##### Project Document

문제정의서(연구개발계획서)

|  |  |
| --- | --- |
| Project Name | 한국어 기반 인공지능 생성 텍스트 탐지 |

17 조

202202501 조은비

202002494 박범창

202002565 좌진우

지도교수: 이종률 교수님 (서명)

Document Revision History

|  |  |  |  |
| --- | --- | --- | --- |
| Rev# | Date | Affected Section | Author |
| 1 | 2025/04/03 | 4, 5 번 문항 작성 | 좌진우 |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

Table of Contents

[1. 연구 개발의 필요성 5](#_Toc194347247)

[2. 연구 개발의 목표 및 내용 5](#_Toc194347248)

[3. 이해당사자 인터뷰/ 설문 인사이트 6](#_Toc194347249)

[4. 기대 효과 및 향후 확장 가능성 6](#_Toc194347250)

[5. 연구 개발의 추진전략 및 방법 6](#_Toc194347251)

[6. AI 도구 활용 정보 7](#_Toc194347252)

[7. 참고문헌(Reference) 7](#_Toc194347253)

# 연구 개발의 필요성

|  |
| --- |
| **작성 요령(제출 시 삭제할 것)** |
| **◦ 연구 개발 과제와 관련된 국내외 연구개발 현황과 문제점 등에 관하여 기술**   * 관심 주제 또는 문제 영역 설명 * 왜 이 문제가 중요한가? (사회적/ 기술적/ 사용자 관점에서)   **◦ 연구개발이 왜 필요한지 그 필요성 및 중요성에 대하여 구체적으로 기술**   * 현장에서 이 문제가 어떻게 나타나는가? * 관련 제도, 서비스, 기술적 상황 간단 조사 |

# 연구 개발의 목표 및 내용

|  |
| --- |
| **작성 요령(제출 시 삭제할 것)** |
| **◦ 본 프로젝트의 궁극적인 연구 개발 목표 및 내용을 기술**   * 본 연구개발 프로젝트가 지향하는 연구개발 목표를 기술 * 연구개발 내용, 연구개발 범위 등에 대하여 자유롭게 기술   **◦ 목표(TO-BE)**   * 프로젝트를 통해 해결하고 싶은 핵심 문제는 무엇인가? * 브레인스토밍 및 기타 시각화 도구 사용 가능   **◦ 아이디어(해결법) 요약**   * 어떤 방법으로 해결을 시도할 것인가? * 사용자 중심에서 어떤 가치를 제공하고 싶은가? * 브레인스토밍 및 기타 시각화 도구 사용 가능 |

브레인스토밍 시각화를 1개 이상 넣어야 합니다.

# 이해당사자 인터뷰/ 설문 인사이트

|  |
| --- |
| **작성 요령(제출 시 삭제할 것)** |
| **◦ 이해당사자 인터뷰/설문 정보**   * 인터뷰/ 설문 정보(기간, 인원 수, 목표, 질문 수, 조사 도구, 수집 방식 등) * 성별, 나이, 직업, 이해당사자 유형 등 조사 방식을 표로 간단히 정리     **◦ 주요 질문 및 응답 요약**   * 질문에 대한 핵심 응답 내용을 요약 정리(중복 내용은 통합 가능)   **◦ 인사이트 정리(요약)**   * 질문을 통해 도출된 공통된 인사이트를 정리 * 시각화 활용 권장(키워드 맵, 표, 그래프 등) |

# 기대 효과 및 향후 확장 가능성

본 연구를 통해 한국어 기반 인공지능 생성 텍스트 탐지 및 가짜 뉴스 판별 기술을 개발함으로써 다음과 같은 효과를 기대할 수 있다.

**⑴ 사용자 관점**

- 인공지능이 생성한 허위 정보로부터 보호받을 수 있으며, 신뢰성 있는 정보 획득 가능

- 가짜 뉴스 탐지 시스템을 통해 올바른 정보 소비 환경 조성

**⑵ 사회적 관점**

- 허위 정보 확산 방지를 통해 사회적 혼란 및 피해 최소화

- 언론 및 공공기관에서 신뢰할 수 있는 정보 제공을 위한 필터링 도구로 활용 가능

- 교육 및 공공 정책 수립에 있어 객관적인 정보 제공

**⑶ 산업적 관점**

- 언론사, SNS 플랫폼, 검색 엔진 등에서 탐지 시스템을 적용하여 콘텐츠 신뢰성 향상

- AI 기반의 탐지 기술을 활용한 신규 비즈니스 모델 개발 가능

**향후 확장 가능성**

**⑴ AI 탐지 모델의 정확도 향상 및 성능 개선**

- 딥러닝 및 최신 AI 탐지 기법을 적용하여 성능을 지속적으로 향상

- 다양한 AI 모델(GPT-4, Claude 3 등)이 생성하는 텍스트 탐지 가능 여부 분석

**⑵ 한국어 데이터셋 확대 및 모델 최적화**

- 다량의 한국어 데이터셋을 확보하여 모델 학습을 고도화

**⑶ 다양한 플랫폼 적용 가능성 검토**

- SNS, 블로그, 뉴스 기사 등 실생활 데이터를 활용하여 탐지 성능 평가

- 각 플랫폼의 특성에 맞춘 탐지 기법 연구

**⑷ 탐지 회피 기법 대응 모델 개발**

- 텍스트 변형, 의미 왜곡 등의 탐지 회피 기법을 분석하여 대응 모델 설계

- 인간이 개입한 허위 정보 생산 방식까지 탐지할 수 있는 기술 연구

# 연구 개발의 추진전략 및 방법

본 연구의 목표는 한국어 기반 인공지능 생성 텍스트 탐지 모델을 개발하고, 가짜 뉴스 및 허위 정보 판별 기술을 고도화하는 것이다. 이를 위해 다음과 같은 전략을 추진한다.

**⑴ 사전 조사 및 인공지능 관련 기초 학습**

- 기존 AI 탐지 모델 및 탐지 기술에 대한 문헌 조사

- 자연어 처리 개념 및 기술 이해

**⑵ 데이터셋 구축 및 탐지 모델 구현**

- 한국어 가짜 뉴스 및 AI 생성 텍스트 데이터셋 수집 및 정제

- SNS, 블로그, 뉴스 기사 등의 데이터를 확보하여 실생활 적용 가능성 분석

- KoBART, KoELECTRA, KLUE-BERT 기반 모델 개발 및 학습

- AI 생성 텍스트 vs 인간 작성 텍스트 차이 학습

**⑶ 가짜 뉴스 탐지 기법 연구**

- 기존 탐지 기법 분석 및 성능 비교

**⑷ 모델 성능 평가 및 개선**

- 실생활에서 발생하는 가짜 뉴스 및 허위 정보 데이터에 대한 성능 테스트

- 지속적인 피드백을 통해 모델 고도화

**목표**

**⑴ 정량적 목표**

- 연구 논문 1편 이상 작성 및 제출

- 파일럿 테스트에서 85% 이상의 탐지 정확도 달성

**⑵ 정성적 목표**

- 가짜 뉴스 탐지 및 AI 생성 텍스트 탐지 기술의 사회적 기여도 분석

# AI 도구 활용 정보

|  |  |
| --- | --- |
| *사용 도구* | *GPT-4, Claude 2.1* |
| *사용 목적* | *인터뷰 질문 초안 작성, 문장 흐름 정리, 사례 리서치 보조* |
| *프롬프트* | * *사용자 경험 중심으로 문제정의서 예시를 보여줘* * *디자인 브레인스토밍 아이디어 10개 제안해줘* |
| *반영 위치* | 1. *인터뷰 질문 목록 (p.5)* 2. *아이디어 설명 문단 정리 (p.6)* |
| *수작업*  *수정* | *있음(논리 보강, 사례 교체 등)* |

# 참고문헌(Reference)

ZhuMingjian 외 9명(2023) “A Million-ScaleBenchmark forDetecting AIGenerated”, 『SCOPUS』,

박소현. (2024). 학생 작문 에세이와 AI 생성 텍스트의 구분을 위한 KoELECTRA 기반 탐지 모델 적용 방법. 한국정보통신학회 종합학술대회 논문집, 28(2), 495-497.

고상훈, 안현철. (2024). 대규모 언어 모델을 활용한 한국어 가짜뉴스 탐지: 한계와 가능성. 지식경영연구, 25(4), 113-127.

박현주, 김병준, 김부근. (2024-06-26). 인공지능 생성 텍스트 탐지 기술의 한국어 적용. 대한전자공학회 학술대회, 제주.

Chang, Y.-H. (2024) “A Study on Deep Learning-Based Detection of AI-Generated News,”Annual Conference of KIPS. 한국정보처리학회, pp. 698–700. doi: 10.3745/PKIPS.Y2024M10A.698.

Wang, X., et al. (2025). Have LLMs reopened the Pandora’s box of AI-generated fake news?Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL).

Li, Y., et al. (2024). MAGE: Machine-generated text detection in the wild.Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL).