

짧은 LLM 생성 댓글 탐지를 위한 Binoculars 확장 실험

문장 길이 보정 · Heterogeneous Multi-Observer · H-MOSP v2

DeepShark Lab 류동훈

2026-10-16

01 연구 배경 및 실험 설계

02 1차 실험 결과

03 F3/H-MOSP 오류 분석 및 수정

04 H-MOSP v2 실험 결과

05 결론 및 향후 연구

발표 목적

이번 발표에서 답할 세 가지 질문

01 짧은 문장에서 왜 탐지 성능이 떨어지는가?

02 어떤 신호를 추가하면 성능이 좋아지는가?

03 여러 신호를 결합하면 얼마나 개선되는가?

1차 비교



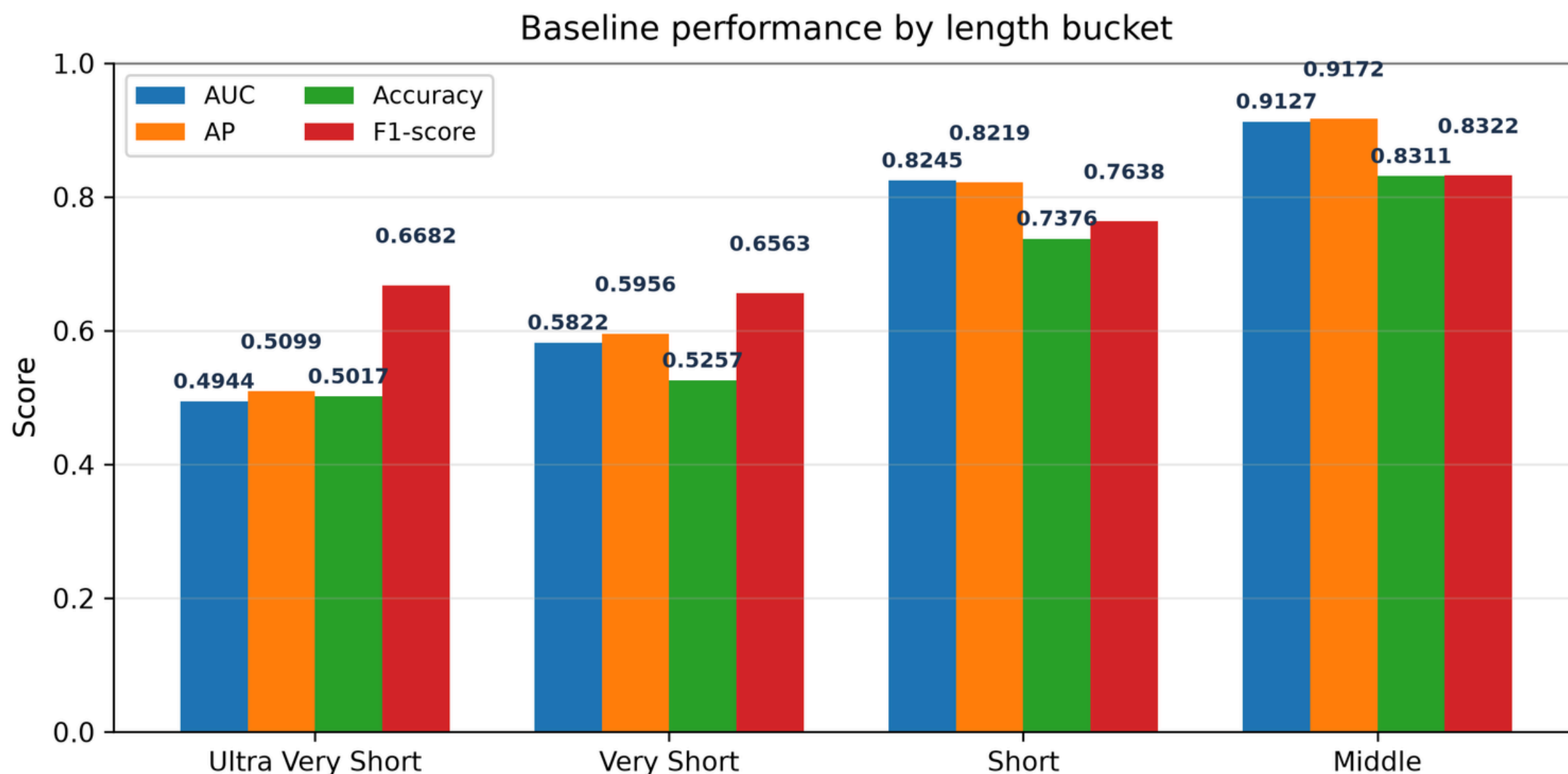
오류 분석·수정



H-MOSP v2

Binoculars의 단문 탐지 한계

현재 데이터에서는 짧은 구간일수록 Baseline 탐지 성능이 낮음



Binoculars 핵심 식

$$B(x) = \frac{NLL_O(x)}{\max(CE_{P \rightarrow O}(x), \varepsilon)}$$

$NLL_O(x)$ Observer의 평균 NLL

$CE_{\{P \rightarrow O\}}(x)$ Performer→Observer cross-entropy

점수가 낮을수록 AI 가능성이 높음

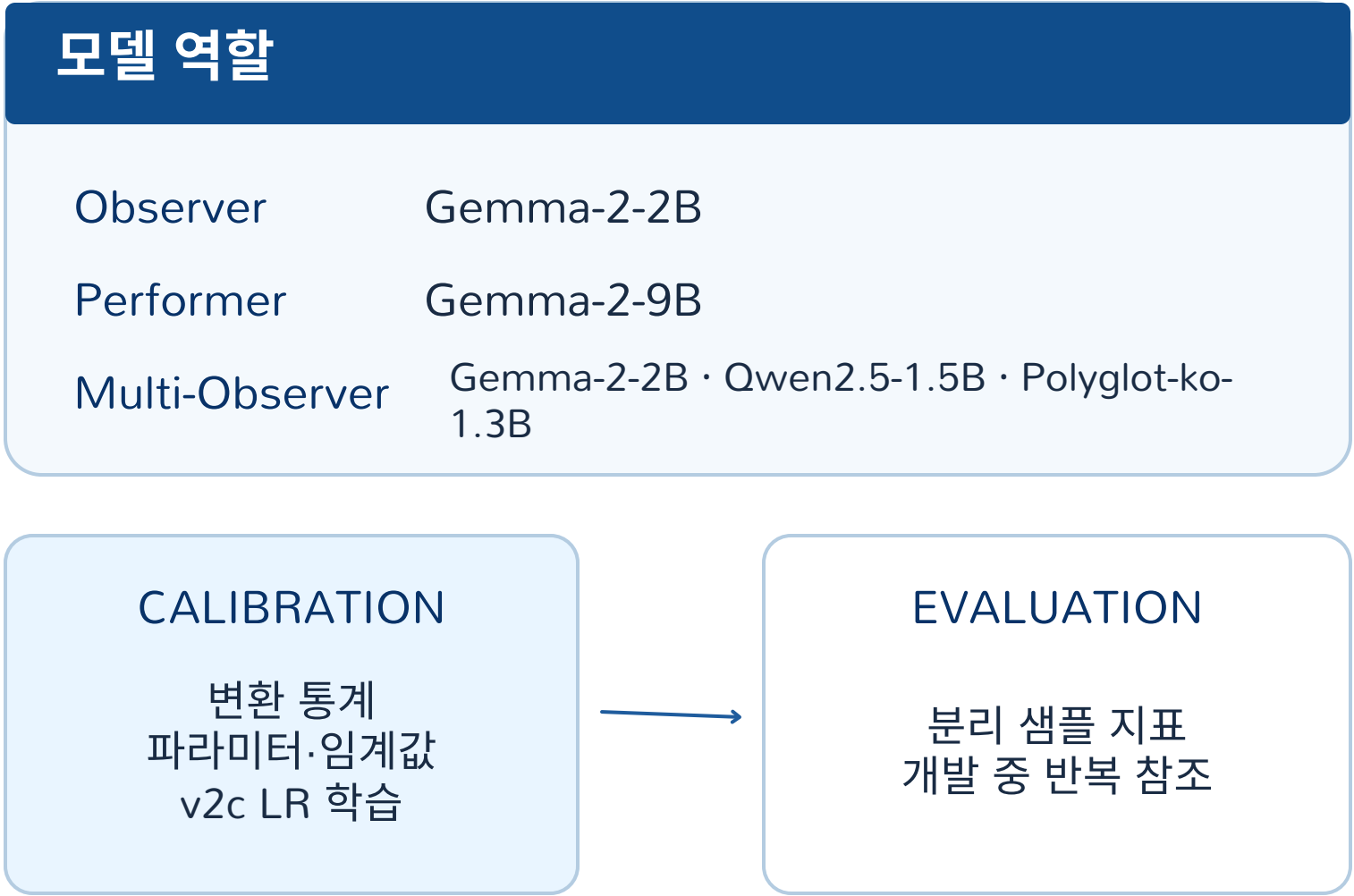
AUC 0.4944 → 0.9127

Ultra Very Short → Middle

데이터 및 실험 환경

동일한 2만 개 샘플과 같은 calibration/evaluation 분할에서 원본·fixed·v2를 비교함

Bucket	Human / AI	Calibration	Evaluation
Ultra Very Short	2,500 / 2,500	1,432	3,568
Very Short	2,500 / 2,500	1,511	3,489
Short	2,500 / 2,500	1,482	3,518
Middle	2,500 / 2,500	1,490	3,510
합계	10,000 / 10,000	5,915	14,085



검증 범위

데이터 생성 조건과 원래 bucket 생성 규칙은 제공 자료로 확인하지 못함 · 독립 최종 test로 표현하지 않음

※sample_id 기반 deterministic hash split · Calibration ≈ 30%

세 가지 확장 방향

먼저 어떤 신호를 추가하는지 세 가지 실험 방향으로 구분함

길이 보정	Performer 신호	Multi-Observer 신호
F1 · F4 · F5 목적 짧은 문장에서 달라지는 길이 효과를 점수에 반영 대표식 $B(x) \log(1 + L)$	F2 목적 Performer의 token entropy를 Baseline에 추가 대표식 $B(x) + \alpha H(x)$	F3 · H-MOSP 목적 여러 Observer의 평균 신호와 disagreement를 결합 대표식 $\mu(x) + \beta d(x)$

기호 설명 | L: 문장 길이 · H(x): Performer 평균 token entropy · μ: Observer 평균 · d: disagreement · α, β: 결합 가중치

Baseline 및 1차 확장 방법 계산식

Baseline과 세 가지 확장 방향의 정확한 계산식을 한 장에서 비교함

Baseline

$$B(x) = \frac{\text{NLL}_O(x)}{\max(\text{CE}_{P \rightarrow O}(x), \varepsilon)}$$

F1 (method)

$$B(x) \log(1 + L)$$

F4

$$\frac{B(x)}{\max(\log(1 + L), \varepsilon)}$$

F5

$$\frac{L}{L + k} B(x)$$

F2 • Performer

$$B(x) + \alpha H(x)$$

F3 • Multi-Observer

$$\mu(x) + \beta d(x)$$

H-MOSP

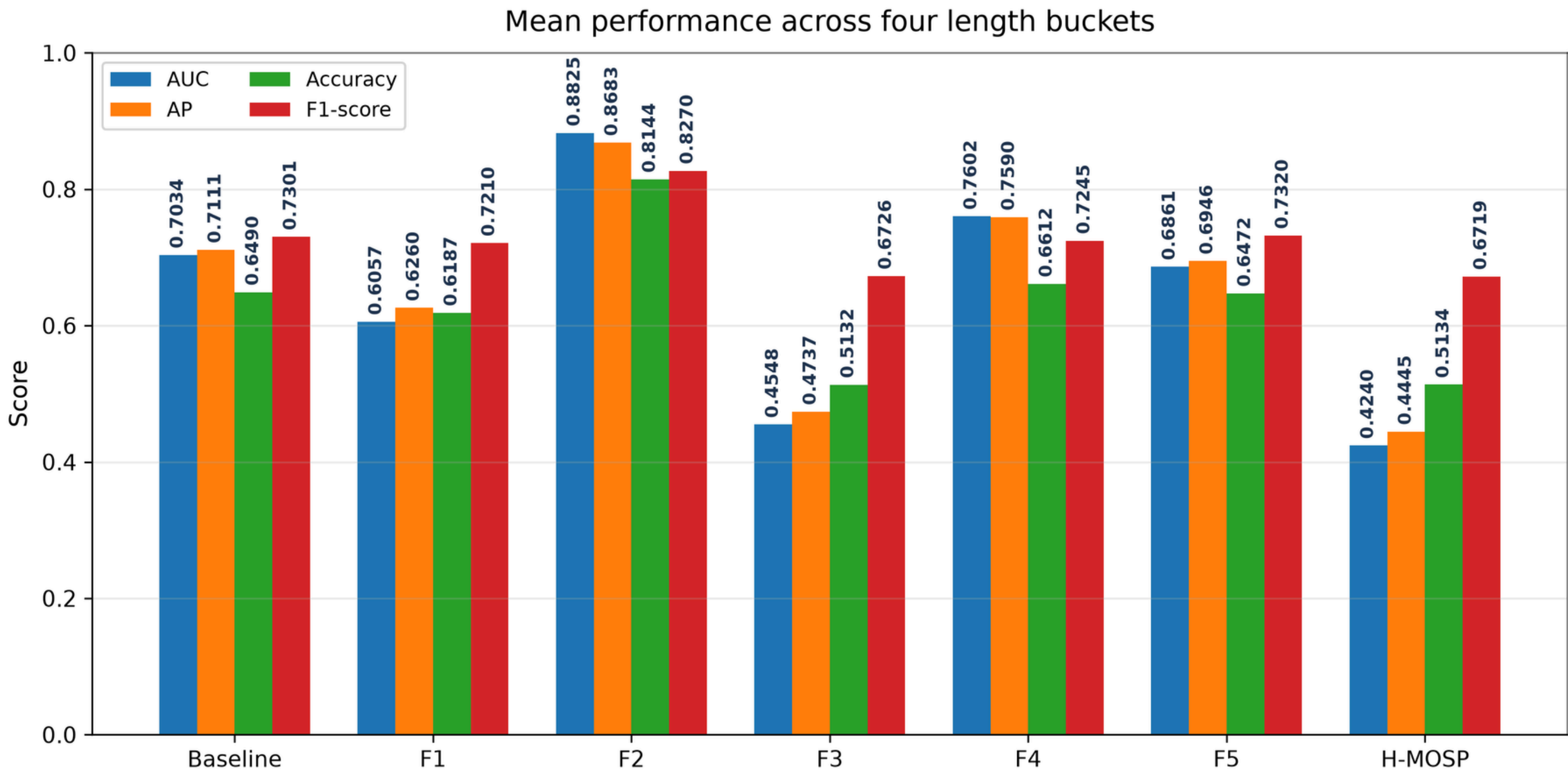
$$Z_c(\mu(x)) + \gamma U(x)$$

기호 설명

k: 길이 보정 상수 · ε: 수치 안정성 상수 · γ: H-MOSP 결합 가중치
Z_c: calibration 통계 기반 robust scaling · U: 양의 disagreement · α, β, γ, k: labeled calibration에서 선택

1차 실험 성능 비교

원본 7방법 중 F2가 평균 AUC 0.8825, 평균 F1-score 0.8270으로 가장 높음



평균 최고 성능

F2

AUC0.8825

F1-score0.8270

평균 AUC
F3 원본 0.4548
H-MOSP 원본 0.4240

해석 범위

Entropy 포함 점수 개선
단독 기여·인과 효과 미검증

Multi-Observer 점수 구성

Observer별 log-ratio에서 평균 신호와 disagreement를 계산한 뒤 robust scaling을 적용함

$r_j(x)$

$$r_j(x) = \log \frac{\max(n_j(x), \epsilon)}{\max(n_P(x), \epsilon)}$$

Observer-Performer log-ratio
이 Observer가 문장을 보는 NLL과 Performer가 보는 NLL의 비율을 로그로 표현한 상대적 반응

$\mu(x)$

$$\mu(x) = \frac{1}{M} \sum_{j=1}^M r_j(x)$$

Observer log-ratio 평균
여러 Observer가 공통적으로 보여주는 평균 반응

$d(x)$

$$d(x) = \sqrt{\frac{1}{M} \sum_{j=1}^M (r_j(x) - \mu(x))^2}$$

Observer 간 disagreement
Observer끼리 얼마나 의견이 다른가

$Z_c(s)$

$$Z_c(s) = \frac{s - \text{median}_{\text{cal}}(s)}{\max(\text{IQR}_{\text{cal}}(s), \epsilon)}$$

calibration 통계 기반 robust scaling
이 값이 calibration 데이터의 일반적인 값에 비해 얼마나 높거나 낮은가

$U(x)$

$$U(x) = \max(Z_c(d(x)), 0)$$

양의 disagreement만 유지
calibration 기준보다 disagreement가 낮은 경우는 무시하고,
평소보다 큰 disagreement만 불확실성으로 남김

기호 설명	$r_j(x)$: j번째 Observer log-ratio · n_j/n_P : Observer/Performer native NLL/byte	M : Observer 개수, $M=3$ ($j=1,2,3$)	$\mu(x)$: Observer log-ratio 평균
	$d(x)$: Observer 간 표준편차, disagreement	$Z_c(\cdot)$: calibration 통계 기반 robust scaling	$U(x)$: 양의 disagreement만 남긴 값

F3/H-MOSP 방향 수정

핵심 변화: + disagreement → - disagreement

원본

낮을수록 AI

$$F3_{\text{original}}(x) = \mu(x) + \beta d(x)$$

$$\text{H-MOSP}_{\text{original}}(x) = Z_c(\mu(x)) + \gamma U(x)$$

+ disagreement

수정

높을수록 AI

$$F3_{\text{fixed}}(x) = \mu(x) - \beta d(x)$$

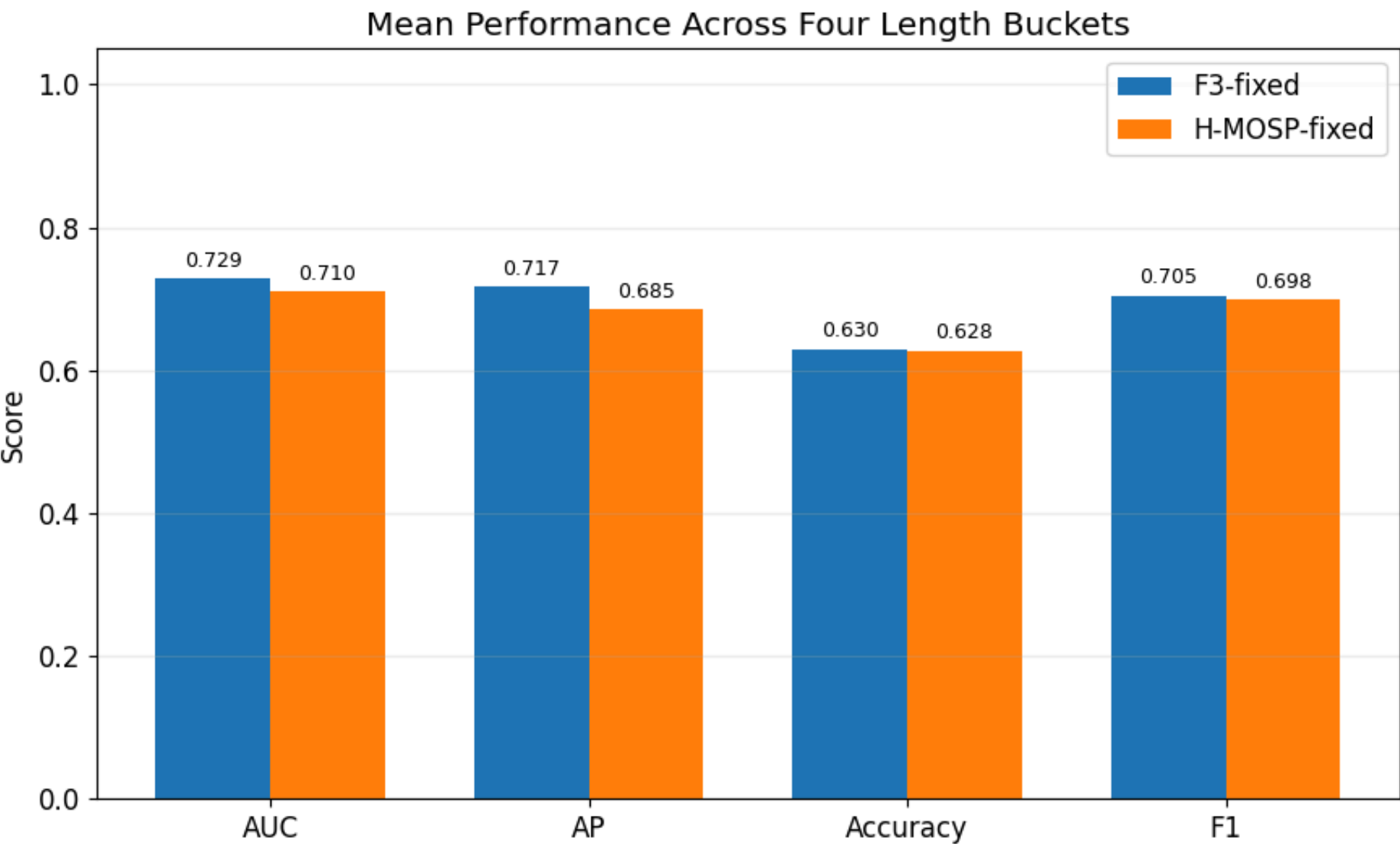
$$\text{H-MOSP}_{\text{fixed}}(x) = Z_c(\mu(x)) - \gamma U(x)$$

- disagreement

+ disagreement → - disagreement · 낮을수록 AI → 높을수록 AI · β, γ 를 calibration에서 재선택

F3-fixed 및 H-MOSP-fixed

방향·결합·calibration 수정 후 평균 성능은 회복했지만 F2 평균에는 미치지 못함



수정 결과

$$F3_{\text{fixed}}(x) = \mu(x) - \beta d(x)$$
$$H\text{-MOSP}_{\text{fixed}}(x) = Z_c(\mu(x)) - \gamma U(x)$$

높을수록 AI

방법	원본 → 수정 평균 AUC	수정 F1-score
F3	0.4548 → 0.7291	0.7048
H-MOSP	0.4240 → 0.7104	0.6982

F2 평균 AUC 0.8825 / F1-score 0.8270

예외: Ultra Very Short의 F1-score는 수정 버전이 F2보다 높음

H-MOSP v2 입력 신호

네 신호를 모두 '높을수록 AI' 방향으로 정렬함

Binoculars

$$e_B(x) = -Z_c(B(x))$$

Baseline의 AI 방향 증거

Performer Entropy

$$e_H(x) = -Z_c(H(x))$$

Performer Entropy 증거

Proxy Mean

$$e_M(x) = Z_c(\mu(x))$$

Multi-Observer 평균 신호

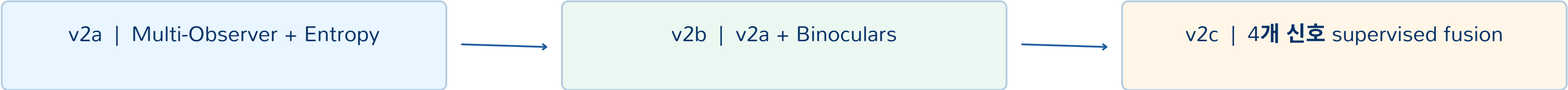
Observer Agreement

$$e_A(x) = -Z_c(d(x))$$

disagreement 반대 방향 · clipping 제거

H-MOSP v2 결합 단계

v2a → v2b → v2c 순서로 신호와 학습 방식을 확장함



v2a

$$S_{v2a}(x) = e_M(x) + \gamma_a e_A(x) + \alpha_a e_H(x)$$

v2b

$$S_{v2b}(x) = e_M(x) + \gamma_b e_A(x) + \alpha_b e_H(x) + \lambda_b e_B(x)$$

v2c

$$u_j(x) = \frac{e_j(x) - \text{mean}_{\text{cal}}(e_j)}{\text{std}_{\text{cal}}(e_j)}$$

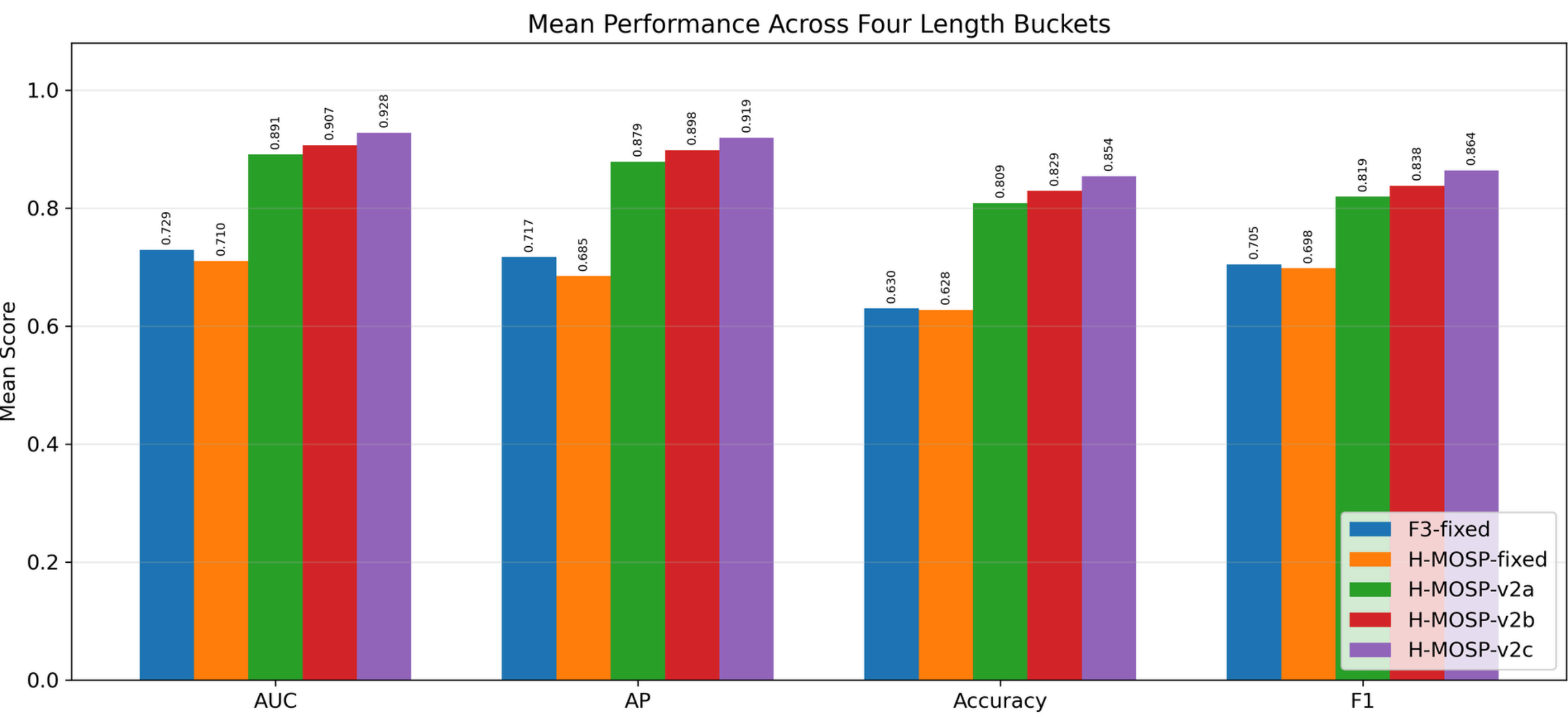
calibration label을 사용해 Logistic Regression을 supervised 학습

$$P(\text{AI} \mid x) = \sigma\left(b + \sum_j w_j u_j(x)\right)$$

기호 설명	γ_a, α_a: v2a 결합 가중치	γ_b, α_b, λ_b: v2b 결합 가중치	u_j(x): 표준화된 j번째 입력 신호
	w_j: Logistic Regression 계수	b: bias	σ: sigmoid 함수

H-MOSP v2 성능 비교

평균 성능은 개선되지만, F1 및 길이별 비교에는 예외가 존재함



4개 길이 bucket별 evaluation 지표의 비가중 평균

평균 비교의 예외

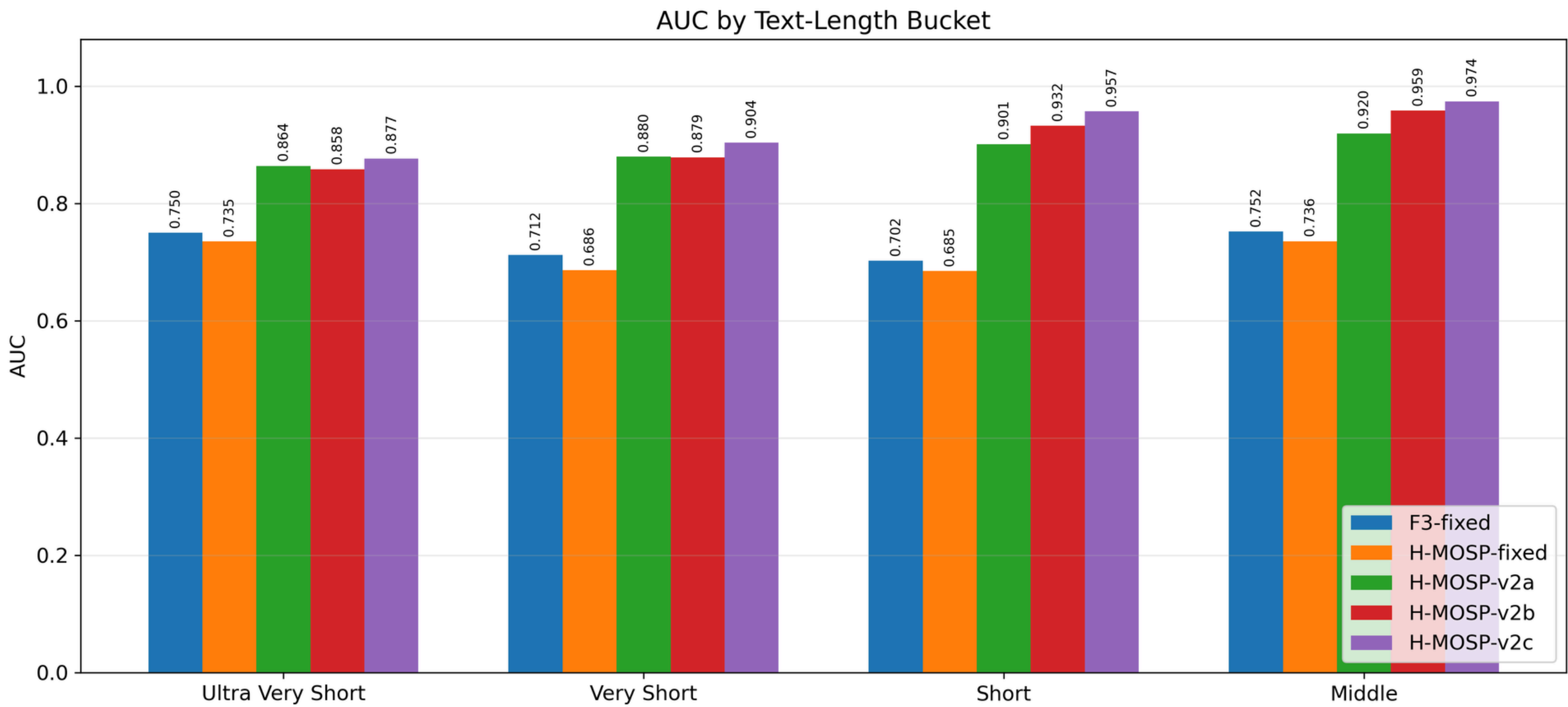
F2 → v2a

평균 AUC
0.8825 → 0.8911 ↑

평균 F1-score
0.8270 → 0.8194 ↓

문장 길이별 성능

v2c는 단문 성능을 개선했지만 길이별 격차와 v2a → v2b의 짧은 구간 예외가 남음



v2a → v2b 예외

Ultra Very Short
AUC 0.8639 → 0.8583 ↓
F1 0.7997 → 0.7957 ↓

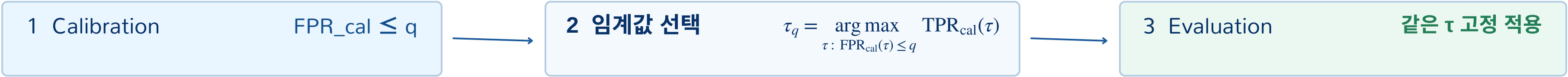
Very Short
AUC 0.8799 → 0.8788 ↓
F1 0.8074 → 0.8051 ↓

평균 개선 ≠ 모든 bucket 개선

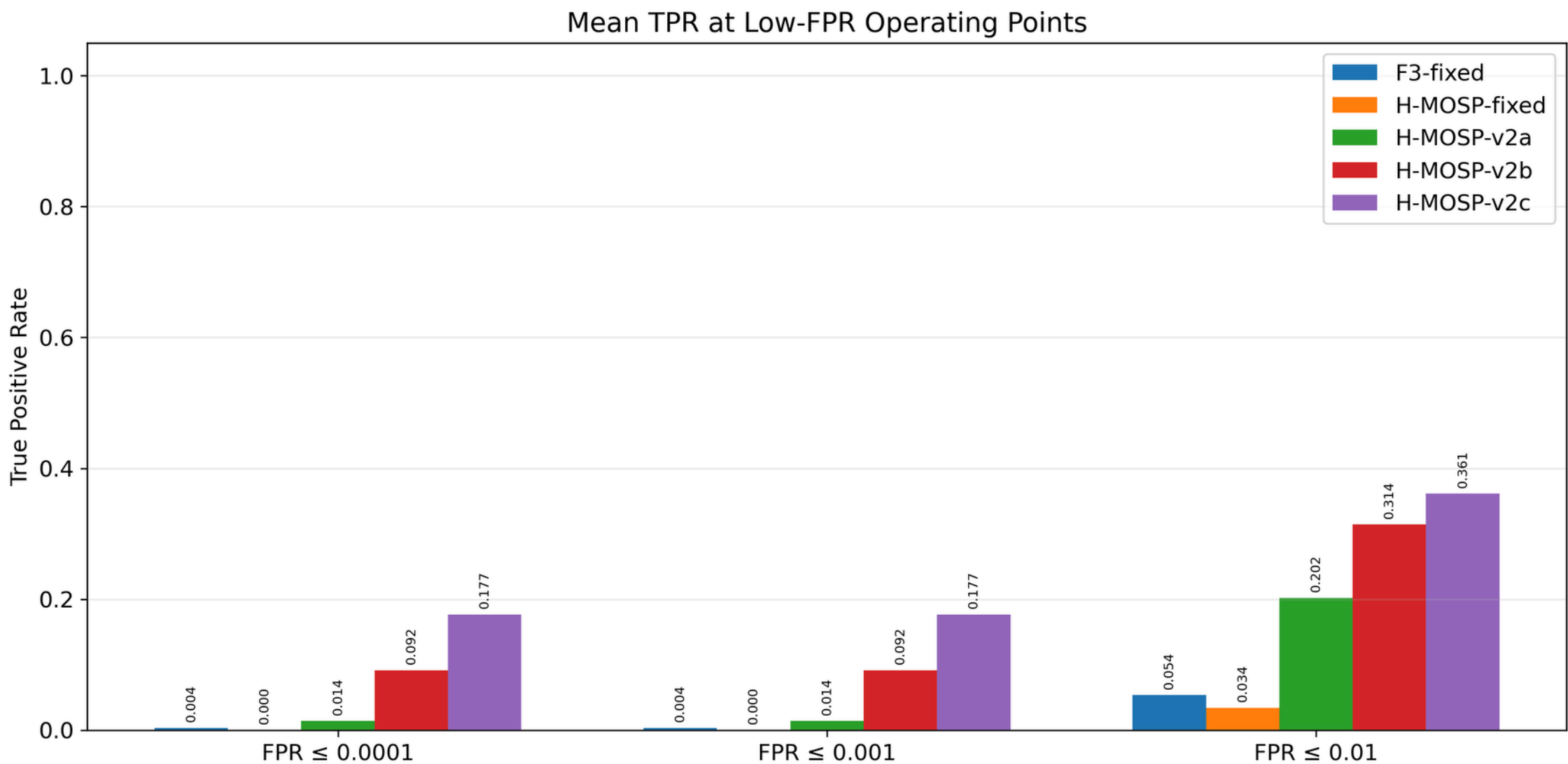
v2c F1-score | Ultra 0.8115 · Very Short 0.8328 · Short 0.8927 · Middle 0.9183

Low-FPR 성능

Calibration에서 목표 FPR 이하 임계값 선택 → Evaluation에 고정 적용



기호 설명 q: calibration 목표 FPR · τ : 선택된 임계값 · TPR: True Positive Rate · FPR: False Positive Rate



v2c · calibration q=1 %

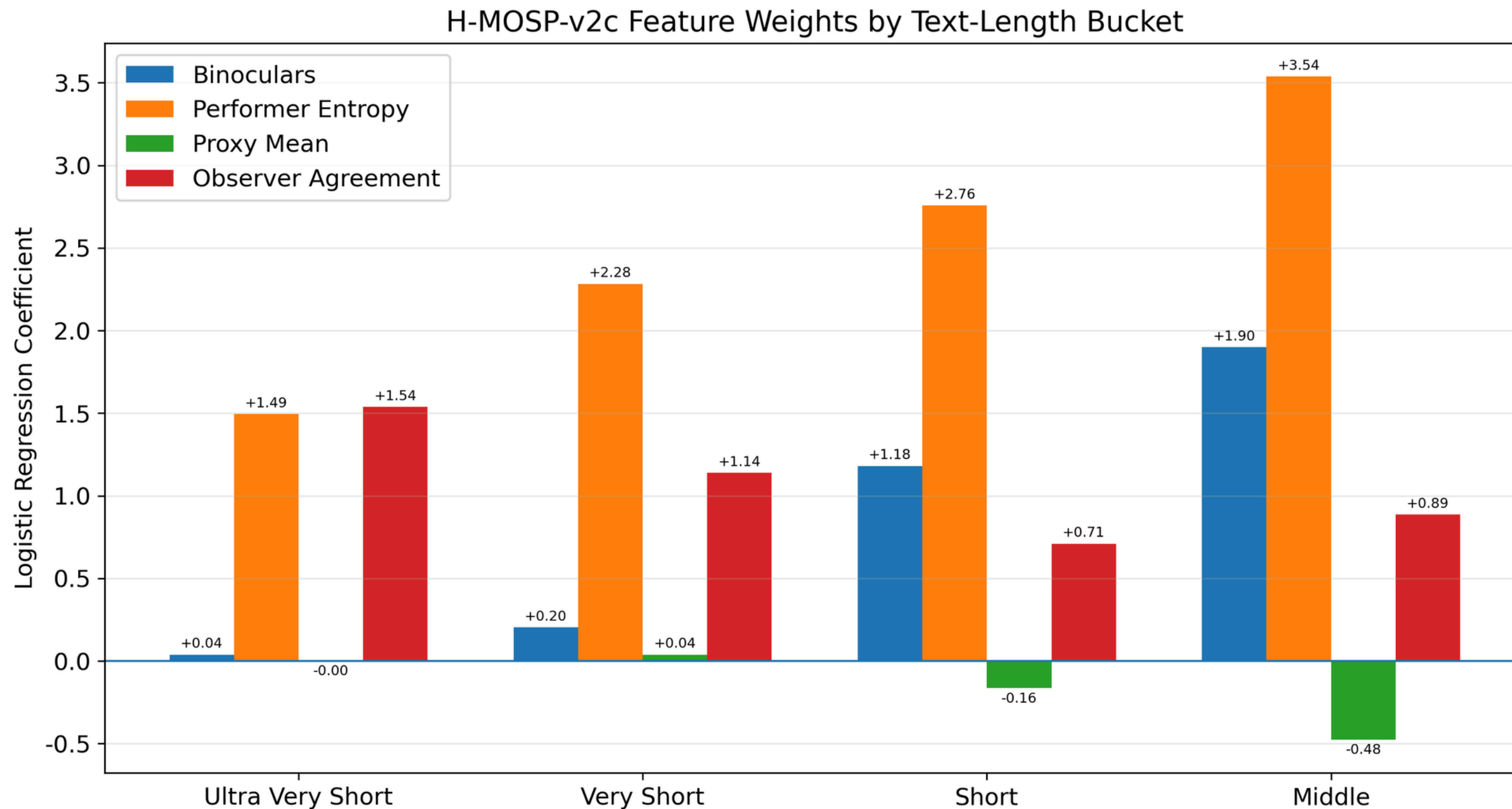
Bucket	Eval TPR	실제 FPR
Ultra Very Short	12.43%	0.85%
Very Short	18.96%	0.76%
Short	45.29%	0.51%
Middle	67.84%	1.83%

evaluation 평균 TPR 36.13%

Middle 실제 FPR은 목표 1% 초과

v2c 특징 계수

길이에 따라 학습된 계수 패턴이 달라짐



읽는 포인트

Binoculars 계수

0.04 → 0.20 → 1.18 → 1.90

계수는 길이 bucket별로
다른 패턴을 보임

주의

표준화된 조건부 LR 계수이며
인과적 feature importance가 아님

보조 수식

$$u_j(x) = \frac{e_j(x) - \text{mean}_{\text{cal}}(e_j)}{\text{std}_{\text{cal}}(e_j)}$$

$$P(\text{AI} | x) = \sigma\left(b + \sum_j w_j u_j(x)\right)$$

Label 사용과 zero-shot 한계

점수·결합 학습과 임계값 선택 단계의 label 사용을 구분함

방법	점수·결합	임계값	핵심 해석
Baseline	고정 점수식	labeled calibration	임계값 선택에 label 사용
v2a / v2b	grid 가중치 선택	labeled calibration	calibration label 사용
v2c	Logistic Regression supervised 학습	같은 calibration	supervised fusion

AUC는 임계값을 사용하지 않는 순위 지표
공통 evaluation을 개발 과정에서 참조

결론 및 향후 연구

확인된 개발 결과를 바탕으로 데이터 규모를 확대해 결과 안정성을 다시 검증함

이번 실험에서 확인

- 짧은 bucket에서 Baseline AUC 저하
- 원본 7방법 중 F2 평균 AUC/F1-score 최고
- F3/H-MOSP 방향·결합 수정 후 성능 회복
- v2c 최고 평균 성능, 단 supervised fusion

향후 연구 핵심

각 길이 bucket
5,000 → 20,000 샘플 확대

확대 후 확인

- 주요 방법 성능 재검증
- 단문 구간 성능 변동 확인
- 동일 수식·평가 프로토콜 유지

핵심 다음 단계 | 각 길이 bucket의 데이터 규모를 4배로 늘려 결과 안정성을 재검증

Q&A