17조
조은비, 박범창, 좌진우

한국어 기반 인공지능 생성 텍스트 탐지

step 1.

연구 배경 및 필요성

생성형 AI의 발전과 영향

GPT, PaLM, LLaMA 등 대규모 언어 모델(LLM) 이 등장하면서 인간 수준의 자연어 처리 가능
누구나 손쉽게 AI를 활용해 텍스트, 이미지, 오디오, 비디오 생성 가능

AI 악용 사례 증가

유명인의 음성을 조작한 딥페이크 오디오,
조작된 가짜 뉴스, 왜곡된 정보 확산
AI가 생성한 가짜 뉴스는 사람 작성 뉴스와 유사하여 판별이 어려움

한국어 탐지 기술의 부족

기존 연구는 영어 중심 → 한국어 기반 탐지 기술은 미흡
한국어 뉴스 및 정보의 신뢰성을 높이기 위한 탐지 모델 필요

step 2.

연구 목표

AI 생성 텍스트
탐지 기술 개발

한국어 기반
AI 생성 텍스트 탐지 모델 구축
가짜 뉴스 및 허위 정보 탐지 기법 적용

한국어 환경에서의
탐지 모델 최적화

기존 영어권 연구의 한계를 극복하고
한국어 환경에 특화된 모델 개발

탐지 기술의
확장성 확보

단순 탐지가 아닌
언론, 교육, 법률 분야 등으로 적용 가능

step 3.

기존 연구 및 관련 연구

제목	주요 내용
Bad Actor, Good Advisor: Exploring the Role of Large Language Models in Fake News Detection (Beizhe Hu, Qiang Sheng, Juan Cao, Yuhui Shi, Yang Li, Danding Wang, Peng Qi)	LLM이 가짜 뉴스 탐지에서 활용될 가능성 연구
인공지능(AI) 윤리 규제 동향 및 표준화 현황 (김혜정, 송현수, 박용주)	AI 윤리와 신뢰성 확보 방안 연구
Combating Disinformation in A Social Media Age (Kai Shu , Amrita Bhattacharjee, Faisal Alatawi, Tahora Nazer, Kaize Ding, Mansooreh Karami, and Huan Liu)	소셜 미디어에서 허위 정보를 탐지하는 방법 연구
대규모 언어 모델을 활용한 한국어 가짜뉴스 탐지: 한계와 가능성 (고상훈, 안현철)	한국어 환경에서 LLM을 활용한 탐지 가능성 연구

step 3.

기존 연구 및 관련 연구

기존 연구의 한계

- 대부분의 연구가 영어 중심
- 한국어 데이터 부족으로 인해 탐지 모델 성능 저하 가능
- 한국어 전용 탐지 모델 필요!

step 4.

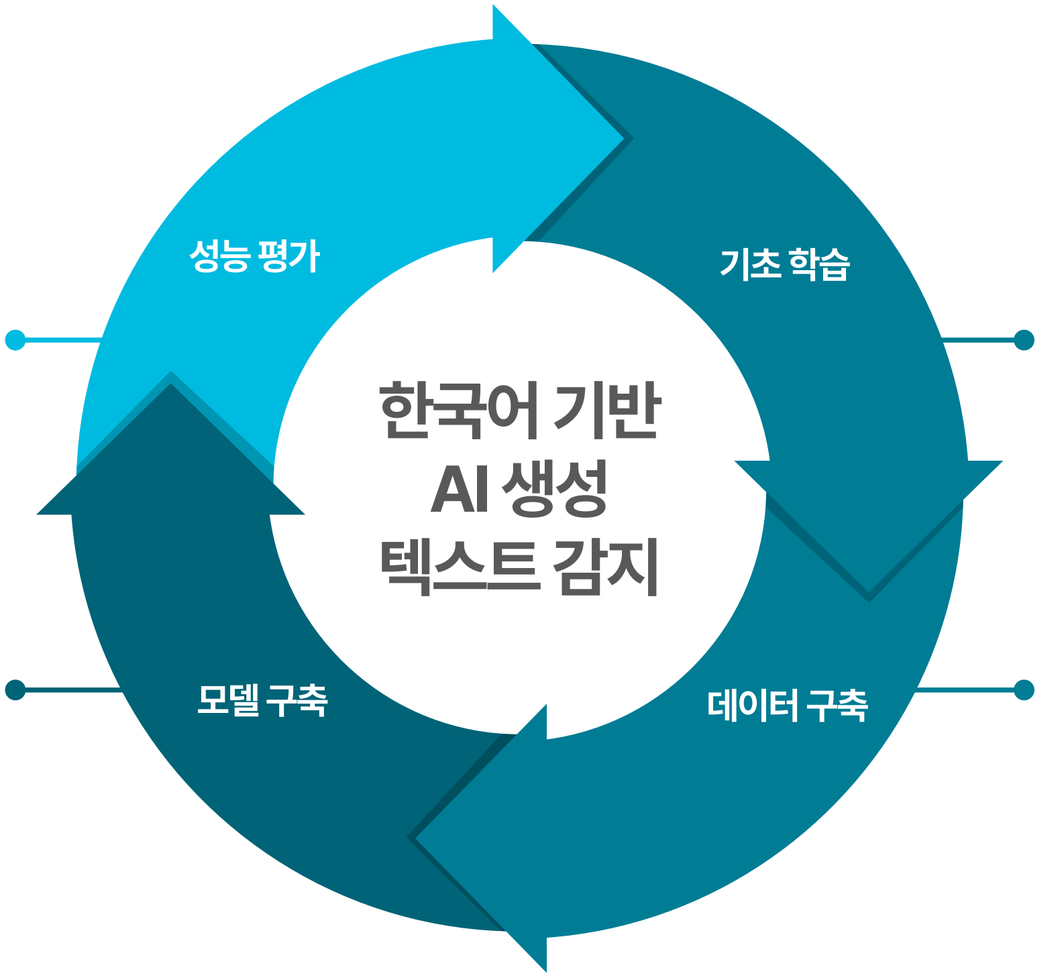
연구 방법론

성능 평가

정확도(Accuracy), 정밀도(Precision),
재현율(Recall) 등의 지표 활용
한국어 데이터에서의 모델 성능 비교 및 개선

모델 구축

BERT, RoBERTa 기반 NLP 모델 활용
머신러닝 기법을 적용하여 AI 생성 텍스트
탐지 모델 개발



기초 학습

LLM 및 NLP 기초 학습
자연어 처리 개념 및 기술 이해

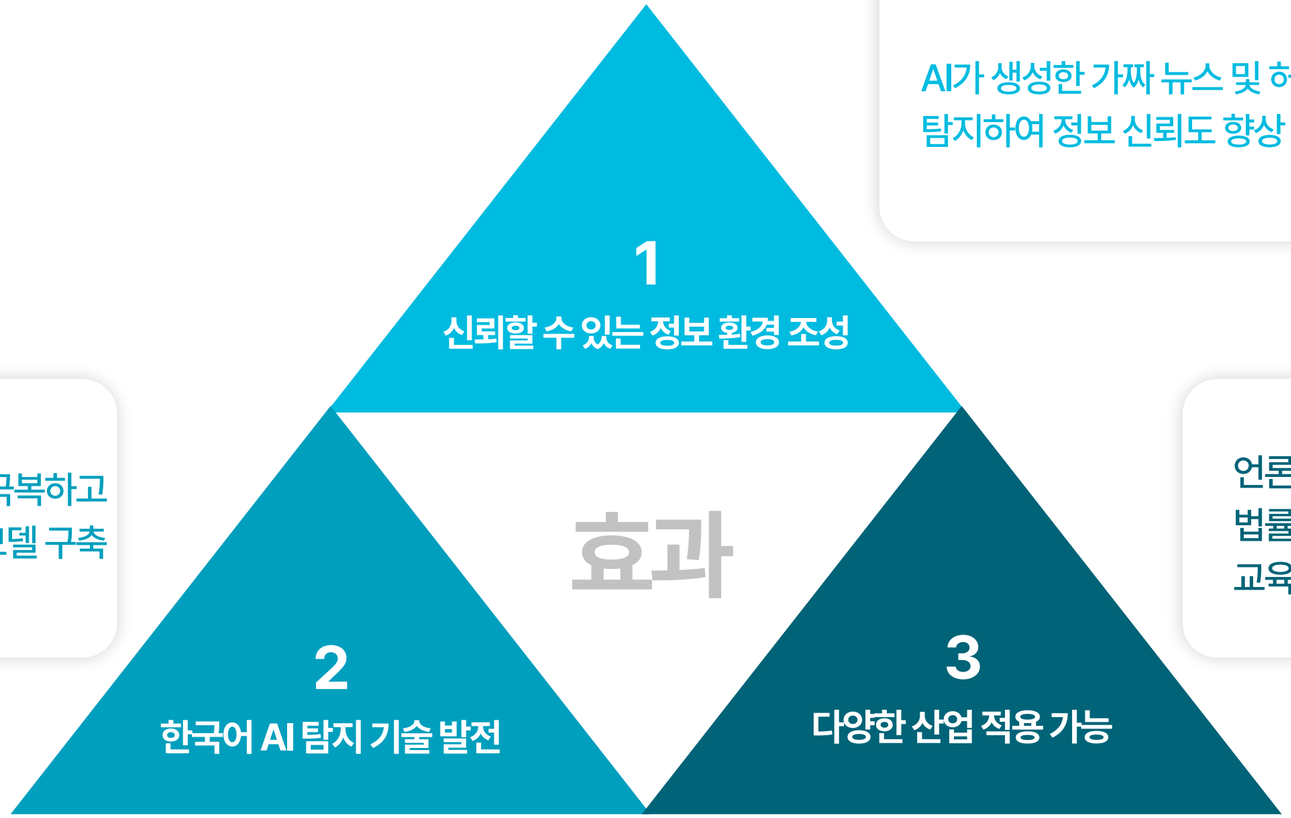
데이터 구축

AI 생성 한국어 텍스트 vs 사람이 작성한
한국어 텍스트 데이터셋 구축
다양한 뉴스, SNS, 공식 문서를 포함한
데이터 수집

step 5

예상 기대 효과

AI가 생성한 가짜 뉴스 및 허위 정보를
탐지하여 정보 신뢰도 향상



기존 영어 중심 연구의 한계를 극복하고
한국어 환경에 맞춘 모델 구축

언론: 뉴스 진위 여부 확인
법률: AI 생성 위조 문서 탐지
교육: 논문 및 과제 표절 탐지

step 6.

연구 일정 및 진행 계획

단계
기간
주요내용

1 LLM 및 NLP 기초 학습

1~2주

자연어 처리 개념 및 기술 이해

2 데이터셋 구축 및 탐지 모델 구현

2~3주

AI 생성 텍스트 데이터 수집 및 모델 학습

3 가짜 뉴스 탐지 기법 연구

2~3주

기존 탐지 기법 분석 및 성능 비교

4 모델 성능 평가 및 개선

1~2주

평가 지표 분석 및 모델 개선

step 7.

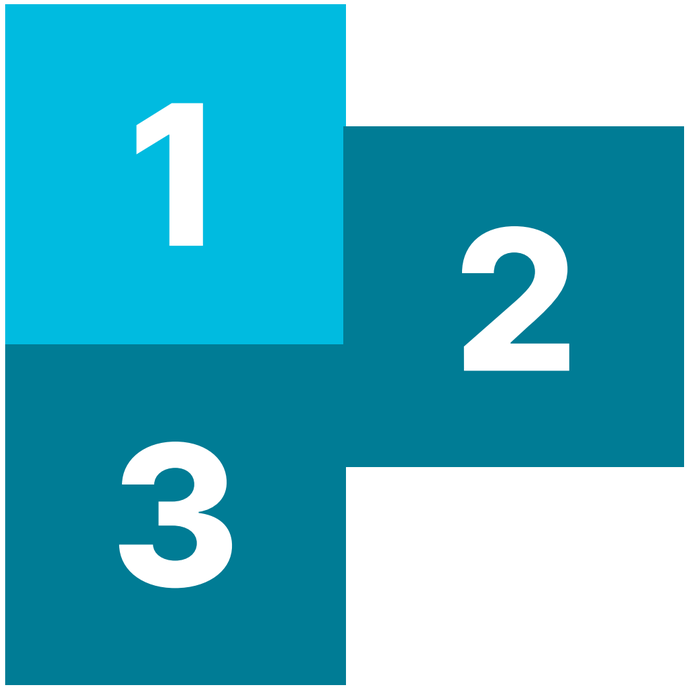
결론 및 향후 과제

연구 요약

- AI 생성 텍스트 탐지의 필요성
- 한국어 환경에서의 탐지 모델 구축
- 탐지 기술의 확장 가능성

연구의 기여

- AI 탐지 기술의 한국어 적용 확대
- 정보 신뢰도 향상에 기여



향후 연구 방향

- AI 탐지 모델의 정확도 향상 및 성능 개선
- 딥러닝 기반 탐지 기술 추가 적용
- 한국어 데이터셋 확대 및 모델 최적화