문장 길이에 따른 LLM 생성 스팸 탐지 효과성 연구

Binoculars 기법을 활용한 *LLM* 생성문 탐지 성능 분석

발표자

류 동훈

DeepShark Lab/Hongik University

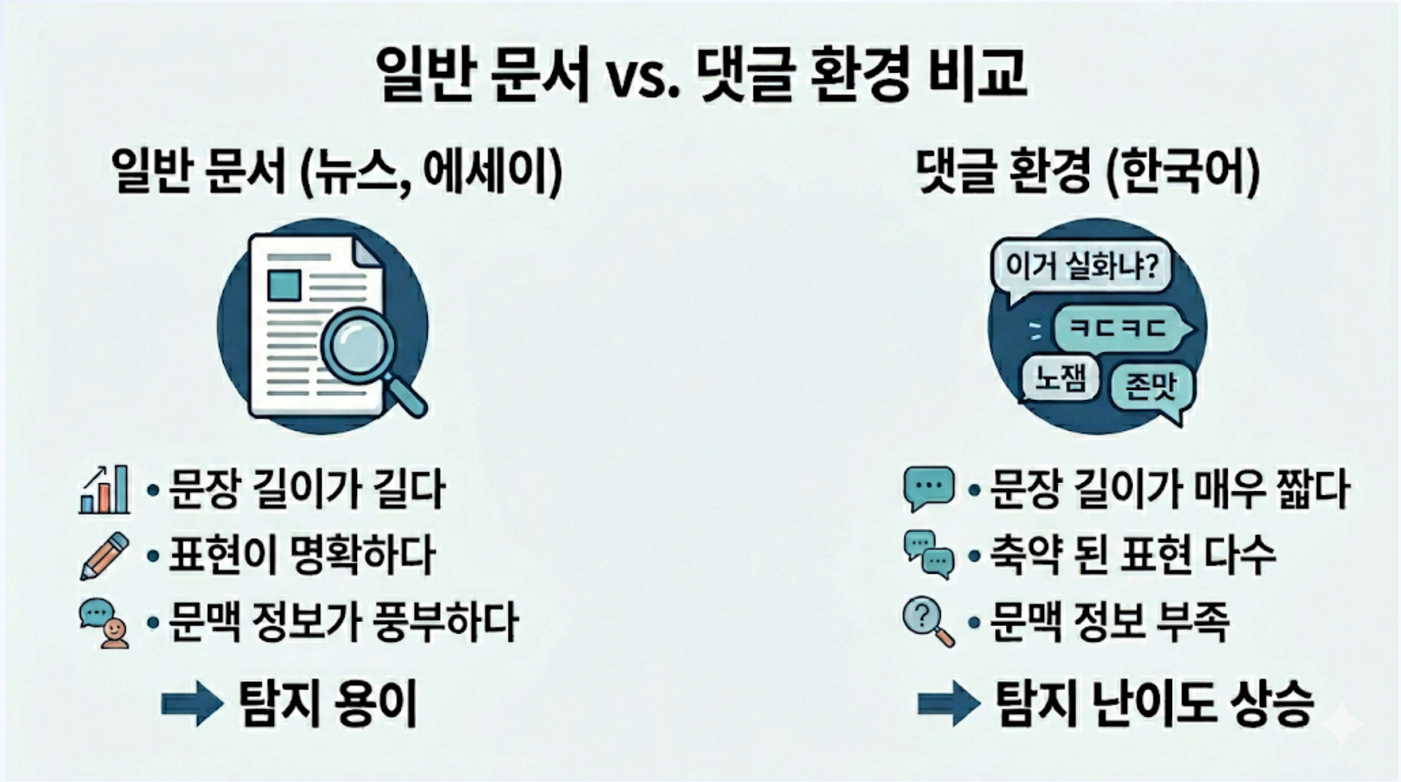
Date:

2026-04-10

목차

연구 배경

- LLM 생성문 탐지는 일반 문서보다 댓글 환경에서 더 어려움
- 댓글은 문장 길이가 매우 짧고, 표현이 축약되어 있으며, 문맥 정보도 부족함
- 따라서 장문 기준에서 유효한 탐지 기법이 짧은 댓글에서도 그대로 동작한다고 보기 어려움



연구 배경

- Binoculars는 zero-shot 기반 탐지 기법으로, perplexity와 cross-perplexity를 함께 활용하는 방식임
- 참조 논문은 주로 뉴스, 에세이, 창작문 등 상대적으로 길이가 있는 텍스트를 중심으로 성능을 평가함
- 본 연구의 관심 대상은 한국어 댓글 데이터이며, 특히 짧은 문장 길이에서 탐지 성능이 언제부터 붕괴하는지를 확인하는 데 있음
- 본 실험은 댓글 기반 데이터에서 문장 길이별 탐지 성능 변화를 확인하고, 어느 길이 구간부터 Binoculars 계열 분석력이 약화되는지를 점검하기 위해 수행함

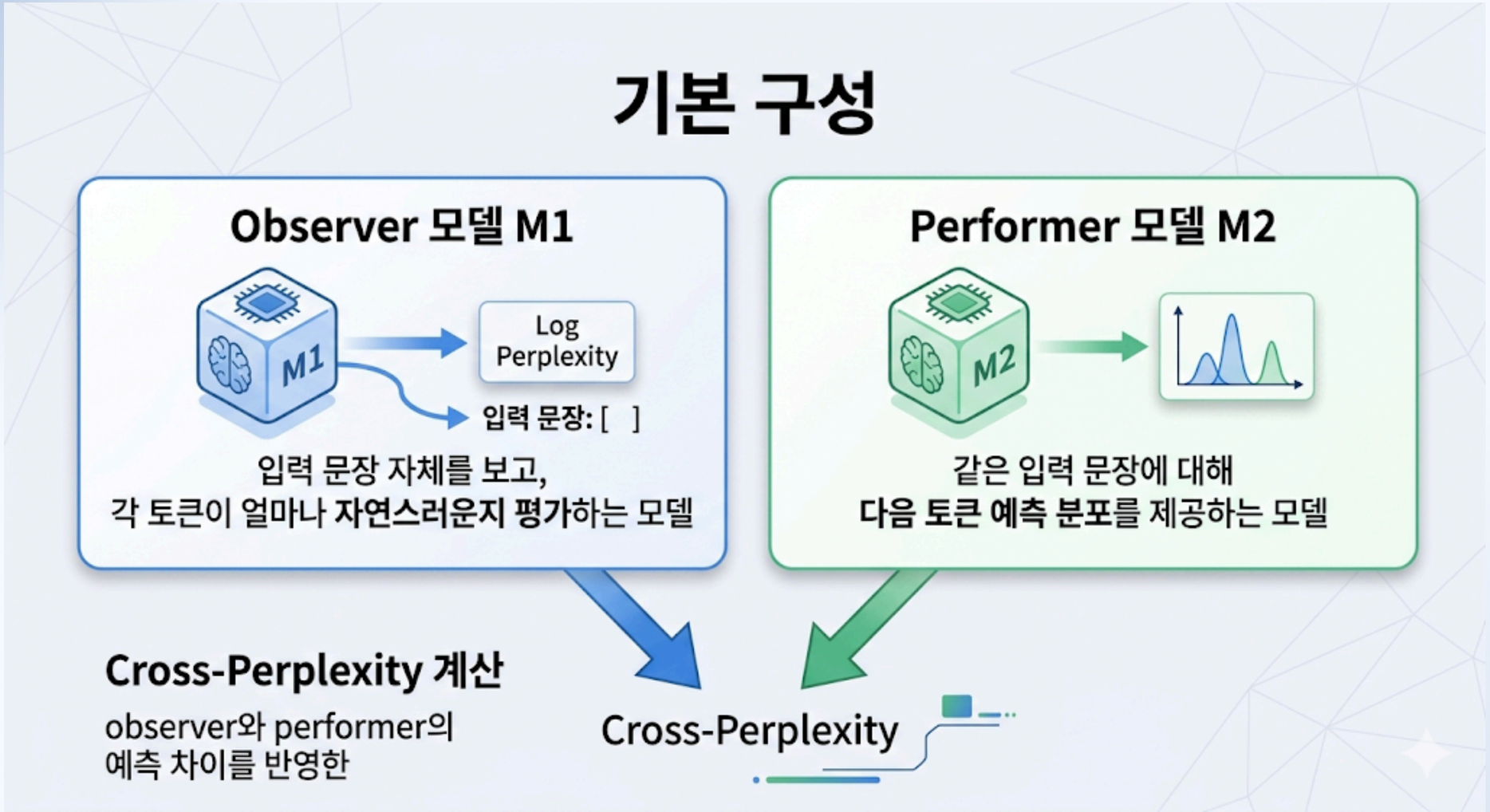
참조 논문

Hans et al., “Spotting LLMs With Binoculars: Zero-Shot Detection of Machine-Generated Text” ICML 2024.

Binoculars 개념

- Binoculars는 두 개의 유사한 언어모델을 함께 사용하여, 입력 문장이 인간 작성문인지 LLM 생성문인지 판별하는 zero-shot 탐지 기법이다.
- 핵심은 단순히 한 모델이 문장을 얼마나 자연스럽게 보는지만 확인하는 것이 아니라, 서로 가까운 두 모델이 그 문장을 얼마나 비슷하게 바라보는지까지 함께 측정하는 데 있다.
- 논문에서는 이를 위해 perplexity와 cross-perplexity를 함께 사용하고, 최종적으로 두 값의 비율을 Binoculars score로 정의한다.

Binoculars 개념



Binoculars 개념

- **Log Perplexity**
  - 문장이 언어모델에게 얼마나 자연스럽게 보이는지를 수치화한 값
  - 입력 문장 s를 토큰 시퀀스 x=(x1,x2,⋯,xL)로 두면, observer 모델 M1의 Log Perplexity는 다음과 같이 정의됨

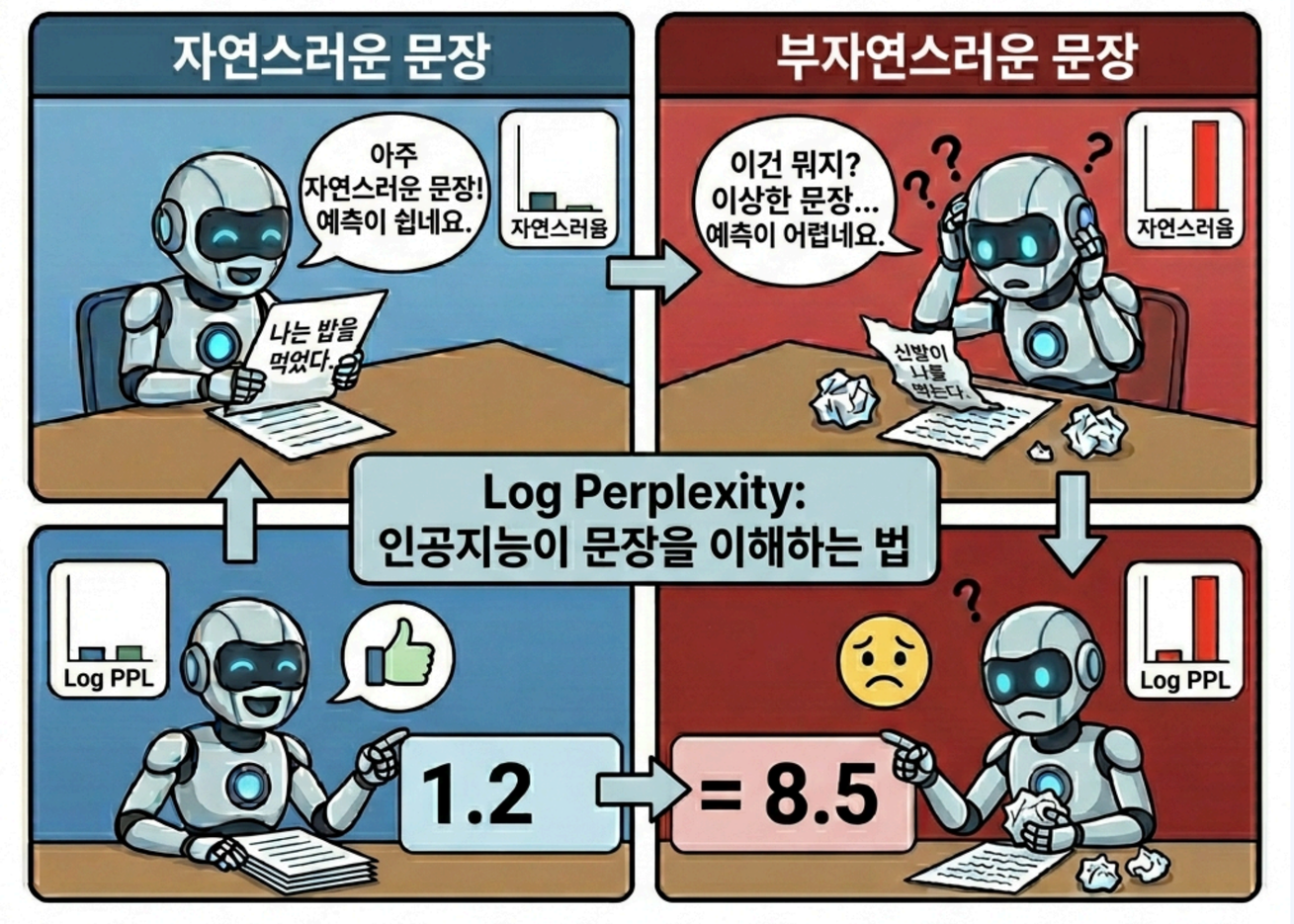
$$\log PPL_{M_1}(s) = -\frac{1}{L}\sum_{i=1}^{L}\log\left(P(x_i \mid x_{0:i-1})\right)$$

  - L: 전체 토큰 길이
  - xi: i번째 실제 토큰
  - P(xi ∣ x0:i−1): 이전 문맥이 주어졌을 때, 모델이 실제 다음 토큰 xi에 부여한 확률
  - 전체 토큰에 대해 평균 음의 로그 확률을 계산한 값

  - 모델이 다음 토큰을 쉽게 예측할수록 Log Perplexity는 작아짐
  - 모델이 문장을 낯설게 느낄수록 Log Perplexity는 커짐
  - 즉, 값이 작을수록 모델 입장에서 자연스러운 문장, 값이 클수록 모델 입장에서 덜 자연스러운 문장으로 해석할 수 있음

Binoculars 개념

- Log Perplexity



Binoculars 개념

- **Cross-Perplexity**
  - Cross-Perplexity는 두 언어모델이 같은 입력 문장을 얼마나 비슷하게 해석하는지를 나타내는 지표
  - Binoculars에서는 observer 모델 M1과 performer 모델 M2를 사용하며, cross-perplexity는 다음과 같이 정의됨

$$\log X\text{-}PPL_{M_1,M_2}(s) = -\frac{1}{L}\sum_{i=1}^{L} M_1(s)_i \cdot \log(M_2(s)_i)$$

- L: 전체 토큰 길이
- M1(s)i: i번째 위치에서 observer 모델이 예측한 다음 토큰 확률분포
- M2(s)i: i번째 위치에서 performer 모델이 예측한 다음 토큰 확률분포
- 두 모델의 다음 토큰 분포 사이의 평균 cross-entropy를 계산한 값

- 두 모델이 비슷한 확률분포를 내면 Cross-Perplexity는 작아짐
- 두 모델이 서로 다르게 반응하면 Cross-Perplexity는 커짐
- 즉, 값이 작을수록 두 모델이 이 문장을 유사하게 이해하고 있음을 의미한다

Binoculars 개념

- Cross-Perplexity

Binoculars 개념

- **Binoculars Score**
  - Binoculars는 입력 문장 s에 대해, Log Perplexity와 Cross-Perplexity의 비율을 최종 탐지 점수로 사용함
  - 최종 점수는 다음과 같이 정의함

$$B_{M_1,M_2}(s) = \frac{\log PPL_{M_1}(s)}{\log X\text{-}PPL_{M_1,M_2}(s)}$$

  - 분자: observer 모델이 입력 문장 자체를 얼마나 자연스럽게 보는지 측정(Log Perplexity)
  - 분모: observer와 performer가 같은 문장을 얼마나 비슷하게 해석하는지 반영(Cross-Perplexity)

  - 이 문장이 그냥 자연스러운 문장인지, 아니면 모델들이 공통적으로 익숙해하는 '모델스러운 문장'인지를 구분하는 점수다.
  - 값이 낮으면 생성문 쪽, 값이 높으면 인간 작성문 쪽으로 해석한다.

실험 설계

논문 재현 실험

- **목적**
  - Binoculars 기법의 기본 동작 여부를 원 논문의 흐름과 유사한 환경에서 확인
  - 이후 한국어 댓글 도메인 확장 실험의 기준 성능 확보
- **데이터**
  - CNN 뉴스 데이터
  - 인간 작성 뉴스 문장과 Falcon 기반 생성문으로 구성
- **생성 모델**
  - Falcon
- **분석 모델**
  - Falcon 기반 Binoculars
- **평가 내용**
  - 기본 탐지 성능 재현 가능성 점검
- **설계 의의**
  - 뉴스 기반 영어 데이터에서 먼저 기초 성능을 확인한 뒤
  - 댓글 기반 한국어 실험으로 확장하기 위한 기준선 실험

실험 설계

- **논문 재현 기반 파생 실험**
  - **목적**
    - 문장 길이에 따라 Binoculars 성능이 어떻게 달라지는지 확인
    - 짧은 문장으로 갈수록 탐지 성능이 약화되는지 점검
  - **데이터**
    - CNN 뉴스 데이터
    - 인간 작성 뉴스 문장과 Falcon 기반 생성문으로 구성
  - **생성 모델**
    - Falcon
  - **분석 모델**
    - Falcon 기반 Binoculars
  - **길이 구간**
    - medium(81 ~ 160 토큰)
    - short(41 ~ 80 토큰)
    - very short(20 ~ 40 토큰)
  - **평가 내용**
    - 길이 구간별 score 분포 비교
    - 길이별 분리 성능 변화 확인
    - 어느 구간에서부터 성능 저하가 나타나는지 분석
  - **설계 의의**
    - 이후 댓글 탐지 실험으로 확장하기 전에
    - 문장 길이 자체가 Binoculars 성능에 주는 영향을 먼저 확인하는 파생 실험

실험 설계

- **Falcon 생성문 → Gemma 기반 분석 모델**
  - **목적**
    - 생성 모델과 분석 모델이 서로 다를 때도 탐지 성능이 유지되는지 확인
    - 동일 모델 계열 의존성을 점검
  - **데이터**
    - CNN 뉴스 데이터
    - Falcon이 생성한 뉴스 문장 사용
  - **생성 모델**
    - Falcon
  - **분석 모델**
    - Gemma 기반 Binoculars
  - **길이 구간**
    - medium(81 ~ 160 토큰)
    - short(41 ~ 80 토큰)
    - very short(20 ~ 40 토큰)
  - **평가 내용**
    - Falcon 생성문에 대한 Gemma 분석 score 분포 확인
    - 인간 뉴스 문장과의 분리 정도 비교
    - 생성 모델–분석 모델 불일치 상황에서의 성능 변화 확인
  - **설계 의의**
    - Binoculars의 교차 모델 일반화 가능성 검토

실험 설계

- **Gemma 생성문 → Falcon 기반 분석 모델**
  - **목적**
    - 반대 방향 교차 분석을 통해 성능 변화 양상 확인
    - 생성 모델과 분석 모델 조합에 따른 편차 검토
  - **데이터**
    - CNN 뉴스 데이터
    - Gemma가 생성한 뉴스 문장 사용
  - **생성 모델**
    - Gemma
  - **분석 모델**
    - Falcon 기반 Binoculars
  - 길이 구간
    - medium(81 ~ 160 토큰)
    - short(41 ~ 80 토큰)
    - very short(20 ~ 40 토큰)
  - **평가 내용**
    - Gemma 생성문에 대한 Falcon 분석 score 분포 확인
    - 인간 뉴스 문장과의 분리 정도 비교
    - Falcon→Gemma와 Gemma→Falcon 결과 비교
  - **설계 의의**
    - 모델 조합 변화에 따른 탐지 성능 차이 확인

실험 설계

EXAONE 생성문 + Falcon/Gemma 분석 모델 비교 (medium / short / very short)

- **목적**
  - 동일한 EXAONE 생성문에 대해 분석 모델을 Falcon과 Gemma로 달리 적용했을 때, 탐지 성능 차이가 어떻게 나타나는지 확인
  - 분석 모델 자체의 차이가 한국어 댓글 환경에서 어떤 영향을 주는지 비교
- **데이터**
  - 한국 커뮤니티 게시글 기반 댓글 데이터
  - 게시글을 프롬프트로 하여 EXAONE이 생성한 댓글 사용
- **생성 모델**
  - EXAONE
- **분석 모델**
  - Falcon 기반 Binoculars
  - Gemma 기반 Binoculars
- **길이 구간**
  - medium(81 ~ 160 토큰)
  - short(41 ~ 80 토큰)
  - very short(20 ~ 40 토큰)
- **평가 내용**
  - 동일 생성문에 대해 Falcon 분석 score와 Gemma 분석 score 비교
  - Human / Machine score 분포 차이 비교
  - 길이 구간별 AUROC, Accuracy, F1 등 성능 비교
  - 어떤 분석 모델이 더 안정적으로 score gap을 확보하는지 확인
- **설계 의의**
  - 한국어 댓글 환경에서 더 적합한 분석 모델 확인

실험 설계

- **EXAONE 생성문 + Gemma 분석 모델 (medium / short / very short / ultra short)**
  - **목적**
    - ultra short 구간을 추가하여 초단문 댓글에서의 분석 한계 확인
  - **데이터**
    - 한국 커뮤니티 게시글 기반 댓글 데이터
    - 게시글을 프롬프트로 하여 EXAONE이 생성한 댓글 사용
  - **생성 모델**
    - EXAONE
  - **분석 모델**
    - Gemma 기반 Binoculars
  - **길이 구간**
    - medium(81 ~ 160 토큰)
    - short(41 ~ 80 토큰)
    - very short(20 ~ 40 토큰)
    - ultra short(5 ~ 19 토큰)
  - **평가 내용**
    - ultra short 추가 시 score 분포 겹침 정도 확인
    - 기존 3구간 실험 대비 성능 하락 폭 비교
    - 초단문 구간의 실질적 판별 가능성 검토
  - **설계 의의**
    - 댓글 기반 탐지의 최소 분석 가능 길이 기준 탐색

사용 모델 및 데이터 설명

- **사용 모델**
  - **Falcon**
    - Falcon-7B 계열 모델
    - Binoculars 논문 재현 실험에서 사용한 기준 모델
    - CNN 뉴스 데이터를 기반으로 생성문 생성,분석 모델 적용, 성능 비교에 활용
    - 의미
      - 영어 뉴스 도메인에서 Binoculars 계열이 기본적으로 동작하는지 확인하는 기준선 모델
      - 이후 Gemma와의 교차 분석 결과를 비교하기 위한 reference 모델

사용 모델 및 데이터 설명

- **사용 모델**
  - **Gemma**
    - Gemma 2 계열 모델을 Binoculars 분석 모델로 사용
    - 실제 사용 구성
      - Tokenizer / 길이 측정: google/gemma-2-9b-it
      - Observer 모델: google/gemma-2-2b
      - Performer 모델: google/gemma-2-9b
    - 활용 방식
      - Falcon 생성문에 대한 교차 분석 모델
      - EXAONE 기반 한국어 댓글 생성문에 대한 주요 분석 모델
      - Gemma 기반 Binoculars 조합으로 인간 댓글과 생성 댓글의 분리 가능성을 확인

사용 모델 및 데이터 설명

- **사용 모델**
  - **EXAONE**
    - 실제 사용 버전
      - exaone3.5:7.8b
    - 활용 방식
      - 한국어 게시글을 입력으로 받아 댓글 생성
      - 생성된 댓글을 이후 Gemma 기반 Binoculars로 분석
    - 의미
      - 영어 뉴스가 아니라 한국어 커뮤니티 댓글 환경을 반영하기 위한 생성 모델
      - 특히 짧은 댓글, 매우 짧은 댓글, 초단문 댓글 구간에서
      - 길이에 따라 탐지 성능이 언제 약화되는지 확인하는 데 사용
    - 선택 이유
      - LG에서 개발한 한국어 특화 모델로 댓글 생성에 적합한 모델로 선정
      - 실제 한국어 게시판 환경과 유사한 생성문 확보를 위한 목적

EXAONE

사용 모델 및 데이터 설명

- **데이터 설명**
  - **CNN 뉴스 데이터**
    - 영어 뉴스 기사 본문 데이터
    - 논문 재현 실험과 Falcon/Gemma 교차 분석 실험에 사용
    - 총 1,873건의 기사로 구성
    - 주요 활용 컬럼
      - article: 인간 작성 기사 본문
      - article_token_len: 기사 길이 정보
      - prompt_50tok: 생성문 생성을 위한 프롬프트 구간
    - 의미
      - 비교적 길고 정제된 문체의 영어 장문 데이터
      - Binoculars 계열의 기준선 성능 확인용 데이터

사용 모델 및 데이터 설명

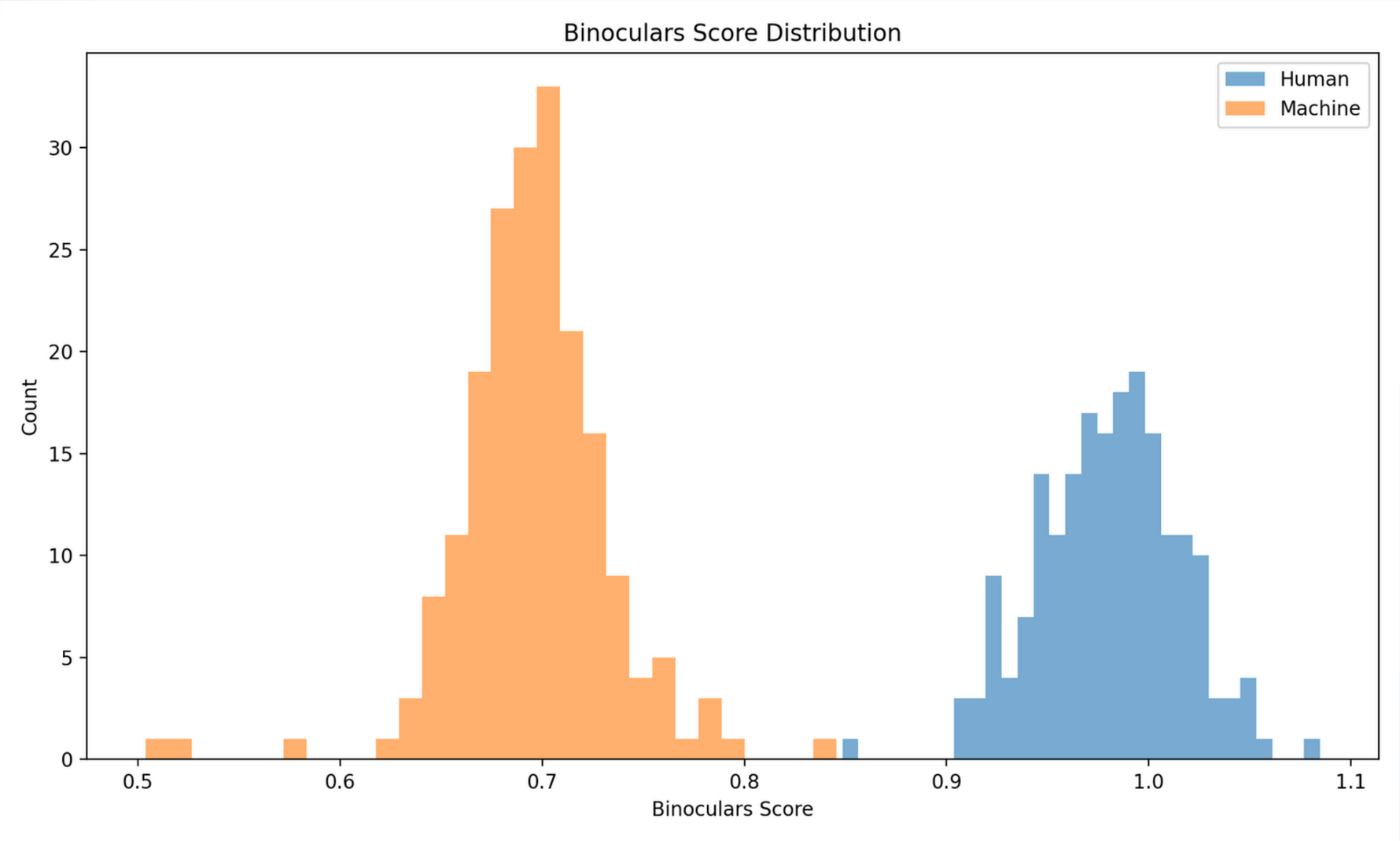
- **데이터 설명**
  - **식물 갤러리 게시글 데이터**
    - 디시인사이드 식물 갤러리 게시글 데이터
    - EXAONE 기반 댓글 생성 실험의 입력 데이터로 사용
    - 총 6,363건의 게시글로 구성
    - 주요 활용 컬럼
      - title: 게시글 제목
      - content: 게시글 본문
      - actual_post_datetime: 게시 시각
      - content_len: 게시글 길이
    - 의미
      - 한국어 커뮤니티 게시판의 실제 게시글 환경을 반영
      - 댓글 생성에 필요한 원문 문맥 데이터

사용 모델 및 데이터 설명

- **데이터 설명**
  - **식물 갤러리 댓글 데이터**
    - 디시인사이드 식물 갤러리 실제 사용자 댓글 데이터
    - 생성 댓글과 비교할 인간 댓글 데이터로 사용
    - 총 40,249건의 댓글로 구성
    - 주요 활용 컬럼
      - comment: 실제 댓글 본문
      - comment_len_with_space: 공백 포함 길이
      - comment_len_no_space: 공백 제외 길이
      - comment_is_ultrashort: 초단문 여부
    - 의미
      - 짧고 비정형적인 실제 한국어 댓글 환경 반영
      - 길이별 탐지 성능 변화와 성능 붕괴 시점 분석에 사용

실험 결과

- 논문 재현 실험

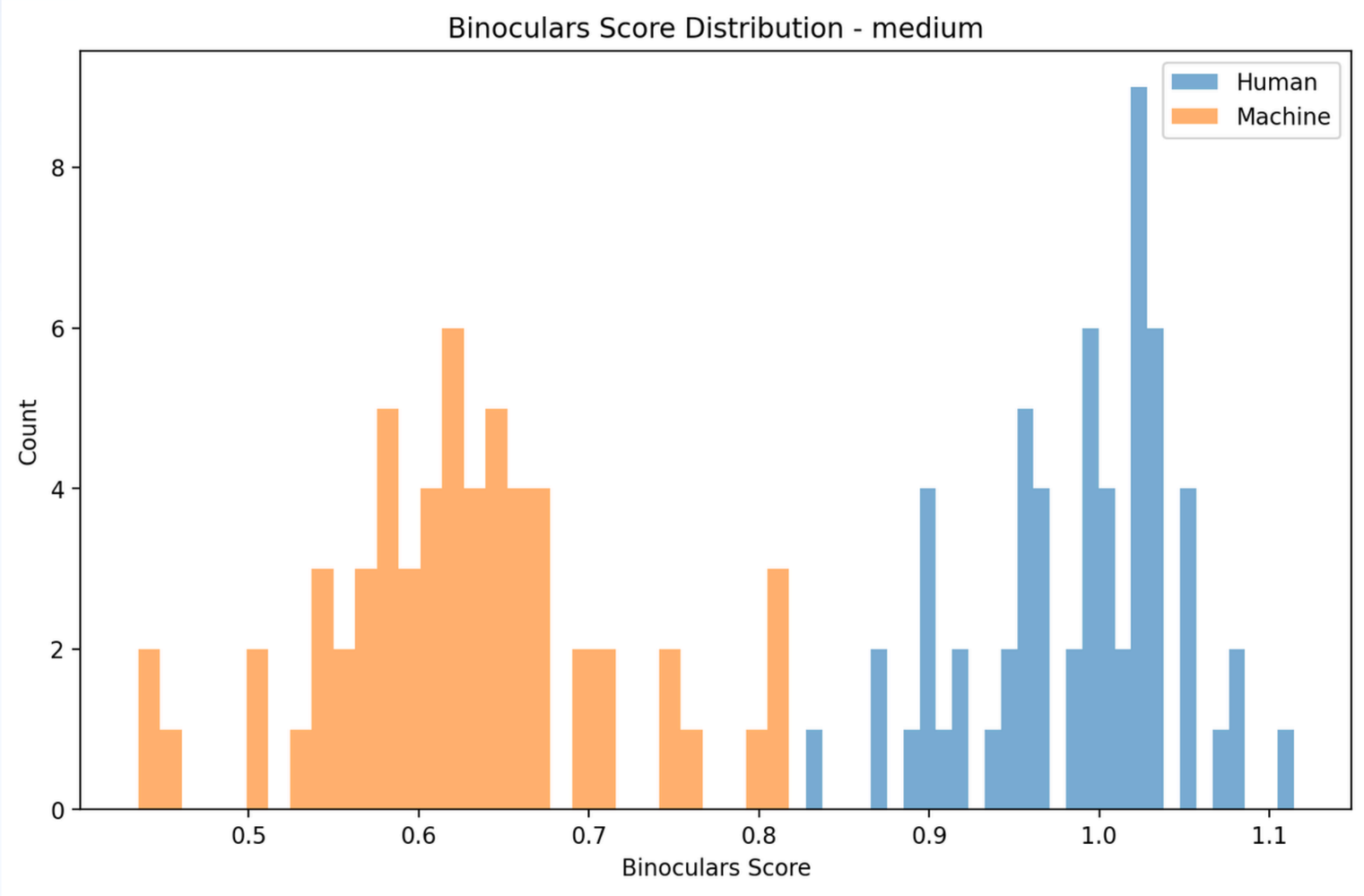

Binoculars Score Distribution
Human
Machine
Count
0
5
10
15
20
25
30
0.5
0.6
0.7
0.8
0.9
1.0
1.1
Binoculars Score

실험 결과

- **논문 재현 실험-종합 해석**
  - Human / Machine score 분포가 전반적으로 명확하게 분리됨
  - Human은 1.0 부근, Machine은 0.7 부근에 집중
  - 분포 간 겹침이 매우 작아 기본적인 재현 가능성 확인
  - 뉴스 도메인에서는 Binoculars가 유효하게 동작함을 확인
  - CNN 뉴스 도메인에서는 Binoculars score만으로도 Human과 Machine이 뚜렷하게 구분되며, 논문 재현의 기본 가능성을 확인하였음

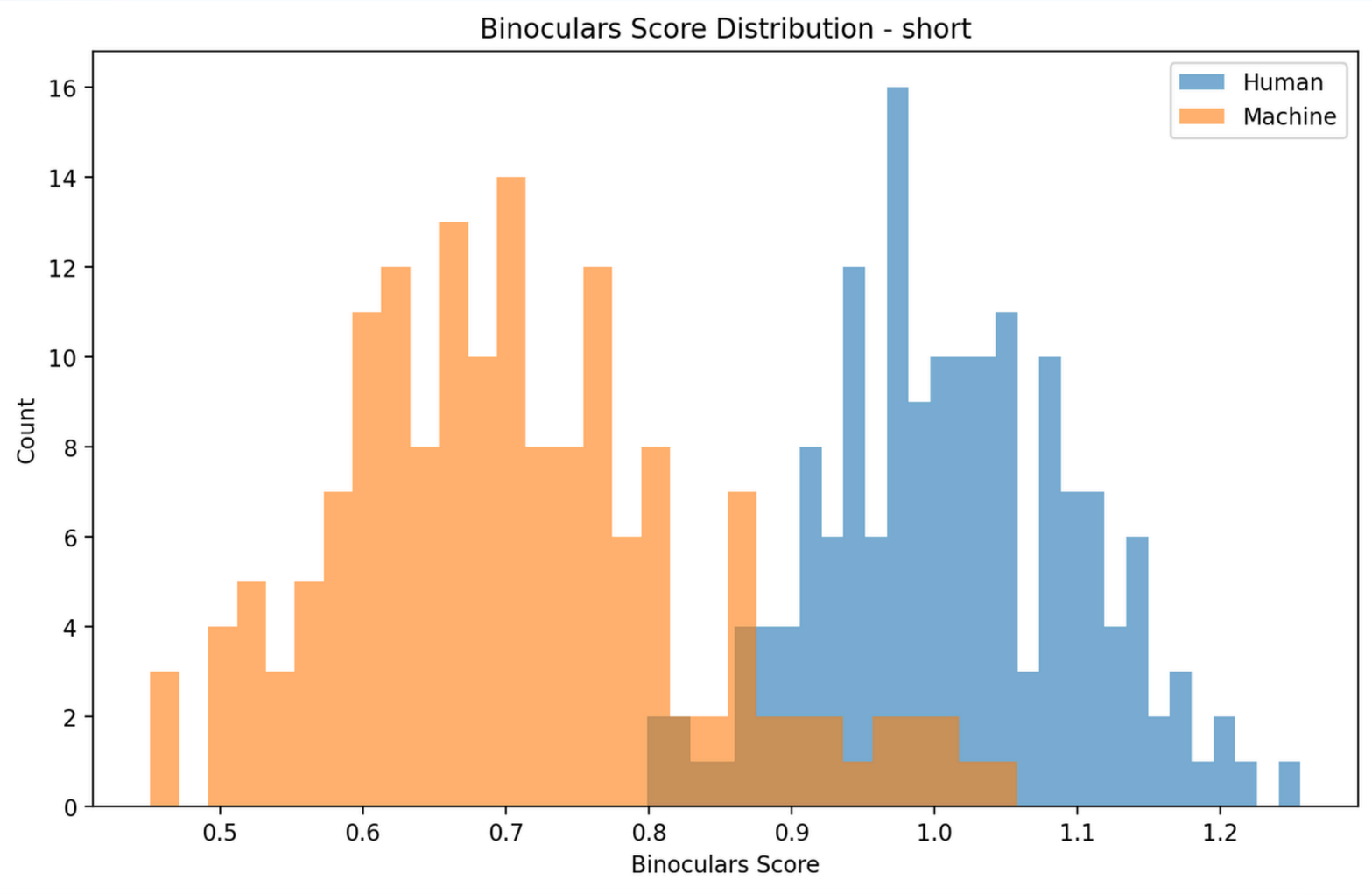
실험 결과

- 논문 재현 파생 실험-medium



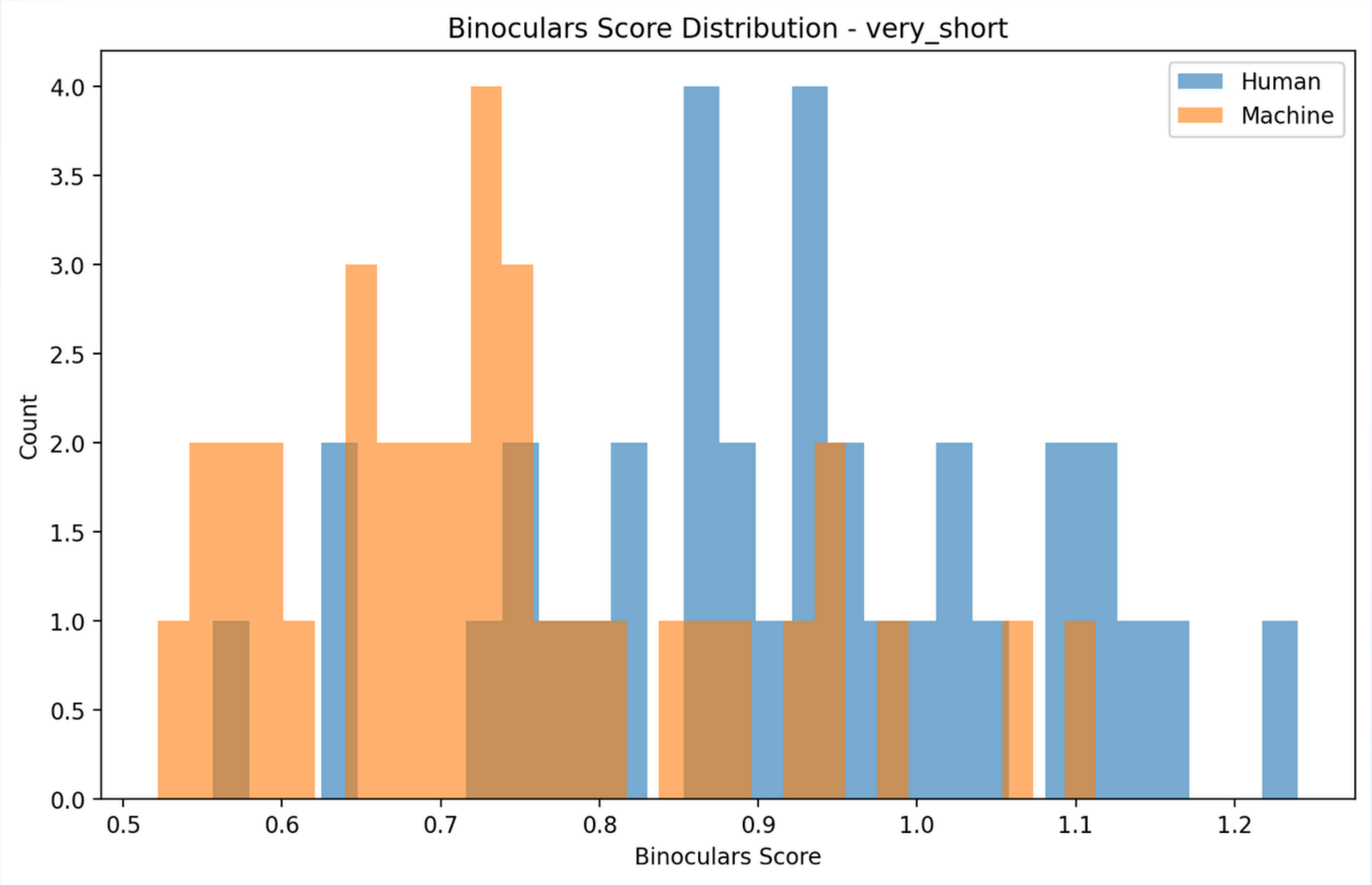
실험 결과

- 논문 재현 파생 실험-short



실험 결과

- 논문 재현 파생 실험-very short

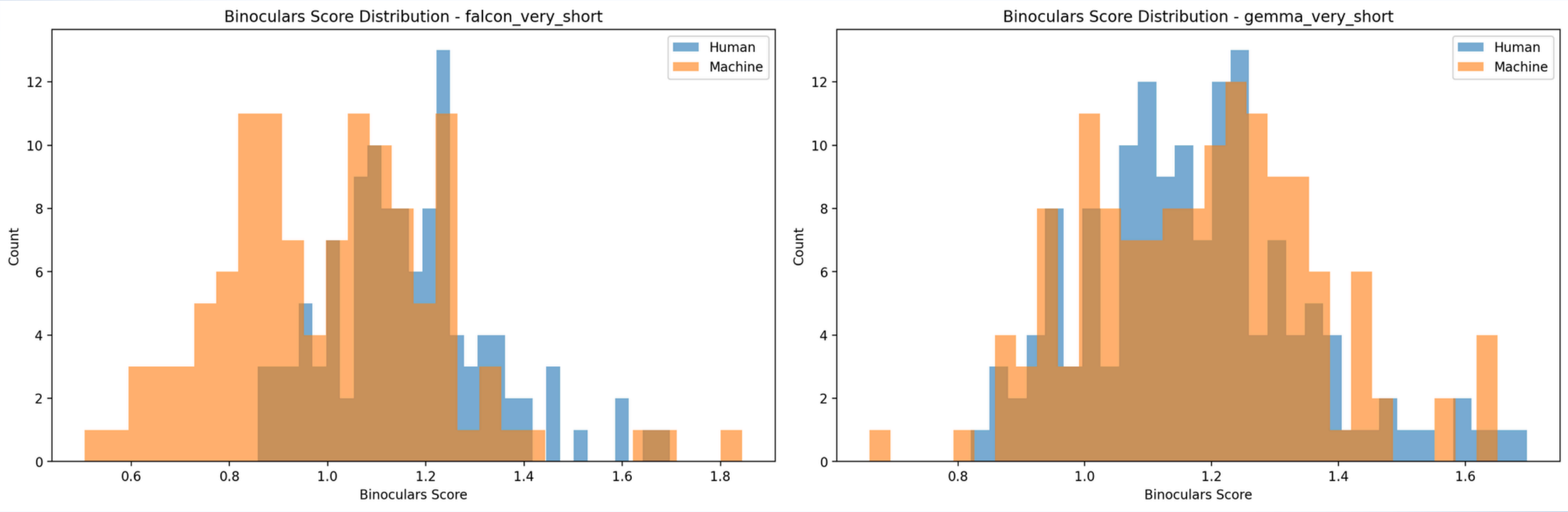


실험 결과

- **논문 재현 파생 실험-종합 해석**
  - 전 구간에서 Human score > Machine score
  - 길이가 길어질수록 score gap 확대
  - VERY SHORT에서 성능 저하가 가장 큼
  - SHORT부터 성능이 크게 개선
  - MEDIUM에서는 거의 완전 분리
  - 즉, Binoculars 성능은 문장 길이에 크게 의존함

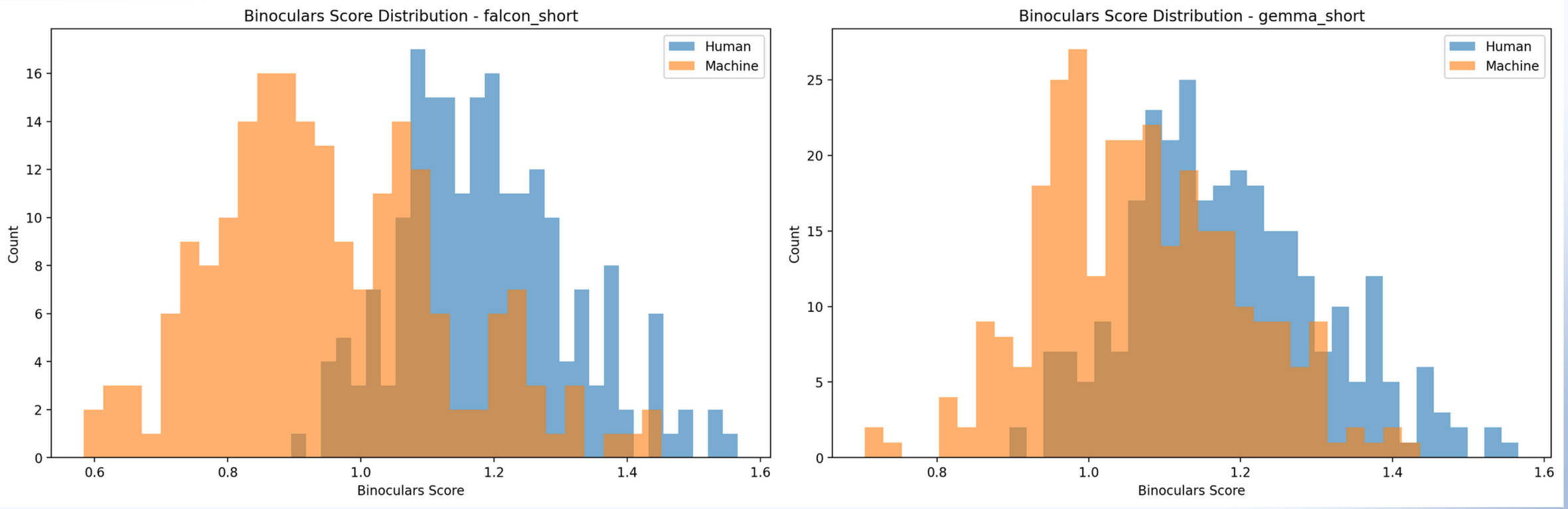
실험 결과

- Falcon 생성문 → Gemma 기반 분석 모델/Gemma 생성문 → Falcon 기반 분석 모델
- very short 비교



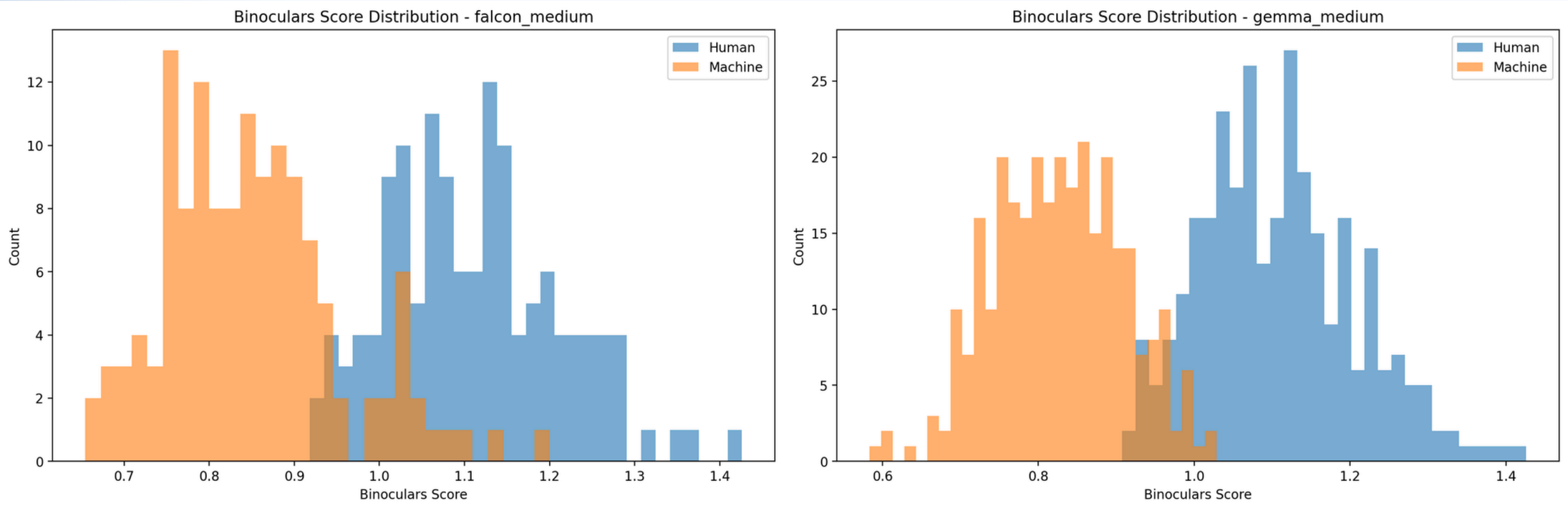
실험 결과

- Falcon 생성문 → Gemma 기반 분석 모델/Gemma 생성문 → Falcon 기반 분석 모델
- short 비교



실험 결과

- Falcon 생성문 → Gemma 기반 분석 모델/Gemma 생성문 → Falcon 기반 분석 모델
- medium 비교

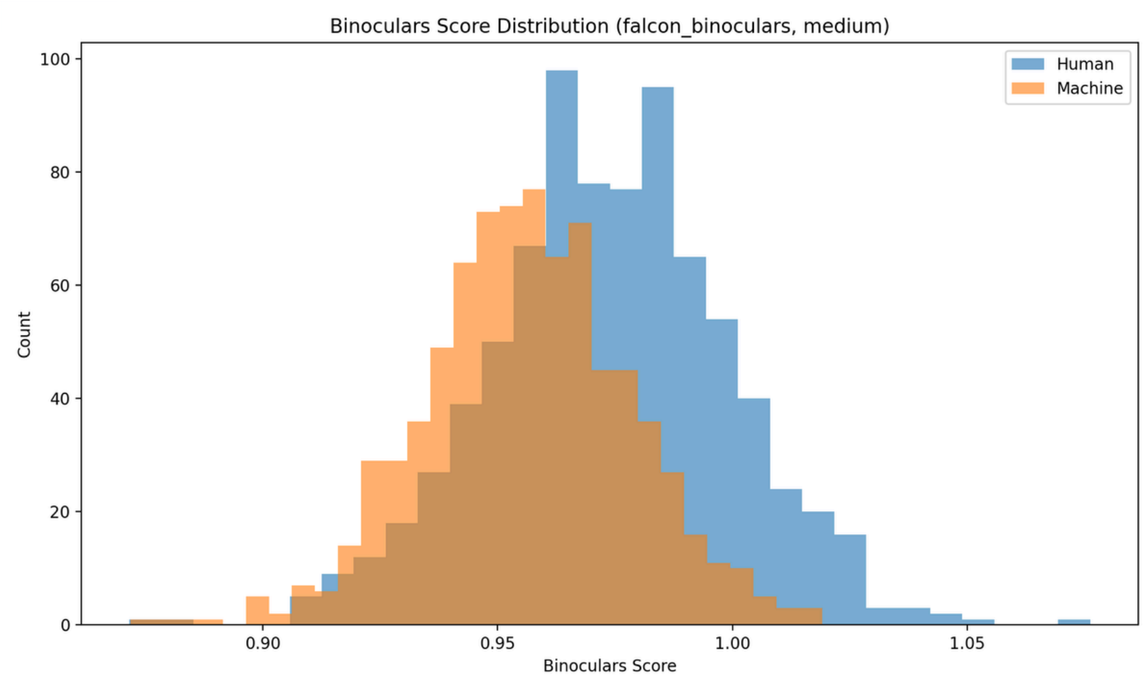


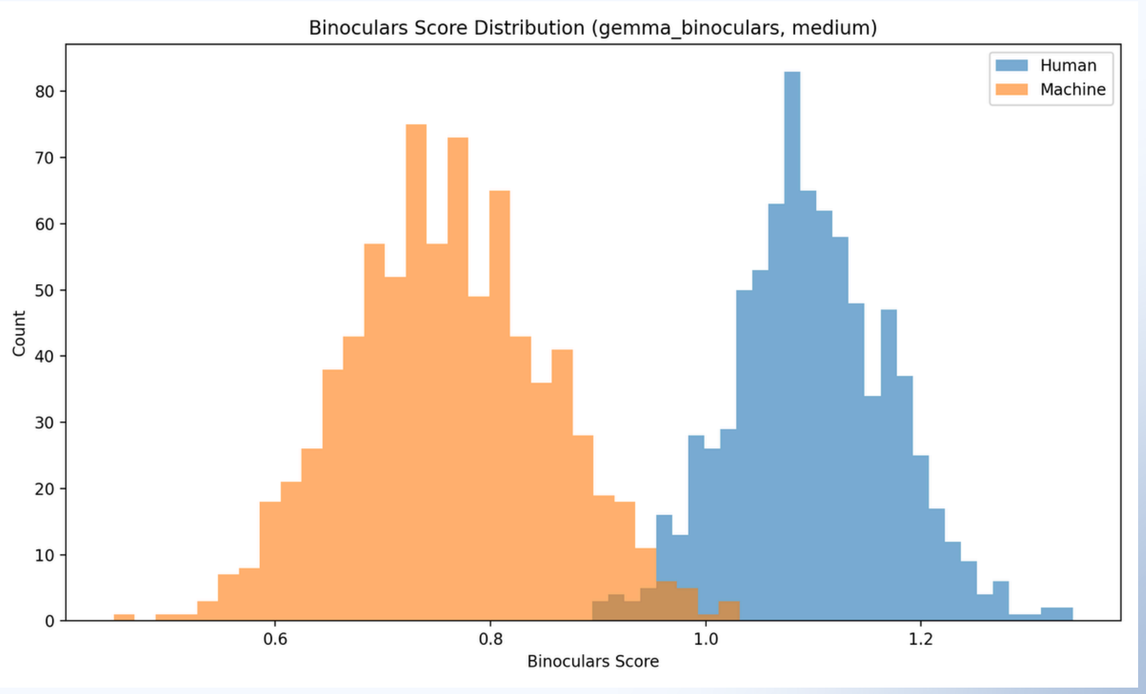
실험 결과

- **Falcon 생성문 → Gemma 기반 분석 모델/Gemma 생성문 → Falcon 기반 분석 모델**
- **종합 해석**
  - 두 조합 모두 길이가 길수록 성능 향상
  - 그러나 전 구간에서 Falcon 생성문 → Gemma 분석이 우세
  - Gemma 생성문 → Falcon 분석은 score gap이 거의 형성되지 않음
  - 성능 차이는 결국 Human–Machine score gap 차이로 설명 가능
  - 즉, 교차 모델 분석은 가능하지만 조합 비대칭성이 크게 나타남
  - Falcon→Gemma 조합만 강하게 동작했고 Gemma→Falcon 조합은 전반적으로 약해 모델 조합 비대칭성이 확인되었음.

실험 결과

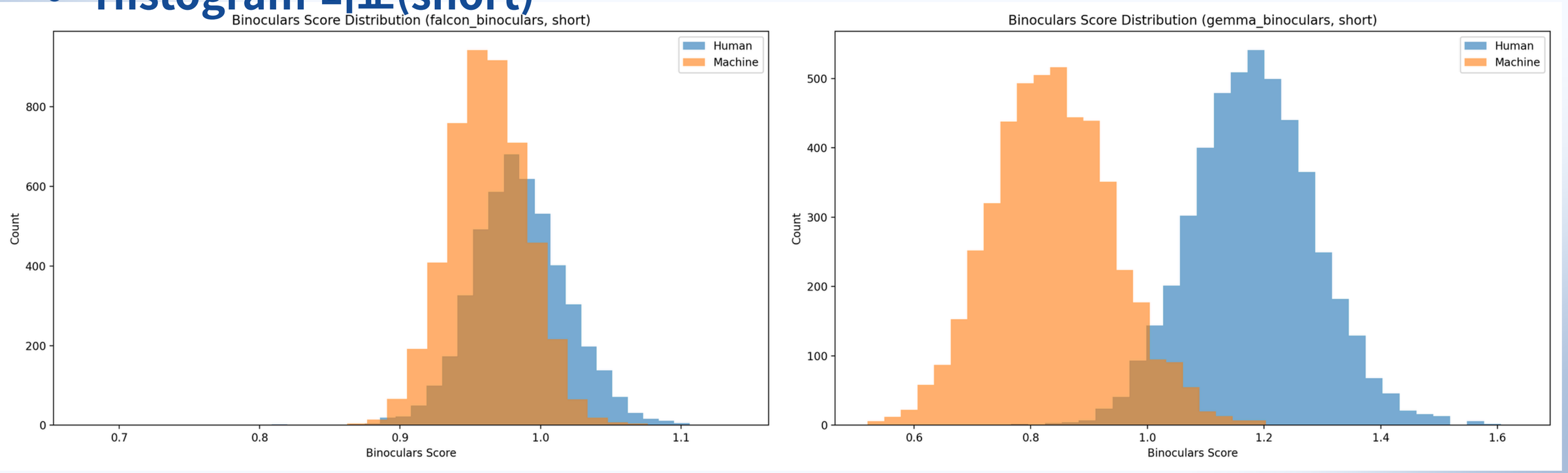
- EXAONE 생성문 + Falcon/Gemma 분석 모델 비교 (medium / short / very short)
- Histogram 비교(medium)





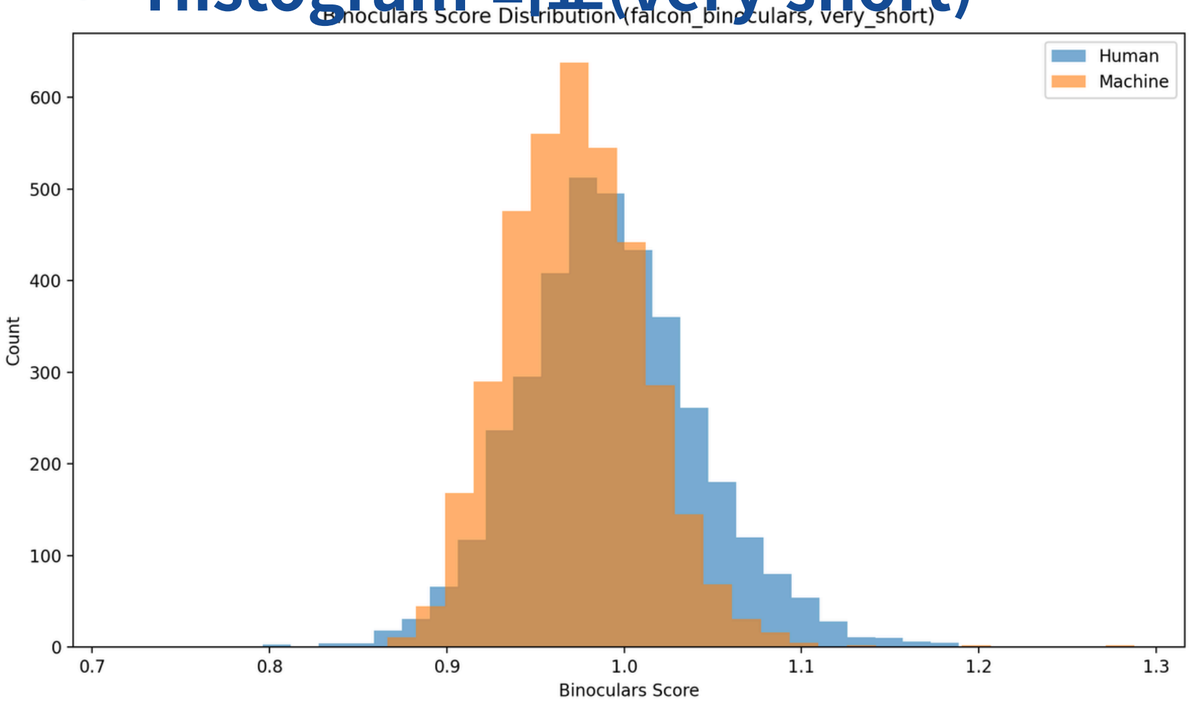
실험 결과

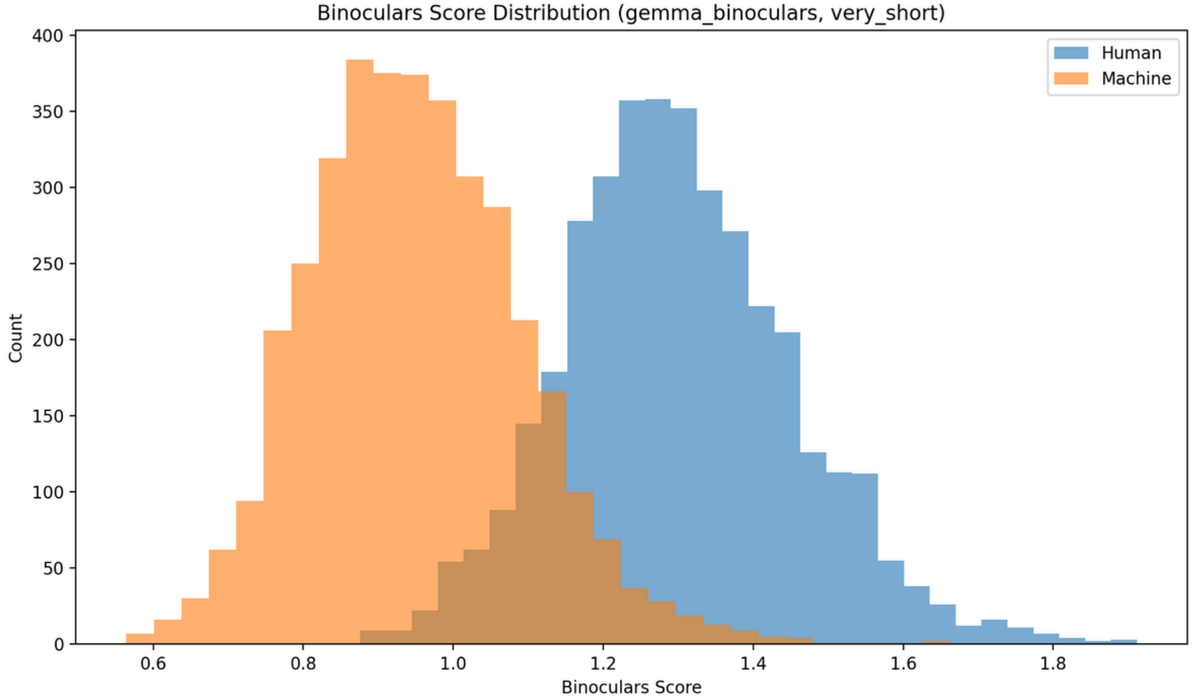
- EXAONE 생성문 + Falcon/Gemma 분석 모델 비교 (medium / short / very short)
- Histogram 비교(short)



실험 결과

- EXAONE 생성문 + Falcon/Gemma 분석 모델 비교 (medium / short / very short)
- Histogram 비교(very short)



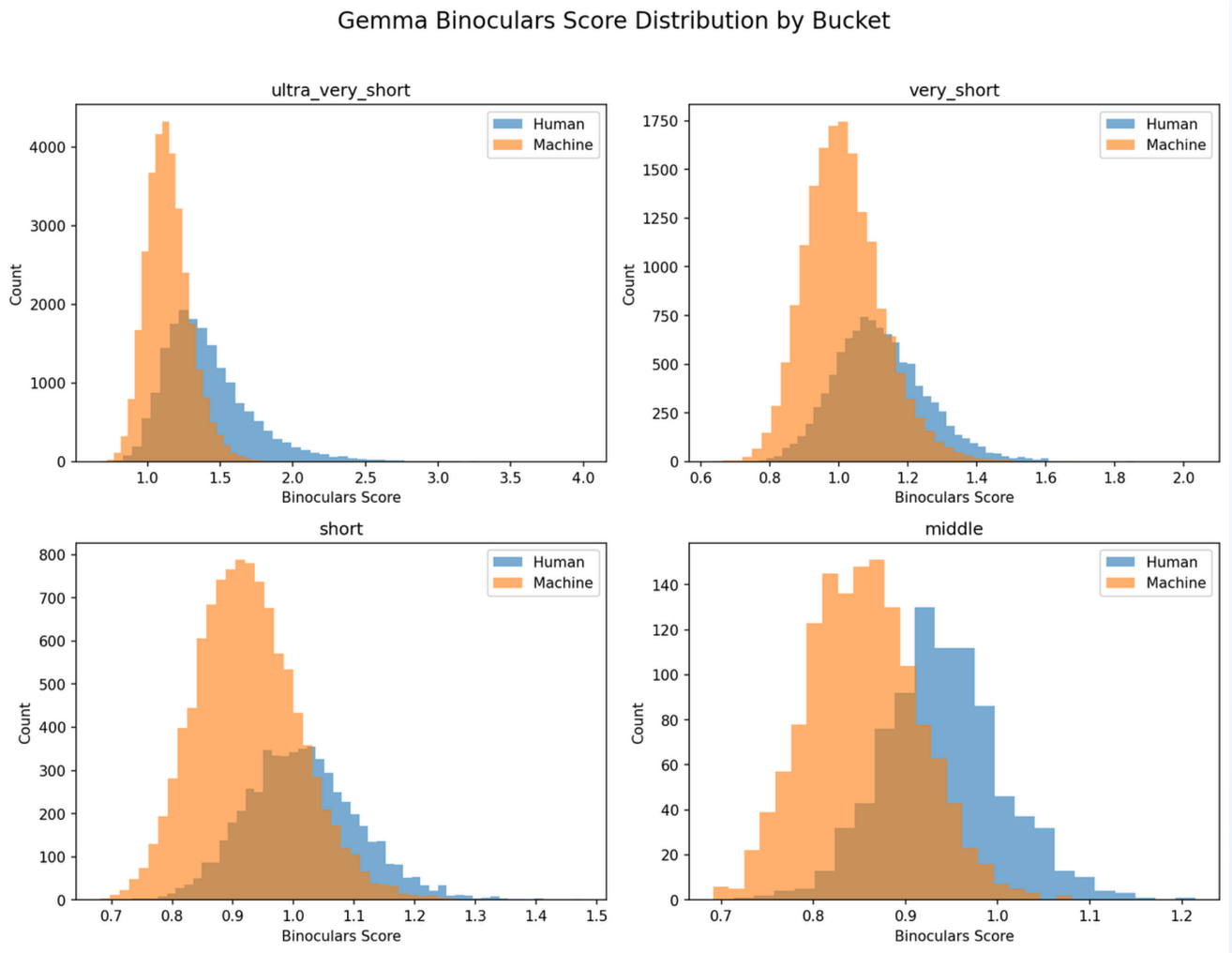


실험 결과

- **EXAONE 생성문 + Falcon/Gemma 분석 모델 비교 (medium / short / very short)**
- **종합 해석**
  - Falcon-binoculars는 EXAONE 생성문에 대해 전반적으로 약한 성능을 보였음
    - 본 실험은 EXAONE 생성문 탐지에서 Gemma 기반 분석 모델이 Falcon 기반 분석 모델보다 훨씬 적합하다는 점을 보여줌

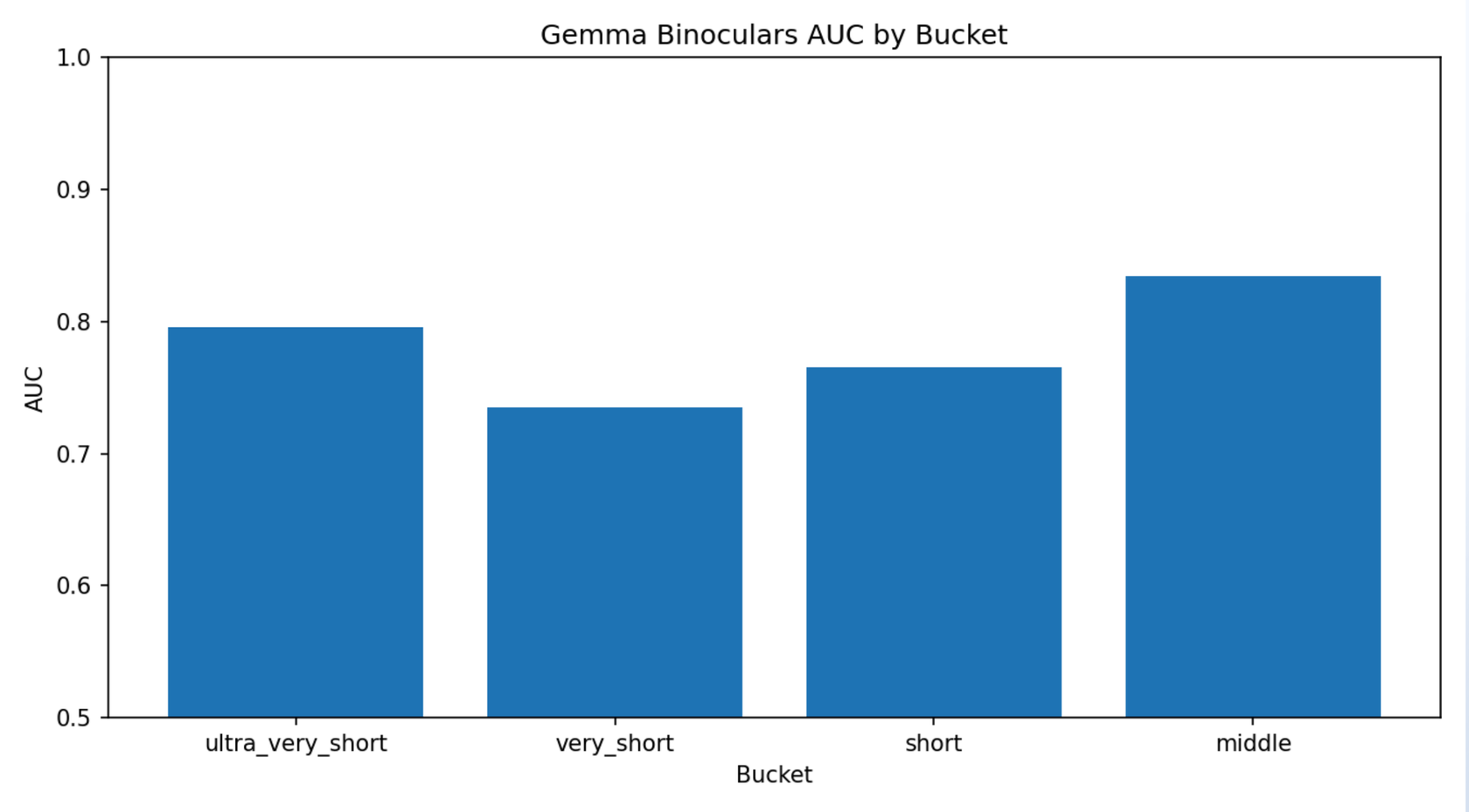
실험 결과

- EXAONE 생성문 + Gemma 분석 모델 (middle / short / very short / ultra short)
- Histogram 비교



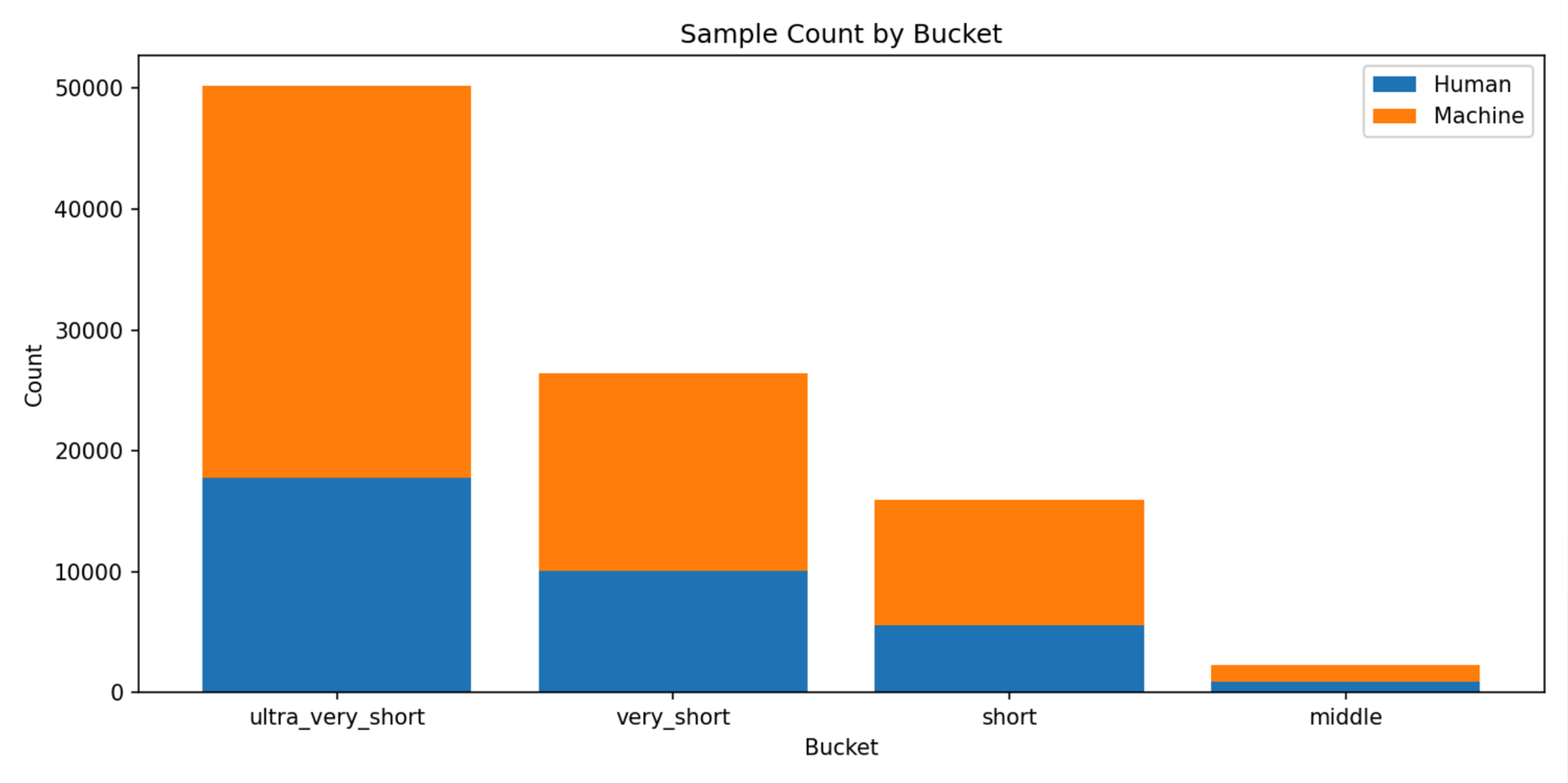
실험 결과

- EXAONE 생성문 + Gemma 분석 모델 (medium / short / very short / ultra short)
- AUC비교



실험 결과

- EXAONE 생성문 + Gemma 분석 모델 (medium / short / very short / ultra short)
- Sample Count



실험 결과

- **EXAONE 생성문 + Gemma 분석 모델 (medium / short / very short / ultra short)**
- **종합 해석**
  - 전체적으로 구간에서 Human / Machine 분포 겹침이 큼
    - AUC가 0.73~0.84 수준에 머물러 전반적으로 낮음
    - 성능이 가장 높은 medium 구간은 표본 수가 가장 적음
    - 실제로 표본이 많이 몰린 ultra_very_short, very_short 구간에서 성능이 충분히 높지 않음
    - 따라서 본 실험은 안정적인 탐지 성능 확보에 실패한 실험으로 판단함

한계

- 토큰 기준 길이 제어로 인해 목표 길이의 생성문 확보가 불안정했음
- 도메인 특화가 적용되지 않아 사람도 쉽게 구분 가능한 생성문이 포함되었음
- 인간 문체와 생성 문체 차이를 충분히 통제하지 못했음
- 버킷별 표본 수 불균형으로 구간별 성능 비교 해석에 한계가 있었음
- 평균 score gap과 실제 분류 성능을 분리해서 해석하지 못한 부분이 있었음

- **본 실험은 Binoculars의 가능성을 확인하는 데는 의미가 있었으나, 길이 제어, 도메인 특화, 문체 통제, 표본 균형, threshold 설정 측면에서 추가 보완이 필요함**
- **후속 실험에서는 이러한 조건을 정교하게 통제하여 실제 탐지 성능과 모델 일반화 가능성을 더 신뢰성 있게 평가할 필요가 있음**

이후 실험 방향 및 파생 실험 제언

- **이후 실험 방향**
  - **생성 데이터 품질 통제 강화**
    - 목표 길이에 맞는 생성문을 더 안정적으로 확보할 수 있도록 생성 조건 재설계
    - 토큰 수 기준뿐 아니라 문장 완결성, 자연스러움, 프롬프트 흔적 여부를 함께 점검
    - 비문, 과도하게 정돈된 문장, 반복 표현 등을 제거하는 품질 검수 단계 추가
  - **도메인 특화 프롬프트 적용**
    - 단순 생성문이 아니라 실제 커뮤니티 문체를 반영한 생성문 구축
    - 게시판별 유행어, 밈, 축약 표현, 반말체, 공격적 표현 등을 반영한 프롬프트 설계
    - 사람도 쉽게 구분 가능한 생성문이 아니라, 실제와 더 유사한 생성문으로 난이도 상향
  - **길이 구간 재정의 및 균형화**
    - ultra_very_short / very_short / short / medium 구간별 데이터 수를 가능한 한 균형 있게 재구성
    - 길이 차이뿐 아니라 정보량과 문장 복잡도도 함께 고려
    - 특히 성능이 낮았던 very short 구간을 중심으로 재실험 수행

이후 실험 방향 및 파생 실험 제언

- **파생 실험 제언**
  - **커뮤니티별 도메인 특화 농도 정량화 실험**
    - **배경**
      - 온라인 커뮤니티는 각기 다른 주제, 사용자 집단, 상호작용 관습을 바탕으로 고유한 언어 분포를 형성함
      - 특히 일부 커뮤니티는 밈, 유행어, 축약 표현, 인터넷 문체 등 일반 텍스트와 구별되는 표현을 반복적으로 사용함
      - 이러한 특성은 해당 커뮤니티 텍스트가 일반 언어 분포로부터 얼마나 이탈해 있는가라는 관점에서 해석할 수 있음
    - **문제의식**
      - 생성문 분석 성능은 단순히 문장 길이나 모델 조합뿐 아니라, 입력 데이터가 얼마나 강한 도메인 특화성을 가지는지에 따라 달라질 가능성이 있음
      - 즉, 도메인 특화 농도가 높은 커뮤니티 텍스트일수록 일반 언어모델 기준에서는 더 낯설고 부자연스럽게 보일 수 있으며, 그 결과 생성문 분석 난이도 또한 달라질 수 있음
      - 따라서 커뮤니티별 언어 분포 차이를 정량화하면, 왜 어떤 도메인에서는 생성문 분석이 더 어렵고, 어떤 도메인에서는 더 쉬운지를 설명할 수 있을 것으로 기대함

이후 실험 방향 및 파생 실험 제언

- **파생 실험 제언**
  - **커뮤니티별 도메인 특화 농도 정량화 실험**
    - **실험 목적**
      - 서로 다른 커뮤니티 텍스트의 도메인 특화 농도를 정량적으로 비교
      - 도메인 특화 농도와 생성문 분석 난이도 사이의 관계를 확인
      - 궁극적으로는 도메인 특화성이 생성문 분석 성능을 설명하는 핵심 변수로 작용하는지 검토
    - **핵심 가설**
      - 도메인 특화 농도가 높을수록 일반 LLM 기준 log-Perplexity가 높아질 가능성이 있음
      - 또한 도메인 특화 문체는 일반적 의미 공간과 다른 분포를 형성하므로, 임베딩 공간에서도 더 독립적인 군집 특성을 보일 가능성이 있음
      - 나아가 이러한 특성은 Cross-Perplexity와 score gap 형성에도 영향을 미쳐, 생성문 분석 난이도 차이로 이어질 수 있음

이후 실험 방향 및 파생 실험 제언

- **파생 실험 제언**
  - **커뮤니티별 도메인 특화 농도 정량화 실험**
    - **분석 방법**
      - log-Perplexity
        - 분석각 커뮤니티 텍스트가 일반 언어모델 기준 얼마나 자연스럽거나 낯설게 보이는지 비교
      - Cross-Perplexity
        - 분석각 커뮤니티 텍스트에 대해 모델 간 예측 일치성이 얼마나 유지되는지 비교
      - 임베딩 기반 분석
        - 커뮤니티별 텍스트의 군집 밀도, 중심 간 거리, 분산 등을 비교
    - **기대 효과**
      - 커뮤니티별 언어 분포 차이를 정량적으로 설명할 수 있음
      - 생성문 분석이 어려운 도메인과 쉬운 도메인을 구조적으로 구분할 수 있음
      - 이후 도메인 적응형 분석 모델이나 도메인 특화 프롬프트 설계의 근거로 활용 가능

Q&A