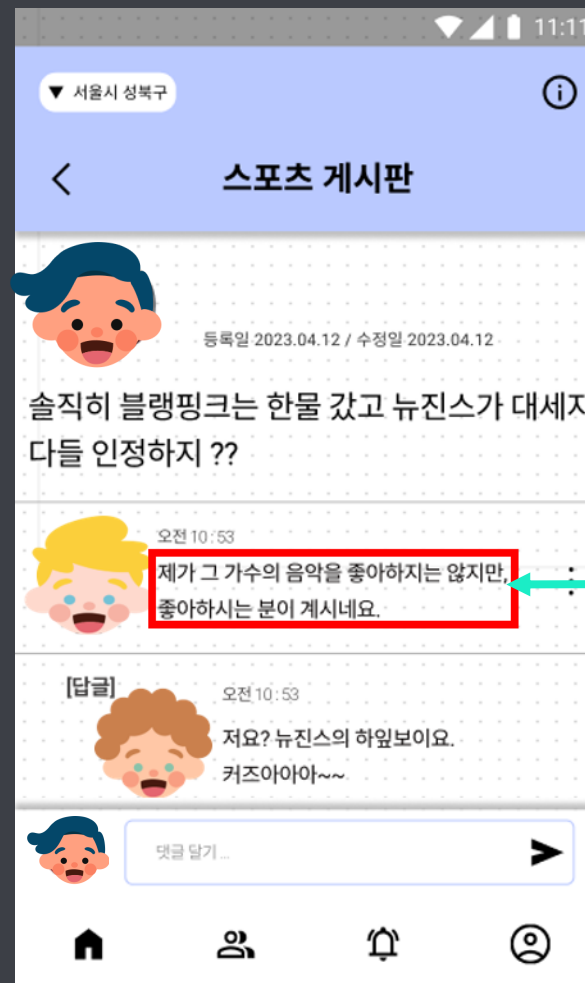
남에게 작성된 댓글 순화 단계

없음



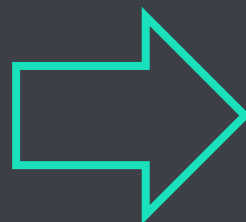
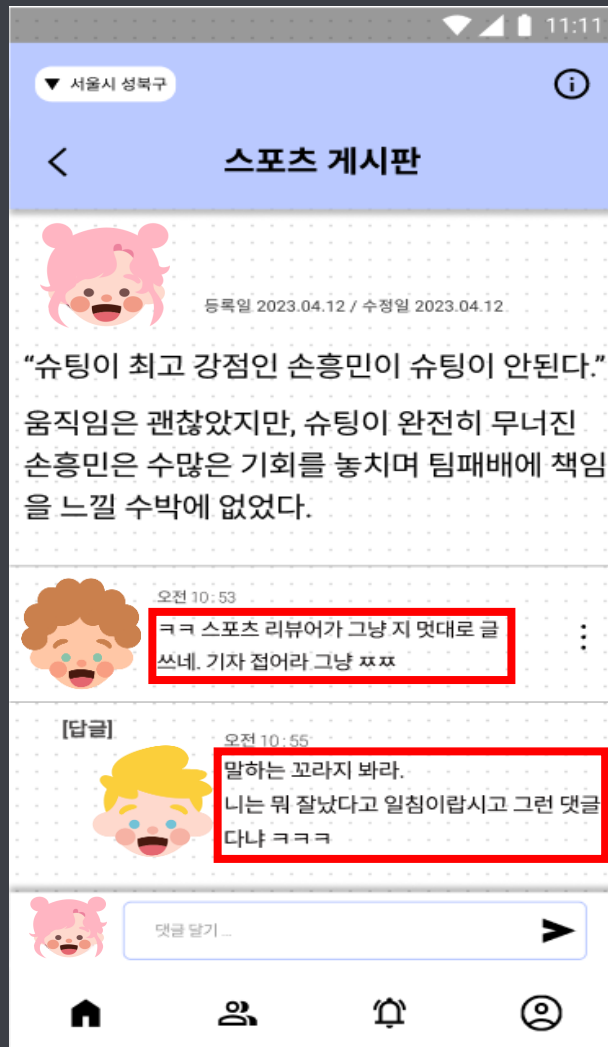
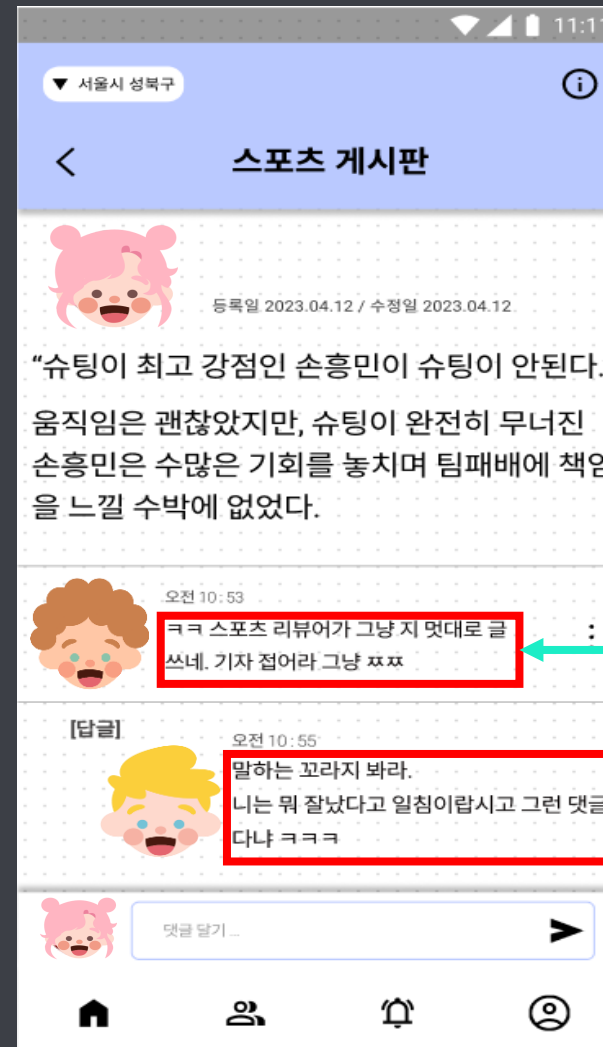
강함

## 원본

김 내로남불  
사용화면

댓글 순화

## 원본

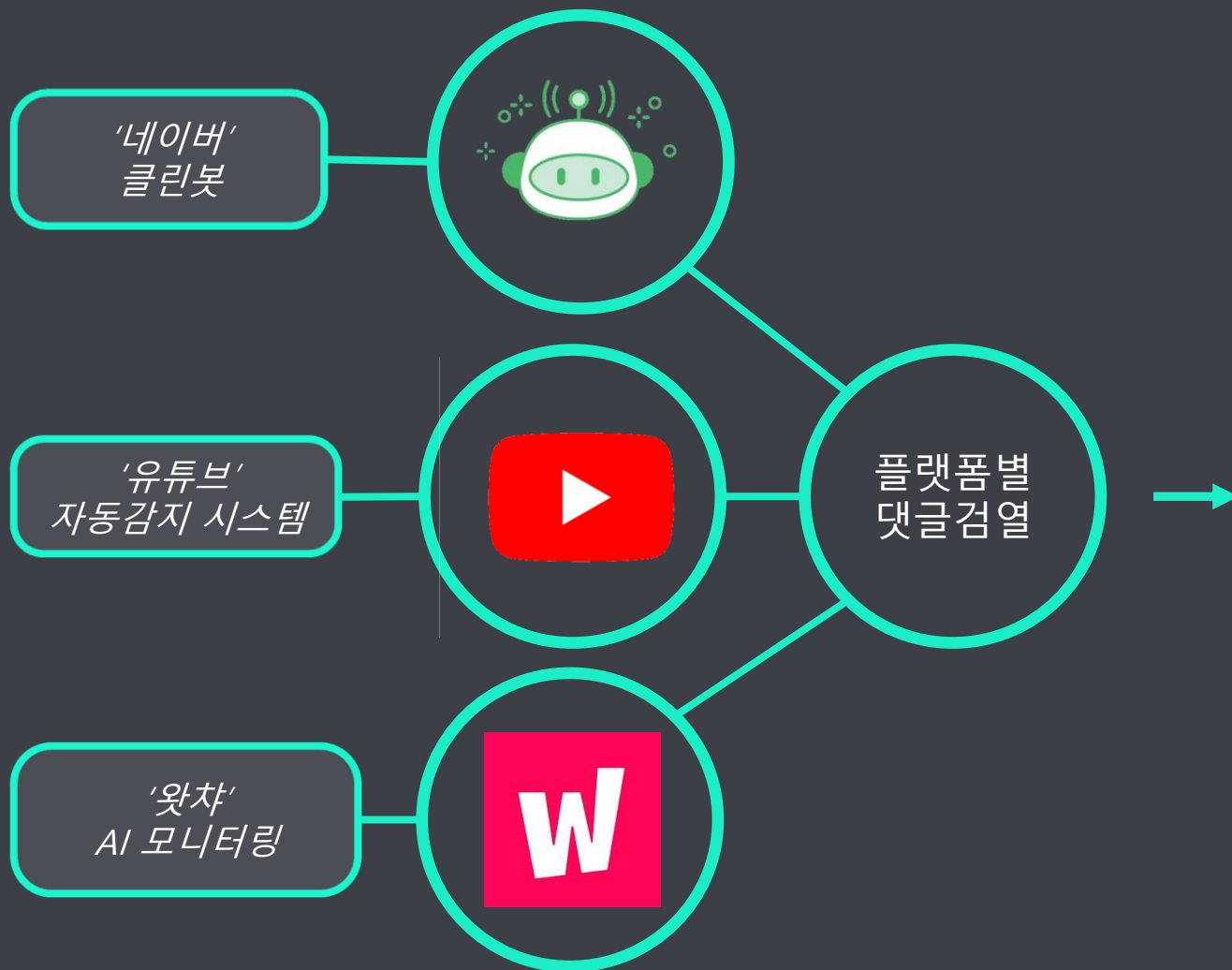
이 피드백  
사용화면

## Differentiation

# Non-AI SVC와의 차별성

- 우리 chat GPT 인공지능을 이용한 댓글 순화 서비스에서는 다양하고 복잡한 문맥과 세부적인 언어적 특징 파악이 가능
  - > 댓글의 전반적인 톤이나 감정을 파악
  - > 순화라는 유연한 대응 방식을 제안

# 기존 댓글 검열 서비스의 한계



- 모든 욕설이나 비하, 혐오 발언 예측하기 어려움.
- 단순한 검열, 삭제로 개인화된 경험을 제공하기 어려움

-> 사용자들의 불만족

“어떤 글의 내용이길래 검열이 됐는지 궁금해!”

## Strength

# 왜 'Chat GPT' 로 접근해야 하는가?

- 댓글들을 순화해주는 기술은 존재하나 해당기술을 사용하여 댓글 순화를 해주는 서비스는 없다
- 심지어 Chat GPT는 인간 피드백 기반 강화 학습을 통해 발전했기 때문에 기존의 GPT, BERT 인공지능 기술보다도 댓글의 의미를 잘 파악한다.
- 댓글 순화 서비스, 댓글 공격성 파악, 말투 변환 및 예측 등 '바글바글' 팀만의 독창적 서비스를 만들 것.


IBM Research(2018). Fighting Offensive Language on Social Media with Unsupervised Text Style Transfer. // text style transfer

Mst Shapna Akter, Hossain Shahriar, Nova Ahmed, Alfredo Cuzzocrea\$Deep Learning(2023). Approach for Classifying the Aggressive Comments on Social Media: Machine Translated Data Vs Real Life Data // gpt-2


# Strength

## 왜 'Chat GPT' 로 접근해야 하는가?

AK "아오 씹ㅂ ㅋㅋㅋㅋㅋㅋ"

 해당 댓글은 욕설이 포함되어 있으며 비판적이거나 부정적인 의도가 없는 유머적인 말로 해석될 수 있습니다. 따라서 우는 이모지 하나로 답변합니다. 🤔

AK "아오 병신 씨발아 답답하게 굴지말고 선택하라고"

 바글바글! 해당 댓글은 비속어와 욕설이 포함되어 있으며, 비판적인 의도가 강합니다. 이러한 언어는 다른 사람에게 상처를 줄 수 있으므로 사용을 삼가해 주시기 바랍니다. 대신 "선택하라는데 어떤 것을 선택해야 할지 도와주시겠어요?"와 같은 존중과 상호작용이 반영된 말씀을 사용하시는 것이 좋습니다.

AK "아...의심해서 죄송합니다.... 역사상 최고, GOAT 이시여..."

 해당 댓글은 비판적인 의도가 없는 것으로 보입니다. 따라서, 웃는 이모지 하나로 대답할 수 있습니다. 😊

AK "아...의심해야만 했습니다... 역사상 최악, JOAT 이시여..."

 바글바글! 해당 댓글은 비판적인 의도가 있으며, 욕설이 포함되어 있습니다. 욕설 부분은 @@@@로 대체하겠습니다.

바글바글! 저희는 다른 분들을 비하하는 발언은 적절하지 않다고 생각합니다. 대신, 어떤 부분에서 문제가 있는지 구체적으로 알려주시면 더 나은 대화가 될 것 같습니다.

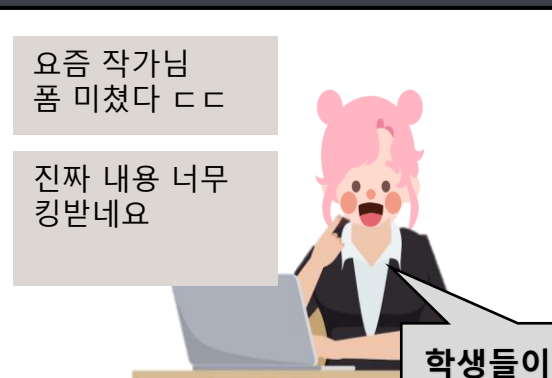
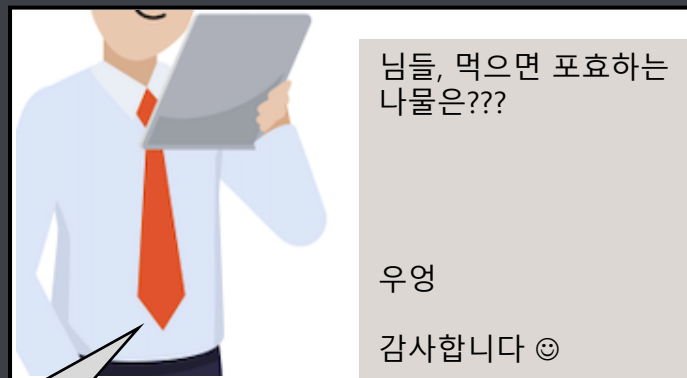
[Regenerate response](#)

- 글의 의도를 파악해 욕설이 들어가더라도 상황과 맥락에 맞는 검열여부를 설정 가능

- 기존 댓글 검열방식으로는 파악하기 어려운 댓글의 의도(비꼬기)를 파악

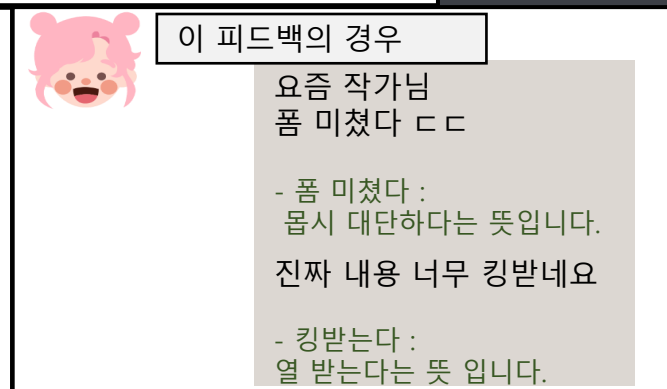
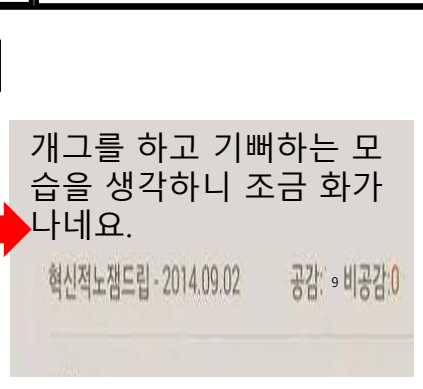
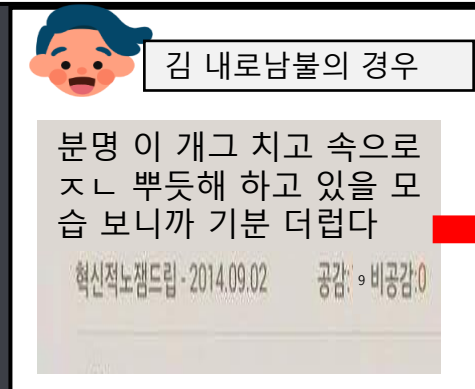
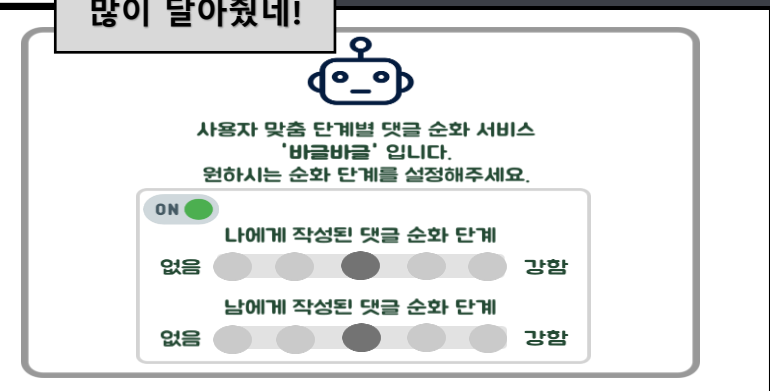
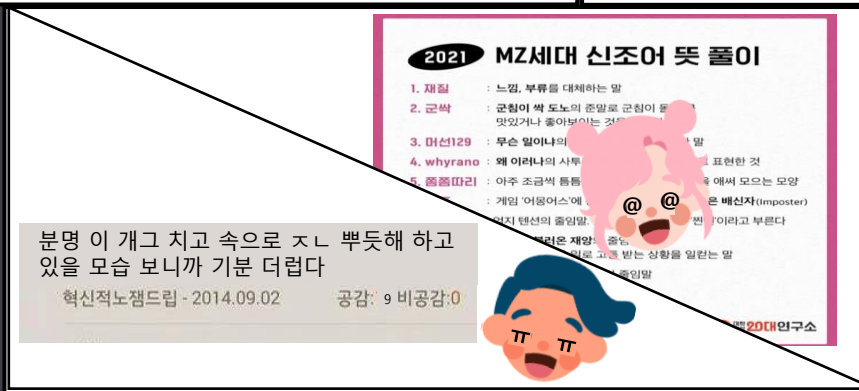


# Storyboard



학생들이 댓글을 많이 달아줬네!

ㅋㅋㅋ 이정도 유머 감각이면 나도 MZ세대지





# Schedule

4

## MID-TERM PRESENTATION

구현하고자 하는 기능의 정의  
와 개발 목표 설정, 중간발표

3.31~4.21

5

## 기능 검증 및 개 선

개발한 서비스, 기능의 품질을  
보증할 수 있도록 팀원들이나  
테스터에게 기능을 테스트 받  
는다.

4.21~5.05

6

## 기능 추가 및 개 선

최소한의 기능뿐만 아니라, 필  
요한 기능을 추가하고 개선하  
는 과정을 거친다.

5.05~5.19

7

## 기능 평가

사용자에게 테스트를 하여 피  
드백 받는 과정을 거친다.

전문가에게 검토 받는 과정을  
거친다.

5.19~6.02

8

## FINAL PRESENTATION

구현된 프로토타입의 시연 및  
최종 발표

6.02~6.16

#바글바글

*Thank you*

Comment purification service using Chat GPT

