



연세대학교
YONSEI UNIVERSITY

Y O N S E I U N I V E R S I T Y

대규모 언어 모델(LLM) 및 이미지 생성 모델의 콘텐츠 신뢰성 확보를 위한 **비가시 워터마킹 알고리즘 및 서비스 개발**

1조

컴퓨터소프트웨어 전공
이태희 유경석 임상환 안태용



CONTENTS

01

연구 배경

02

연구의 필요성 및 목표

03

연구 방향 제안

04

기대 효과

05

연구 일정

06

References

01. 연구 배경

1) 생성형 인공지능의 부적절한 활용 사례

1-1) 가짜뉴스 생성

- 최근 인공지능을 이용하여 가짜뉴스가 생성 및 유포되는 사례가 급증함.
- 2023년 5월 펜타곤이 폭발하는 사진이 유포되어 사회적 혼란을 야기한 바 있으며, 이후 이는 인공지능에 의해 생성된 가짜 사진임이 밝혀짐.

펜타곤 폭발 가짜사진¹⁾ ▶



1-2) 디지털 성범죄

- 생성형 인공지능을 악용한 성적 허위영상물을 제작 및 유포하는 사례가 빈번히 발생하고 있어 사회적 문제로 대두되고 있음

* 1) 트위터 캡처

01. 연구 배경

1) 생성형 인공지능의 부적절한 활용 사례

1-3) 저작권 분쟁

- 생성형 인공지능이 저작권 보호를 받는 타인의 창작물을 학습하는 사례가 존재하며, 이로 인한 저작권 관련 법적 분쟁이 제기되고 있음.
- Stability AI가 허가 없이 게티이미지 소유의 이미지 수백만 장을 AI 학습에 활용하여, 게티이미지가 해당 회사를 상대로 저작권 침해 소송 제기한 바 있음.

1-4) 인공지능 기술을 활용한 표절

- 학술 논문, 자기소개서, 기사 등에 생성형 인공지능을 활용하는 사례가 증가하면서, 이로 인한 표절 문제가 대두되고 있음.
- 온라인 에세이 제출 플랫폼 및 표절 탐지 전문 서비스 기업인 Turnitin에 따르면, 2023년 4월 이후 플랫폼에 제출된 2억개의 논문 중 11% 이상의 논문에 AI 생성 콘텐츠가 최소 20% 이상 포함됨을 발표.

01. 연구 배경

2) 생성형 인공지능 콘텐츠에 워터마크 표기 요구

2-1) 워터마크 표기의 의의

- 이미지, 텍스트 등 다양한 타입의 콘텐츠에 워터마크를 삽입함으로써 콘텐츠의 출처 및 소유권 확인에 활용 가능.
- 생성형 인공지능 콘텐츠에 워터마크를 표기할 경우 콘텐츠의 투명성과 신뢰성을 향상시킬 수 있음.

2-2) 인공지능 생성물에 대한 표시 의무화 추세

- 미국, EU, 중국 등 주요 국가에서는 인공지능으로 생성된 콘텐츠를 식별할 수 있도록 하는 표시를 의무화하는 관련 정책들을 발표하고 있음.
- 내년 1월 시행 예정인 '인공지능 발전과 신뢰 기반 조성 등에 관한 기본법(AI 기본법)'에서는, 인공지능사업자가 생성형 인공지능을 이용한 제품 또는 서비스가 해당 인공지능에 기반하여 운용된다는 사실을 이용자에게 사전에 고지하여야 한다고 규정하고 있음.
- 이와 같은 사회, 산업계의 요구에 부합하여, 여러 기업들은 생성형 인공지능 콘텐츠에 워터마크를 적용하기 위한 기술을 개발 중에 있음.

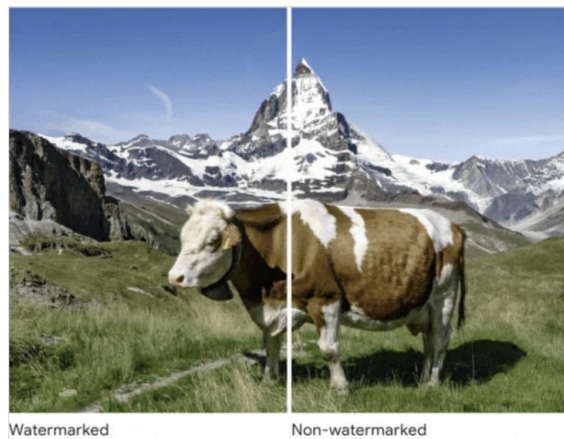
3) 인공지능 생성물에 워터마크 적용 사례

3-1) 이미지 생성물 워터마크 활용 사례

- 가시성 워터마크: 이미지에 사람이 인지 가능한 워터마크를 삽입하여 인공지능 생성물임을 표기.
- 비가시성 워터마크: 이미지 픽셀 등 하위 단위에 워터마크를 삽입하여 사람은 인지 불가능하나 프로그램으로 판독 가능한 워터마크를 삽입.



가시성 워터마크 예시¹⁾
(SK텔레콤의 에이닷)



비가시성 워터마크 예시²⁾
(구글 딥마인드의 'SynthID')

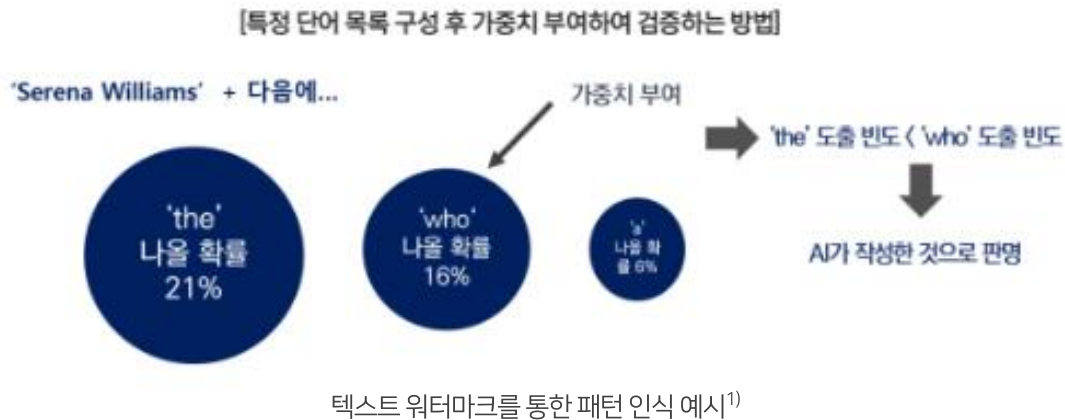
* 1), 2) 한국정보통신기술협회, 인공지능(AI) 워터마크 기술 동향 보고서

01. 연구 배경

3) 인공지능 생성물에 워터마크 적용 사례

3-2) 텍스트 생성물 워터마크 활용 사례

- 오픈 AI는 ChatGPT가 생성한 텍스트에 비가시성 워터마크 기술의 적용 여부를 검토 중.
- 구글 딥마인드는 제미니(Gemini)가 단어를 선택해 문서를 생성할 때 편향을 주는 방식으로 워터마크를 삽입하는 SynthID 알고리즘을 적용.
- 다만, 현재 텍스트 생성물 워터마킹 기술은 워터마크의 제거나 위변조와 같은 공격에 취약할 것이라는 지적이 제기됨.



* 1) 한국저작권위원회, "오픈 AI, 텍스트 워터마크 도구 개발 그러나 공개에는 신중", 저작권 이슈 브리프, 2024-9-2호

02. 연구의 필요성 및 목표

1) 본 연구의 필요성



02. 연구의 필요성 및 목표

2) 연구 목표



대규모 언어 모델(LLM)과
이미지 생성 모델에 의해 생성된
콘텐츠의 진위와 출처를
효과적으로 검증하기 위해,
보이지 않는 방식의
워터마킹 알고리즘을 제시하고자 함.



콘텐츠 내에
비가시성 워터마크를 삽입하고,
이를 효율적으로 추출 및 식별할 수 있는
알고리즘을 제공하고자 함.



비가시성 워터마킹 방법을
웹 기반 서비스로
제공하고자 함.

03. 연구 방향 제안

1) 이론적 배경

- ☑ 비가시 워터마킹 기술은 AI가 생성한 콘텐츠의 신뢰성을 보장하기 위해 필수적인 기술로 떠오르고 있음.
- ☑ 대규모 언어 모델(LLM)과 이미지 생성 모델에서 생성된 콘텐츠는 쉽게 조작될 수 있으며, 이에 따라 출처를 명확히 하고 변조 여부를 검증할 수 있는 기술이 필요.

1-1) 기존 연구와 한계



A | 이미지 워터마킹 (WAM: Watermark Anything Model)

- 기존 이미지 워터마킹 기법은 전체 이미지에 적용되었지만, 특정 영역만 변조된 경우 신뢰성이 떨어지는 문제가 있음.
- 이를 해결하기 위해, WAM은 이미지 내 워터마크가 포함된 영역을 감지하고, 해당 영역에서 식별 메시지를 추출하는 방법을 제안함



B | 텍스트 워터마킹

- LLM이 생성한 텍스트의 진위 여부를 판별하기 위해, 특정 워드 리스트("그린 리스트")를 활용한 비가시 워터마킹이 연구되고 있음.
- LLM이 워터마크된 텍스트를 생성하면, 감지 알고리즘을 통해 해당 문장이 AI가 생성한 것인지 판별할 수 있음.
- 주요 기법으로 확률적 워드 선택, 통계적 감지 기법(z-score 기반 판별) 등이 있음.



C | 워터마킹의 신뢰성과 내구성

- 워터마킹이 인간이나 다른 AI에 의해 변형될 경우, 감지가 어려워질 가능성이 있음.
- 그러나 실험 결과, 800개 이상의 토큰이 주어질 경우 인간이 변조한 문장에서도 워터마킹을 감지할 수 있음.
- AI가 생성한 텍스트와 인간이 생성한 텍스트가 혼합된 경우에도 일정 비율 이상 워터마크된 텍스트가 포함되어 있다면 감지가 가능함.

03. 연구 방향 제안

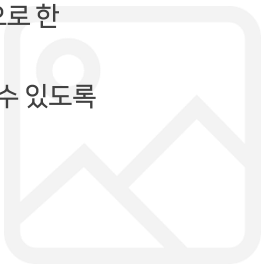
2) 연구 수행 내용

- ☑ 본 연구에서는 비가시 워터마킹 알고리즘을 개발하고, 웹 기반 검증 서비스를 구축하는 것을 목표로 함.
이를 위해 다음과 같은 연구 단계를 수행할 예정임.

2-1) 비가시 워터마킹 알고리즘 개발

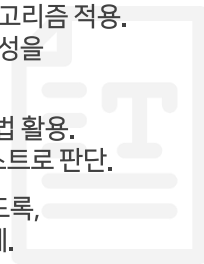
이미지 콘텐츠

- WAM(Watermark Anything Model)을 기반으로 한 비가시 이미지 워터마킹 모델 구현.
- 이미지 내 특정 영역에서만 워터마크를 검출할 수 있도록 DBSCAN 기반 탐색 기법 적용.
- JPEG 압축, 크롭, 이미지 합성(Splicing) 등의 변조 공격에 대한 내구성 실험 진행.



텍스트 콘텐츠

- LLM이 생성하는 텍스트에 대한 소프트 워터마킹 알고리즘 적용.
초록 토큰의 등장 확률을 제어하여 워터마크의 강건성을 개선하고 자연스러운 텍스트 생성 유도.
- AI가 생성한 문장을 판별하는 z-score 기반 감지 기법 활용.
초록 토큰의 등장이 랜덤하지 않은 경우 AI 생성 텍스트로 판단.
- 파라프레이징(Paraphrasing) 공격을 방어할 수 있도록,
일정 토큰 수 이상 포함된 경우 감지 가능하도록 설계.

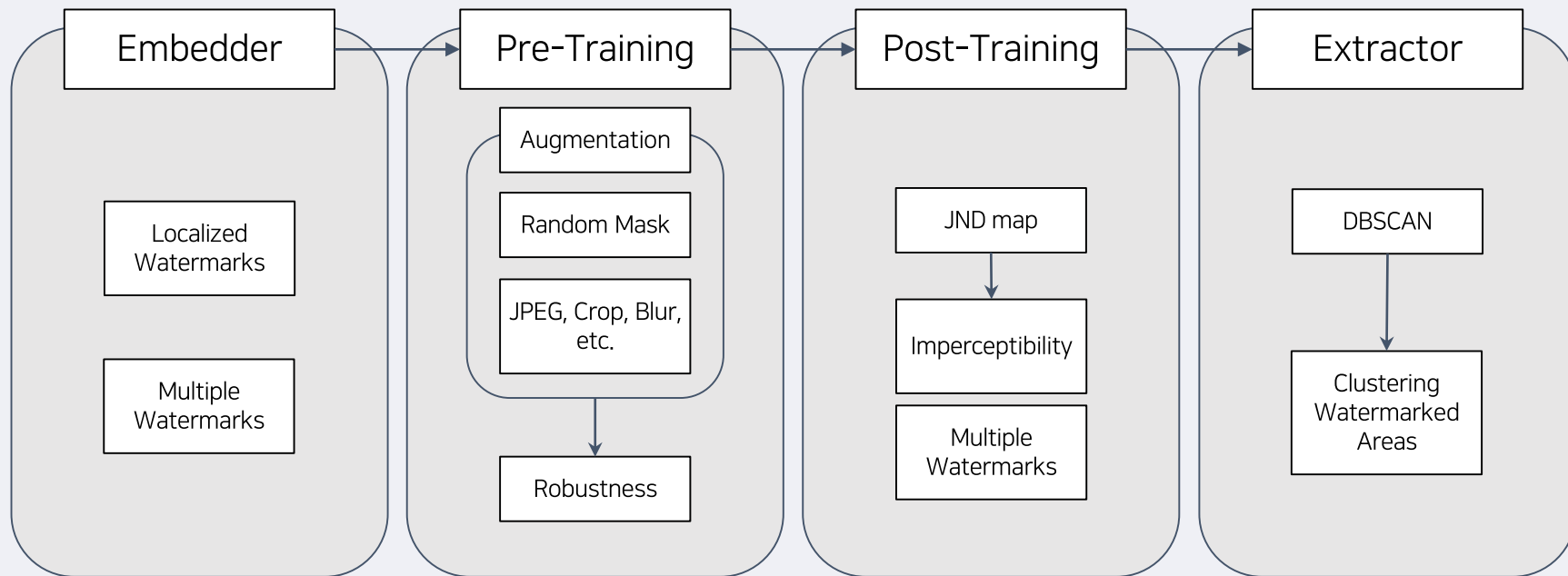


2-2) 성능 평가 및 구현

- 기존 워터마킹 기법과 비교하여 감지 성능을 평가하고 다양한 변형, 조작 등에서 워터마킹 복원 및 검출율을 측정.
- 사용자가 텍스트/이미지를 업로드하면 워터마킹 여부를 자동 분석하는 웹 애플리케이션 형태로 구현.

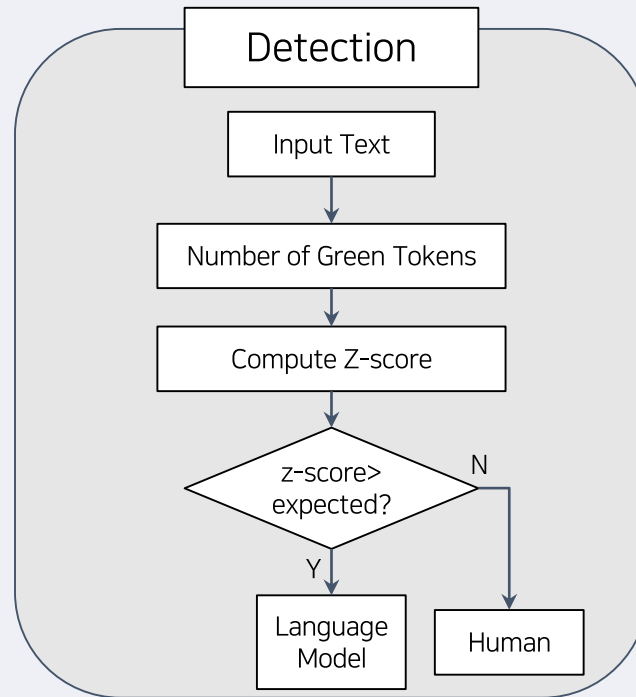
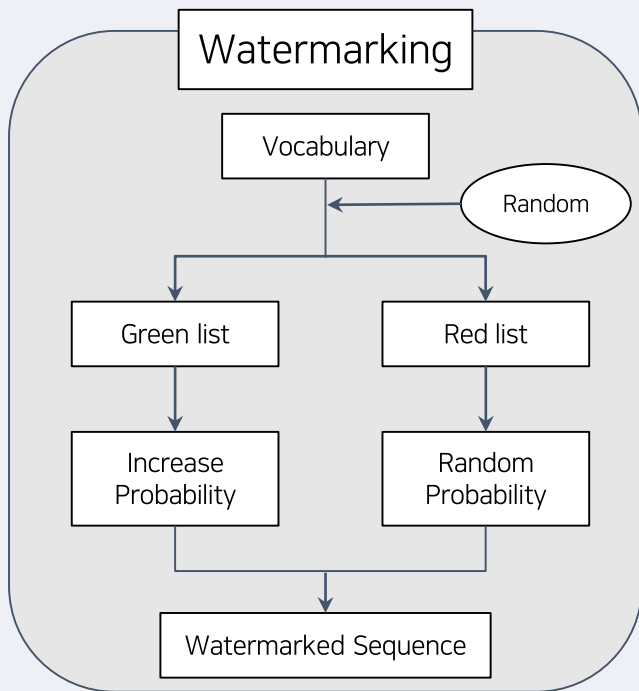
3) Model Layout Diagram

3-1) 비가시 이미지 워터마킹 모델



3) Model Layout Diagram

3-2) 비가시 텍스트 워터마킹 모델



03. 연구 방향 제안

4) 웹서비스 와이어프레임

Invisible Watermarking for AI Content Authentication

Text

Image

Source

Watermarked Result

Drag & drop an image here
or click to upload.

Embed
Watermark

Your watermarked image
will appear here.

Watermark text


Invisible Watermarking for AI Content Authentication

Text


Image

Source

Watermarked Result



Embed
Watermark



Watermark text

Details in the Image

Original image

Image with localized watermark

Difference image between original and localized watermark

Position of the watermark

Predicted watermark position

Clusters

03. 연구 방향 제안

4) 웹서비스 와이어프레임

Invisible Watermarking for AI Content Authentication

Text Image

Prompt

Explain LLM concisely.

Embed Watermark

Text without watermark

A Large Language Model (LLM) is a computer program used in the field of Natural Language Processing (NLP). These models have the ability to engage in conversations in a...

Text with watermark

A Large Language Model (LLM) is a widely used generative language model in Natural Language Processing (NLP) that has the ability to understand diverse...

Invisible Watermarking for AI Content Authentication

Text Image

A Large Language Model (LLM) is a widely used generative language model in Natural Language Processing (NLP) that has the ability to understand diverse...

This text is **watermarked** with 99% confidence.

Detect Watermark

Details in the text


prediction	True
confidence	0.99999999
num_tokens_scored	115
num_green_tokens	57
green_fraction	0.49565217
z_score	6.08371549
p_value	5.87144e-10

4) 웹서비스 디자인 시안



Invisible Watermarking for AI Content Authentication

Embeds invisible watermarks into content generated by LLMs and image models and verifies their presence.






Click to choose file or drag and drop.
We recommend high-quality .jpg, .png files less than 20MB.

Your [watermarked image](#)
will appear here.

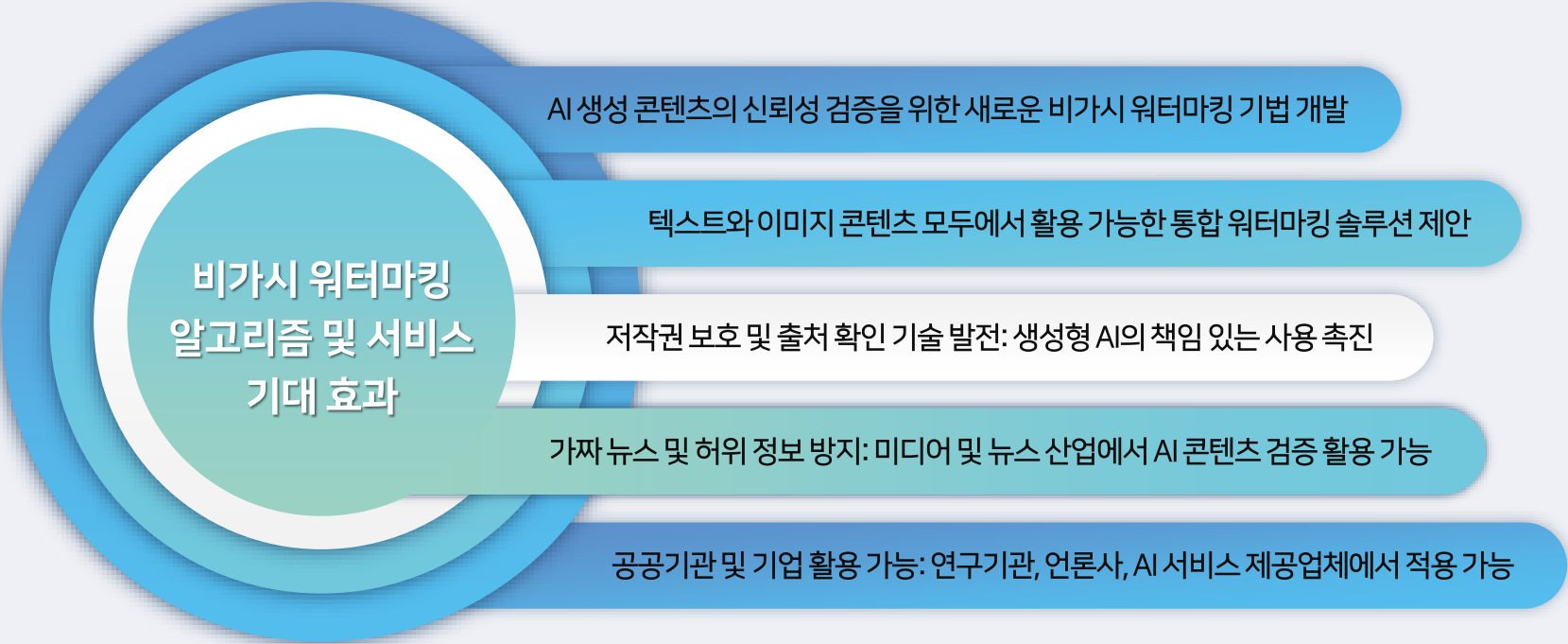
Text to embed

Embed Watermark

Sample images



04. 기대효과



비가시 워터마킹 알고리즘 및 서비스 기대 효과

AI 생성 콘텐츠의 신뢰성 검증을 위한 새로운 비가시 워터마킹 기법 개발

텍스트와 이미지 콘텐츠 모두에서 활용 가능한 통합 워터마킹 솔루션 제안

저작권 보호 및 출처 확인 기술 발전: 생성형 AI의 책임 있는 사용 촉진

가짜 뉴스 및 허위 정보 방지: 미디어 및 뉴스 산업에서 AI 콘텐츠 검증 활용 가능

공공기관 및 기업 활용 가능: 연구기관, 언론사, AI 서비스 제공업체에서 적용 가능

05. 연구 일정

구분	날짜	비고
연구 제안서 미팅	3월 20일 (목)	추가 미팅 진행 여부 및 일정 확인
연구 제안서 완성	4월 6일 (일)	
개발 시작	4월 7일 (월)	
중간 보고	4월 24일 (목)	추가 미팅 진행 여부 및 일정 확인
개발 완료	6월 8일 (일)	
최종 보고서 작성 완료	6월 15일 (일)	
최종 발표	6월 19일 (목)	

06. References

- John Kirchenbauer et al., 2024, A Watermark for Large Language Models
- John Kirchenbauer et al., 2024, On The Reliability Of Watermarks For Large Language Models
- TomSander et al., 2024, Watermark Anything with Localized Messages
- 과학기술정보통신부·한국정보통신기술협회, 2025, 인공지능(AI) 워터마크 기술 동향 보고서



연세대학교
YONSEI UNIVERSITY

감사합니다