##### Project Document

Research Proposal

|  |  |
| --- | --- |
| 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 | 2025/03/13 | 연구개요서 초안 작성 | 좌진우 |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

Table of Contents

목차

[1. 연구 주제 이름 5](#_Toc192519163)

[2. 연구 배경 및 관련 연구 5](#_Toc192519164)

[3. 프로젝트 수행자의 의도 5](#_Toc192519165)

[4. 탐구 내용 및 기대 결과 5](#_Toc192519166)

[5. 프로젝트 관련 학습 계획 6](#_Toc192519167)

[6. 연구 일정 계획 6](#_Toc192519168)

# 연구 주제 이름

**한국어 기반 인공지능 생성 텍스트 탐지**

# 연구 배경 및 관련 연구

대규모 언어 모델(Large Language Models, LLM)은 방대한 규모의 데이터를 학습하고, 인간 수준의 언어 이해 능력을 갖추어 다양한 분야에서 인상적인 능력을 보여주고 있다(Beizhe Hu et al., 2024). LLM은 GPT, PaLM, LLaMA 등의 대규모 텍스트 데이터에 대해 훈련된 Transformer 언어 모델이며, 이미 다양한 산업에서 유용하게 사용되고 있다.

최근 생성형 AI를 통해 누구나 손쉽게 텍스트, 이미지, 오디오 및 비디오 등을 제작할 수 있다. 유명인의 이미지를 악의적으로 편집 및 배포할 수 있으며, 텍스트, 이미지, 오디오 등을 조합해 새로운 동영상을 만들 수 있다. 이러한 다양한 AI 기술의 악용 사례는 사회적 혼란을 초래하게 된다(김혜정, 송현수, 박용주, 2024).

생성형 AI의 발전은 정보 생산의 효율성을 높이는 동시에, 가짜 뉴스와 허위 정보의 확산을 가속화하는 요인이 되고 있다. AI가 생성한 가짜 뉴스는 사람이 작성한 기사와 문체나 표현 방식이 유사하여, 기존의 탐지 기법만으로는 판별이 어려운 경우가 많다(Shu et al., 2020).

또한, 고상훈(2024)에 따르면, 대부분의 기존 연구가 영어 환경에 치중되어 있어 한국어 기반 플랫폼에서 발생하는 가짜 뉴스와 허위 정보를 효과적으로 탐지하지 못하고 있다.

이에 본 연구에서는 이러한 문제점들을 해결하기 위해 한국어 기반에서의 인공지능이 생성한 텍스트를 탐지하고, 나아가 가짜 뉴스와 허위 정보들을 판별하는 모델을 구축하고, 성능을 확장시킬 수 있는 방안을 제시하고자 한다.

# 프로젝트 수행자의 의도

생성형 AI의 발전으로 인해 AI가 생성한 텍스트가 인간이 작성한 글과 구별하기 어려워지는 문제가 점점 더 부각되고 있다. 특히, 한국어 기반의 가짜 뉴스와 허위 정보는 기존 탐지 기술이 영어권을 중심으로 개발된 한계로 인해 효과적으로 탐지되지 않는 문제가 있다. 이러한 문제를 해결하기 위해, 한국어 환경에 최적화된 AI 생성 텍스트 탐지 모델을 구축하고자 한다. AI가 생성한 가짜 뉴스와 허위 정보가 검증 없이 확산될 경우, 정치·사회적 혼란을 초래하고 신뢰성을 훼손할 가능성이 높기 때문이다. 따라서, AI 기반 탐지 기술을 활용해 생성된 텍스트의 출처를 판별하고, 신뢰도를 평가하는 기술을 구축함으로써, 보다 투명하고 신뢰할 수 있는 정보 환경을 조성하는 것이 본 연구의 목표이다.

# 탐구 내용 및 기대 결과

프로젝트가 진행되는 동안 단계적으로 해결할 문제는 다음과 같다.

- 한국어 기반의 인공지능이 생성한 텍스트를 감지하는 모델 구축

- AI가 생성한 텍스트와 사람이 작성한 한국어 텍스트에 대한 데이터셋 구축

- BERT, RoBERTa 등 NLP 모델 및 머신러닝 기법을 적용하여 모델 구축

- 해당 모델에 대한 성능 평가

- 단지 인공지능이 생성한 텍스트를 탐지하는 것만이 아닌 다양한 분야로 확장 가능

- 가짜 뉴스, 허위 정보, AI 리뷰 판별 등

- 언론, 교육, 법률 분야에도 활용 가능

- 현재까지의 AI 탐지 기술이 주로 영어권에 집중된 문제를 해결하고, 한국어 환경에서의 탐지 모델 연구를 활성화하는 데 기여할 것으로 기대

- 이후 생성형 AI가 더욱 발전할 것을 고려하여, 지속적으로 보완 및 개선할 수 있는 탐지 시스템 제안

# 프로젝트 관련 학습 계획

|  |  |  |
| --- | --- | --- |
| 학습할 내용 | 기간 | 역할 분담 |
| LLM 및 NLP 기초 이론 학습 | 1~2주 | 팀원 전원 |
| 데이터셋 구축 및 탐지 모델 구현 연습 | 2~3주 | 팀원 전원 |
| 가짜 뉴스 및 허위 정보 판별 기법 학습 | 2~3주 | 팀원 전원 |
| 모델 성능 평가 방법 및 정량적, 정성적 분석 기법 학습 | 1~2주 | 팀원 전원 |

# 연구 일정 계획

|  |  |  |
| --- | --- | --- |
| 조사할 내용 | 기간 | 역할 분담 |
| AI 생성 텍스트 탐지 관련 기존 연구 및 한국어 데이터셋 조사 | 2주 | 팀원 전원 |
| 가짜 뉴스 및 허위 정보 탐지 기법 분석 및 기존 모델 비교 연구 | 3주 | 팀원 전원 |
| 실험을 위한 평가 지표 선정 및 성능 분석 방법 조사 | 2주 | 팀원 전원 |

**Related Work Summary Table**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 번호 | 연구 제목(저자) | 저널/컨퍼런스(연도) | 주요 내용 요약 | 주요 인사이트 |
| 1 | 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) | The Thirty-Eighth AAAI Conference on Artificial Intelligence (AAAI-24) | LLM이 가짜 뉴스 탐지에서 어떤 역할을 할 수 있는지를 탐구 | LLM은 가짜 뉴스 탐지에 있어 훌륭한 조언자의 역할을 수행 가능함. |
| 2 | 인공지능(AI) 윤리 규제 동향 및 표준화 현황 (김혜정, 송현수, 박용주) | 인공지능윤리연구 (2024) | 인공지능 윤리 규제 동향 및 AI 윤리 표준화의 전반적인 현황 | 인공지능, 인공지능윤리, 신뢰성, 인공지능 윤리 표준 |
| 3 | Combating Disinformation in A Social Media Age ( Kai Shu , Amrita Bhattacharjee, Faisal Alatawi, Tahora Nazer, Kaize Ding, Mansooreh Karami, and Huan Liu) | arXiv preprint arXiv:2007.07388 (2020) | 소셜 미디어 시대에 퍼지는 허위 정보와 그에 대응하기 위한 전략과 방법을 탐구 | 허위 정보의 효과적인 방지 |
| 4 | 대규모 언어 모델을 활용한 한국어 가짜뉴스 탐지: 한계와 가능성 (고상훈, 안현철) | 지식경영연구 (2024) | 한국어 환경에서의 LLM 가짜뉴스 탐지 연구의 확장 | 가짜 뉴스 탐지뿐 아니라 다양한 언어, 문화적 맥락에서의 LLM 활용 가능성 제시 |