EMNLP 2026 · 논문 리서치 발표

TCD

문맥 의존성을 활용한 LLM 생성 텍스트 제로샷 탐지

TCD: Empowering Zero-Shot Detection of LLM-Generated Text using Token Contextual Dependency

논문 저자 Shihan Ma · Quan Wang | EMNLP 2026

발표 Deep Shark Lab 석사과정 류동훈

CONTENTS

목차

01	연구 배경 및 문제 정의	04–07
02	제안 기법 : TCD	08–13
03	실험 설정 및 주요 결과	14–17
04	강건성 분석	18–23
05	개입 · 제거 실험 및 추론 효율	24–26
06	연구의 한계와 결론	27–28
Q&A		29

TCD 의 핵심 연구 문제

모든 토큰이 LLM 생성 여부를 판별하는 데 똑같이 중요한가 ?

01	기존 집계 방식	토큰별 검출 통계의 균등 평균
02	핵심 관찰	저엔트로피 토큰에서 높은 판별력 관찰
03	TCD 접근법	Attention Entropy · DLS · Reweighting
04	검출 성능 개선	평균 AUROC +3.13 / +2.32 / +2.23%p

기존 제로샷 검출기와 TCD 의 차별점

Fast-DetectGPT

조건부 확률 곡률에 따른
토큰별 검출 통계

Binoculars

두 언어모델의 예측 통계 비교

DNA-DetectLLM

언어모델 간 예측 차이를
활용하는 검출 통계

공통 한계 — **토큰별 판별 정보량을 고려하지 않는 균등 집계**

TCD 의 차별점 — **Attention Entropy 에 따라 토큰별 가중치를 적용**

기존 검출기의 가정 : 모든 토큰은 동등한가 ?

기존 방식 : 모든 토큰에 동일한 기여도 1/N 적용



동일한 가중치가 실제 검출 기여도까지 같다는 뜻은 아니다 .

핵심 문제 : 검출 신호가 약한 토큰까지 균등 평균에 포함되면 유용한 신호가 희석될 수 있음

핵심 가설 : 문맥 의존성과 검출 정보량

낮은 엔트로피

Attention 이 일부 선행 토큰에 집중
→ 높은 문맥 의존성
→ 더 강한 검출 신호 가설

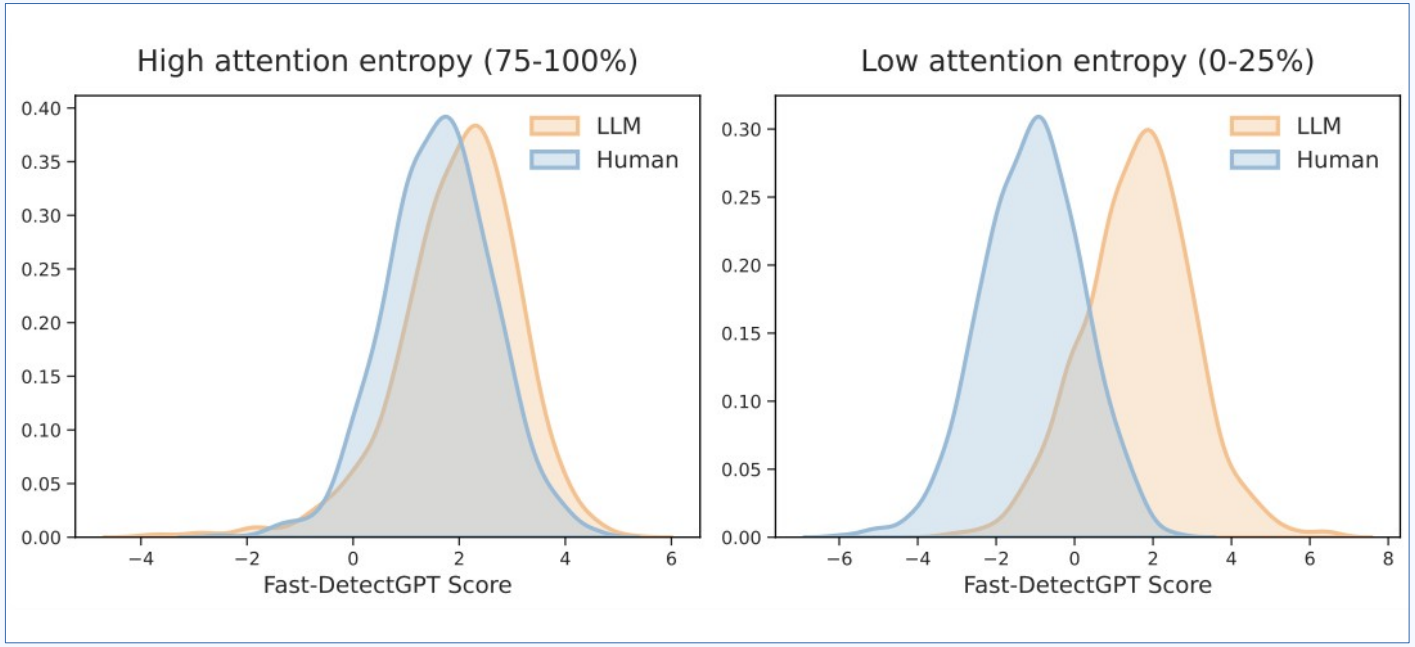
높은 엔트로피

Attention 이 넓은 문맥에 분산
→ 낮은 문맥 의존성
→ 상대적으로 약한 검출 신호 가설

“토큰 위치별로 검출 정보량이 불균등하게 분포한다 .”

핵심 가설에 대한 실험적 근거

RealDet 2,000 개 균형 샘플 · Fast-DetectGPT 점수 분포



고엔트로피 상위 25%: 분포가 크게 겹침

저엔트로피 하위 25%: 분포가 뚜렷하게 분리됨

TCD 프레임워크 : 3 단계 처리 과정

01 어텐션 엔트로피

모델 레이어별 토큰 엔트로피 산출

→

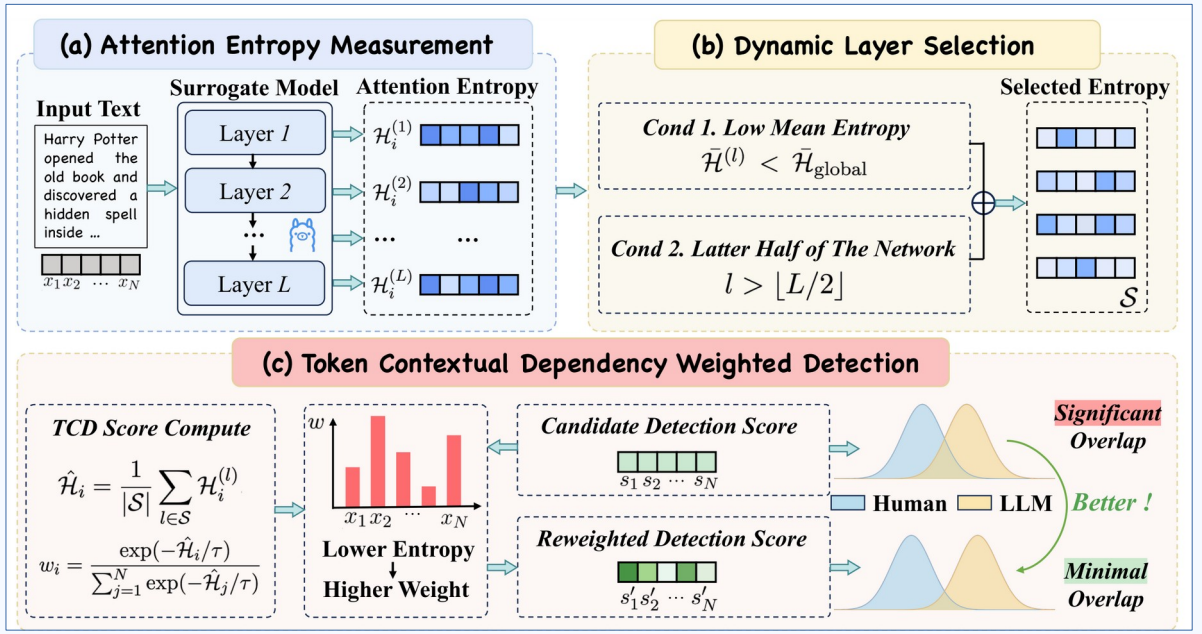
02 동적 레이어 선택

정보성이 높은 후반부 레이어 선택

→

03 토큰별 가중치 적용

낮은 엔트로피 토큰에 더 큰 가중치



기존 방식 : 토큰 점수의 균등 집계

$$S(x) = f\left(\frac{1}{N}\sum_{i=1}^{N} s_i\right)$$

$x = \{x_1, x_2, \ldots, x_n\}$

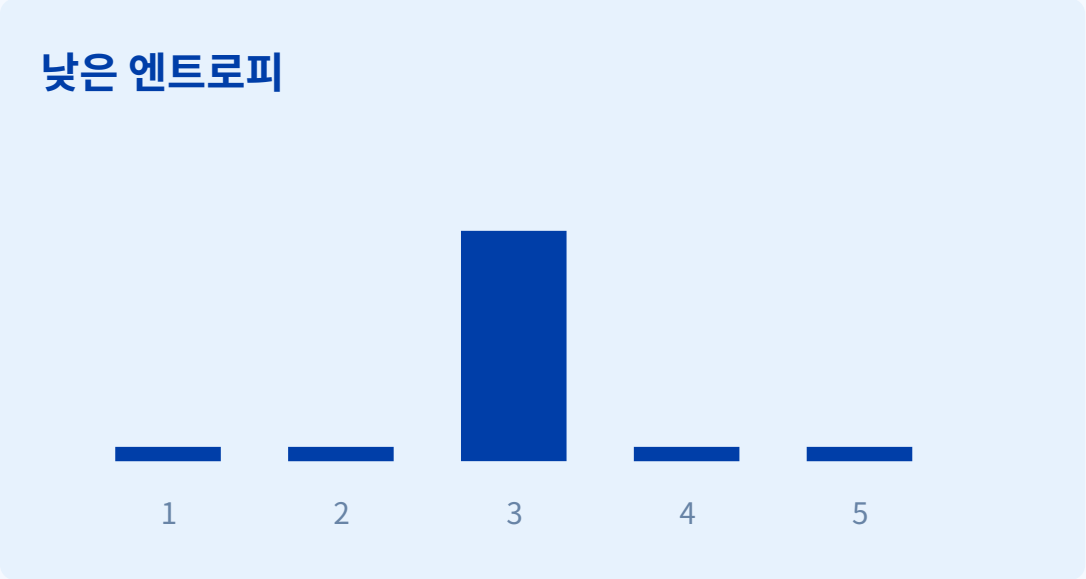
s_i: 토큰 i 의 검출 통계

N: 토큰 수

$f(\cdot)$: 검출기별 변환 함수

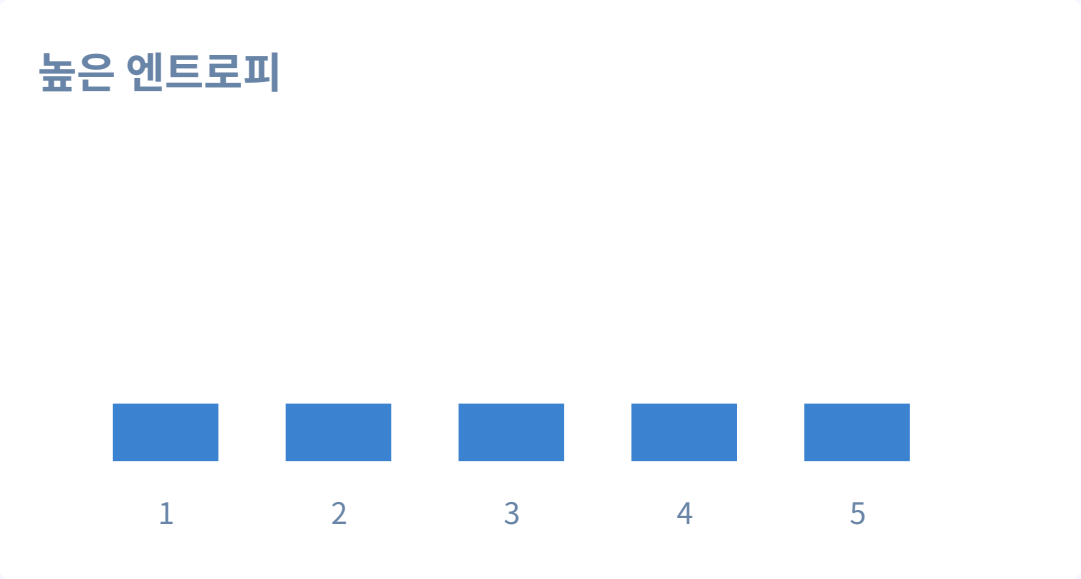
TCD 가 바꾸는 것은 s_i 자체보다 , 각 s_i 를 결합하는 가중치다 .

Attention Entropy: 집중과 분산



일부 문맥 토큰에 집중

높은 문맥 의존성



관심이 여러 토큰에 분산

낮은 문맥 의존성

위 그래프는 개념 설명용 가상 Attention 분포

Attention Entropy 계산과 레이어 선택

토큰별 Attention Entropy

$$H_i^{(l)} = -\frac{1}{H}\sum_{h=1}^{H}\sum_{j=1}^{N} A_{h,i,j}^{(l)} \log A_{h,i,j}^{(l)}$$

동적 레이어 선택 (DLS)

$$\bar{H}^{(l)} = \frac{1}{N}\sum_{i=1}^{N} H_i^{(l)}$$

Condition 1 · 평균 엔트로피가 전체 레이어 평균보다 낮아야 함

Condition 2 · 네트워크 후반부 레이어 (l > ⌊L/2⌋) 만 유지

엔트로피에서 토큰 가중치로

1 · 선택된 레이어의 엔트로피 집계

$$\hat{H}_i = \frac{1}{|S|} \sum_{l \in S} H_i^{(l)}$$

선택된 레이어의 엔트로피 평균

2 · 온도 조절 Softmax

$$w_i = \frac{\exp(-\hat{H}_i/\tau)}{\sum_{j=1}^{N} \exp(-\hat{H}_j/\tau)}$$

낮은 엔트로피 → 높은 가중치

온도 파라미터 τ 가중치 집중도 조절 · 기본값 $\tau = 0.7$

TCD 가중치 적용 전후의 차이

$$S'(x) = f\left(\sum_{i=1}^{N} s'_i\right) = f\left(\sum_{i=1}^{N} w_i \cdot s_i\right)$$

가상 예시 : 5 개 토큰의 중요도가 서로 다를 때



토큰 수는 동일하지만 검출 점수에 반영되는 기여율은 달라진다 .

실험 환경 및 평가 설정

주요 벤치마크 · 균형 표본 : 각 데이터셋 Human 1,000 / LLM 1,000

RealDet

다중 도메인과 생성 모델을 포함한 현실적 평가

RAID

도메인 · 생성기 · 공격 · 디코딩 강건성 평가

M4

다중 생성기 · 도메인 · 언어를 포함한 평가

Detection Models

Fast-DetectGPT · Binoculars · DNA-DetectLLM

Surrogate Model

Llama-3.2-1B-Instruct / Base 1B

Evaluation

AUROC · AUPR · 기본 $\tau = 0.7$

NVIDIA A40 48GB · Max input 512 tokens · Batch 8

주요 실험 결과 : 평균 AUROC 개선

TCD 적용에 따른 평균 AUROC 증가량



재가중치만으로 세 검출기의 평균 AUROC 가 모두 개선

원문 표 1 · Avg. 열

Avg.	
AUROC	AUPR
84.81	88.40
85.76	88.99
83.09	85.70
85.87	87.56
87.63	91.85
90.76	**93.22**
3.13	*1.37*
87.94	91.79
90.26	**92.70**
2.32	*0.91*
88.09	91.65
90.32	**92.68**
2.23	*1.03*

주요 실험 결과 : 데이터셋별 개선 폭

Methods	RealDet		RAID		M4		Avg.	
	AUROC	AUPR	AUROC	AUPR	AUROC	AUPR	AUROC	AUPR
Zero-Shot Detectors								
Likelihood	88.09	91.75	77.51	83.93	88.81	89.51	84.81	88.40
LogRank	88.91	92.25	78.46	83.95	89.91	90.76	85.76	88.99
Entropy	87.35	90.52	75.20	79.41	86.72	87.17	83.09	85.70
LRR	88.87	91.31	80.01	80.64	88.74	90.72	85.87	87.56
Fast-DetectGPT	90.54	93.55	82.80	89.75	89.54	92.25	87.63	91.85
+TCD (Ours)	**93.56**	**95.10**	**87.25**	**91.44**	**91.48**	**93.12**	**90.76**	**93.22**
(Diff)	*3.02*	*1.55*	*4.45*	*1.69*	*1.94*	*0.87*	*3.13*	*1.37*
Binoculars	91.60	94.15	83.10	89.94	89.12	91.28	87.94	91.79
+TCD (Ours)	**93.02**	**94.51**	**86.88**	**91.19**	**90.90**	**92.39**	**90.26**	**92.70**
(Diff)	*1.42*	*0.36*	*3.78*	*1.25*	*1.78*	*1.11*	*2.32*	*0.91*
DNA-DetectLLM	91.86	94.22	82.16	88.90	90.25	91.84	88.09	91.65
+TCD (Ours)	**93.36**	**94.81**	**85.24**	**89.51**	**92.37**	**93.70**	**90.32**	**92.68**
(Diff)	*1.50*	*0.59*	*3.08*	*0.61*	*2.12*	*1.86*	*2.23*	*1.03*

원문 표 1: RealDet / RAID / M4 / Avg. · AUROC 와 AUPR

정밀도–재현율 기반 성능 (AUPR) 개선

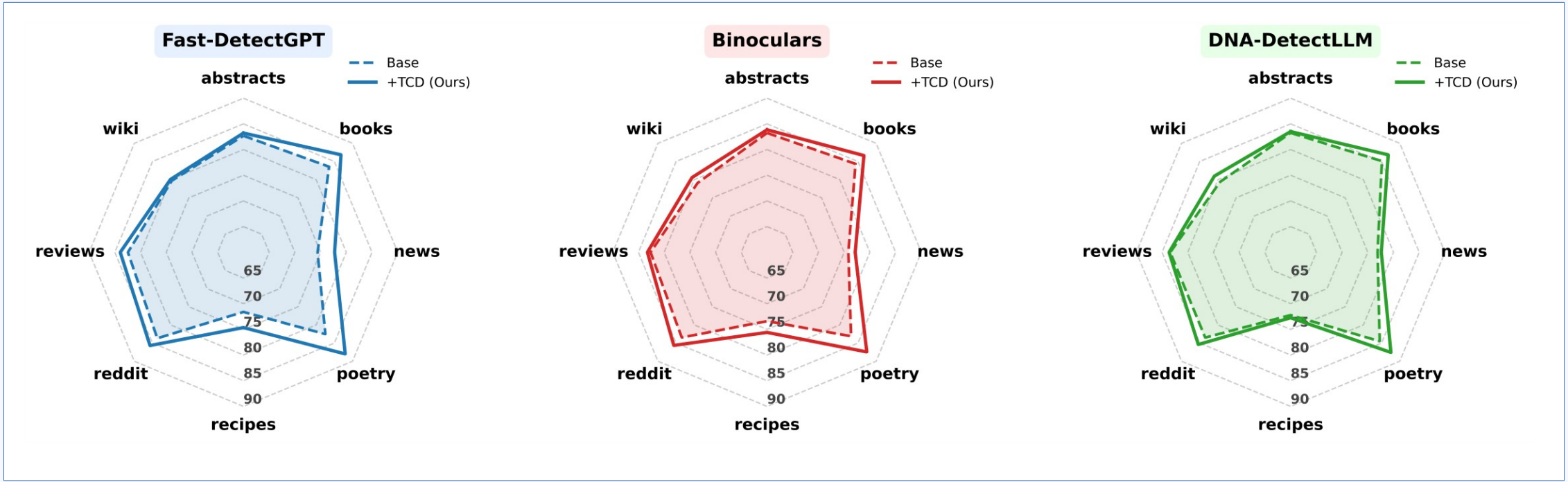
원문 표 1 의 TCD 적용 전후 비교 구간 (원본 크롭)

Fast-DetectGPT	90.54	93.55	82.80	89.75	89.54	92.25	87.63	91.85
+TCD (Ours)	**93.56**	**95.10**	**87.25**	**91.44**	**91.48**	**93.12**	**90.76**	**93.22**
(Diff)	*3.02*	*1.55*	*4.45*	*1.69*	*1.94*	*0.87*	*3.13*	*1.37*
Binoculars	91.60	94.15	83.10	89.94	89.12	91.28	87.94	91.79
+TCD (Ours)	**93.02**	**94.51**	**86.88**	**91.19**	**90.90**	**92.39**	**90.26**	**92.70**
(Diff)	*1.42*	*0.36*	*3.78*	*1.25*	*1.78*	*1.11*	*2.32*	*0.91*
DNA-DetectLLM	91.86	94.22	82.16	88.90	90.25	91.84	88.09	91.65
+TCD (Ours)	**93.36**	**94.81**	**85.24**	**89.51**	**92.37**	**93.70**	**90.32**	**92.68**
(Diff)	*1.50*	*0.59*	*3.08*	*0.61*	*2.12*	*1.86*	*2.23*	*1.03*

평균 AUPR 증가 : Fast-DetectGPT +1.37%p · Binoculars +0.91%p · DNA-DetectLLM +1.03%p

도메인 간 일반화 성능

RAID 8 개 도메인 · 각 도메인에 복수 생성 모델 포함



Fast-DetectGPT: 시 (Poetry) +5.46%p · 도서 (Books) +3.28%p

인간 ·AI 혼합 텍스트 탐지 : HART

HART: 세 가지 텍스트 작성 수준에 대한 AUROC 평가

Methods	Level-1	Level-2	Level-3
Fast-DetectGPT	81.31	69.00	84.59
+TCD (Ours)	**83.02**	**71.93**	**88.65**
(Diff)	*1.71*	*2.93*	*4.06*
Binoculars	81.68	70.10	86.43
+TCD (Ours)	**83.08**	**72.05**	**89.08**
(Diff)	*1.40*	*1.95*	*2.65*
DNA-DetectLLM	81.86	71.43	87.28
+TCD (Ours)	**83.80**	**72.81**	**89.05**
(Diff)	*1.94*	*1.38*	*1.77*

Level 1

AI 관여 여부

Level 2

기반 작성 주체

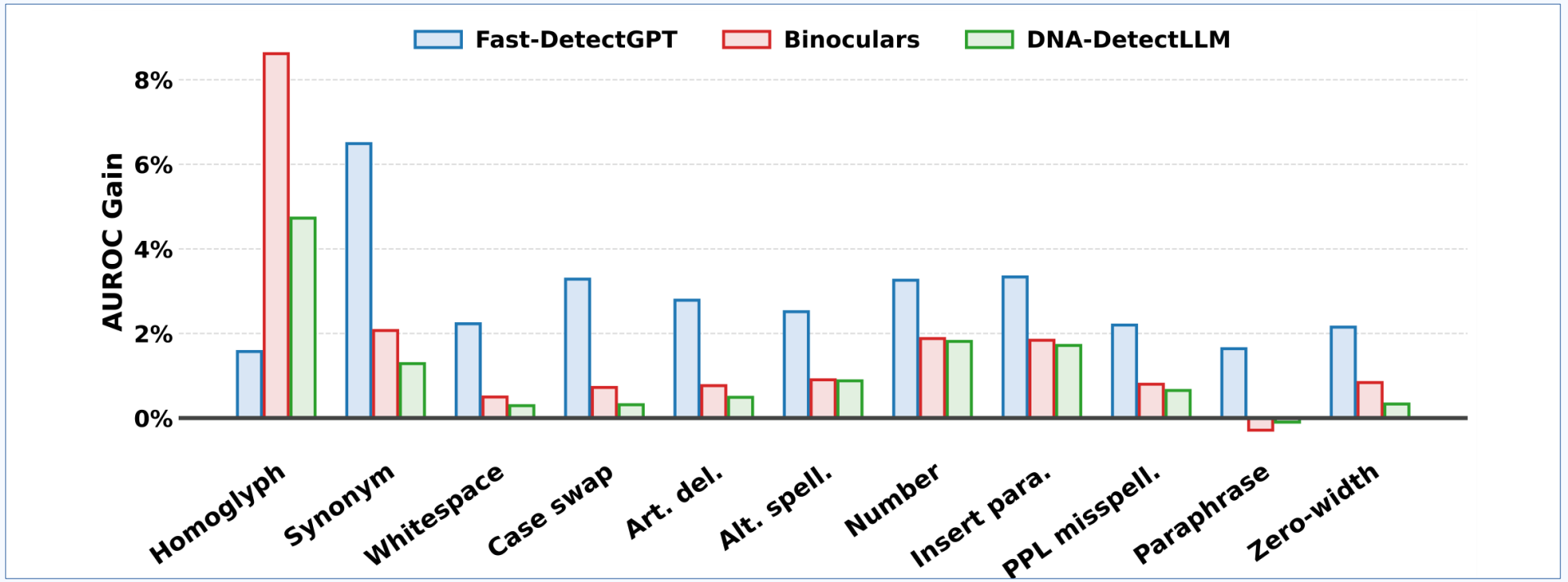
Level 3

완전 AI 생성 여부

Level 2 가 모든 검출기에서 가장 어려운 평가 조건

적대적 텍스트 변형에 대한 강건성

RAID · 11 개 공격 유형에서 기반 검출기 대비 AUROC 차이

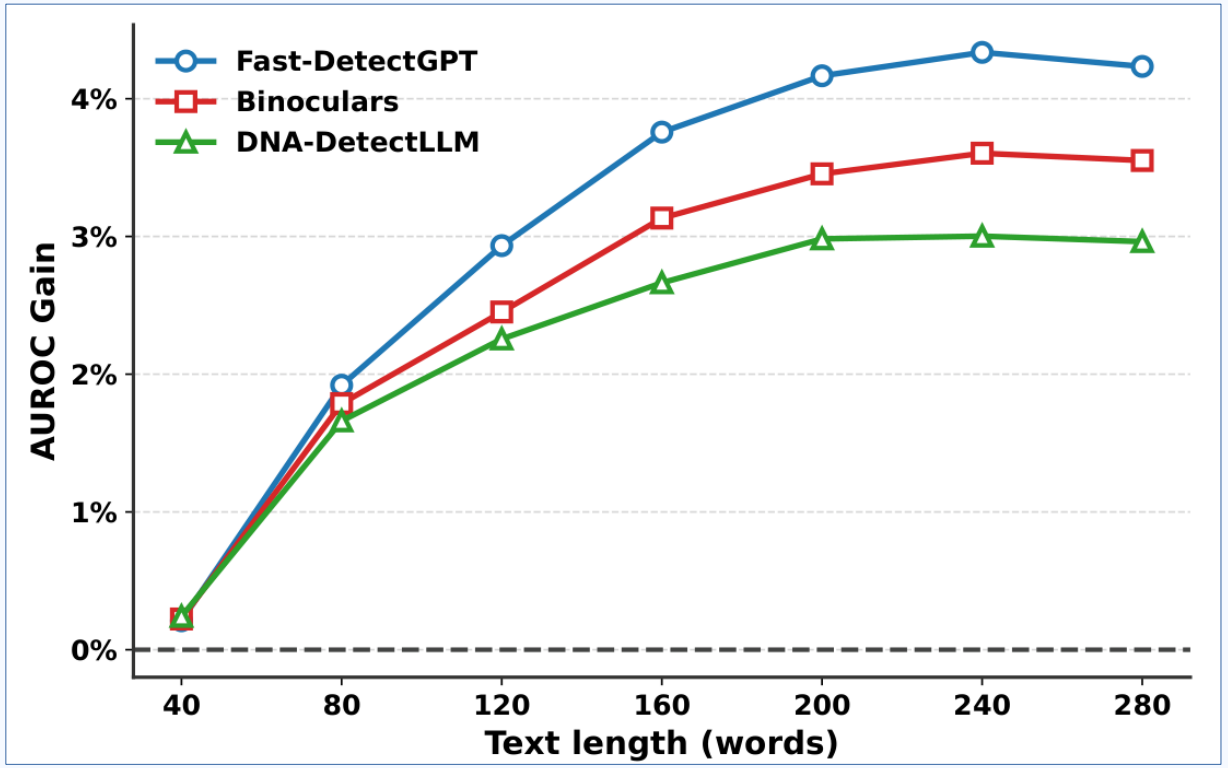


유사 문자 / 동의어 치환 : 큰 개선 폭

의역 (Paraphrase): 성능 예외 존재

텍스트 길이에 따른 강건성

RAID 샘플을 40~280 단어로 절단 · AUROC 개선 폭



모든 시험 길이에서 개선 : 다만 최단 40 단어 구간에서 개선 폭이 작음

대리 모델에 따른 성능 민감도

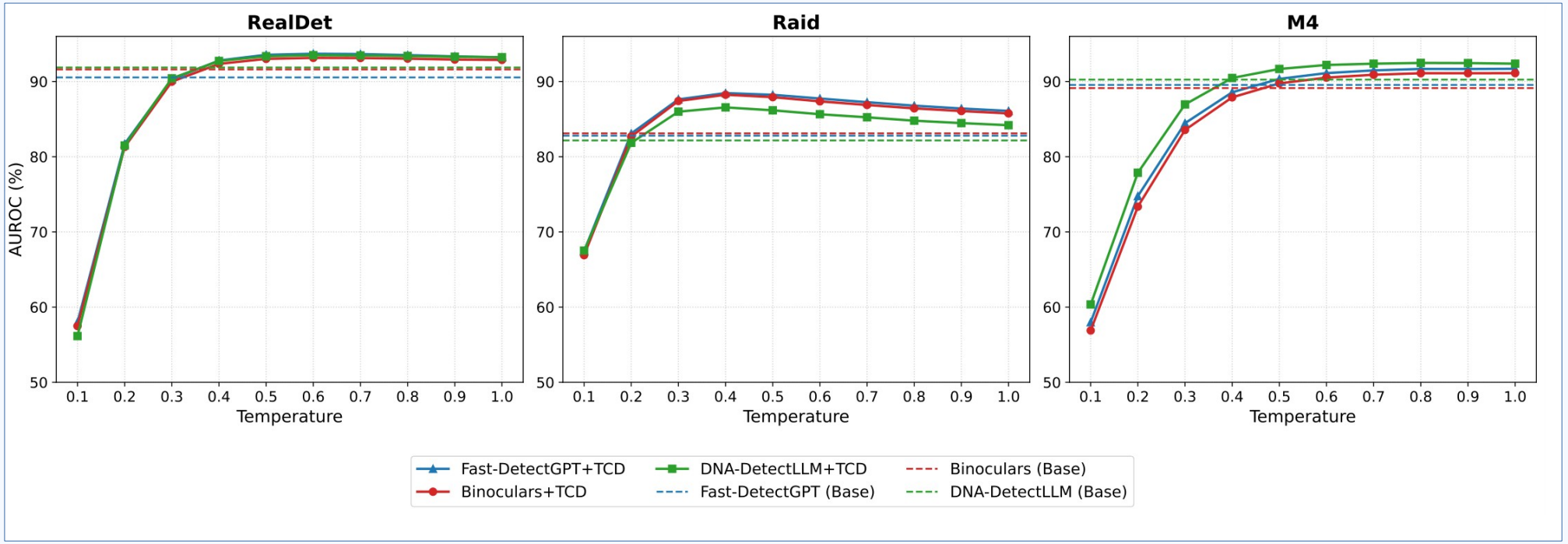
원문 표 3: 모델 계열별 평균 AUROC 개선량

Model series	Fast.	Bino.	DNA.
Llama-3.2-1B family	3.13	2.32	2.23
Llama-3.2-3B family	2.28	1.57	1.79
Gemma-2-2B family	1.97	1.46	2.38

Llama-3.2 와 Gemma-2 계열 모두에서 TCD 개선 확인

Softmax 온도 변화에 따른 성능

Softmax 온도 파라미터 τ: 0.1~1.0



τ = 0.7

기본 설정

추가적인 가중치 집중도 하이퍼파라미터

매우 낮은 τ

소수 토큰에 가중치 과집중 가능

개입 실험 : 토큰 가중치는 실제로 유효한가 ?

Configuration	Fast.	Bino.	DNA.
Base detector (no TCD)	87.63	87.94	88.09
Base detector +TCD	**90.76**	**90.26**	**90.32**
(a) Mask 30% low-entropy tokens	79.00	83.40	83.99
(b) Mask 30% high-entropy tokens	90.10	89.68	89.72
(c) Permute entropy weights	86.81	87.10	87.52
(d) Invert entropy weights	81.40	84.05	84.74
(e) Ablate 30% low-entropy heads	90.59	90.09	90.20
(f) Ablate low-entropy layers	88.35	88.47	88.90

저엔트로피 토큰 제거 및 가중치 반전 시 성능이 크게 저하

DLS 구성요소 제거 실험 (Ablation)

Variant		Fast.	Bino.	DNA.
Cond.1	*Cond.2*			
Uniform (no TCD)		87.63	87.94	88.09
✗	✗	89.48	89.30	89.53
✗	✓	90.29	89.86	89.97
✓	✗	90.40	89.99	90.03
✓	✓	**90.76**	**90.26**	**90.32**

+1.85%p

Uniform → 전체 레이어 재가중치

+1.28%p

전체 레이어 → DLS 선택

Reweighting 이 주효하고 DLS 가 추가 개선

추론 시간 분석 : 성능 향상의 비용

Methods	Time (s)
Fast-DetectGPT	0.070
+TCD (Ours)	0.145
Binoculars	0.372
+TCD (Ours)	0.468
DNA-DetectLLM	0.669
+TCD (Ours)	0.720

추가 추론 시간

0.05–0.10 s

/ sample

추가 모델 호출 없이
Attention 출력 활성화

추론 지연 증가 : Fast-DetectGPT 2.07× · Binoculars 1.26× · DNA-DetectLLM 1.08×

연구의 한계 (Limitations)

모델 접근성 제약

- 대리 모델은 Transformer 기반이어야 함
- Layer 별 Attention Weight 에 접근할 수 있어야 함
- Logprobs 만 제공하는 상용 API 에는 적용 불가

계산 비용 증가

- Attention Entropy 계산으로 추론 오버헤드 발생
- 추가 모델 호출 및 학습 파라미터는 필요하지 않음
- 엄격한 추론 지연 조건에서는 비용이 제약이 될 수 있음

논문 결론

01 **토큰별 검출 신호는 균등하지 않음**

토큰별 문맥 의존성과 정보량이 균등하지 않음

02 **엔트로피 기반 가중치의 효과**

저엔트로피 토큰에 높은 가중치 부여

03 **기존 검출기에 적용 가능하지만 비용 발생**

기존 검출기에 적용 가능하나 Attention 접근 및 추론 비용 필요

평균 AUROC 개선 : +3.13 / +2.32 / +2.23 %p

Q&A

경청해 주셔서 감사합니다 .

논문 : Ma & Wang · TCD · EMNLP 2026

발표 : Deep Shark Lab 석사과정 류동훈

https://github.com/RecursionM/TCD