Binoculars 탐지 한계 분석 및 Multi-Observer 기반 확장 실험

발표자

류 동훈

DeepShark Lab/Hongik University

Date:

2026-05-22

목차

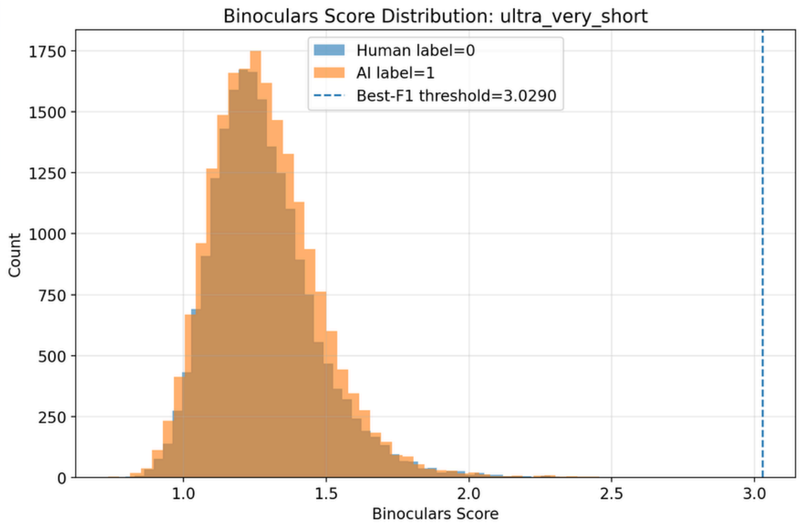
기존 Binoculars의 단문 탐지 한계

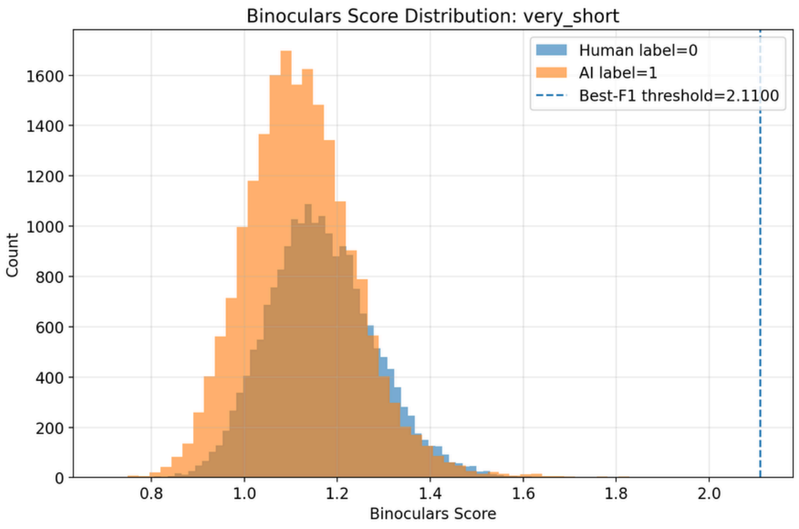
- 기존 Binoculars 기법을 짧은 댓글 데이터에 적용
- 특히 very short, ultra very short 구간에서 성능 붕괴 양상 확인
- Human 댓글과 AI 생성 댓글의 score distribution이 충분히 분리되지 않음
- 댓글 길이가 짧아질수록 PPL / Cross-PPL 기반 탐지 신호 약화

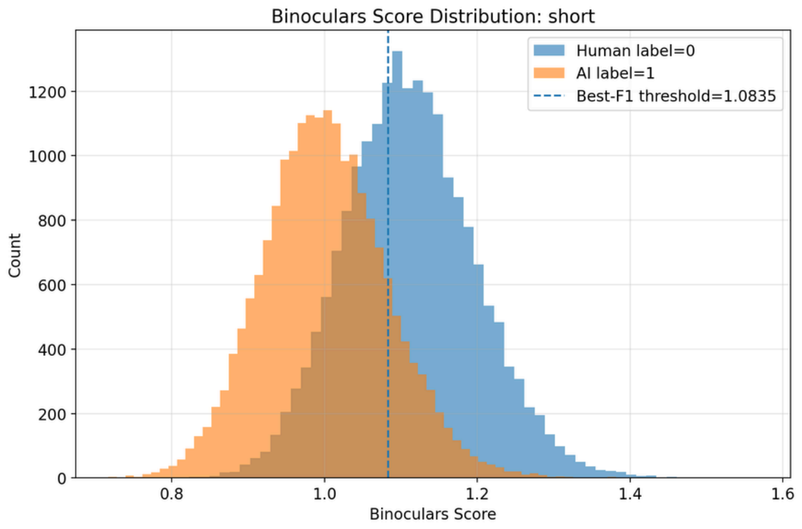
Reason

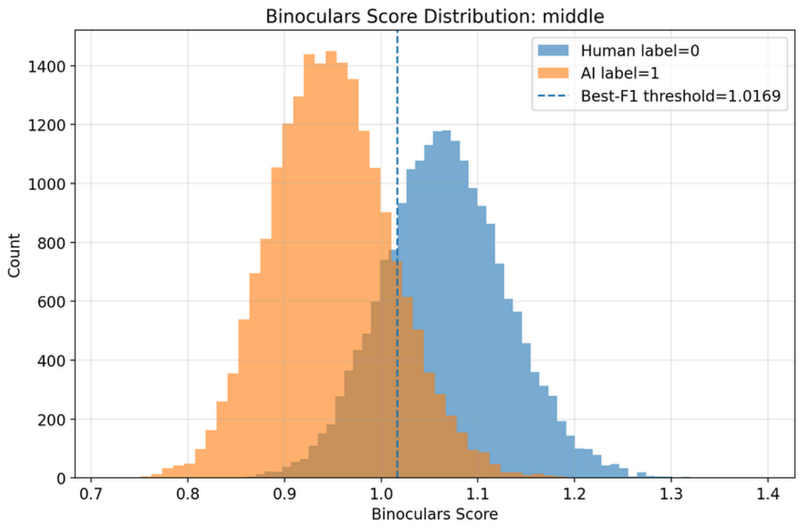
- 짧은 댓글은 문체적 특징이 충분히 드러나지 않음
- 토큰 수가 적어 일부 단어가 전체 score에 과도한 영향을 줌
- 댓글 특유의 구어체, 줄임말, 비문, 감탄사 등이 PPL을 불안정하게 만듦
- Human 댓글도 언어모델 기준에서는 부자연스럽게 평가될 수 있음
- 단일 observer 사용 시 특정 모델의 토크나이저·언어 분포 편향이 그대로 반영됨

기존 BINOCULARS의 단문 탐지 한계









기존 BINOCULARS의 단문 탐지 한계

구간	토큰 수	탐지 특성
ultra_very_short	5–20	분포 완전 중첩
very_short	21–40	분포 중첩 심함
short	41–80	일부 탐지 가능성 확인
middle	81–160	상대적으로 안정적인 탐지 가능

Multi-Observer 기반 1차 실험 설계

단일 Observer 구조의 문제점

- 기존 Binoculars는 하나의 Observer 모델과 하나의 Performer 모델을 사용
- Observer가 계산한 PPL과 Performer 기준 Cross-PPL의 관계를 기반으로 탐지
- 단문/초단문 댓글에서는 하나의 Observer score가 쉽게 **불안정**해질 수 있음
- 특정 모델의 토크나이저, 학습 데이터, 언어 분포 편향이 score에 직접 반영됨

$$B_{O,P}(X) = \frac{\log \mathrm{PPL}_O(X)}{\log \text{X-PPL}_{O,P}(X)}$$

- 기존 구조는 **하나**의 Observer 판단에 크게 의존함
- 단문 댓글에서는 이 단일 score의 신뢰도가 낮아질 수 있음

Multi-Observer 기반 1차 실험 설계

Multi-Observer 구조 도입

- 단일 Observer 의존성을 줄이기 위해 Observer를 **여러 개**로 확장
- 기존 Binoculars의 M1,M2 구조는 유지
- 단, M1에 해당하는 Observer를 O1, O2, O3...On으로 다중화
- 동일한 댓글을 여러 Observer 모델이 각각 평가
- Performer P는 고정
- 동일한 댓글 x를 여러 Observer가 각각 평가
- 각 Observer별 Binoculars score를 계산한 뒤 결합하여 최종 score 산출

$$\mathrm{O} = \{O_1, O_2, \ldots, O_n\}$$

$$B_{O_i,P}(x) = \frac{\log \mathrm{PPL}_{O_i}(x)}{\log \mathrm{X\text{-}PPL}_{O_i,P}(x)}$$

$$S_{\mathrm{OP}}(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{\log \mathrm{PPL}_{O_i}(x)}{\log \mathrm{X\text{-}PPL}_{O_i,P}(x)}$$

Multi-Observer 기반 1차 실험 설계

Observer / Performer 모델 구성

역할	모델	기반 토크나이저	사용 목적
Observer 1	google/gemma-2-2b	SentencePiece 계열 subword tokenizer	Baseline Observer로 사용. 기존 Gemma 기반 Binoculars score의 기준점 확보
Observer 2	Qwen2.5-1.5B	BPE 계열 multilingual tokenizer	Gemma와 다른 모델 계열의 반응 비교. Observer 다양성 확보
Observer 3	polyglot-ko-1.3b	GPT-NeoX 계열 BPE tokenizer	한국어 특화 모델 반응 확인. 한국어 댓글 표현에 대한 민감도 비교
Performer	google/gemma-2-9b	SentencePiece 계열 subword tokenizer	Cross-PPL 계산 기준. Performer를 고정하여 Observer 변화의 영향 비교

Multi-Observer 기반 1차 실험 설계

실험 데이터 및 길이 버킷 구성

Bucket	토큰 수	단어 개수 평균 값	AI 작성 댓글 개수	사람 작성 댓글 개수	총 데이터셋 개수
ultra_very_short	5–20	4.85	20,000	20,000	40,000
very_short	21–40	10.35	20,000	20,000	40,000
short	41–80	18.44	20,000	20,000	40,000
middle	81–160	34.01	20,000	20,000	40,000
Total	-	-	**80,000**	**80,000**	**160,000**

결과 해석

Ratio 계열

방식	계산 방식	목적
Ratio of Means	Observer PPL 평균 후 ratio 계산	PPL 변동성 완화
Mean of Ratios	Observer별 ratio 계산 후 평균	Observer별 반응 차이 반영

Ratio of Means

$$S_{\mathrm{RoM}}(x) = \frac{\frac{1}{n}\sum_{i=1}^{n} \log \mathrm{PPL}_{O_i}(x)}{\log \text{X-PPL}_{O,P}(x)}$$

Mean of Ratios

$$S_{\mathrm{MoR}}(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{\log \mathrm{PPL}_{O_i}(x)}{\log \text{X-PPL}_{O_i,P}(x)}$$

결과 해석

Variance-Aware

- Ratio 계열 방식의 한계를 보완하기 위해 Observer 간 분산 정보 반영
- 단순 평균 score뿐 아니라, Observer들이 서로 얼마나 다르게 반응했는지 측정
- 동일한 댓글에 대해 모델별 score 차이가 클 경우 이를 추가 탐지 신호로 활용
- 단문 댓글에서는 평균값보다 **모델 간 반응 차이**가 더 유의미한 신호가 될 수 있음

Observer별 Binoculars Score

$$B_i(x) = \frac{\log \mathrm{PPL}_{O_i}(x)}{\log \mathrm{X\text{-}PPL}_{O_i,P}(x)}$$

평균 계산

$$\mu_B(x) = \frac{1}{n} \sum_{i=1}^{n} B_i(x)$$

표준 편차

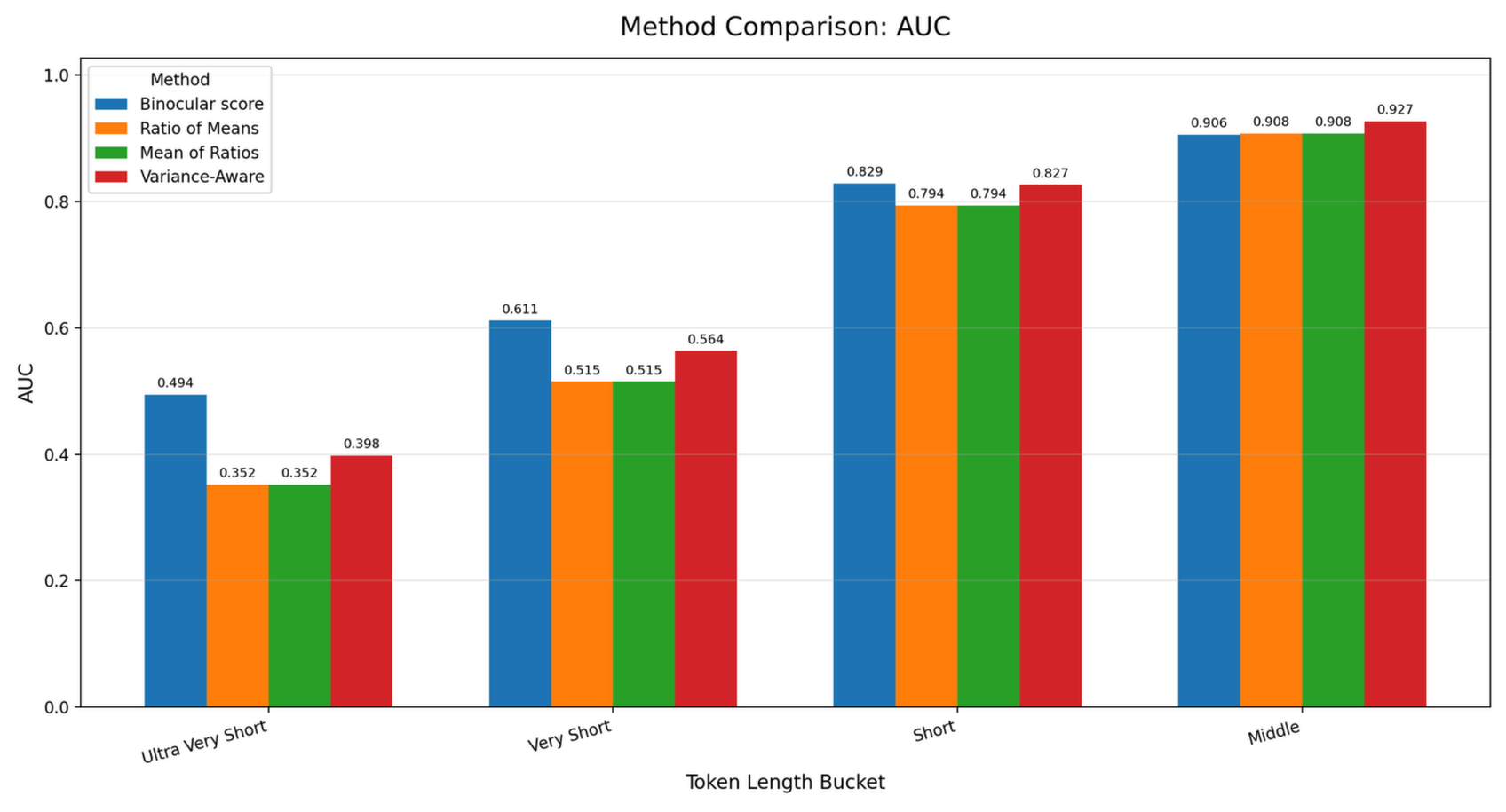
$$\sigma_B(x) = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (B_i(x) - \mu_B(x))^2}$$

Variance-Aware Score

$$S_{\mathrm{VA}}(x) = \mu_B(x) + \lambda \sigma_B(x)$$

결과 해석

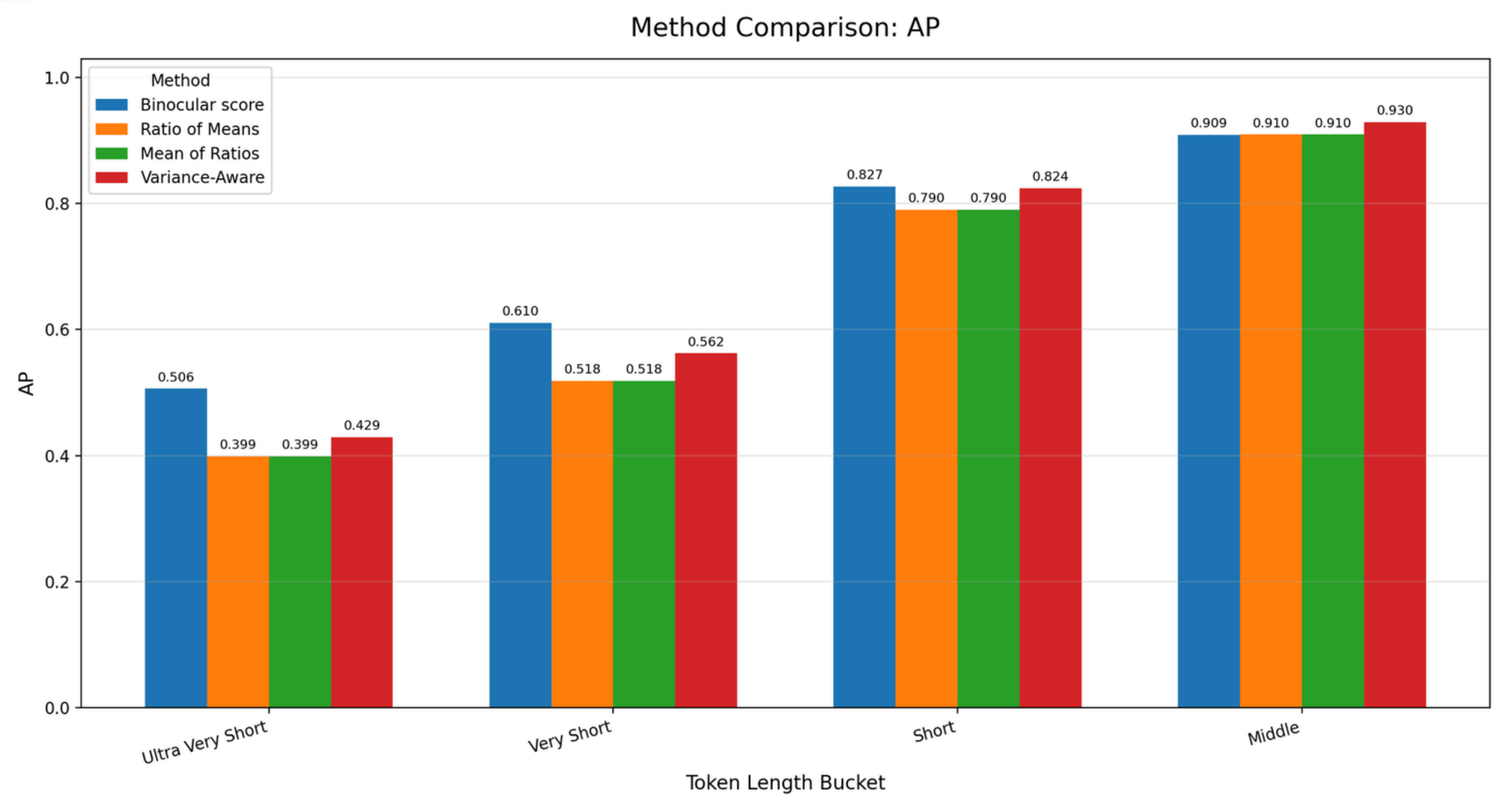
1차 실험 결과 요약: AUC 기준



- Variance-Aware는 Ratio 계열보다 낮지만, 초단문에서는 기존 Binocular score를 넘지 못함

결과 해석

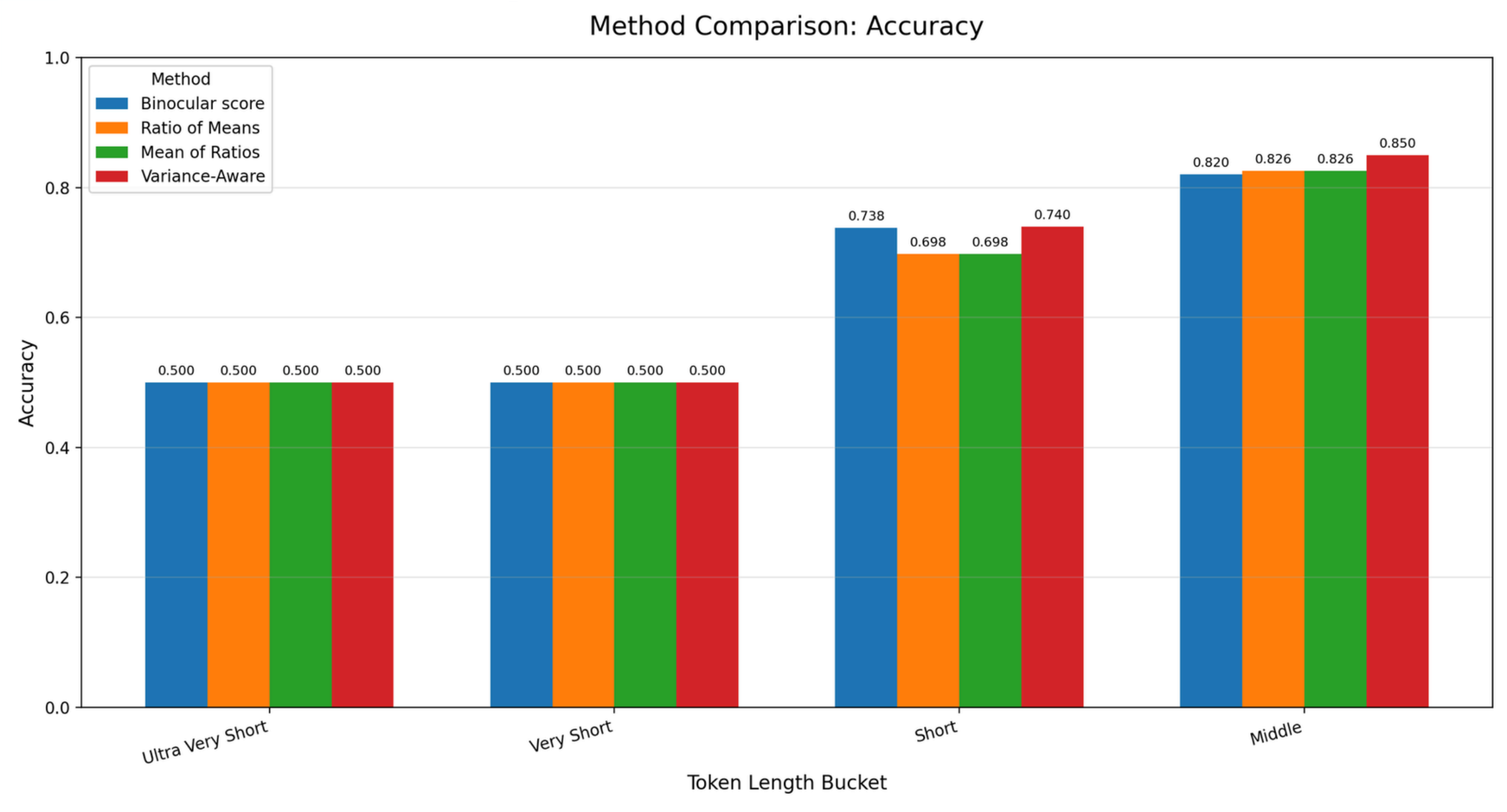
1차 실험 결과 요약: AP 기준 성능 비교



- Variance-Aware는 중·장문 구간에서 유효하지만, 초단문 보완 효과는 제한적

결과 해석

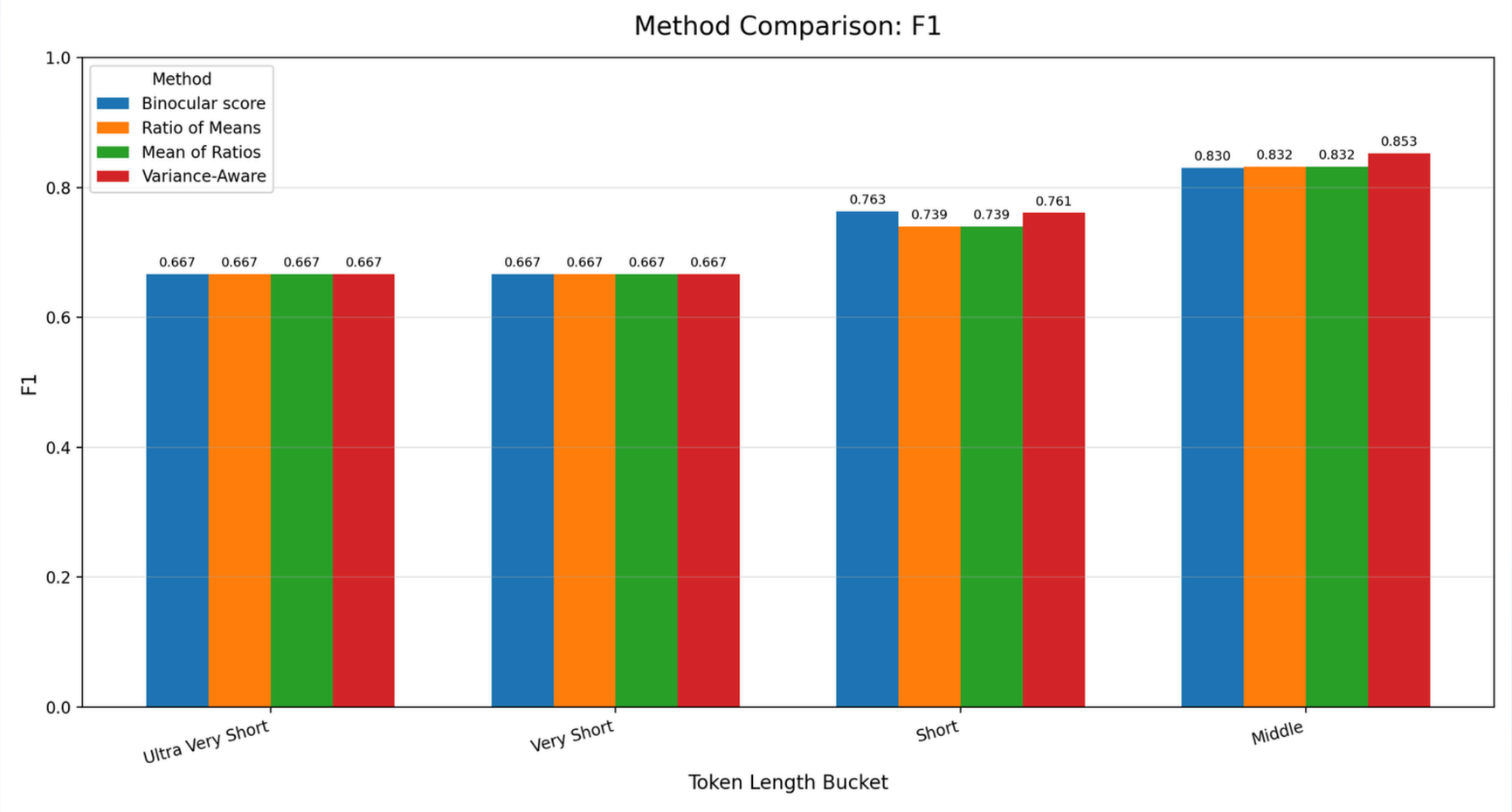
1차 실험 결과 요약: Accuracy 기준



- 40토큰 이하에서는 score 방식과 무관하게 threshold 분류가 붕괴

결과 해석

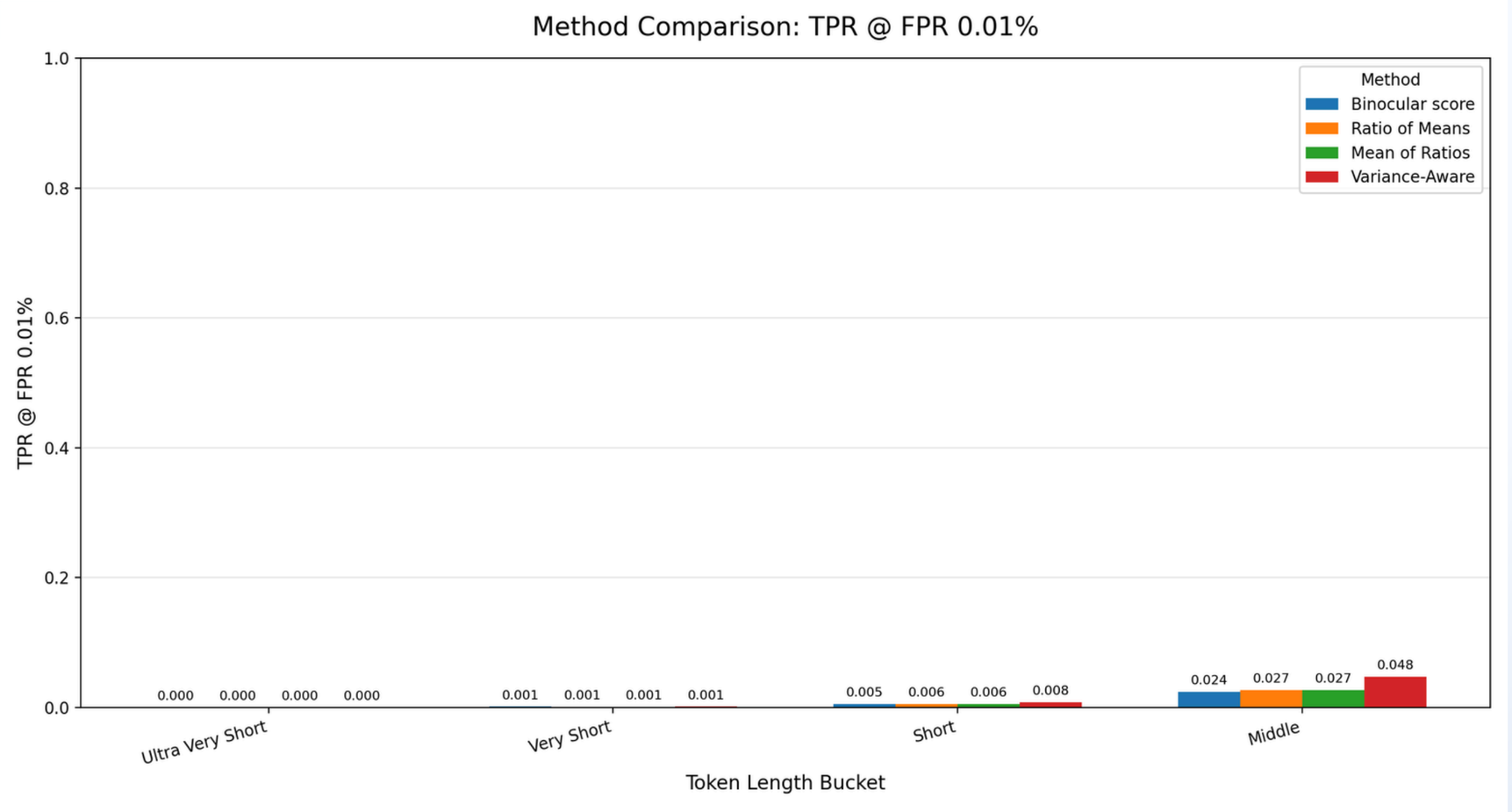
1차 실험 결과 요약: F1 기준



- 초단문 성능은 F1보다 AUC, Accuracy, Low-FPR TPR 중심으로 해석 필요

결과 해석

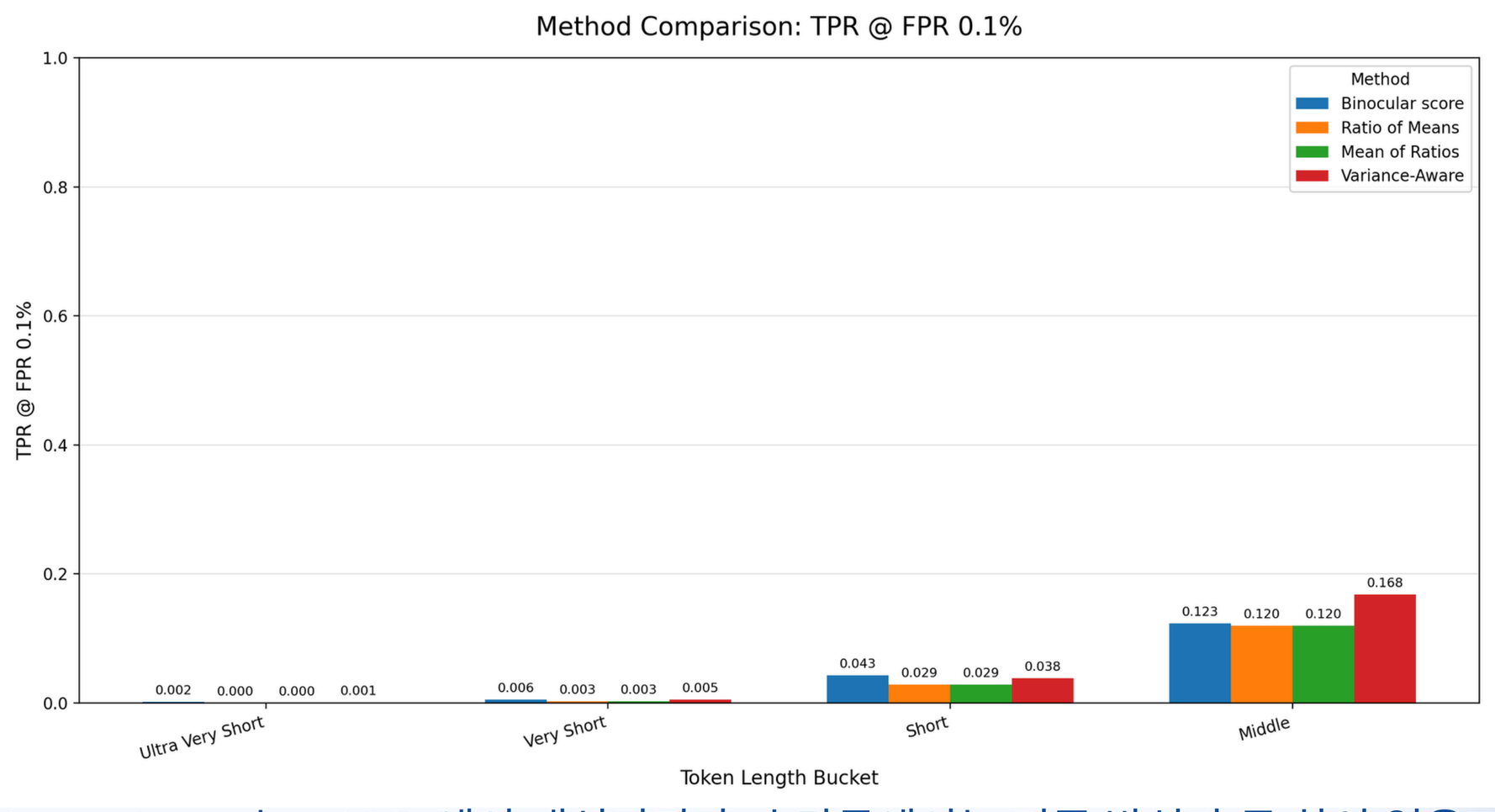
Low-FPR 결과: TPR @ FPR 0.01%



- FPR 0.01% 조건에서는 모든 방식이 실용적 탐지율을 확보하지 못함

결과 해석

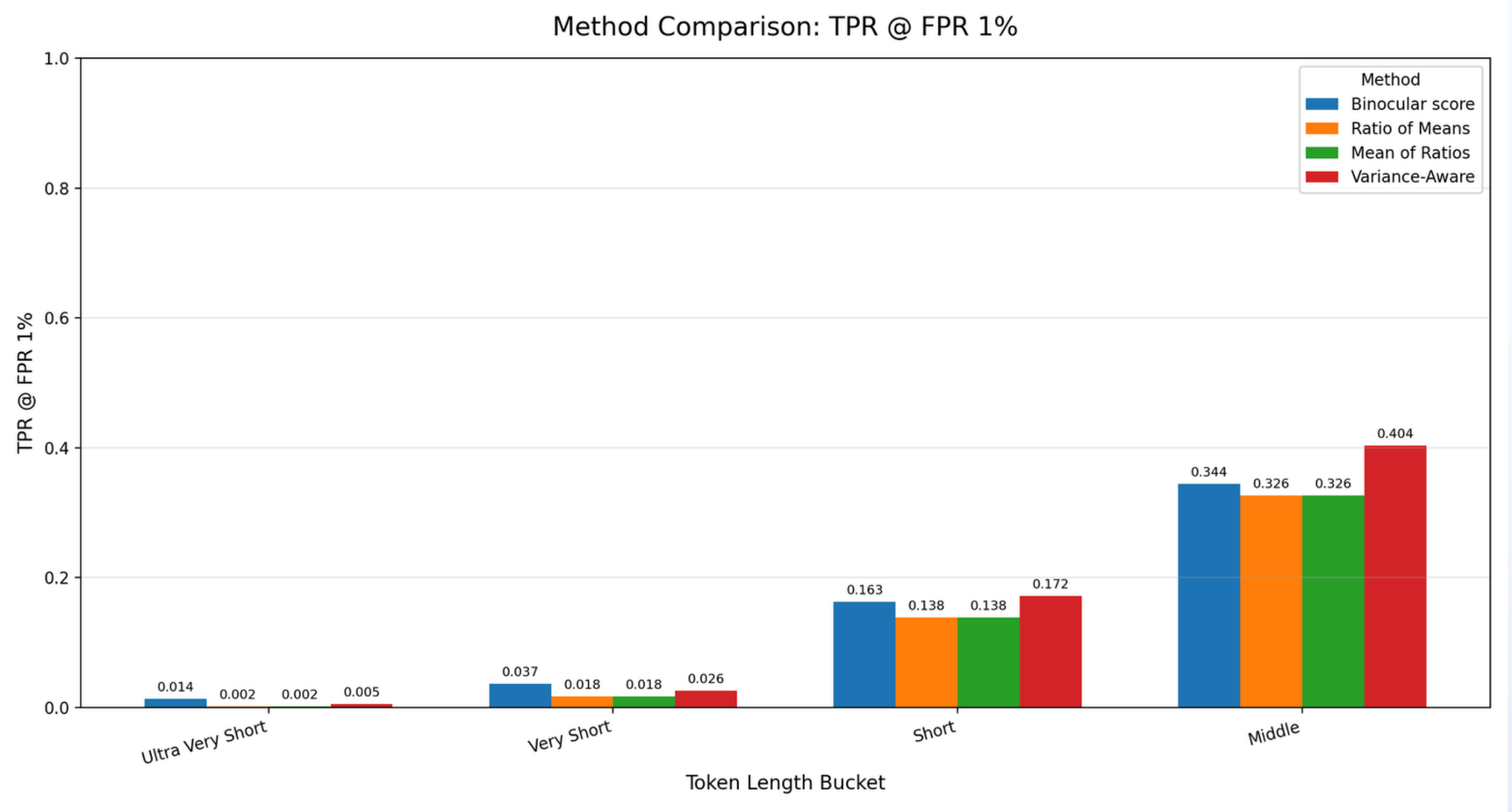
Low-FPR 결과: TPR @ FPR 0.1%



- Variance-Aware는 middle에서 개선되지만, 초단문에서는 기존 방식과 큰 차이 없음

결과 해석

Low-FPR 결과: TPR @ FPR 1%



- Variance-Aware는 short/middle에서 효과가 있으나, 초단문 탐지 한계는 유지

결과 해석

결과 해석 종합

- 기존 Binocular score는 **초단문**에서 상대적으로 안정적
- Ratio 계열은 기존보다 **성능 저하**
- Variance-Aware는 Ratio 계열 대비 **개선**
- short / middle에서 Variance-Aware **가능성 확인**
- ultra / very short에서는 **기존 Binocular score를 넘지 못함**

- **모든 방식에서 초단문 hard prediction 한계 확인**

진행 중인 Binoculars 변형 기법

- 단순 Multi-Observer 결합만으로는 초단문 score 불안정성 해결 어려움
- Binoculars score를 변형하는 후속 실험 진행 중
- 핵심 방향은 **길이 보정**, **Entropy 반영**, **Observer 간 반응 차이**, **Score 신뢰도 보정**

구분	변형 방향	목적
F1	Length-aware / LTV	댓글 길이 보정
F2	Entropy Gap	모델 예측 확신도 반영
F3	Observer Disagreement	Observer 간 반응 차이 반영
F4	Length Normalized Log Ratio	길이별 score scale 정규화
F5	Bayesian Reliability	짧은 댓글 score 신뢰도 보정

진행 중인 Binoculars 변형 기법

F1:Length-aware Binoculars: 댓글 길이 보정

- 단문/초단문 댓글은 토큰 수가 적어 score 변동성이 큼
- 일부 단어가 전체 PPL / X-PPL에 과도한 영향
- 기존 Binoculars score에 댓글 길이 정보를 반영
- 길이가 짧을수록 score 신뢰도를 낮게 보정
- 초단문 구간의 **불안정한** score 완화 목적

$$B(x) = \frac{\log \mathrm{PPL}(x)}{\log \text{X-PPL}(x)}$$

기존 Binoculars score

$$g(L(x)) = \log(1 + L(x))$$

길이보정값

길이가 늘어날수록 보정값은 커지지만
너무 급격하게 커지지는 않게 하기 위함.

$$S_{\mathrm{F1}}(x) = B(x) \cdot g(L(x))$$

제안식

댓글 길이를 score에 반영해서 초단문에서 흔들리는 Binoculars score를 보정하는 방식

진행 중인 Binoculars 변형 기법

F2:Entropy Gap: 모델 예측 확신도 반영

- 기존 Binoculars는 PPL / X-PPL 비율에 집중
- 하지만 단문 댓글에서는 ratio만으로 **탐지 신호**가 부족할 수 있음
- Performer의 다음 토큰 예측 분포를 **entropy**로 측정
- 모델이 얼마나 확신 있게 다음 토큰을 예측했는지 반영
- LLM 생성 댓글과 Human 댓글의 예측 확신도 차이 활용

$$H_P(x_t) = -\sum_{v \in V} p_P(v \mid x_{<t}) \log p_P(v \mid x_{<t})$$

토큰 시점 t의 entropy

$$H_P(x) = -\frac{1}{T} \sum_{t=1}^{T} \sum_{v \in V} p_P(v \mid x_{<t}) \log p_P(v \mid x_{<t})$$

문장 전체의 Entropy

$$S_{\mathrm{F2}}(x) = B(x) + \alpha H_P(x)$$

제안식

PPL 비율만 보지 않고, 모델이 다음 토큰을 얼마나 확신했는지도 함께 보는 방식

진행 중인 Binoculars 변형 기법

F3:Observer Disagreement: 모델 간 반응 차이 반영

- Variance-Aware 실험에서 모델 간 반응 차이의 가능성 확인
- 각 Observer가 동일 댓글에 대해 산출한 score 차이를 직접 반영
- 단순 평균이 제거하는 모델 간 disagreement를 탐지 신호로 사용
- Human 댓글과 AI 댓글이 Observer 간 반응에서 다른 패턴을 보일 수 있음
- Variance-Aware를 수식 확장 방향으로 정리한 방식

$$B_i(X) = \frac{\log \mathrm{PPL}_{O_i}(X)}{\log \text{X-PPL}_{O_i,P}(X)}$$

Observer별 score

$$D(X) = \sigma_B(X) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(B_i(X) - \mu_B(X))^2}$$

Disagreement 정의

$$S_{\mathrm{F3}}(X) = \mu_B(X) + \beta D(X)$$

제안식

여러 Observer가 같은 댓글을 얼마나 다르게 판단하는지를 탐지 신호로 쓰는 방식

진행 중인 Binoculars 변형 기법

F4:Length Normalized Log Ratio: 길이별 score 정규화

- 댓글 길이에 따라 Binoculars score scale이 달라질 수 있음
- 초단문에서는 적은 토큰 수 때문에 score가 과도하게 흔들림
- 기존 log ratio를 길이 함수로 정규화
- 길이가 다른 댓글 간 score 비교 가능성 개선
- 길이 버킷 간 편차 완화 목적

$$S_{\mathrm{F4}}(x) = \frac{\log \mathrm{PPL}(x)}{\log \mathrm{X\text{-}PPL}(x)} \Big/ \log(1 + L(x))$$

제안식

기존 Binoculars score를 댓글 길이로 나누어 길이별 score scale을 맞추는 방식

진행 중인 Binoculars 변형 기법

F5: Bayesian Reliability: score 신뢰도 보정

- 모든 댓글의 Binoculars score를 동일하게 신뢰하기 어려움
- 5–20토큰 댓글과 80–160토큰 댓글의 score 신뢰도는 다름
- 댓글 길이가 짧을수록 score 반영 비중을 낮춤
- 길이가 충분할수록 기존 Binoculars score를 더 신뢰
- 초단문 구간의 과도한 오판 완화 목적

$$R(x) = \frac{L(x)}{L(x) + k}$$

Reliability weight

$$S_{\mathrm{F5}}(x) = \frac{L(x)}{L(x) + k} \cdot \frac{\log \mathrm{PPL}(x)}{\log \text{X-PPL}(x)}$$

제안식

댓글 길이에 따라 Binoculars score를 얼마나 믿을지 다르게 정하는 방식

진행 중인 Binoculars 변형 기법

요약

구분	변형 기법	핵심 반영 요소	목적
F1	Length-aware Binoculars	댓글의 토큰 길이를 PPL 계산 또는 score 보정에 직접 반영	초단문 댓글에서 일부 토큰이 전체 score를 과도하게 흔드는 문제를 완화
F2	Entropy Gap	Performer 모델의 다음 토큰 예측 분포와 entropy 값을 추가 신호로 활용	모델이 얼마나 확신 있게 예측했는지를 반영해 PPL / X-PPL ratio만으로 부족한 탐지 신호를 보완
F3	Observer Disagreement	여러 Observer가 동일 댓글에 대해 보이는 PPL / X-PPL 반응 차이를 반영	Human 댓글과 AI 댓글이 모델 간 반응 차이에서 다른 패턴을 보이는지 확인
F4	Length Normalized Log Ratio	기존 Binoculars log-ratio를 댓글 길이 함수로 정규화	길이 구간별 score scale 차이를 줄여 서로 다른 길이의 댓글을 더 안정적으로 비교
F5	Bayesian Reliability	댓글 길이에 따라 산출된 score의 신뢰도를 다르게 부여	정보량이 부족한 초단문 score는 낮게 신뢰하고, 길이가 충분한 댓글 score는 더 강하게 반영

Q&A