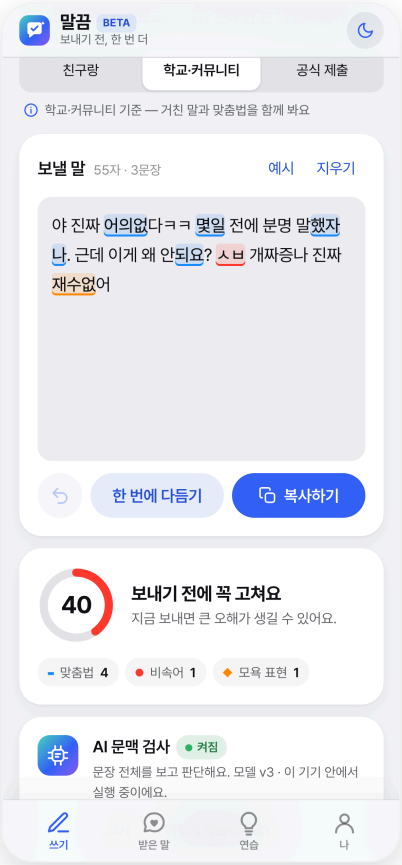
AI 활용 사례 소개서

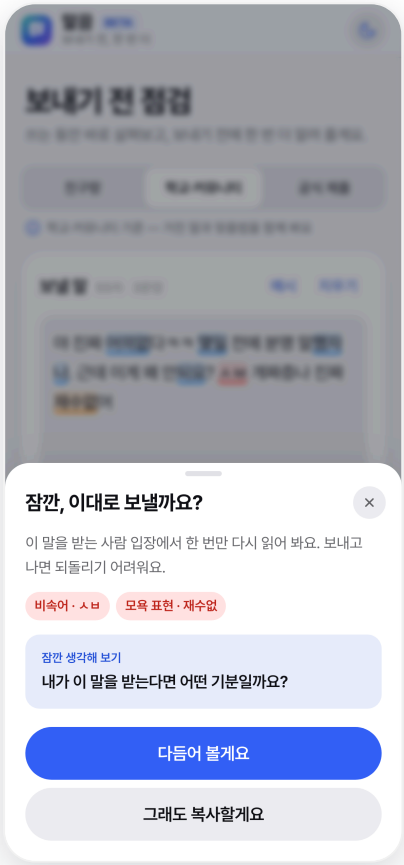
보내기 전, 한 번 더

청소년 말 습관을 돕는 기기 안 AI '말끔'

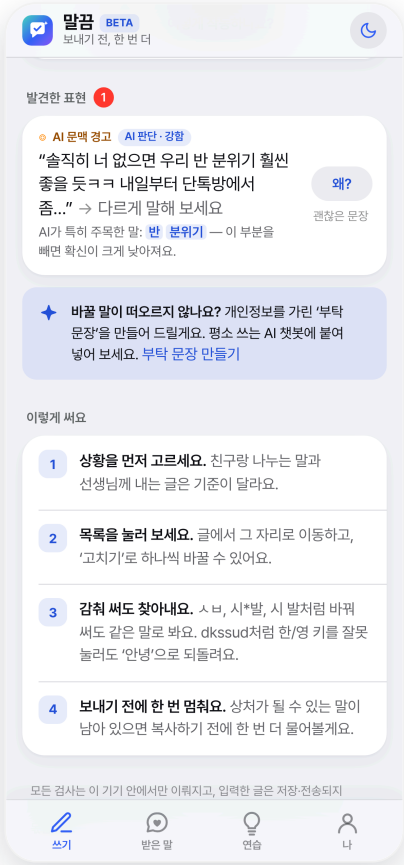
메시지를 보내기 전에 상처가 될 수 있는 말과 맞춤법을 스스로 알아차리도록, AI 코딩 도구로 웹앱을 만들고 직접 학습시킨 한국어 AI를 휴대폰 안에서 돌려 **쓴 글이 밖으로 한 글자도 나가지 않게** 한 사례입니다.



① 쓰는 동안 알려 줘요
바꿔 쓴 욕·맞춤법까지



② 보내기 전에 멈춰요
복사하기 전에 한 번 더



③ 욕 없는 상처는 AI가
휴대폰 안에서 실행

욕 없이 상처 주는 문장 탐지
6.8% → **71.8%**
규칙만 → 규칙+AI (검증 103문장)

휴대폰에서 도는 AI 크기
14.1MB
원본 55MB를 int8로 경량화

밖으로 나가는 글
0건
서버 전송 없음 · 보안 정책으로 외부 통신 차단

자동 검증 통과
133/133
규칙 검사 문장 · 토크나이저 일치 5,561/5,561

한 줄 요약 — 신고·차단처럼 말이 전달된 *뒤에* 작동하는 방법 대신, 보내기 *전에* 스스로 고치게 합니다. 규칙 검사(즉시·설명 가능)와 한국어 AI(문맥)를 나눠 쓰고, AI도 기기 안에서만 돌립니다.

바로 써 보기 **malkkeum-beta.vercel.app** · 설치·가입·비용 없음 · 제출자 김유환 · 2026년 9월

1. 문제 인식

왜 '보내기 전'이어야 할까

청소년 자녀가 학교 활동에서 '단톡방·SNS 댓글의 거친 말'을 주제로 고민하는 것을 옆에서 보며 시작했습니다.

42.7%

사이버폭력을 경험한 청소년 비율(가해·피해 포함). 언어폭력이 가장 많았습니다.
방송통신위원회·한국지능정보사회진흥원, 「2024년 사이버폭력 실태조사」(2025.3 발표)

37.8% · 7.0%

학교폭력 피해 유형 중 언어폭력·사이버폭력 비중(피해응답률 2.9%, 역대 최고).
교육부, 「2026년 1차 학교폭력 실태조사」(2026.9 발표)

지금 방법으로는 풀리지 않는 세 가지

① 전부 '보낸 뒤'에 작동

신고·차단·삭제는 상처가 전달된 다음입니다. 보내기 직전에 한 번 멈춰 주는 장치가 없었습니다.

② 단어 목록으로는 못 잡음

'ㅅ*', '시*발', 'tlqkf'처럼 바꿔 쓰고, "너 없으면 분위기 좋을 듯"처럼 욕 없이 상처 주는 말이 많습니다. 실제로 규칙 검사기만으로는 이런 문장의 **6.8%**만 찾았습니다.

③ 대화를 서버로 보내면 안 됨

청소년의 사적인 대화를 외부 AI 서버로 보내 검사하는 것은 개인정보 면에서 맞지 않습니다.

그래서 정한 성공 기준

기준	어떻게 확인하나	현재
보내기 직전에 스스로 멈추게 한다	'보내기 전 한 번 더' 알림 뒤 다시 고친 비율	베타 테스트로 측정
욕 없는 상처 문장도 찾는다	학습에 쓰지 않은 검증 문장으로 탐지율 비교	6.8% → 71.8%
멀쩡한 말은 괜히 막지 않는다	정상 문장 오경보율 · '괜찮은 말' 제보	10.8% (검증 120문장)
쓴 글이 밖으로 나가지 않는다	서버 전송 코드 없음 + 보안 정책(CSP) connect-src 'self'	0건
누구나 바로 쓴다	설치·가입·비용 없이 주소로 접속, 오프라인 동작	충족

기존 방법과 비교

방법	언제 작동	바꿔 쓴 욕	욕 없는 상처	대화가 밖으로	비용·설치
신고·차단·삭제 기능	보낸 뒤	–	–	신고 시 전송	무료
금칙어 필터	보낼 때(차단)	일부	못 잡음	서비스마다 다름	서비스 내장
서버형 AI 검사	보낼 때	잡음	잡음	**서버로 전송**	유료·가입인 경우 많음
말끔	**보내기 전(알림)**	**잡음**	**AI가 잡음**	**전송 없음**	**무료·설치 없음**

일반적인 방식끼리의 비교이며 특정 서비스를 평가한 것이 아닙니다.

원칙 — 말끔은 막는 도구가 아니라 알려 주는 도구입니다. 멀쩡한 말에 자꾸 경고가 뜨면 아무도 쓰지 않으므로, 애매할수록 '한 번 더 볼까요?' 수준으로 부드럽게 알리고 마지막 판단은 쓰는 사람에게 맡깁니다.

2. AI 활용 과정

AI로 만들고, AI를 기기 안에 넣었습니다

1 문제·규칙 정리 — Claude Code와 대화
"보내기 전에 알려 주되 막지 않는다" 같은 원칙과 기능을 말로 정리하면, AI가 설계와 코드를 제안했습니다.

2 규칙 검사기와 사전
바꿔 쓴 욕을 같은 말로 보는 정규화, 오탐 예외('시발점' 등), 인용·교육·게임 문맥 보정. 용어 231개·맞춤법 91개 — AI 초안을 자동 테스트 133개로 검증.

3 학습 데이터 준비
공개 데이터 K-MHaS·BEEP! 뉴스 댓글 86,558문장 + AI(Claude)가 쓴 청소년 말투 예시 1,013문장.

4 한국어 AI 미세조정
KoELECTRA-small(약 1,400만 매개변수)을 '상처 줄 수 있는 문장' 판별기로 학습. 노트북 CPU로 약 20분.

5 평가와 보정
학습에 쓰지 않은 검증 문장으로 성능을 재고, 오경보가 적도록 경고 기준(임계값 0.7)을 정했습니다.

6 경량화·기기 안 탑재
int8 양자화로 55MB → 14.1MB. WebAssembly로 브라우저 안에서 실행 — 서버 없이 동작.

7 베타 테스트
글 내용 없이 숫자만 모으는 '사용 기록 공유'로 효과를 측정하고, '괜찮은 말' 제보로 사전·모델을 고칩니다.

도구를 고른 이유

도구	한 일 / 고른 이유
Claude Code 생성형 AI 코딩	기획·코드·사전 초안·테스트·학습 스크립트. 대화로 만들고, 만든 코드를 AI가 설명해 줘 직접 고칠 수 있었습니다.
KoELECTRA-small 한국어 언어모델	한국어 전용이고 작아서 휴대폰에서 돕니다. 큰 챗봇형 AI는 서버로 글을 보내야 하고 느려서 제외했습니다.
K-MHaS · BEEP! 공개 데이터	한국어 악성 댓글 연구용 공개 데이터(CC BY-SA 4.0). 출처와 라이선스를 앱에 표기했습니다.
ONNX Runtime Web 실행 엔진	설치 없이 브라우저에서 AI를 돌립니다 (WebAssembly).

규칙과 AI의 역할 분담

규칙(1차) 즉시·설명 가능 — 노골적인 욕을 97.9% 잡음
AI(2차) 문장 전체의 뜻 — 규칙이 못 잡은 문장에만 표시해 알림 피로를 줄임
설명 판단 강도(보통·강함) + '왜?'(한 어절씩 빼 보며 주목한 말 표시)

막혔던 곳과 해결

- 애플 GPU(MPS)에서 학습 손실이 0.69에 고정 → CPU로 비교 실험해 원인을 확인하고 전환
- 첫 모델이 청소년 대화에 약함(F1 0.57) → 오답 유형을 분석해 문장 보강, 두 번 재학습(F1 0.79)
- 평가 중 규칙 검사기 버그 발견: '조정·조절'을 욕으로, '죽었네'를 위협으로 인식 → 음절 경계 검사 추가

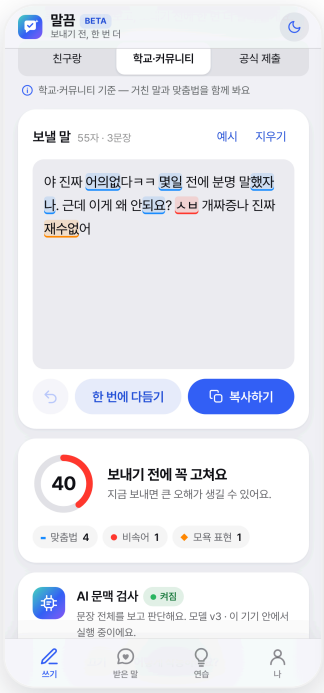
사람이 한 일, AI가 한 일

단계	AI(Claude)가 한 일	사람(제출자)이 한 일
기획	기능·화면 구성안, 기술 선택지 제안	문제 정의, '막지 않고 알린다' 같은 원칙 결정, 최종 채택
사전·데이터	용어·맞춤법 초안, 청소년 말투 예시 문장 작성	무엇을 '상처 주는 말'로 볼지 기준 결정, 결과 확인
개발	코드 작성, 오류 원인 분석, 자동 테스트 작성	직접 써 보며 불편한 점 지시, 결과 확인
AI 학습·평가	학습·변환·평가 스크립트, 지표 계산	평가 기준·경고 기준값 결정, 오답 검토, 공개 범위 판단
서류	초안 작성	사실 확인·수정, 제출

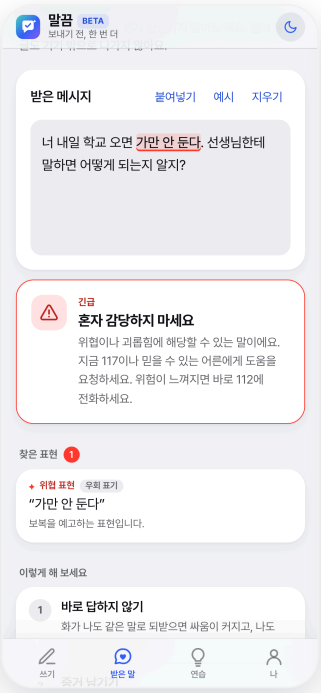
AI 활용 범위 — 코드·사전 초안·청소년 말투 예시·학습 스크립트·서류 초안은 Claude로 작성했고, 문제 정의·원칙과 기능 결정·결과 확인·최종 판단은 제출자가 했습니다.

3. 결과

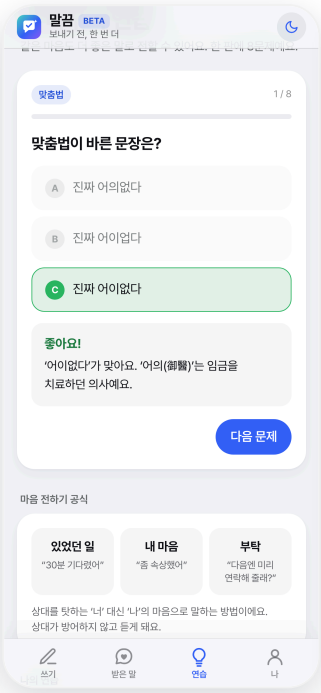
말끔 베타 — 네 가지 화면



보내기 전 점검
상황 3단계 · 한 번에 다듬기 · 클린 지수



받은 말 살펴보기
위험도 4단계 · 117·1388 연결



말 바꾸기 연습
나 전달법 · 높임말 · 안전 대처



나의 기록
숫자만 저장 · 레벨 · 배지

쓰는 사람을 위한 기능

- **상황별 기준** 친구랑 / 학교·커뮤니티 / 공식 제출
- **우회 표기·한영 오타** ㅅ*·시*발 인식, dkssud → 안녕
- **보내기 전 한 번 더** 복사 전 '받는 사람 입장' 질문
- **AI에게 안전하게 부탁하기** 연락처·학교명을 가린 부탁 문장
- **괜찮은 말 제보** 잘못 걸린 말은 내 기기에서 표시 안 함

받는 사람·어른을 위한 기능

- **받은 말 점검** 안전·주의·위험·긴급 4단계와 대처 4단계
- **도움 창구** 117 · 1388 · 109 · 디지털성범죄피해자지원센터 · 112
- **기록 메모** 신고·상담용 메모(날짜·받은 곳·찾은 표현)
- **선생님·보호자 안내** 수업 활용법과 한계
- **설치형 웹앱** 홈 화면 추가 · 오프라인 · 다크 모드 · 접근성

이렇게 씁니다

1. 쓴다
보낼 말을 말끔에 쓰거나 붙여 넣고, 상황(친구랑·학교·공식)을 고릅니다.

2. 알려 준다
거친 말·맞춤법은 즉시, 욕 없는 상처는 AI가 문장 단위로 표시합니다.

3. 멈춘다
상처가 될 말이 남아 있으면 복사 전에 "받는 사람이라면?"을 묻습니다.

4. 고쳐 보낸다
제안된 말로 바꾸거나 직접 고친 뒤, 복사·공유로 원래 앱에 보냅니다.

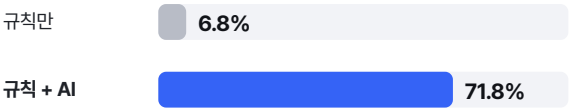
베타 테스트 — 청소년·보호자 대상 7일 베타 테스트를 준비했습니다. 글 내용 없이 숫자만 모으는 '사용 기록 공유'로 '보내기 전 한 번 더' 알림 뒤 다시 고친 비율, 다듬은 표현 수, 잘못 걸림 제보를 측정해 2026년 10월 초 결과를 추가합니다.

4. 성능과 검증

AI를 더하니 무엇이 달라졌나

학습·기준값 보정에 쓰지 않은 청소년 대화 검증 문장 223개(상처 주는 문장 103 · 정상 문장 120)로 같은 조건에서 비교했습니다. '학교·커뮤니티' 기준, AI 경고 기준값 0.7.

상처 주는 문장을 찾아낸 비율 (높을수록 좋음)



정상 문장을 잘못 표시한 비율 (낮을수록 좋음)

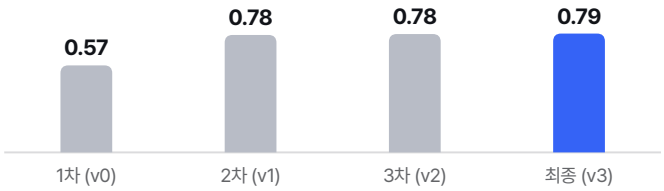


AI 모델 성능 (경고 기준값에서)

평가 자료	F1	정밀도	재현율
뉴스 댓글 시험셋 (K-MHaS, 21,939문장)	0.83	91.3%	76.6%
외부 평가셋 (APEACH, 3,770문장)	0.63	93.9%	47.8%
청소년 대화 검증 1차 (142문장)	0.76	88.9%	66.7%
청소년 대화 검증 2차 (81문장 · 최종 학습 전 작성)	0.84	81.8%	87.1%

F1 = 정밀도와 재현율의 조화평균(1에 가까울수록 좋음). 노골적 욕 문장(48개)은 규칙이 97.9%를 잡습니다.

모델 개선 과정 — 청소년 대화 검증 F1



1차는 뉴스 댓글 위주라 청소년 말투에 약했습니다. 오답 유형(암시적 위협, 비꼬기, '사진'·'못 말려' 오경보)을 보강했고, 최종판은 F1을 유지하면서 정상 문장 오경보를 16.7% → 10.8%로 줄였습니다.

자동 검증

항목	방법	결과
규칙 검사 회귀 테스트	걸리면 안 되는 정상 문장 59개 + 걸려야 하는 문장 74개	133/133
토크나이저 일치	브라우저(자바스크립트)와 학습(파이썬)의 문장 쪼개기 결과 대조, 이모지·한자 포함	5,561/5,561
경량화 손실	int8 모델과 원본 모델의 확률 차이(검증 문장 1,157개 평균)	0.015
휴대폰 화면·오류	아이폰 크기(390px) 에뮬레이션 · 콘솔 오류 · 보안 정책 위반 · 가로 넘침	0건

틀린 사례도 공개합니다

놓친 문장 (상처 주는데 표시 못 함)

- 살 좀 빼고 나서 말해
- 너 옆에 있으면 내가 연예인 같아 보여서 좋아 ㅋㅋ
- 이마 넓은 거 가리려고 앞머리 내렸냐

잘못 표시한 문장 (정상인데 표시함)

- 너 사진 보내 줘 앨범 만들려고 ㅎㅎ
- 사진 찍어 줄게 저기 서 봐
- 너 사진 진짜 분위기 있다 프사 바꿔

정직하게 밝혀 둘 점 — 검증 문장은 학습에 쓰지 않았지만, 제작 과정에서 모델의 오답 유형을 본 뒤 작성했습니다. 표본(223문장)도 작습니다. 실제 성능은 베타 테스트의 '괜찮은 말'·'못 잡은 말' 제보로 다시 확인합니다.

5. 윤리·책임 · 6. 따라 하기와 확산

안전하게, 그리고 누구나

개인정보

- 쓴 글·받은 메시지는 기기 안에서만 검사하고 **저장·전송하지 않음**
- 브라우저 보안 정책(CSP `connect-src 'self'`)으로 **외부 서버 통신 자체를 차단**, 광고·분석 도구 없음
- 기기에 남는 것은 숫자 기록·설정·사용자가 고른 허용 단어뿐, 언제든 삭제
- 베타 결과도 글 없이 숫자 요약만, 본인이 직접 복사해 공유

편향·오류 대응

- '장애인·외국인·여성' 같은 집단을 중립적으로 말한 정상 문장 9개를 따로 점검 — 잘못 표시 0개(0%)
- 막지 않고 알리기, AI 판단 강도와 근거('왜?') 표시, 판단은 사용자 — 정확한 확률처럼 보이지 않게 %는 쓰지 않음
- 잘못 걸림은 '괜찮은 말'로 즉시 끄고, 제보로 사전·모델 보완
- 위급하면 판정과 상관없이 117·112 안내

한계

새로 생긴 은어, 긴 맥락의 괴롭힘, 반어법은 놓칠 수 있고 멀쩡한 말이 걸릴 수도 있습니다. 말끔은 어른의 관심과 학교의 대응을 대신하지 않습니다.

다른 사람도 따라 할 수 있게

1. **문제와 규칙을 AI와 대화로 정리**
   Claude Code 같은 AI 코딩 도구로 '우리 상황' 규칙을 만듭니다.
2. **공개 데이터 + 우리 문장으로 소형 모델 학습**
   노트북 CPU로 20분이면 됩니다.
3. **경량화해 브라우저에서 실행**
   서버 없이 무료로 배포합니다.
4. **베타 테스트로 고치기**
   잘못 걸린 말 제보를 모아 사전·모델을 보완합니다.

확산 계획

- **학교** 국어 '언어 예절', 정보 '디지털 윤리' 수업·학급 단톡방 규칙 만들기
- **청소년 기관** 상담·예방교육에서 '받은 말' 점검과 도움 창구 안내
- **다른 분야** 사전·데이터만 바꿔 직장 메신저·다문화 가정 한국어 학습 등으로

앞으로의 계획

시기	할 일	확인 지표
2026년 9~10월	7일 베타 테스트(청소년·보호자) · 잘못 걸림 제보로 사전·모델 보완	멈춘 뒤 다시 고친 비율 · 오경보 제보 수
2026년 11월~	학급 단위 수업 적용(언어 예절·디지털 윤리) · 선생님 의견 반영	수업 후 설문 · 연습 정답률 변화
이후	소스와 모델 공개 · 다른 상황(직장 메신저 등) 사전 확장	재사용 사례 수

출처와 라이선스

항목	출처	라이선스
기반 모델	KoELECTRA-small-v3 (Park, 2020)	Apache-2.0
학습 데이터	K-MHaS (Lee et al., COLING 2022) · BEEP! (Moon et al., SocialNLP 2020)	CC BY-SA 4.0
평가 데이터	APEACH (Yang et al., 2022)	CC BY-SA 4.0
미세조정 모델	말끔 (위 데이터로 학습)	CC BY-SA 4.0 공개
실행 엔진 · 글꼴	ONNX Runtime Web (Microsoft) · Pretendard (길형진)	MIT · OFL 1.1
통계	방송통신위원회·NIA 「2024 사이버폭력 실태조사」, 교육부 「2026년 1차 학교폭력 실태조사」	공공 발표자료