Zero-Shot Detection of LLM-Generated Text using Temperature Sensitivity 논문 리뷰

LLM 생성문과 사람 작성문은 확률분포의 온도 변화에 서로 다르게 반응하는가?

Shixuan Ma et al.

ACL 2026 Main Conference, Long Paper

발표자

류 동훈

DeepShark Lab/Hongik University

Date:

2026-08-14

목차

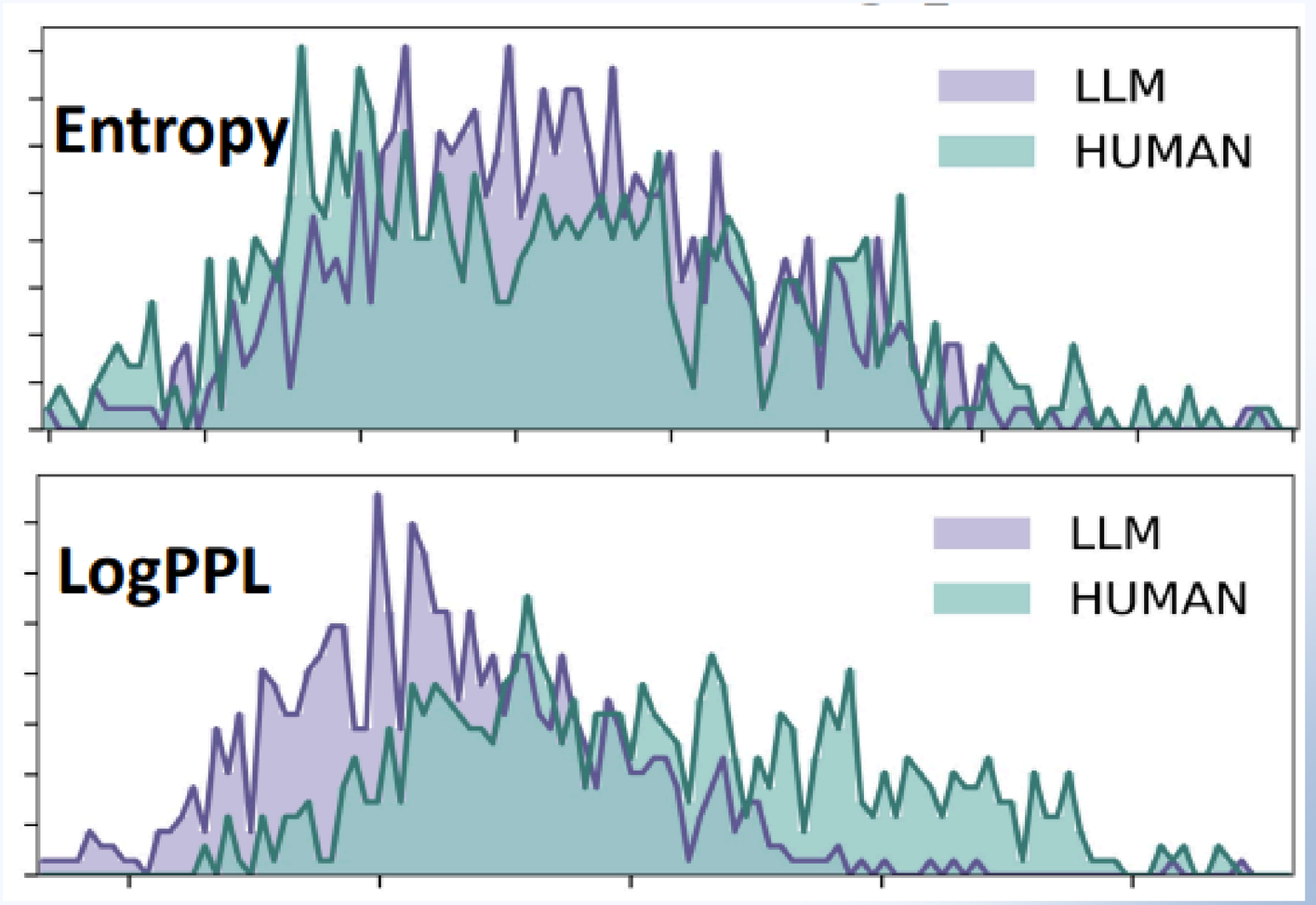
기존 탐지 방법의 한계

- **기존 제로샷 탐지 방식**
  - 대리 모델(탐지용 LLM)의 확률분포에서 로그 퍼플렉시티, 엔트로피, 순위 등의 통계값을 추출
  - 추출된 통계값을 기준으로 사람 작성문과 LLM 생성문을 구분
-
- **문제점**
  - 언어모델의 출력은 어휘 전체에 대한 **고차원 확률분포**
  - 기존 방식은 이를 하나의 통계값으로 **축약**하는 과정에서 **정보 손실** 발생
  - 서로 다른 확률분포를 가진 사람 작성문과 LLM 생성문도 **유사한 통계값을 가질 수 있음**

기존 탐지 방법의 한계

- **X:** 탐지 통계값
- **Y:** 표본 분포

- 기존 통계량에서는 사람 작성문과 LLM 생성문의 분포가 상당 부분 중첩



핵심 아이디어: 확률분포의 반응

핵심 아이디어 — 확률분포의 '값'이 아닌 '반응'을 측정

- 기존 방식
  - 하나의 확률분포에서 로그 퍼플렉시티, 엔트로피 등의 통계값을 계산
  - 확률분포의 상태를 **정적인 값**으로 평가

- 논문의 접근 방식
  - **Normalized Temperature Sensitivity**(NTS): 정규화 온도 민감도
  - LLM에서의 Temperature: **무작위성(randomness)과 창의성**을 조절하는 설정값

  - 언어모델의 온도 T를 변화시켜 확률분포의 집중도를 조절
  - 같은 입력문이 서로 다른 온도에서 얼마나 다르게 평가되는지 측정
  - 이 **변화량**을 LLM 생성문 탐지에 활용

핵심 아이디어: 확률분포의 반응

온도(Temperature, T)

- LLM의 토큰 확률분포의 집중도와 무작위성을 조절하는 매개변수

$$p_T(w_j) = \frac{\exp\left(\ell_j / T\right)}{\sum_{k=1}^{V} \exp\left(\ell_k / T\right)}$$

- T<1
  - 고확률 토큰 집중
  - 더 결정적
- T=1
  - 기본 확률분포
- T>1
  - 확률분포 평탄화
  - 더 다양한 선택

wj: j번째 후보 토큰

ℓj: 해당 토큰의 로짓

pT(wj): 온도 T에서 해당 토큰의 확률

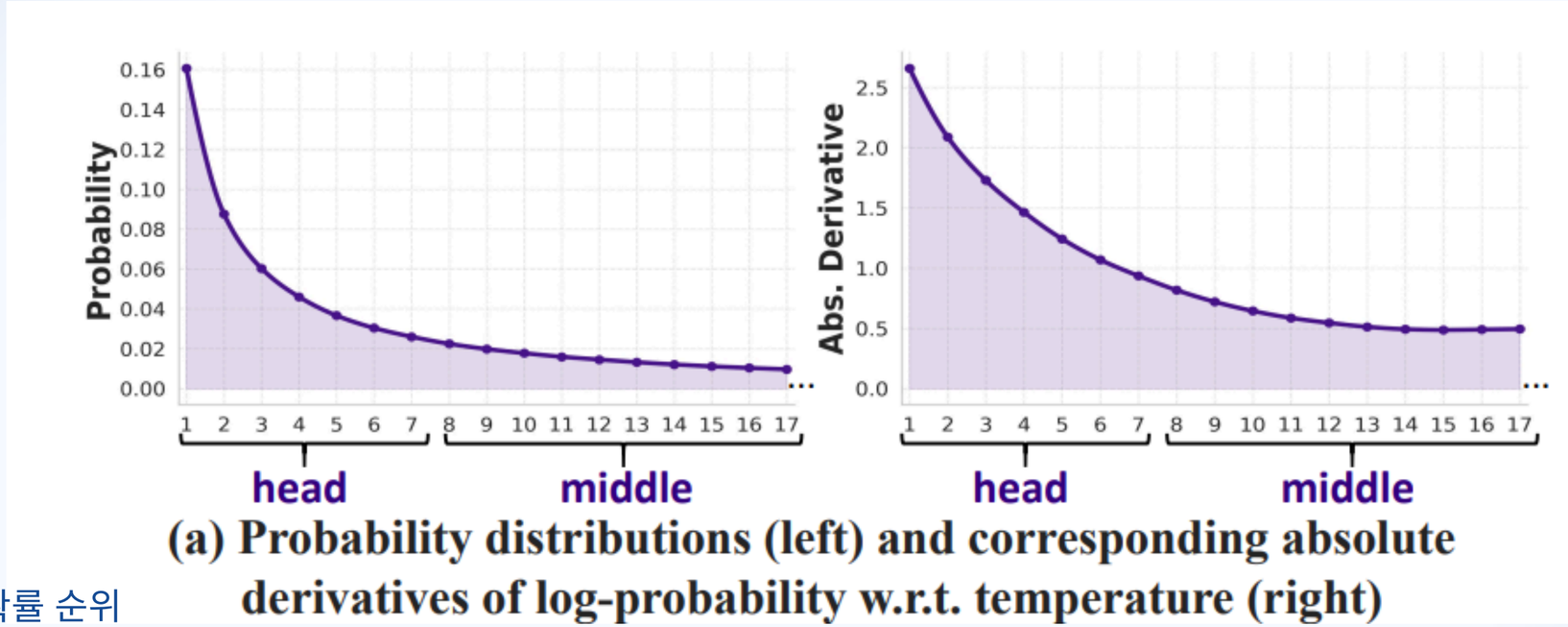
V: 모델이 선택할 수 있는 전체 토큰 후보의 수 (Vocabulary크기)

k: 전체 어휘집에 대한 합산 인덱스

핵심 아이디어: 확률분포의 반응

토큰 확률과 온도 민감도의 관계

- 토큰을 확률이 높은 순서로 정렬하여 확률분포 확인
- Head 영역의 토큰일수록 온도 변화에 따른 로그 확률 변화량이 크게 나타남
- 토큰의 확률분포상 위치에 따라 온도 민감도에 차이 발생



(a) Probability distributions (left) and corresponding absolute derivatives of log-probability w.r.t. temperature (right)

X축: 토큰 확률 순위

왼쪽 Y축: 토큰 확률

오른쪽 Y축: 온도 변화에 대한 로그 확률 변화량

고확률 토큰일수록 온도 변화에 더 민감한 경향

핵심 아이디어: 확률분포의 반응

온도 민감도의 수학적 근거

- 온도 변화에 대한 실제 토큰의 로그 확률 변화량을 미분으로 표현

$$\left| \frac{\partial \log p(w_{\text{label}})}{\partial T} \right| = \left| \frac{\hat{\ell} - \ell_{\text{label}}}{T^2} \right|$$

- wlabel:입력 문장에서 실제로 등장한 토큰
- ℓlabel:실제 등장 토큰의 로짓
- ℓ^:전체 후보 토큰의 기대 로짓
- T:확률분포를 조절하는 온도

$$\left| \ell - \ell_{\text{label}} \right| \uparrow \quad \Rightarrow \quad \text{Temperature Sensitivity} \uparrow$$

- 실제 토큰 로짓과 기대 로짓의 차이가 클수록
  - 온도 변화에 따른 로그 확률 변화량 증가
  - **온도 민감도 증가**

핵심 아이디어: 확률분포의 반응

온도 민감도를 이용한 LLM 생성문 탐지

- LLM 생성문
  - Greedy, Top-k, Top-p 등의 디코딩 방식을 통해 생성
  - 상대적으로 고확률 토큰을 선택하는 경향
  - 확률분포의 Head 영역 토큰 비중 증가
  - → 높은 온도 민감도
- 사람 작성문
  - 특정 디코딩 알고리즘의 제약 없이 단어 선택
  - 대리 모델 기준 Middle 영역의 토큰도 상대적으로 많이 포함
  - → 상대적으로 낮은 온도 민감도

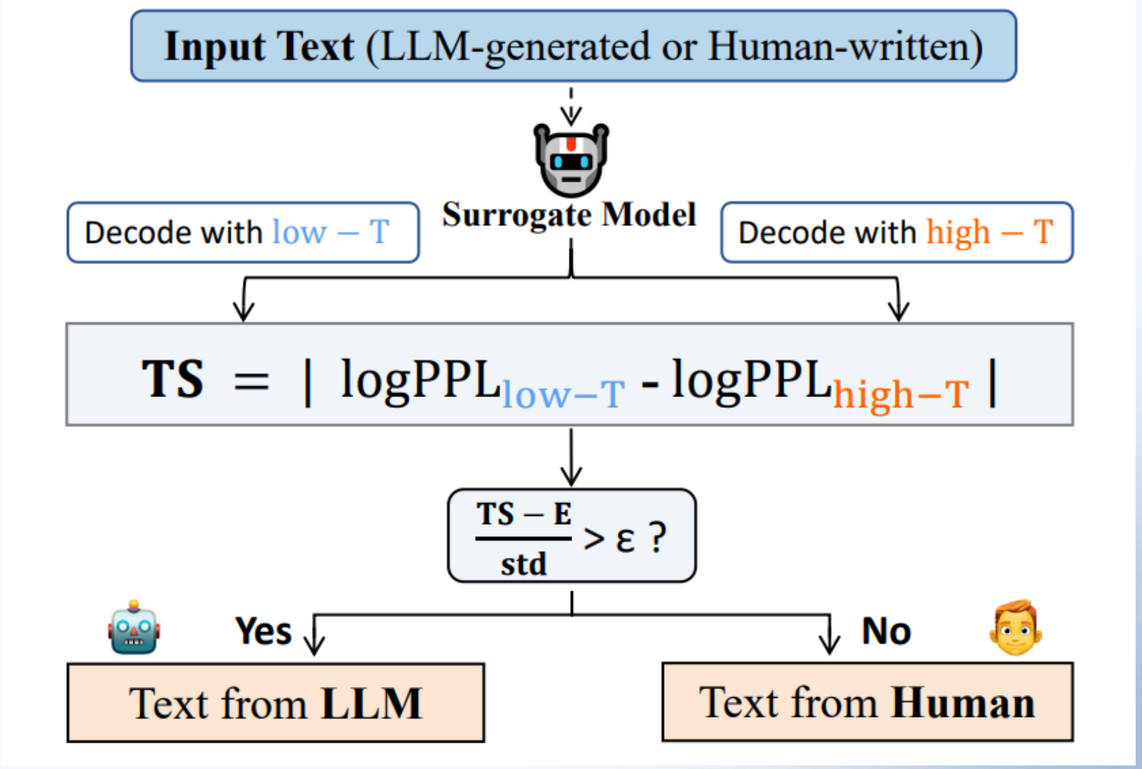
LLM 생성문
고확률 토큰 ↑ → 온도 민감도 ↑
사람 작성문
중간 확률 토큰 ↑ → 온도 민감도 ↓

핵심 아이디어: 확률분포의 반응

Temperature Sensitivity를 이용한 NTS 계산

- 동일 입력을 대리 모델에 입력
- 낮은 온도와 높은 온도에서 각각 LogPPL 계산
- 두 값의 절대 차이를 온도 민감도(TS)로 정의

- Raw TS는 도메인·생성 모델에 따라 최적 임계값이 달라질 수 있음
- 평균과 표준편차를 이용해 TS를 정규화
- 최종 NTS 점수 계산



기존 방식과의 차이점

방법	주요 탐지 신호	대리 모델(LLM 모델)
Entropy	확률분포의 엔트로피	1
LogPPL	로그 퍼플렉시티	1
Log-Rank	실제 토큰의 확률 순위	1
LRR	Likelihood / Log-Rank 관계	1
Fast-DetectGPT	조건부 확률 곡률	2*
Binoculars	정규화 LogPPL	2*
IRM	암묵적 보상 점수	2*
NTS	**온도 변화에 대한 민감도**	**1**

실험

실험 환경 및 공통 설정

- **대리 모델**: Falcon-7B
- **온도 설정**
  - T.low=0.7
  - T.high=1.4
- **평가 지표**
  - AUROC
  - F1 Score
- **비교 방법**
  - Zero-shot detector 7종
    - Entropy/LogPPL/LogRank/LRR/Fast-DetectGPT/Binoculars/IRM
  - 추가적으로 supervised detector와의 비교도 수행
- **평가 데이터셋**:RAID, HART, EvoBench

실험

Robust AI Detection

- 다중 도메인·다중 생성 모델 환경에서 탐지 일반화 성능 평가
- Human-written: 2,000개
- LLM-generated: 2,000개

Domain	Language	Dev	Test
News	English	500	500
Books	English	500	500
Wiki	English	500	500
Abstracts	English	500	500
Reddit	English	500	500
Recipes	English	500	500
Poetry	English	500	500
Reviews	English	500	500

- 8개 도메인News
- 11개 생성 원본 모델의 LLM 텍스트 사용

- **도메인 일반**
  - 뉴스에서 잘 되는 탐지기가 Reddit·논문·리뷰에서도 작동하는가?
- **생성 모델 일반화**
  - 특정 LLM의 생성문에만 특화되지 않고 다른 LLM도 탐지하는가?
- NTS가 **다양한 텍스트 출처와 다양한 생성 원본 모델에서도 일관된 탐지력을 갖는지 평가**

실험

HART (Hierarchical AI Risk in Text Creation)

- AI 개입 수준에 따른 탐지 평가
- Human-written: 4,000개
- LLM-generated: 4,000개
- 4개 도메인
- 7개 생성 원본 모델
- 5개 언어 환경 평가

Domain	Language	Length	Dev	Test
Student essay	English	241 words	2K	2K
Arxiv Intro	English	410 words	2K	2K
Creative Wrting	English	345 words	2K	2K
CC News	English	148 words	2K	2K
CC News	Chinese	590 chars	2K	2K
CC News	French	258 words	2K	2K
CC News	Spanish	285 words	2K	2K
CC News	Arabic	152 words	2K	2K

- 단순한 Human vs 완전 AI 생성문뿐 아니라 **AI가 글을 수정·보조한 상황까지 평가**
- Level이 **3 → 2 → 1**로 내려갈수록 AI로 판정하는 범위가 넓어짐

실험

HART (Hierarchical AI Risk in Text Creation)

구분	AI로 판정하는 범위	의미
Level 3 — Relaxed	완전한 LLM 생성문	**처음부터 끝까지 AI가 생성했는가?**
Level 2 — Medium	내용 자체가 LLM에 의해 생성된 글	**글의 핵심 내용을 AI가 생성했는가?**
Level 1 — Strict	AI가 개입한 모든 글	**작성 과정에 AI가 조금이라도 개입했는가?**

실험

EvoBench 기반 추가 분석 데이터

- 발전하는 LLM 환경을 고려한 **LLM 생성 텍스트 탐지 벤치마크**
- 해당 논문에서는 EvoBench 전체가 아닌 **일부 데이터**만 활용
- 사용 데이터
  - XSum — 뉴스 요약 텍스트
  - Writing — 일반 글쓰기 텍스트
- 영어(English) 데이터 사용
- 각 데이터셋 **1,200개** 샘플을 실험에 활용
- RAID·HART와 달리 **NTS의 세부 특성을 분석하기 위한 추가 실험에 활용**

- **NTS의 온도 및 대리 모델 설정에 대한 강건성을 추가적으로 분석하기 위해 활용**

실험 결과 및 성능 평가

RAID — 다중 도메인 탐지 성능

Detector	News	Books	Wiki	Abstr.	Reddit	Recipes	Poetry	Reviews	ALL	F1
Zero-Shot Detectors										
Entropy (Mitchell et al., 2023)	0.545	0.654	0.624	0.504	0.685	0.582	0.637	0.642	0.585	0.67
LogPPL (Bao et al., 2025)	0.644	0.725	0.701	0.680	0.725	0.627	0.706	0.698	0.663	0.66
Log-Rank (Bao et al., 2025)	0.666	0.745	0.719	0.701	0.735	0.645	0.725	0.716	0.681	0.67
LRR (Su et al., 2023)	0.750	0.816	0.804	0.771	0.779	0.669	0.776	0.773	0.746	0.70
Fast-DetectGPT (Bao et al., 2024) †	0.761	0.845	0.803	0.821	0.794	0.749	0.818	0.810	0.800	0.76
Binoculars (Hans et al., 2024) †	0.768	0.850	0.804	0.826	0.811	0.759	0.826	0.812	0.807	0.77
IRM (Liu et al., 2025) †	0.715	0.712	0.573	0.705	0.647	0.758	0.759	0.660	0.690	0.67
NTS (ours)	**0.844**	**0.888**	**0.844**	**0.897**	**0.841**	**0.844**	**0.862**	**0.831**	0.856	0.78

- NTS: **AUROC 0.856 / F1 0.78**로 전체 최고 성능
- Binoculars 대비 **AUROC +4.9%p**
- 다중 도메인·생성 모델 환경에서 높은 일반화 성능 확인

실험 결과 및 성능 평가

HART 성능 평가

Detector	Level-3 Detection Task					Level-2 Detection Task					Level-1 Detection Task				
	ArXiv	Writ.	News	**ALL**	**F1**	ArXiv	Writ.	News	**ALL**	**F1**	ArXiv	Writ.	News	**ALL**	**F1**
Zero-Shot Detectors															
Entropy (Mitchell et al., 2023)	0.660	0.667	0.573	0.656	0.68	0.665	0.529	0.712	0.577	0.67	0.639	0.464	0.687	0.529	0.67
LogPPL (Bao et al., 2025)	0.850	0.810	0.733	0.799	0.75	0.485	0.438	0.596	0.473	0.67	0.530	0.625	0.407	0.576	0.67
Log-Rank (Bao et al., 2025)	0.874	0.813	0.762	0.814	0.77	0.460	0.441	0.571	0.465	0.67	0.542	0.611	0.418	0.573	0.67
LRR (Su et al., 2023)	0.909	0.797	0.841	0.840	0.78	0.616	0.551	0.556	0.573	0.67	0.576	0.558	0.520	0.568	0.67
Fast-DetectGPT (Bao et al., 2024) †	0.877	0.840	0.850	0.862	0.81	0.688	0.692	0.717	0.711	0.68	0.769	0.740	0.711	0.778	0.72
Binoculars (Hans et al., 2024) †	0.882	0.847	**0.866**	0.870	**0.83**	0.715	0.693	0.698	0.711	0.69	0.769	0.740	0.717	0.780	**0.73**
IRM (Liu et al., 2025) †	0.788	0.712	0.718	0.732	0.70	0.778	0.663	0.750	0.697	0.69	0.750	0.636	0.724	0.692	0.67
NTS (ours)	**0.915**	**0.875**	0.853	0.874	0.81	**0.840**	**0.817**	**0.780**	0.794	**0.72**	**0.818**	**0.802**	**0.726**	0.801	**0.73**

- NTS AUROC: **Level 3 0.874 / Level 2 0.794 / Level 1 0.801**
- 모든 AI 개입 수준에서 높은 성능 유지
- 특히 Level 2에서 Binoculars 대비 **+8.3%p**

실험 결과 및 성능 평가

강건성 및 일반화

Detector	Level-3 Detection Task					Level-2 Detection Task					Level-1 Detection Task				
	English	French	Spanish	Chinese	Arabic	English	French	Spanish	Chinese	Arabic	English	French	Spanish	Chinese	Arabic
Zero-Shot Detectors															
Fast-DetectGPT (Bao et al., 2024) †	0.850	0.865	0.830	0.867	0.590	0.688	0.773	0.695	0.836	0.466	0.711	**0.750**	0.732	0.835	0.432
Binoculars (Hans et al., 2024) †	**0.866**	0.881	0.846	0.869	0.571	0.697	**0.777**	0.708	0.838	0.450	0.716	0.747	0.740	**0.838**	0.424
IRM (Liu et al., 2025) †	0.718	0.605	0.577	0.621	0.378	0.750	0.638	0.547	0.579	0.436	0.724	0.595	0.540	0.576	0.391
NTS (ours)	0.853	**0.890**	**0.859**	**0.878**	**0.651**	**0.780**	0.776	**0.738**	**0.872**	**0.563**	**0.726**	0.741	**0.755**	0.832	**0.547**

- **5개 언어**에서 NTS의 다국어 일반화 성능 평가
- 다양한 언어·AI 개입 수준에서도 경쟁력 있는 성능 유지
- 공격 및 생성 조건 변화에서도 비교적 높은 강건성 확인

실험 결과 및 성능 평가

온도 및 대리 모델 분석

(T_{low}, T_{high})	(1.05,1.10)	(1.00,1.1)	(0.95,1.15)	(0.90,1.2)	(0.85,1.25)	(0.80,1.30)	(0.75,1.35)	(0.70,1.40)	(0.65,1.45)
NTS (ours)	0.979	0.973	0.973	0.975	0.977	0.980	0.982	0.980	0.963

Table 14: AUROC of NTS under different combinations of (low-T, high-T).

(Surroage Model)	Llama2-7b	Llama3.1-8b	Falcon-7b	Falcon-7b-ins	GPT-Neo-2.7b	GPT2-xl
NTS (ours)	0.921	0.954	0.980	0.767	0.890	0.731

Table 15: AUROC of NTS under different surrogate models.

- 다양한 온도 조합에서도 성능이 비교적 안정적으로 유지
- 기본 설정: Tlow=0.7,Thigh=1.4
- 대규모 **Base LLM**에서 높은 성능, Instruction-tuned 모델에서는 성능 저하

실험 결과 및 성능 평가

계산 효율성

Detector	Time (s)	Space (GB)
LogPPL (Bao et al., 2025)	0.105	18.25
IRM (Liu et al., 2025) †	0.210	36.14
NTS (ours)	0.106	18.81
(*Absolute* ↑)	**0.01%**	**3.00%**

- NTS: **0.106초 / 18.81GB**
- LogPPL과 거의 동일한 계산 비용
- **1개 대리 모델 + 1회 Forward Pass**로 높은 계산 효율성 확보

실험 결과 및 성능 평가

실험 결과 및 성능 평가 요약

- **RAID**: NTS AUROC **0.856**, Binoculars 대비 **+4.9%p**
- **HART**: 모든 Level에서 높은 성능, Level 2에서 **+8.3%p 향상**
- **강건성**: 다국어·Perturbation 환경에서도 경쟁력 있는 성능 유지
- **하이퍼파라미터**: 다양한 온도 조합에서도 성능이 비교적 안정적
- **효율성: 1개 대리 모델 + 1회 Forward Pass**, LogPPL과 유사한 계산 비용

한계 및 연구 적용 방안

연구의 한계

- **대규모 Base 대리 모델에 대한 의존성**
  - 소형 모델에서는 성능 감소
  - Instruction-tuned 모델과의 호환성 저하
- **High-Temperature Perturbation에 취약**
  - 높은 온도로 생성·변형된 텍스트에서 탐지 성능 감소

향후 연구 적용 방안

- **초단문 환경에 대한 검증 부족**
  - 매우 짧은 댓글 수준의 텍스트 성능은 확인되지 않음
- **한국어 성능 미검증**
  - 다국어 실험에 한국어가 포함되지 않음

Q&A