J. Internet Comput. Serv.
ISSN 1598-0170 (Print) / ISSN 2287-1136 (Online)
http://www.jics.or.kr

LLM은 LLM이 생성한 텍스트를 어느 정도 탐지하는가?☆

How Well Can LLMs Detect Text Generated by LLMs?

류 동 훈[1] 노 기 섭[1*]
Dong-Hoon Ryu Giseop Noh

요 약

본 논문은 한국어 커뮤니티 댓글 환경에서 LLM 생성 댓글에 대한 LLM 기반 탐지 성능과 low-FPR 조건에서의 검출력이 댓글 길이에 따라 어떻게 달라지는지를 실험적으로 분석하였다. 이를 위해 한국어 커뮤니티의 실제 게시글과 댓글을 수집하여 인간 작성 댓글 데이터를 구성하고, 게시글 맥락을 바탕으로 LLM 생성 댓글을 구축하였다. 이후 댓글을 토큰 수 기준의 길이 구간으로 나누고, 각 구간에서 인간 작성 댓글과 LLM 생성 댓글의 분리 성능을 비교하였다. 실험 결과, 일정 길이 이상의 댓글에서는 LLM 기반 탐지가 비교적 안정적인 성능을 보였으나, 댓글 길이가 짧아질수록 탐지 성능이 저하되었으며, 특히 초단문 구간에서는 low-FPR 조건에서의 검출력이 크게 낮아졌다. 본 논문의 기여는 1) 한국어 댓글 환경에서 LLM 생성문 탐지 성능을 길이 구간별로 실증적으로 분석하였다는 점, 2) 실제 게시글 맥락을 반영한 LLM 생성 댓글을 기반으로 실험을 수행하였다는 점, 3) 평균 성능뿐 아니라 low-FPR 조건에서의 검출력을 함께 평가해야 함을 제시하였다는 점이다. 분석 결과는 댓글, 채팅, 짧은 리뷰와 같은 short-form 텍스트 환경에서 LLM 생성문 탐지 기법을 평가할 때 길이 특성과 오탐 통제 조건을 함께 고려해야 함을 입증하였다.

☞ 주제어 : LLM, Binoculars, Zero-shot, 댓글, 길이, Low-FPR

ABSTRACT

This paper empirically analyzes how LLM-based detection performance and low-FPR detection capability for LLM-generated comments vary by comment length in a Korean online community. Human-written comments were collected from the community, and LLM-generated comments were created from real post contexts. The comments were grouped by token length to evaluate the separation performance between human-written and LLM-generated comments in each range. Results show that detection remains stable for longer comments but degrades as comments become shorter, with low-FPR detection capability dropping sharply in ultra-short ranges. The main contributions are: 1) empirical analysis of LLM-generated text detection across length ranges in a Korean comment environment, 2) context-aware LLM-generated comments based on real posts, and 3) evidence that low-FPR detection capability should be evaluated together with average performance. These findings show that, in short-form text environments such as comments, chats, and brief reviews, both length characteristics and false positive constraints should be considered when evaluating LLM-generated text detection methods.

☞ keyword : LLM, Binoculars, Zero-shot, Comments, Length, Low-FPR

1. 서 론

최근 대규모 언어모델(Large Language Model, LLM)은 인간이 작성한 텍스트와 유사한 수준의 자연스럽고 유창한 문장을 생성할 수 있을 정도로 빠르게 발전하고 있다. 그에 따라 인간 작성문과 LLM 생성문을 구별하는 일은 점차 어려워지고 있으며, 생성문 탐지는 중요한 과제로 자리 잡고 있다[1]. 이러한 변화는 온라인 정보 환경의 신뢰성과도 직접 맞닿아 있다. LLM은 자동 생성 댓글, 홍보성 게시글과 같은 짧은 텍스트를 대량으로 생산하는 데 활용될 수 있으며, 이는 온라인 정보 생태계의 신뢰를 훼손할 수 있다[2].

댓글은 온라인 커뮤니티와 게시판 등에서 빈번하게 생산·소비되는 반응 텍스트라는 점에서 영향력이 크다. 댓글은 단순한 반응을 넘어 동조와 반박, 감정 표현, 밈과 유행어의 재생산을 통해 집단적 분위기와 담론의 흐름 형성에 관여하며, 담론 확산과 국가적 여론 형성에 영향을 미칠 수 있다[3, 4].

댓글 데이터는 기사, 에세이, 일반 문서와는 다른 언어

1 Dept. of Software and Communication Engineering, Hongik University, Sejong, 30016, Korea
* Corresponding author (kafa46@hongik.ac.kr)
[Received 04 May 2026, Reviewed 13 May 2026(R2 16 July 2026), Accepted 24 July 2026]
☆ 본 연구는 과학기술정보통신부/정보통신산업진흥원 세종 SW 융합클러스터2.0 사업의 지원으로 수행되었음 (과제번호. PJTS080126041007000100030200).

적 특성을 지닌다. 댓글은 대체로 짧고 비격식적이며, 축약어와 구어체, 인터넷식 문체가 빈번하게 사용된다[3]. 이러한 short-form, informal text 환경에서는 기존 탐지 기법이 활용해 온 어휘적·구문적 단서가 충분히 드러나지 않으며, 실제 뉴스 댓글의 평균 길이도 51자, 11단어 수준에 머문다는 선행 연구가 있다[3]. 이는 장문 중심으로 설계된 기존 탐지 방식과 실제 댓글 환경 사이에 차이가 있음을 보여주지만 이에 대한 실증 연구는 부족하다.

LLM이 생성한 문장을 LLM으로 탐지하려는 연구가 진행되었지만 지금까지 제안된 접근법에 대한 실증적 연구는 부족하였다. 본 논문에서는 실제 데이터를 수집하고 LLM을 탐지하는 LLM에 대한 성능을 실험적으로 검증하였다. 다양한 실험 결과 단문/초단문 상황에서 성능이 붕괴하는 현상을 입증하였으며 향후 연구에 필요한 다양한 요소를 도출하였다.

본 논문의 기여는 다음과 같다. 첫째, 본 논문은 한국어 커뮤니티 댓글 환경에서 LLM 기반 생성문 탐지 성능이 텍스트 길이에 따라 어떻게 달라지는지를 실험적으로 분석하였다. 둘째, 실제 커뮤니티 댓글과 게시글 기반 LLM 생성 댓글을 함께 구성하고, 토큰 수 기준의 길이 구간별로 탐지 성능을 비교함으로써 middle과 short 구간에서는 비교적 높은 성능이 유지되지만 very_short 이하의 짧은 구간에서는 성능 저하가 뚜렷하게 나타남을 확인하였다. 셋째, 낮은 false positive rate 조건에서는 초단문 구간의 검출력이 크게 약화되어, 평균 성능만으로 댓글형 생성문 탐지의 실제 적용 가능성을 판단하기 어렵다는 점을 보였다. 이를 통해 본 논문은 댓글, 채팅, 짧은 리뷰와 같은 short-form 텍스트 환경에서는 길이 구간별 성능과 low-FPR 조건을 함께 고려한 평가가 필요하다는 점을 제시한다.

2. 관련 연구

2.1 LLM 생성문 탐지

LLM 생성문 탐지는 주어진 텍스트가 인간에 의해 작성되었는지, 혹은 언어모델에 의해 생성되었는지를 판별하는 문제로 이해할 수 있다[1]. 최근 대규모 언어모델의 생성 품질이 빠르게 향상되면서 인간 작성문과 생성문 사이의 경계가 점차 흐려지고 있으며, 이에 따라 생성문 탐지는 정보 환경의 신뢰성과 직결되는 문제로 다뤄지고 있다[2].

현재 사용되는 탐지 방식은 크게 통계 기반 접근, 분류기 기반 접근, 워터마킹 등으로 나눌 수 있다. 통계 기반 접근은 퍼플렉서티, 로그 확률, 토큰 순위와 같은 신호를 이용해 생성문의 확률적 특성을 포착하며, 분류기 기반 접근은 인간 작성문과 생성문을 함께 학습한 모델을 활용한다. 워터마킹은 생성 단계에서 특정 패턴을 삽입하여 사후 검출이 가능하다[1].

LLM 생성문 탐지를 위해 통계 기반 접근, 분류기 기반 접근, 워터마킹, zero-shot 방식 등이 제안되어 왔다[1]. 이 가운데 DetectGPT는 생성 모델의 로그 확률 곡률을 활용하는 zero-shot 방식을 제안하였고[5], Binoculars는 두 개의 사전학습 언어모델 간 관점 차이를 활용하여 추가 학습 없이 생성 여부를 판별하는 방법으로 제시되었다[6]. 이후 Fast-DetectGPT는 DetectGPT의 계산 부담을 줄이는 방향을 제시하였으며[7], 최근에는 평균 성능뿐 아니라 낮은 false positive rate 조건에서의 안정성 역시 중요한 평가 기준으로 다뤄지고 있다[8].

2.2 Zero-shot 탐지

Zero-shot 계열 탐지 방식은 별도의 탐지기 학습 없이 생성문을 판별한다는 점에서 실용성이 크다. 새로운 생성 모델이 등장할 때마다 라벨 데이터를 다시 구축하거나 분류기를 재학습할 필요가 없다. DetectGPT는 생성 모델의 로그 확률 곡률을 이용하여 생성 여부를 판별하는 대표적인 zero-shot 방식이며[5], Fast-DetectGPT는 이를 보다 효율적으로 계산할 수 있도록 확장한 방법이다[7].

Binoculars는 두 개의 사전학습 언어모델을 함께 사용하는 zero-shot 탐지 방식이다[6]. 한 모델은 입력 텍스트의 log-perplexity를 계산하고, 다른 모델은 다음 토큰 분포를 제공한다. 이 둘의 관계를 cross-perplexity와 결합하여 탐지 신호로 사용한다. 최종적으로 log-perplexity와 cross-perplexity의 비율을 Binoculars score로 정의하여 활용한다. 자세한 수식은 [6]을 참조한다. Binoculars는 Binoculars score를 활용함으로써 프롬프트 의존성을 줄이고, 별도의 학습 데이터 없이도 LLM의 생성문을 LLM이 판별할 수 있는 방법이다.

최근 zero-shot 탐지에서는 평균 성능뿐 아니라 낮은 false positive rate 조건에서의 안정성도 중요한 평가 기준으로 다뤄지고 있다[8]. False positive가 높으면 인간 작성문을 생성문으로 잘못 판별하는 비율이 커지기 때문이다. 따라서 Binoculars와 같은 zero-shot 탐지 방식을 댓글 환경에 적용할 때에도, 평균적인 분리 성능과 함께 낮은 false positive rate 조건에서의 검출력을 함께 검토할 필요

가 있다[6, 8].

3. LLM 기반 탐지의 한계와 실증 필요성

댓글은 기사, 에세이, 일반 문서와는 다른 언어적 조건에서 생산되는 텍스트다. 대체로 길이가 짧고, 비격식적 표현과 축약어, 구어체, 인터넷식 문체가 자주 나타난다[3]. 또한 문법적 완결성보다 즉시성과 반응성이 우선되는 경우가 많아, 게시글 본문이나 선행 댓글의 맥락을 전제로 한 반응문 형태로 나타나기도 한다. 이 때문에 장문에서 드러나는 어휘적·구문적 단서가 댓글에서는 충분히 드러나지 않을 수 있다.

이러한 특성은 LLM이 생성한 텍스트를 LLM으로 탐지하는 문제의 난도를 높인다. 댓글은 short-form, informal text에 해당하며, 기존 탐지 기법이 활용해 온 어휘적·구문적 복잡성이 충분하지 않다[3].

결국 댓글 환경에서의 생성문 탐지는 짧은 길이, 높은 비격식성, 강한 맥락 의존성으로 인해 일반 장문 텍스트와 동일한 방식으로 다루는 것은 불합리하다. 따라서 장문 환경에서 높은 성능을 보인 zero-shot 탐지 방식이라 하더라도, 댓글과 같은 짧고 비정형적인 텍스트 조건에서 기존의 LLM 탐지 기술이 댓글과 같은 실제 환경에서도 같은 수준의 성능을 유지하는지 검증이 필요하다.

본 논문에서는 실증적 연구를 통해 LLM이 생성한 댓글을 LLM이 탐지하는 성능에 대한 검증을 수행하고 보완해야 할 기술적 요소를 식별하였다.

4. 데이터셋 및 실험

4.1 인간 데이터 구축

본 논문은 실제 커뮤니티 댓글과 LLM 생성 댓글을 함께 사용하여 생성문 탐지 성능을 비교하였다. 인간 작성 댓글 데이터는 대한민국의 식물 관련 익명 커뮤니티 사이트에서 수집한 게시글과 댓글로 구성하였으며, 게시글은 생성 댓글 구축을 위한 입력 맥락으로, 댓글은 인간 작성 댓글로 사용하였다. 인간 작성 댓글의 순도를 높이기 위해 수집 대상은 2023년 이전에 작성된 게시글과 댓글로 한정하였다. 이는 생성형 AI의 대중적 활용이 본격화되기 이전 시기의 댓글을 사용함으로써, 수집 데이터에 LLM 생성문이 포함되었을 가능성을 낮추고 인간 작성 댓글 중심의 비교 기준을 확보하기 위한 것이다.

타겟 커뮤니티로 선정한 이유는 해당 사이트가 익명 커뮤니티이며, 다양한 반응 양식과 표현을 포함하는 댓글 표본을 관측하였기 때문이다. 또한 다른 익명 커뮤니티 사이트에 비해 욕설이나 혐오 표현의 비중이 상대적으로 낮게 관측되어, 생성 모델의 안전 정책이 출력에 과도하게 개입하지 않는 조건에서 게시글 기반 댓글 생성을 수행하기에 적절하다고 판단하였다.

수집 데이터는 분석에 적합한 자연어 댓글과 게시글만 남기기 위해 정제하였다. 먼저 전체 완전 중복문(exact duplicate)은 동일한 텍스트가 데이터셋에 반복 포함되는 경우를 의미하며, 중복 데이터 사용을 줄이기 위해 제거하였다. 게시글과 댓글은 텍스트 내용이 완전히 동일한 경우를 중복 항목으로 판단하였으며, 동일한 내용이 반복 포함될 경우 특정 표현이 과대표집되어 데이터 분포를 왜곡할 수 있으므로 분석 대상에서 제외하였다.

과도한 반복 표현은 동일 문자, 음절, 짧은 구문 또는 기호가 비정상적으로 반복되는 경우로 정의하였다. 구체적으로 고유 문자 비율이 낮거나, 일정 길이의 문자열이 반복적으로 등장하는 경우를 제거 대상으로 보았다. 이는 동일 표현을 반복적으로 나열한 도배성 글이 데이터셋에 포함되는 것을 방지하기 위한 것이다.

기호 또는 초성만으로 구성된 항목도 제거하였다. 예를 들어 ㅇㅇ, ㄹㅇ, ㄴㄴ, ㅇㅋ와 같은 짧은 반응 표현이나, ㅋㅋ, ㅎㅎ, ㅠㅠ, !!!, ~~~ 등 초성 또는 기호만으로 이루어진 텍스트는 문장 구조와 의미 정보가 부족하다고 판단하였다. 이러한 항목은 댓글 형식으로 존재하더라도 인간 작성 댓글의 언어적 특성을 비교하기 위한 분석 단위로는 적절하지 않아 제외하였다.

스팸·광고성 키워드를 포함한 항목은 일반 사용자 댓글과 작성 목적 및 언어 분포가 다르므로 분석 대상에서 제외하였다. 제거 기준에는 광고, 홍보, 업체, 예약, 추천코드, 가입코드, 바카라, 카지노, 토토, 판매, 구매, 대행, 무료체험, 수익 등 상업적·홍보성 표현이 포함되었다. 또한 특정 도메인이나 외부 서비스 링크가 포함된 경우 역시 일반적인 반응 댓글과 구분하여 제거하였다.

게시글의 경우 URL을 포함한 항목도 추가로 제거하였다. 구체적으로 http://, https://, www.와 같은 외부 링크 표현이 포함된 경우를 제거 대상으로 정의하였다. URL 중심 게시글은 본문 내용보다 외부 링크에 정보가 의존하는 경우가 많으며, 게시글 자체만으로 댓글 생성에 필요한 입력 맥락을 충분히 제공하지 못할 수 있다. 따라서 게시글 기반 LLM 생성 댓글 구축의 입력 자료로 사용하기 어렵다고 판단하였다.

전처리 결과 게시글은 6,555개 중 329개가 제거되어

최종 6,226개로 구성되었으며, 댓글은 42,777개 중 5,204개가 제거되어 최종 37,573개를 인간 데이터 구축에 사용하였다. 이는 수집된 원자료에서 분석 목적에 부합하지 않는 항목을 제외하고, 게시글과 댓글의 대응 관계를 유지한 상태에서 정리한 결과이다. 데이터 수집 및 전처리 결과는 표 1과 같다.

(표 1) 데이터 전처리 요약
(Table 1) Preprocessing Summary

구분	대상수	제외/제거수	최종수
게시글 정제	6,555	329	6,226
댓글 정제	42,777	5,204	37,573

게시글에서는 과도한 반복 표현 235개, URL 포함 항목 54개, 스팸·광고성 키워드 포함 항목 35개, 전체 완전 중복문 5개를 제거하였다. 게시글 제거 사유별 분포는 표 2와 같다.

(표 2) 게시글 제거 사유
(Table 2) Removed Post Reasons

제거 사유	개수
과도한 반복	235
URL 포함	54
스팸 및 광고 키워드 포함	35
전체 완전 중복문(exact duplicate)	5
합계	329

(표 3) 댓글 제거 사유
(Table 3) Removed Comment Reasons

제거 사유	개수
전체 완전 중복문(exact duplicate)	2,783
과도한 반복	2,270
기호·초성만 존재	100
스팸 및 광고 키워드 포함	51
합계	5,204

댓글에서는 전체 완전 중복문 2,783개, 과도한 반복 표현 2,270개, 기호 또는 초성만으로 구성된 항목 100개, 스팸·광고성 키워드 포함 항목 51개를 제거하였다. 댓글 제거 사유별 분포는 표 3과 같다.

4.2 LLM 생성 댓글 구축

본 연구의 목표는 LLM의 단순한 짧은 문장 생성 여부를 비교하는 것이 아니라, LLM이 게시글을 참고하여 생성한 댓글과 인간이 직접 작성한 댓글을 비교하는 상황을 만들고 이를 LLM이 탐지하는 성능을 분석하는 것이다.

생성 댓글은 실제 게시글을 입력으로 사용하여 LLM으로 구축하였다. 타겟 사이트에서 수집한 게시글을 프롬프트의 핵심 맥락으로 제시한 뒤, 해당 게시글에 반응하는 댓글을 생성하도록 모델에 요청하였다. 생성 길이는 토큰 단위로 제한하였고, 각 댓글은 사전에 정의한 길이 구간 가운데 하나에 속하도록 생성하였다.

이러한 구성은 생성 댓글을 맥락과 무관한 독립 문장 집합으로 만들기보다, 실제 커뮤니티 환경에서 나타나는 반응형 댓글 형식에 가깝게 구성하기 위한 것이다.

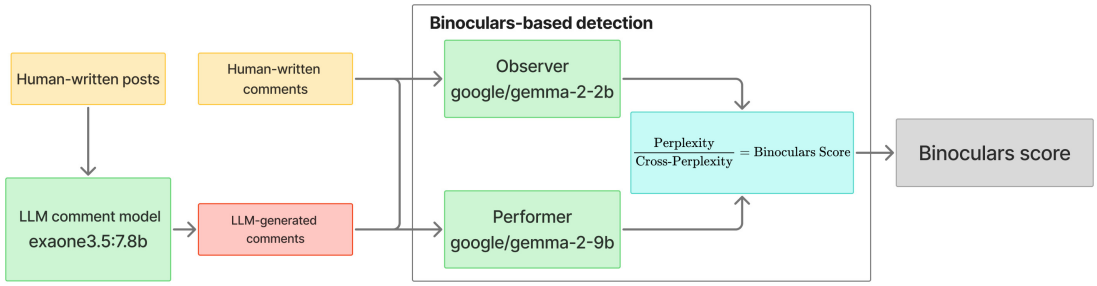
생성된 댓글 역시 인간 작성 댓글과 동일한 기준으로 정리하였으며, 댓글 형식에서 크게 벗어나거나 입력 문맥을 과도하게 반복하는 경우는 제외하였다.

4.3 댓글 생성 LLM과 탐지용 LLM 선정

댓글을 생성하는 LLM은 EXAONE 3.5 7.8B 모델을 사용하였다. 이 모델은 instruction-tuned 구조를 바탕으로 한국어와 영어를 함께 지원하며, 2.4B, 7.8B, 32B 모델군 가운데 7.8B는 성능과 운용 부담 사이의 균형을 고려하기에 적절한 규모에 해당한다. 또한 긴 문맥 처리와 실제 사용 환경을 고려한 설계를 갖추고 있어, 게시글을 입력으로 받아 짧은 반응문을 반복적으로 생성하는 조건에 비교적 잘 부합한다. 한국어 게시글을 기반으로 다수의 댓글을 생성해야 한다는 점을 고려하여, 한국어 문맥 반영 능력과 생성 안정성을 기준으로 해당 모델을 생성 모델로 선정하였다.

LLM 생성 댓글 탐지에는 Binoculars를 적용하였으며, Observer 모델은 google/gemma-2-2b, Performer 모델은 google/gemma-2-9b로 설정하였다. 두 모델은 동일한 Gemma 2 계열에 속하면서도 규모 차이를 가지므로, 기본적인 언어 처리 구조와 토큰화 기준의 일관성을 유지하면서도 모델 규모 차이에서 비롯되는 확률적 관점 차이를 비교할 수 있다. 이러한 특성은 두 모델의 출력 차이를 이용해 인간 작성 댓글과 LLM 생성 댓글을 구분하는 본 연구의 설정에 적합하다.

본 실험에서는 Gemma 2 계열의 동일한 토큰화 기준을 사용하여 모든 입력 댓글을 토큰 시퀀스로 변환하였



(그림 1) 댓글 생성 및 탐지를 위한 LLM 구조도
(Figure 1) LLM Architecture for Comment Generation and Detection

다. 이후 각 댓글에 대해 Observer 모델의 log-perplexity와 Performer 모델을 이용한 log-cross-perplexity를 계산하고, 두 값의 비율을 Binoculars score로 사용하였다. 산출된 Binoculars score는 인간 작성 댓글과 LLM 생성 댓글을 구분하기 위한 기본 판별 점수로 활용하였다. 댓글 생성 모델과 탐지 모델 간의 관계는 그림 1과 같다.

4.4 성능 측정을 위한 버킷 분류

댓글 길이에 따른 탐지 성능 차이를 확인하기 위해 전체 댓글을 토큰 수 기준으로 네 개의 버킷으로 정의하였다. 본 논문에서 버킷은 댓글을 일정한 토큰 수 범위에 따라 묶은 길이 기반 분석 단위를 의미한다. 즉, 각 댓글은 산정된 토큰 수에 따라 하나의 버킷에만 배정되며, 이후 탐지 성능은 버킷별로 집계·비교된다. 댓글 길이는 Hugging Face Transformers의 AutoTokenizer를 통해 Gemma 2 계열 토크나이저를 로드한 뒤, 특수 토큰을 제외한 인코딩 결과의 길이로 산정하였다.

Gemma 계열 모델은 Gemini와의 호환성을 위해 SentencePiece 토크나이저의 부분집합을 사용하며, 숫자 분리, 공백 보존, 미등록 토큰에 대한 byte-level encoding을 포함한다[9]. SentencePiece는 신경망 기반 텍스트 처리를 위해 제안된 언어 독립적 서브워드 토크나이저 및 디토크나이저로, 사전 단어 분리 없이 원문 문장에서 직접 서브워드 단위 분할을 수행할 수 있다[10].

따라서 본 연구에서의 댓글 길이는 글자 수나 어절 수가 아니라, 언어모델의 입력 처리 단위에 따른 토큰 수를 의미한다.

다만 이러한 길이 산정 방식은 사용한 토크나이저에 따라 달라질 수 있다. 동일한 댓글이라도 다른 언어모델의 토크나이저를 사용할 경우 토큰 수와 길이 구간 배정이 달라질 수 있으며, 한국어 댓글의 축약어, 초성 표현, 이모티콘, 비표준 표기 등은 이러한 차이를 더 크게 만들 수 있다. 따라서 초단문 구간의 탐지 성능은 텍스트 길이뿐 아니라 토큰 분할 방식의 영향도 함께 고려하여 해석할 필요가 있다.

버킷은 LLM으로 댓글을 생성할 때 최소 길이 조건과 길이에 따른 탐지 난이도를 고려하여 설정하였다. 댓글 생성 과정에서 최소 생성 길이를 5 토큰으로 두었기 때문에 가장 짧은 버킷의 하한을 5 토큰으로 설정하였고, 첫 번째 버킷의 상한은 그 4배인 20 토큰으로 정하였다. 이후 버킷은 21~40 토큰, 41~80 토큰, 81~160 토큰으로 구성하여 길이가 단계적으로 약 2배씩 증가하도록 하였다. 이는 토큰 수가 적은 영역에서는 작은 길이 차이도 탐지 성능에 영향을 줄 수 있는 반면, 토큰 수가 증가할수록 활용 가능한 문맥 정보가 상대적으로 많아진다는 점을 반영한 것이다. 각 버킷의 정의는 표 4와 같다.

(표 4) 댓글 길이 버킷 정의
(Table 4) Definition of Comment Length Buckets

버킷	토큰 수 범위
ultra_very_short	5~20
very_short	21~40
short	41~80
middle	81~160

표 4의 기준에 따라 인간 작성 댓글과 LLM 생성 댓글을 버킷별로 배정하였다. 이에 따라 구간별 댓글 예시를 표로 제시하였다. 사람이 작성한 댓글은 표 5에, LLM이 작성한 댓글은 표 6에 표기하였다.

4.5 실험 환경 및 적용 설정

실험은 길이 구간별로 수행하였다. 각 버킷에서는 전체 후보 문장 집합으로부터 인간 작성 댓글 800개와 EXAONE 생성 댓글 1,200개를 무작위로 추출하여 하나의 샘플 집단을 구성하였고, 이러한 집단 구성을 3회 반복하였다. 이후 버킷별 3개 샘플 집단에 동일한 Binoculars 설정을 적용하여 평가를 수행하였다.

이러한 반복 추출은 특정 표본 구성에 따라 결과가 과도하게 좌우되는 것을 완화하고, 길이 구간별 성능 경향이 일관되게 나타나는지를 확인하기 위한 것이다. 각 샘플 집단의 모든 댓글에 대해 Binoculars score를 산출한 뒤, 이를 기준으로 인간 작성 댓글과 생성 댓글의 분포를 비교하였다. 최종 성능 값은 각 버킷에서 얻어진 3회 반복 실험 결과의 평균값으로 정리하였다.

4.6 평가 지표

평가에는 AUC, Average Precision, Accuracy, F1-score, Precision, Recall을 사용하였다. AUC는 임계값 변화 전반에서 인간 작성 댓글과 LLM 생성 댓글이 얼마나 잘 분리되는지를 나타내며, Average Precision은 precision-recall 관계를 기반으로 한 분리 성능을 보여준다. Accuracy는 전체 분류 정확도를 의미하고, F1-score는 precision과 recall의 균형을 반영한다. Precision은 LLM 생성 댓글로 판별된 결과가 실제로 생성 댓글일 비율을, Recall은 실제 LLM 생성 댓글 가운데 탐지된 비율을 의미한다.

본 연구에서 임계값은 Binoculars score를 이용하여 인간 작성 댓글과 LLM 생성 댓글을 구분하기 위한 판별 기준값을 의미한다. 각 댓글에 대해 Observer 모델의 log-perplexity와 Performer 모델을 이용한 log-cross-perplexity를 계산한 뒤, 두 값의 비율을 Binoculars score로 산출하였다. 본 실험에서는 인간 작성 댓글을 음성 클래스(label 0), LLM 생성 댓글을 양성 클래스(label 1)로 설정하였다.

평가 과정에서는 LLM 생성 댓글이 더 높은 판별 점수를 갖도록 Binoculars score의 방향을 조정하였다. 구체적으로 machine_score=-Binoculars score를 사용하였으며, machine score가 임계값 이상인 경우 해당 댓글을 LLM

(표 5) 버킷별 사람 작성 댓글 예시
(Table 5) Examples of Human-Written Comments by Bucket

버킷	토큰 수	예시 댓글
ultra_very_short	9	온도의 문제인가?ㅜㅠ
	15	장미들 진짜 강한 거 같음 이 추위를…
very_short	25	ㅋㅋ집에 이쁜토분 새로 들였는데 이따 집가서 심고 올릴겡ㅋ
	36	흙 배합따라 잘 안먹힐 수도 있거든. 걍 화분 들어보거나 손가락으로 찔러보는게 젤 정확
short	44	아 페페 귀여운데… 아 귀여운데 고부기페페 래드배리도 좋고 진짜 괘니 추천받았다 다 좋다 ㅋㅋㅋ아씨
	51	그넘 마자.. 미련이 남아서 또 올림.. 아니 계속 같은 곳에 있었는데 며칠 사이 요래됨 급 추워진 것도 아니고 최근에 날 풀렸는데 흙흙
middle	81	그 일반 사각 베타 하는곳 보다 넓게 잇어여..원형이라 왜곡이 있어서 그런데… 글구 안에 수초로 빽빽하게 해놔서 깊히 못들어가도 상류에서만 놉니다 … 완전 핀 다 녹던애 댈구와서 나름 핀도 다 살고 통통해요ㅠ
	111	호접이면 화분수태넣어 키워도 괜찮고 부작해도되고 심지어 반수경해도 됨 화분수태는 확실히 물주는 주기 길어져서 편한데 가끔 과습되면 골로갈 수 있고 부작은 간지나는데 전반적으로 관리하기 쉽지않은 듯 반수경은 뿌리 과습올일은 없는데 물시중들기가 쫌 빡셈
out_of_range	4	로즈마리
	164	글게..집집마다 다를거같음. 베란다가 통으로 빠진 집, 베란다와 방이 분리된집 이렇게 다를거 같고.. 베란다 분리된 집도 바깥쪽 창가랑 방쪽에 가까운 창가랑 온도차이도 크더라고. 어제 영하 11도일때 방쪽 창가 온도가 8.9도였음. 그 전날에 영하 13도일땐 바깥쪽 창가 온도는 영상 2도 찍었구. 베란다에서 식물을 어디에 두냐에 따라서도 월동이 가능할지,불가능할지 갈릴것 같아ㅇㅇ

(표 6) 버킷별 LLM 생성 댓글 예시
(Table 6) Examples of LLM-Generated Comments by Bucket

버킷	토큰 수	예시 댓글
ultra_very_short	14	추억 속 웃음 소리가 들리는 것 같아요
	15	뉴기니 고스트 꽃 사진 찾기 쉽지 않네요
very_short	25	저도 공감해요! 집에서 편안하게 쉬는 게 제일 힐링이죠.
	36	괌에서는 식물 수입 규정을 꼭 확인해보세요! 다육이는 비교적 관리가 쉬우니 좋은 선택일 수 있어요.
short	44	분갈이 잘 하셨네요! 세미분재로 변화시키는 것도 멋진 아이디어네요. 식물이 더 건강해지길 바라요!
	50	아이크소서, 식물도 아프면 걱정되네. 3호기는 괜찮아지려나? 다른 방법으로 혹시 도움이 될지 찾아보는 건 어떨까?
middle	81	정말 답답하시겠네요. 소통이 잘 안 되면 이렇게 스트레스 받게 되는 거죠. 다음엔 좀 더 자세한 사진이나 상태 설명을 덧붙여 보시는 건 어떨까요? 이해를 돕고 불필요한 오해를 줄일 수 있을 것 같아요.
	110	상태가 많이 안 좋다니 안타깝네요. 온도 조절이 중요하겠네요. 혹시 가능하다면, 식물 주변에 보온 커버를 덮어주거나 따뜻한 곳으로 옮겨보는 건 어떨까요? 셀러와 연락해서 조언을 구하는 것도 방법일 것 같아요. 회복될 가능성이 있다면 조금만 더 노력해보는 건 어떨까요?

생성 댓글로, 임계값 미만인 경우 인간 작성 댓글로 판별하였다.

AUC와 Average Precision은 임계값 변화 전반의 분리 성능을 평가하는 지표이므로 특정한 하나의 임계값에 의존하지 않는다. 반면 Accuracy, F1-score, Precision, Recall은 이진 분류 결과를 필요로 하므로 machine score에 임계값을 적용하여 산출하였다. 본 연구에서는 각 샘플 집단에서 가능한 machine score 값을 후보 임계값으로 사용하였다. Accuracy는 Accuracy가 가장 높게 나타나는 임계값에서 산출하였고, F1-score, Precision, Recall은 F1-score가 가장 높게 나타나는 임계값을 기준으로 산출하였다. 최종 결과는 각 길이 버킷에서 수행한 3회 반복 실험 결과의 평균값으로 정리하였다.

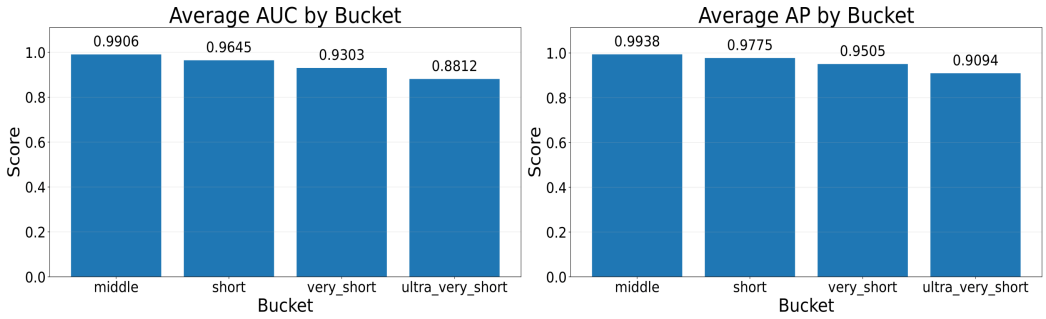
또한 평균 지표만으로는 실제 적용 가능성을 충분히 설명하기 어렵기 때문에, 낮은 false positive rate 조건에서의 true positive rate도 함께 측정하였다. Low-FPR 분석에서는 인간 작성 댓글의 machine score 분포를 기준으로 목표 FPR 조건에 해당하는 임계값을 설정하였다. 이후 해당 임계값에서 LLM 생성 댓글이 생성문으로 검출되는 비율을 TPR로 계산하였다. 본 연구에서는 0.01%, 0.1%, 1%의 목표 FPR 조건에서 TPR을 사용하였으며, 최종 결과는 각 길이 버킷별 3개 샘플 집단에서 얻어진 반복 실험 결과의 평균값으로 정리하였다.

5. 실험 결과 및 분석

5.1 길이 구간별 전체 성능 비교

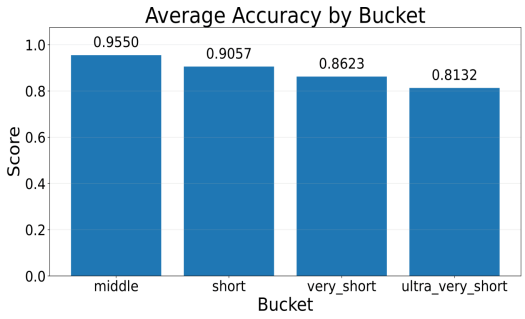
그림 2는 길이 구간별 평균 탐지 성능을 비교한 결과를 보여준다. 댓글 길이가 짧아질수록 전반적인 탐지 성능이 감소하는 경향이 확인되었으며, 모든 지표는 공통적으로 middle에서 가장 높고 ultra_very_short로 갈수록 점차 감소하였다. middle 구간에서는 AUC 0.9906, Average Precision 0.9938, Accuracy 0.9550, F1-score 0.9621로 매우 높은 성능을 보였다. short 구간에서도 AUC 0.9645, Average Precision 0.9775, Accuracy 0.9057, F1-score 0.9206으로 비교적 안정적인 수준이 유지되었다. 반면 very_short와 ultra_very_short 구간에서는 AUC, Average Precision, Accuracy, F1-score가 모두 감소하였으며, ultra_very_short 구간에서 가장 낮은 성능을 나타냈다.

성능 지표는 middle에서 short, very_short, ultra_very_short로 이동할수록 일관되게 하락하는 양상을 보였다. 이는 댓글 길이가 짧아질수록 탐지에 활용 가능한 문체적·확률적 단서가 줄어들고, 그 결과 인간 작성 댓글과 생성 댓글을 구분하는 능력이 약화됨을 시사한다. 다만 이러한 차이는 텍스트 길이 자체의 영향과 함께 생성 댓글의 도메인 자연성 차이가 일부 함께 작용한 결과일 가능성도 있으므로, 길이 구간별 성능 변화는 댓글형 환경에서 Binoculars의 적용 민감도를 보여주는 결과로 해석할 필요가 있다.

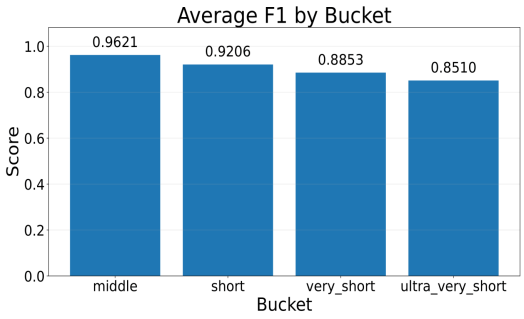


(a) Average AUC by Bucket

(b) Average AP by Bucket



(c) Average Accuracy by Bucket



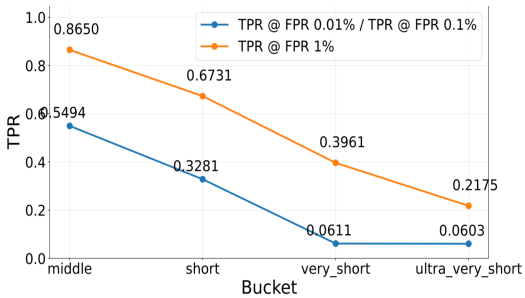
(d) Average F1 by Bucket

(그림 2) 길이 구간별 전체 성능 비교 (AUC, AP, Accuracy, F1)
(Figure 2) Overall Detection Performance by Length Bucket(AUC, AP, Accuracy, F1)

5.2 Low-FPR 조건에서의 성능 변화

낮은 false positive rate 조건에서의 탐지 성능은 평균 성능보다 더욱 큰 폭의 차이를 보였다. TPR@1% FPR 기준으로 middle 구간은 0.8650, short 구간은 0.6731로 나타났으나, very_short 구간에서는 0.3961, ultra_very_short 구간에서는 0.2175로 크게 감소하였다. 또한 0.01% 및 0.1% FPR 조건에서는 middle 구간의 TPR이 0.5494, short 구간의 TPR이 0.3281로 나타난 반면, very_short와 ultra_very_short 구간에서는 각각 0.0611, 0.0603 수준에 머물렀다.

그림 3은 낮은 FPR 조건에서 길이에 따른 성능 차이가 더욱 뚜렷하게 확대됨을 보여준다. middle과 short 구간은 1% FPR 조건에서 비교적 높은 TPR을 유지하지만, very_short 이하에서는 검출력이 급격히 약화된다. 특히 0.01%와 0.1% FPR 조건에서는 very_short 및 ultra_very_short 구간의 TPR이 극히 낮은 수준에 머물러, 초단문 댓글 환경에서는 오탐을 엄격하게 통제할수록 실질적인 검출이 매우 어려워짐을 알 수 있다.



(그림 3) 길이 구간별 낮은 오탐 조건에서의 탐지 성능 비교
(Figure 3) Detection Performance under Low False Positive Rate Conditions by Length Bucket

5.3 결과 해석

이상의 결과는 Binoculars 기반 탐지 성능이 댓글 길이에 따라 일관된 차이를 보이며, 특히 짧은 구간으로 갈수록 그 한계가 더욱 뚜렷해짐을 보여준다. middle과 short 구간에서는 AUC, Average Precision, Accuracy, F1-score가 전반적으로 높은 수준을 유지하였고, 낮은 false positive rate 조건에서도 일정 수준의 검출력이 확보되었다. 반면 very_short 구간부터는 평균 성능이 점차 감소할 뿐 아니라, low-FPR 조건에서의 true positive rate가 더 큰 폭으로 낮아지는 양상이 확인되었다. ultra_very_short 구간에서는 이러한 경향이 더욱 두드러져, 평균 지표상 일정 수준의 분리 가능성이 남아 있더라도 실제 운영 조건에 가까운 낮은 오탐 환경에서는 실질적인 검출력이 크게 제한될 수 있음을 보여준다.

이는 댓글과 같은 short-form 텍스트 환경에서 생성문 탐지 성능을 평가할 때 평균적인 분리 성능만으로는 충분하지 않음을 시사한다[3, 8]. 실제 적용 환경에서는 인간 작성 댓글을 생성문으로 잘못 분류하는 비율을 낮게 유지하는 것이 중요하므로, low-FPR 조건에서 충분한 검출력이 확보되는지가 함께 검토되어야 한다[6, 8]. 본 실험 결과는 Binoculars가 일정 길이 이상의 댓글에서는 비교적 안정적인 탐지 성능을 보일 수 있으나, very_short 이하의 짧은 구간에서는 평균 성능과 운영 관점의 성능 사이에 차이가 커진다는 것을 실험적으로 확인하였다.

다만 본 연구의 결과는 Binoculars를 대상으로 한 실증 결과이므로, 이를 zero-shot 탐지 기법 전반의 일반적 한계로 단정하기는 어렵다. 그러나 DetectGPT, Fast-DetectGPT, Binoculars와 같은 zero-shot 계열 탐지 방식은 입력 텍스트의 확률적·분포적 신호에 의존한다는 공통점이 있다. 따라서 초단문 댓글처럼 토큰 수와 문맥 단서가 제한적인 환경에서는 다른 zero-shot 탐지 방식에서도 유사한 성능 민감도가 나타날 가능성이 있다. 이러한 가능성을 검증하기 위해서는 후속 연구에서 다양한 zero-shot 탐지 기법과의 비교가 필요하다.

6. 결론 및 향후 연구

6.1 결론

본 논문은 한국어 커뮤니티 댓글 환경에서 LLM 생성 댓글에 대한 LLM 기반 탐지 성능과 low-FPR 조건에서의 검출력이 댓글 길이에 따라 어떻게 변화하는지를 분석하였다. 이를 위해 실제 커뮤니티 게시글과 댓글을 기반으로 인간 작성 댓글과 LLM 생성 댓글 데이터를 구성하고, 토큰 수 기준으로 길이 구간을 나누어 탐지 성능을 비교하였다.

실험 결과, middle과 short 구간에서는 검증 대상인 Binoculars가 전반적으로 높은 분리 성능을 유지하였으며, 평균 지표 기준으로도 비교적 안정적인 탐지 결과를 보였다. 반면 very_short 구간부터는 평균 성능이 점차 감소하였고, ultra_very_short 구간에서는 성능 저하가 더욱 뚜렷하게 나타났다. 특히 낮은 false positive rate 조건에서는 이러한 차이가 더 크게 확대되었으며, 초단문 구간으로 갈수록 실제 운영 조건에 가까운 환경에서의 검출력이 약화되는 양상을 확인하였다.

본 결과는 생성문 탐지 성능을 단일한 평균 지표만으로 평가하는 방식에 한계가 있음을 보여준다. 댓글형 생성문 탐지에서는 전체 평균 성능뿐 아니라, 길이 구간별 성능과 low-FPR 조건에서의 검출력을 함께 검토할 필요가 있다. 또한 댓글, 채팅, 짧은 리뷰와 같이 길이가 짧고 비정형적인 텍스트를 대상으로 할 경우, 장문 중심 환경에서 보고된 탐지 성능을 그대로 일반화하기 어렵다는 점을 확인하였다.

다만 본 연구는 단일 한국어 커뮤니티 도메인에서 수집한 댓글과 특정 생성 모델 및 탐지 모델 조합을 대상으로 수행되었으므로, 본 연구의 결과를 모든 한국어 댓글 환경이나 모든 LLM 생성문 탐지 상황으로 일반화하는 데에는 한계가 있다.

6.2 향후 연구

LLM 생성 댓글 탐지 능력 향상을 위해 필요한 추가 연구는 다음과 같이 정리할 수 있다. 첫째, 본 연구의 실험은 각 길이 구간에서 인간 작성 댓글 800개와 생성 댓글 1,200개를 기준으로 수행되었으므로, 데이터 규모와 표본 구성 측면에서 추가 검증이 필요하다. 향후 연구에서는 길이 구간별 표본 수를 확대하고, 다양한 게시글·댓글 조합을 포함하여 결과의 안정성을 검토할 필요가 있다.

둘째, 본 연구는 단일 한국어 커뮤니티 도메인과 특정 생성 모델 및 탐지 모델 조합을 대상으로 수행되었다는 점에서 일반화에 한계가 있다. 정치, 사회, 엔터테인먼트, 제품 리뷰 등 도메인에 따라 댓글의 길이, 어휘, 감정 표현, 밈 사용 양상은 달라질 수 있으며, 이러한 차이는 인간 작성 댓글과 LLM 생성 댓글의 분포 차이에 영향을 줄 수 있다. 따라서 후속 연구에서는 다양한 도메인의 댓

글 데이터를 대상으로 본 연구에서 관찰된 길이별 탐지 성능 저하 경향이 동일하게 나타나는지 검증할 필요가 있다. 또한 EXAONE 외의 상용 LLM이나 다른 오픈소스 LLM을 사용하여 댓글을 생성할 경우, 생성문의 문체와 분포가 달라질 수 있으며, Binoculars 외에 Fast-DetectGPT와 같은 다른 zero-shot 탐지 기법을 적용할 경우에도 길이에 따른 성능 저하 양상이 달라질 가능성이 있다. 이에 따라 다양한 생성 모델과 탐지 기법을 포함한 비교 실험이 필요하다.

셋째, 본 연구의 생성 댓글은 실제 게시글 맥락과 토큰 수 기준을 반영하여 구성되었으나, 일부 문장이 실제 커뮤니티 댓글에 비해 다소 인위적으로 형성되었을 가능성이 있다. 또한 생성 과정에서 커뮤니티 특유의 은어, 밈, 축약 표현, 구어체 표현, 반응 관습과 같은 도메인 특성이 충분히 반영되지 못하였을 가능성이 있다. 후속 연구에서는 커뮤니티 문체를 강화한 생성 조건을 추가하여, 일반 생성 조건과 문체 강화 생성 조건 사이에서 탐지 성능이 어떻게 달라지는지 비교할 필요가 있다. 마지막으로, 서로 다른 토크나이저와 언어모델 조합을 적용하여 초단문 댓글에서 토큰 분할 방식이 탐지 성능에 미치는 영향도 함께 분석할 필요가 있다.

참고문헌(Reference)

[1] J. Wu, S. Yang, R. Zhan, Y. Yuan, L. S. Chao, and D. F. Wong, “A Survey on LLM-Generated Text Detection: Necessity, Methods, and Future Directions,” Computational Linguistics, vol.51, no.1, pp.275-338, Mar. 2025. https://doi.org/10.1162/coli_a_00549

[2] W. Meng, J. Harvey, J. Goulding, C. J. Carter, E. Lukinova, A. Smith, P. Frobisher, M. Forrest, and G. Nica-Avram, “Large Language Models as ‘Hidden Persuaders’: Fake Product Reviews are Indistinguishable to Humans and Machines,” arXiv preprint, arXiv:2506.13313, 2025. https://doi.org/10.48550/arXiv.2506.13313

[3] W. Go, H. Kim, A. Oh, and Y. Kim, “XDAC: XAI-Driven Detection and Attribution of LLM-Generated News Comments in Korean,” in Proc. of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp.22728-22750, 2025. http://doi.org/10.18653/v1/2025.acl-long.1108

[4] C. Rodríguez, “Beyond the Meme: Far-Right Radicalism and Its Online Propaganda,” Revista Internacional de Estudios sobre Terrorismo, vol.V, no.12, pp.20-32, 2024. https://voxpol.eu/file/beyond-the-meme-far-right-radicalism-and-its-online-propaganda/

[5] E. Mitchell, Y. Lee, A. Khazatsky, C. D. Manning, and C. Finn, “DetectGPT: Zero-Shot Machine-Generated Text Detection using Probability Curvature,” in Proc. of the 40th International Conference on Machine Learning, vol.202, pp.24950-24962, 2023. https://dl.acm.org/doi/10.5555/3618408.3619446

[6] A. Hans, A. Schwarzschild, V. Cherepanova, H. Kazemi, A. Saha, M. Goldblum, J. Geiping, and T. Goldstein, “Spotting LLMs With Binoculars: Zero-Shot Detection of Machine-Generated Text,” in Proc. of the 41st International Conference on Machine Learning, vol.235, pp.17519-17537, 2024. https://dl.acm.org/doi/10.5555/3692070.3692768

[7] G. Bao, Y. Zhao, Z. Teng, L. Yang, and Y. Zhang, “Fast-DetectGPT: Efficient Zero-Shot Detection of Machine-Generated Text via Conditional Probability Curvature,” in Proc. of the Twelfth International Conference on Learning Representations, 2024. https://doi.org/10.48550/arXiv.2310.05130

[8] X. Zhu, Y. Ren, Y. Cao, X. Lin, F. Fang, and Y. Li, “Reliably Bounding False Positives: A Zero-Shot Machine-Generated Text Detection Framework via Multiscaled Conformal Prediction,” in Proc. of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp.12298-12319, 2025. http://doi.org/10.18653/v1/2025.acl-long.601

[9] Gemma Team, “Gemma: Open Models Based on Gemini Research and Technology,” arXiv preprint, arXiv:2403.08295, 2024. http://doi.org/10.48550/arXiv.2403.08295

[10] T. Kudo and J. Richardson, “SentencePiece: A Simple and Language Independent Subword Tokenizer and Detokenizer for Neural Text Processing,” in Proc. of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pp.66-71, 2018. http://doi.org/10.18653/v1/D18-2012

◐ 저 자 소 개 ◑

류 동 훈(Dong-Hoon Ryu)
2025년 홍익대학교 소프트웨어융합학과(공학사)
2025년 3월～현재 홍익대학교 대학원 소프트웨어융합학과 석사과정
관심분야 : LLM, LLM 탐지, 생성형 AI
E-mail : ryudong1104@g.hongik.ac.kr

노 기 섭(Giseop Noh)
1998년 공군사관학교 산업공학과(공학사)
2009년 콜로라도대학교 대학원 컴퓨터학과(공학석사)
2014년 서울대학교 대학원 컴퓨터공학과(공학박사)
2016년 12월～2018년 2월 공군사관학교 전산학과 조교수
2018년 3월～2025년 2월 청주대학교 인공지능SW학과 조교수
2025년 3월～현재 홍익대학교 SW융합학과 조교수
관심분야 : 자연어처리, 신경망 이론, 강화학습
E-mail : kafa46@hongik.ac.kr