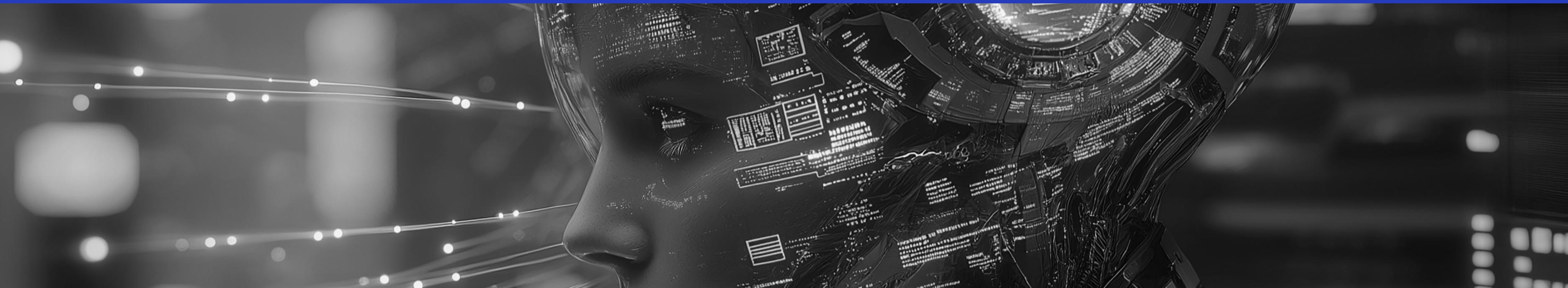


LLM의 내부 레이어 표현을 활용한 환각 탐지: 머신러닝 기반 실증 연구

D조 | 박제우, 박윤서, 허지원, 서정원



문제 배경 및 연구 질문

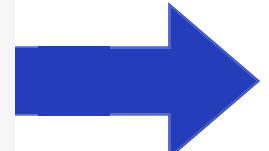
LLM은 QA·요약·추론에서 뛰어난 성능
그러나 **hallucination**은 여전히 치명적 문제

기존 접근

- output을 다시 LLM으로 평가
- heuristic / rule-based filtering

한계

- 추론 비용 증가
- 모델 종속성 큼
- LLM 내부 표현이 환각 정보를 담고 있는지에 대한 실증 부족



RQ.

**LLM의 레이어 표현을 FEATURE로 사용할 때,
머신러닝 분류기를 통한 환각 평가는
어느 정도까지 가능한가?**

선행연구 분석

내부 표현 기반 환각 탐지:

LLM Internal States Reveal Hallucination Risk Faced With a Query (arXiv 2024)

- 출력 이전, hidden states만으로 환각 위험 예측 가능
- 간단한 probing classifier로 ~84% 정확도
- 환각은 출력의 문제가 아니라 질의 처리 단계의 내부 표현에서 이미 드러남

- LLM을 representation encoder로 보는 관점 공유
- 본 연구는 '위험 예측'이 아니라 실제 hallucination 여부 판별로 문제 확장

레이어별 정보 활용:

Hallucination Detection with Internal Layers of LLMs (arXiv 2025)

- 여러 레이어 hidden states를 가중합하여 환각 탐지
- HalluEval, TruthfulQA 등에서 성능 향상

- 본 연구는 레이어별 표현 자체의 판별력을 먼저 검증
- 어느 레이어가 유효한지에 대한 기초 실증 역할

레이어 간 정보 흐름:

Detecting Hallucination via Layer-wise Information Flow (EMNLP 2025)

- 환각은 모델이 사용할 수 있는 정보가 내부적으로 충분히 전달되지 못해서 발생
- 환각은 '틀린 정보 생성'이 아니라 '정보 전달 실패 상태에서의 추론 결과'

- 본 연구는 first, middle, last 레이어의 표현을 비교
- 환각 관련 정보가 어느 단계의 표현에서 가장 잘 드러나는지 실증 시도

사용 데이터셋

HaluEval

- EMNLP 2023 공개 벤치마크
- 약 35,000건
- QA, Summarization 등 다양한 생성 시나리오
- Context + Question + Answer
→ hallucination 여부 이진 분류

선택 이유

- hallucination이 명시적으로 라벨링
- 내부 representation 분석에 적합

{'ID', 'user_query', 'chatgpt_response', 'hallucination', 'hallucination_spans'}

{'knowledge': 'Jasminum multipartitum (Starry Wild Jasmine or Imfohlafohlane) is a species of jasmine, in the family Oleaceae, that is native to Mozambique, Zimbabwe, Swaziland and South Africa.Zimbabwe (), officially the Republic of Zimbabwe, is a landlocked country located in southern Africa, between the Zambezi and Limpopo Rivers.',

'question': 'Jasminum multipartitum is native to what landlocked country located in southern Africa, between the Zambezi and Limpopo Rivers?',

'answer': 'Jasminum multipartitum is native to a territory that spans across multiple southern African countries, including Zimbabwe.',

'hallucination': 'yes',

'task': 'qa_samples'},

{'knowledge': 'Audrey Kathryn Lindvall (August 11, 1982 - August 2, 2006) was an American model. She was the sister of supermodel Angela Lindvall, and the former face of Coach and Ann Taylor.Angela Lindvall (born January 14, 1979) is an American supermodel and actress. Lindvall was discovered by an IMG scout when she was 14 years old, and immediately signed with IMG New York.',

'question': "At which age Audrey Kathryn Lindvall's super model sister was discovered by IMG scout?",

'answer': '14 years old',

'hallucination': 'no',

'task': 'qa_samples'},

방법론 개요

Pipeline
✓ STEP 1. Context + Q + A → 사전학습 LLM 입력
✓ STEP 2. 특정 layer hidden representation 추출
✓ STEP 3. representation을 feature로 사용
✓ STEP 4. ML 분류기로 hallucination 이진 분류

LLM Backbone
1. Mistral 7B <ul style="list-style-type: none">상대적으로 짧고 밀도 높은 표현attention 구조가 효율적으로 설계됨불필요한 redundancy가 적음
2. Qwen2 7B <ul style="list-style-type: none">다양한 데이터로 학습 (다국어·지식·코드)표현 공간이 넓고 정보가 많이 담김
3. LLaMA <ul style="list-style-type: none">baseline으로 흔히 사용레이어별 정보 분화가 비교적 명확
4. Falcon <ul style="list-style-type: none">비교적 generation fluency에 초점후반 레이어의 출력 신호가 강함

ML Classifier
1. Logistic Regression (L2)
2. Logistic Regression (L1)
3. SVM Linear
4. Random Forest
5. XGBoost

Condition
First/Middle/Last layer
feature hidden size = 4096
PCA: None/Fixed/Variance

실험 결과: Mistral

ML 모델	PCA 여부	PCA Param	Accuracy	AUROC	Layer
Logistic Regression (L2)	none		0.688179	0.764516	first
Logistic Regression (L1)	none		0.690382	0.767145	first
SVM Linear	none		0.6876	0.762869	first
Random Forest	none		0.586693	0.63734	first
XGBoost	none		0.638856	0.70049	first
Logistic Regression (L2)	none		0.665604	0.727926	last
Logistic Regression (L1)	none		0.667748	0.729558	last
SVM Linear	none		0.66459	0.726427	last
Random Forest	none		0.585678	0.630982	last
XGBoost	none		0.624511	0.681947	last
Logistic Regression (L2)	none		0.728548	0.80511	middle
Logistic Regression (L1)	none		0.729852	0.807732	middle
SVM Linear	none		0.727678	0.803982	middle
Random Forest	none		0.632132	0.698331	middle
XGBoost	none		0.687194	0.768606	middle

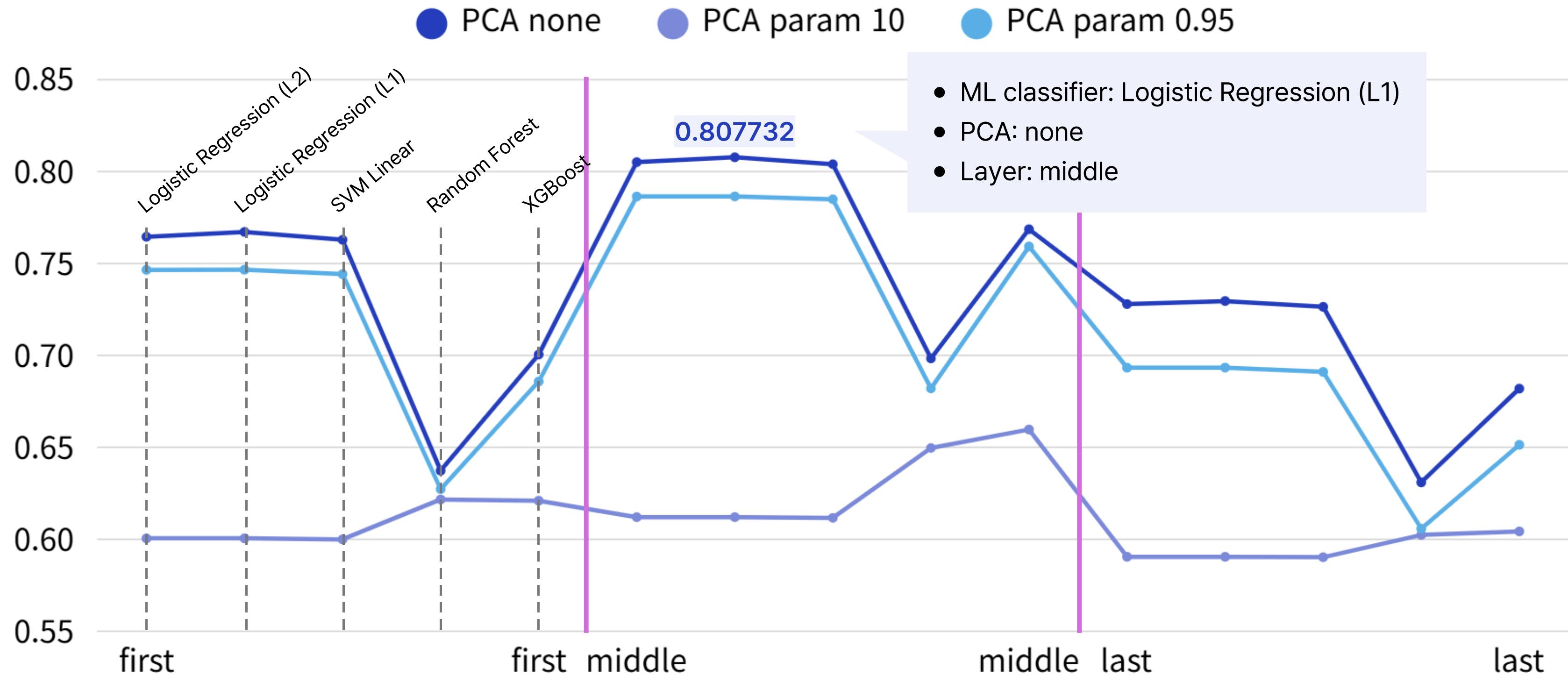
실험 결과: Mistral

ML 모델	PCA 여부	PCA Param	Accuracy	AUROC	Layer
Logistic Regression (L2)	fixed	10	0.561451	0.600536	first
Logistic Regression (L1)	fixed	10	0.561422	0.600544	first
SVM Linear	fixed	10	0.560379	0.59999	first
Random Forest	fixed	10	0.575159	0.62168	first
XGBoost	fixed	10	0.573681	0.620996	first
Logistic Regression (L2)	fixed	10	0.554844	0.590514	last
Logistic Regression (L1)	fixed	10	0.555047	0.590519	last
SVM Linear	fixed	10	0.554438	0.590293	last
Random Forest	fixed	10	0.562495	0.602373	last
XGBoost	fixed	10	0.566523	0.604312	last
Logistic Regression (L2)	fixed	10	0.56774	0.612074	middle
Logistic Regression (L1)	fixed	10	0.568001	0.612081	middle
SVM Linear	fixed	10	0.56745	0.6117	middle
Random Forest	fixed	10	0.595792	0.649638	middle
XGBoost	fixed	10	0.602805	0.659715	middle

실험 결과: Mistral

ML 모델	PCA 여부	PCA Param	Accuracy	AUROC	Layer
Logistic Regression (L2)	variance	0.95	0.674559	0.746506	first
Logistic Regression (L1)	variance	0.95	0.674646	0.746582	first
SVM Linear	variance	0.95	0.673603	0.744172	first
Random Forest	variance	0.95	0.548816	0.62733	first
XGBoost	variance	0.95	0.627148	0.685822	first
Logistic Regression (L2)	variance	0.95	0.635552	0.69333	last
Logistic Regression (L1)	variance	0.95	0.635349	0.693375	last
SVM Linear	variance	0.95	0.634625	0.691035	last
Random Forest	variance	0.95	0.555366	0.605781	last
XGBoost	variance	0.95	0.605211	0.651397	last
Logistic Regression (L2)	variance	0.95	0.706842	0.786399	middle
Logistic Regression (L1)	variance	0.95	0.707045	0.786456	middle
SVM Linear	variance	0.95	0.707016	0.784903	middle
Random Forest	variance	0.95	0.616107	0.682003	middle
XGBoost	variance	0.95	0.68192	0.75926	middle

실험 결과: Mistral AUROC



실험 결과: Qwen2

ML 모델	PCA 여부	PCA Param	Accuracy	AUROC	Layer
Logistic Regression (L2)	none		0.692961	0.770171	first
Logistic Regression (L1)	none		0.694729	0.772657	first
SVM Linear	none		0.673892	0.746761	first
Random Forest	none		0.603733	0.656741	first
XGBoost	none		0.6481	0.717347	first
Logistic Regression (L2)	none		0.666908	0.735134	last
Logistic Regression (L1)	none		0.668357	0.737409	last
SVM Linear	none		0.647984	0.706098	last
Random Forest	none		0.589938	0.634555	last
XGBoost	none		0.633871	0.693627	last
Logistic Regression (L2)	none		0.72507	0.803387	middle
Logistic Regression (L1)	none		0.727852	0.806264	middle
SVM Linear	none		0.716666	0.790966	middle
Random Forest	none		0.640942	0.713865	middle
XGBoost	none		0.704958	0.79067	middle

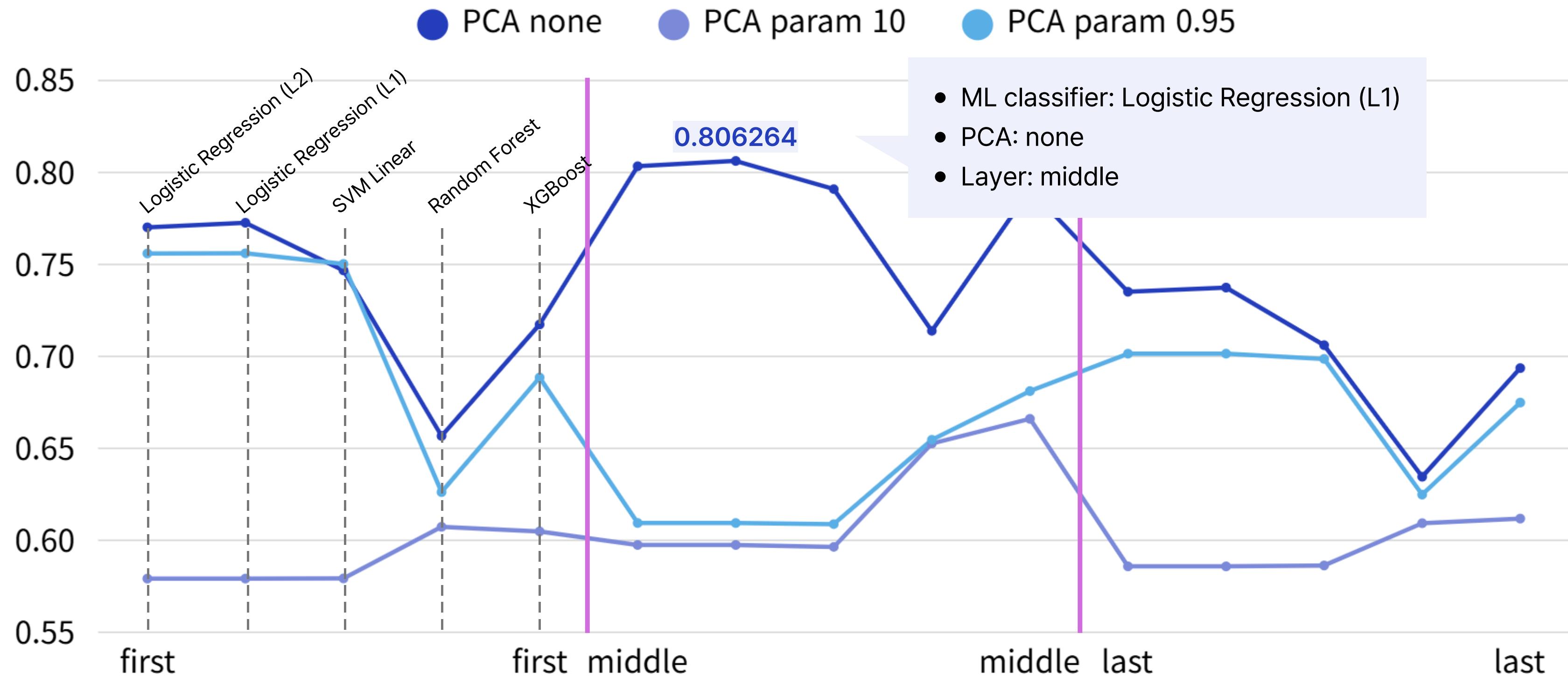
실험 결과: Qwen2

ML 모델	PCA 여부	PCA Param	Accuracy	AUROC	Layer
Logistic Regression (L2)	fixed	10	0.545976	0.579162	first
Logistic Regression (L1)	fixed	10	0.545947	0.579183	first
SVM Linear	fixed	10	0.545947	0.579274	first
Random Forest	fixed	10	0.567566	0.607337	first
XGBoost	fixed	10	0.566059	0.604819	first
Logistic Regression (L2)	fixed	10	0.552902	0.585821	last
Logistic Regression (L1)	fixed	10	0.553192	0.585813	last
SVM Linear	fixed	10	0.553279	0.586302	last
Random Forest	fixed	10	0.568291	0.609298	last
XGBoost	fixed	10	0.574	0.611794	last
Logistic Regression (L2)	fixed	10	0.558959	0.597471	middle
Logistic Regression (L1)	fixed	10	0.558901	0.597471	middle
SVM Linear	fixed	10	0.556988	0.596352	middle
Random Forest	fixed	10	0.599444	0.652633	middle
XGBoost	fixed	10	0.608746	0.666036	middle

실험 결과: Qwen2

ML 모델	PCA 여부	PCA Param	Accuracy	AUROC	Layer
Logistic Regression (L2)	variance	0.95	0.681688	0.755961	first
Logistic Regression (L1)	variance	0.95	0.681717	0.756047	first
SVM Linear	variance	0.95	0.67737	0.750257	first
Random Forest	variance	0.95	0.555192	0.626204	first
XGBoost	variance	0.95	0.631408	0.688522	first
Logistic Regression (L2)	variance	0.95	0.640971	0.701462	last
Logistic Regression (L1)	variance	0.95	0.640826	0.701491	last
SVM Linear	variance	0.95	0.639609	0.698638	last
Random Forest	variance	0.95	0.576318	0.624841	last
XGBoost	variance	0.95	0.620831	0.674881	last
Logistic Regression (L2)	variance	0.95	0.56945	0.609401	middle
Logistic Regression (L1)	variance	0.95	0.569566	0.609393	middle
SVM Linear	variance	0.95	0.568667	0.608796	middle
Random Forest	variance	0.95	0.599733	0.654651	middle
XGBoost	variance	0.95	0.619034	0.681134	middle

실험 결과: Qwen2 AUROC



실험 결과: LLaMA

ML 모델	PCA 여부	PCA Param	Accuracy	AUROC	Layer
Logistic Regression (L2)	none		0.68565794	0.761611676	first
Logistic Regression (L1)	none		0.685744859	0.762140561	first
SVM Linear	none		0.684469763	0.759917755	first
Random Forest	none		0.585678205	0.636386407	first
XGBoost	none		0.634624761	0.696310844	first
Logistic Regression (L2)	none		0.672414249	0.740944126	last
Logistic Regression (L1)	none		0.673109806	0.742486851	last
SVM Linear	none		0.671197052	0.73939273	last
Random Forest	none		0.59376366	0.638798046	last
XGBoost	none		0.630799644	0.694226649	last
Logistic Regression (L2)	none		0.742747974	0.828165646	middle
Logistic Regression (L1)	none		0.745153242	0.830116235	middle
SVM Linear	none		0.742226366	0.82643014	middle
Random Forest	none		0.649462422	0.720678749	middle
XGBoost	none		0.702176234	0.786682031	middle

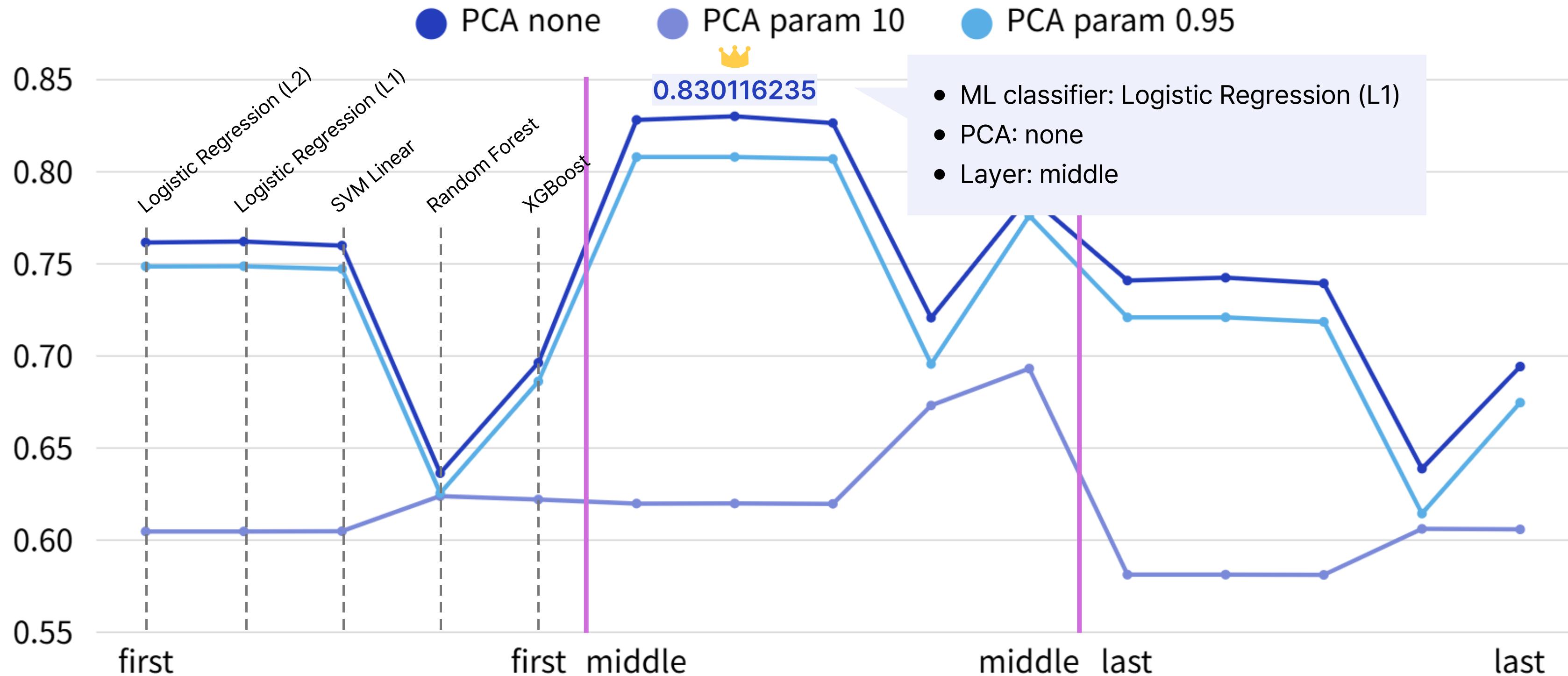
실험 결과: LLaMA

ML 모델	PCA 여부	PCA Param	Accuracy	AUROC	Layer
Logistic Regression (L2)	fixed	10	0.565943167	0.604720899	first
Logistic Regression (L1)	fixed	10	0.565972148	0.60471179	first
SVM Linear	fixed	10	0.566146026	0.604835856	first
Random Forest	fixed	10	0.57808553	0.623889113	first
XGBoost	fixed	10	0.575738338	0.622037783	first
Logistic Regression (L2)	fixed	10	0.546468785	0.581238491	last
Logistic Regression (L1)	fixed	10	0.546410831	0.581236149	last
SVM Linear	fixed	10	0.546381838	0.581098796	last
Random Forest	fixed	10	0.565102751	0.606092904	last
XGBoost	fixed	10	0.567826726	0.605822701	last
Logistic Regression (L2)	fixed	10	0.573072009	0.619853723	middle
Logistic Regression (L1)	fixed	10	0.573187931	0.619878021	middle
SVM Linear	fixed	10	0.573187931	0.619727497	middle
Random Forest	fixed	10	0.61326661	0.673079211	middle
XGBoost	fixed	10	0.624365942	0.693129887	middle

실험 결과: LLaMA

ML 모델	PCA 여부	PCA Param	Accuracy	AUROC	Layer
Logistic Regression (L2)	variance	0.95	0.67447183	0.748630349	first
Logistic Regression (L1)	variance	0.95	0.674413874	0.748742885	first
SVM Linear	variance	0.95	0.674413869	0.747186701	first
Random Forest	variance	0.95	0.548381536	0.625442184	first
XGBoost	variance	0.95	0.628800002	0.686178949	first
Logistic Regression (L2)	variance	0.95	0.656562455	0.720917932	last
Logistic Regression (L1)	variance	0.95	0.656475519	0.720980013	last
SVM Linear	variance	0.95	0.653925318	0.718506968	last
Random Forest	variance	0.95	0.559770454	0.614412535	last
XGBoost	variance	0.95	0.62372848	0.67464231	last
Logistic Regression (L2)	variance	0.95	0.720433531	0.807994585	middle
Logistic Regression (L1)	variance	0.95	0.720636384	0.808026757	middle
SVM Linear	variance	0.95	0.720317573	0.806912202	middle
Random Forest	variance	0.95	0.625728128	0.695565517	middle
XGBoost	variance	0.95	0.697394722	0.776138136	middle

실험 결과: LLaMA AUROC



실험 결과: Falcon

ML 모델	PCA 여부	PCA Param	Accuracy	AUROC	Layer
Logistic Regression (L2)	none		0.69333731	0.769796673	first
Logistic Regression (L1)	none		0.698727617	0.777101449	first
SVM Linear	none		0.691627518	0.768112423	first
Random Forest	none		0.570521864	0.612682721	first
XGBoost	none		0.611557158	0.662690784	first
Logistic Regression (L2)	none		0.650737409	0.710126858	last
Logistic Regression (L1)	none		0.652505187	0.712196102	last
SVM Linear	none		0.651374964	0.709807211	last
Random Forest	none		0.57385454	0.613637098	last
XGBoost	none		0.603095088	0.653129266	last
Logistic Regression (L2)	none		0.710725235	0.782041422	middle
Logistic Regression (L1)	none		0.712956696	0.785281888	middle
SVM Linear	none		0.710261538	0.780970508	middle
Random Forest	none		0.61564323	0.673942562	middle
XGBoost	none		0.666502315	0.740769586	middle

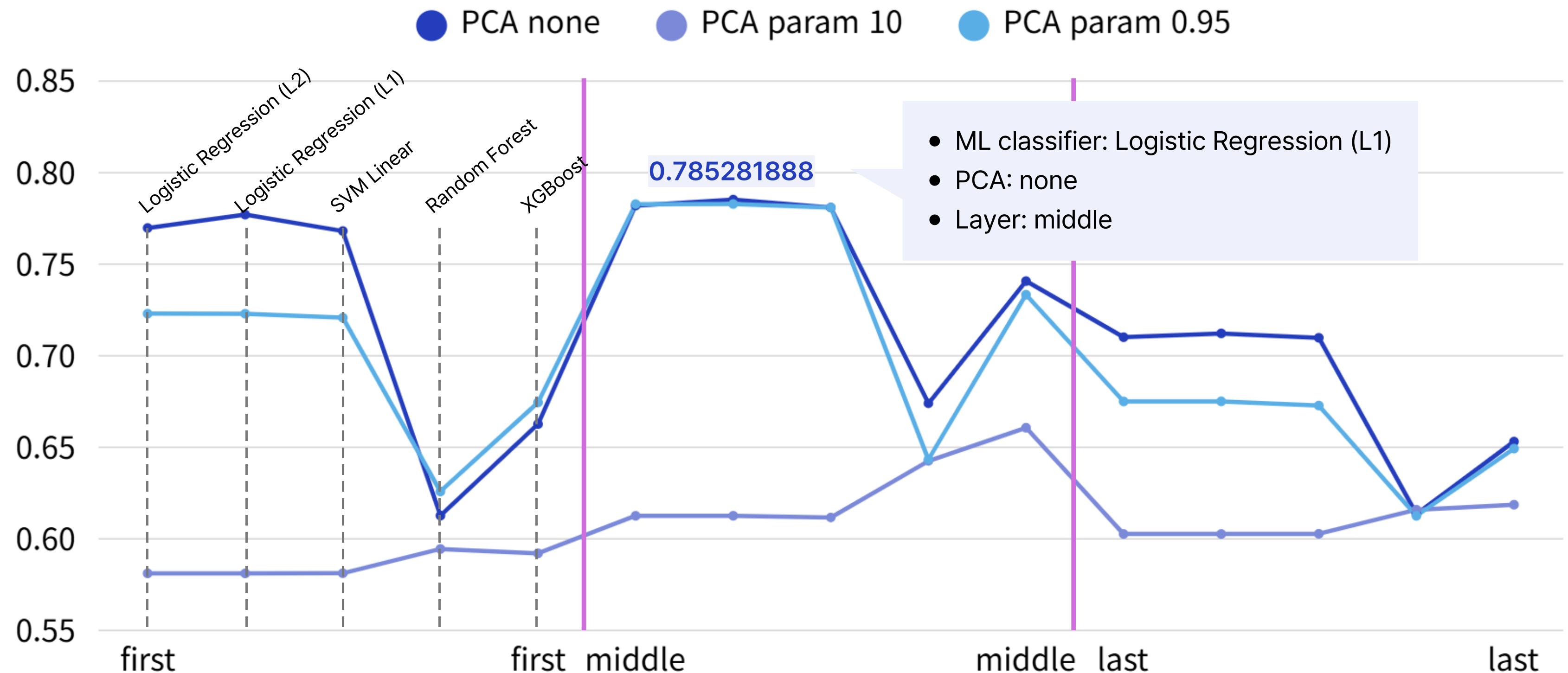
실험 결과: Falcon

ML 모델	PCA 여부	PCA Param	Accuracy	AUROC	Layer
Logistic Regression (L2)	fixed	10	0.547715056	0.581134015	first
Logistic Regression (L1)	fixed	10	0.547686073	0.581112181	first
SVM Linear	fixed	10	0.548265652	0.581206344	first
Random Forest	fixed	10	0.558205597	0.594472378	first
XGBoost	fixed	10	0.556205937	0.592063051	first
Logistic Regression (L2)	fixed	10	0.567826938	0.602637615	last
Logistic Regression (L1)	fixed	10	0.567855921	0.602642191	last
SVM Linear	fixed	10	0.567450189	0.6027463	last
Random Forest	fixed	10	0.574173325	0.615913438	last
XGBoost	fixed	10	0.576491747	0.618623857	last
Logistic Regression (L2)	fixed	10	0.572869391	0.612577232	middle
Logistic Regression (L1)	fixed	10	0.572898371	0.612597004	middle
SVM Linear	fixed	10	0.571275528	0.611622776	middle
Random Forest	fixed	10	0.587764822	0.642508568	middle
XGBoost	fixed	10	0.602921203	0.660683945	middle

실험 결과: Falcon

ML 모델	PCA 여부	PCA Param	Accuracy	AUROC	Layer
Logistic Regression (L2)	variance	0.95	0.65705505	0.722995981	first
Logistic Regression (L1)	variance	0.95	0.656678314	0.722942454	first
SVM Linear	variance	0.95	0.655084414	0.720752847	first
Random Forest	variance	0.95	0.548816142	0.625876072	first
XGBoost	variance	0.95	0.620424914	0.674519585	first
Logistic Regression (L2)	variance	0.95	0.620656642	0.675023522	last
Logistic Regression (L1)	variance	0.95	0.620859497	0.675099712	last
SVM Linear	variance	0.95	0.620134973	0.672811822	last
Random Forest	variance	0.95	0.556003099	0.612567058	last
XGBoost	variance	0.95	0.602138609	0.649291756	last
Logistic Regression (L2)	variance	0.95	0.703943962	0.782823755	middle
Logistic Regression (L1)	variance	0.95	0.704349676	0.782900698	middle
SVM Linear	variance	0.95	0.703364365	0.780984785	middle
Random Forest	variance	0.95	0.585214588	0.643303023	middle
XGBoost	variance	0.95	0.664473934	0.733353246	middle

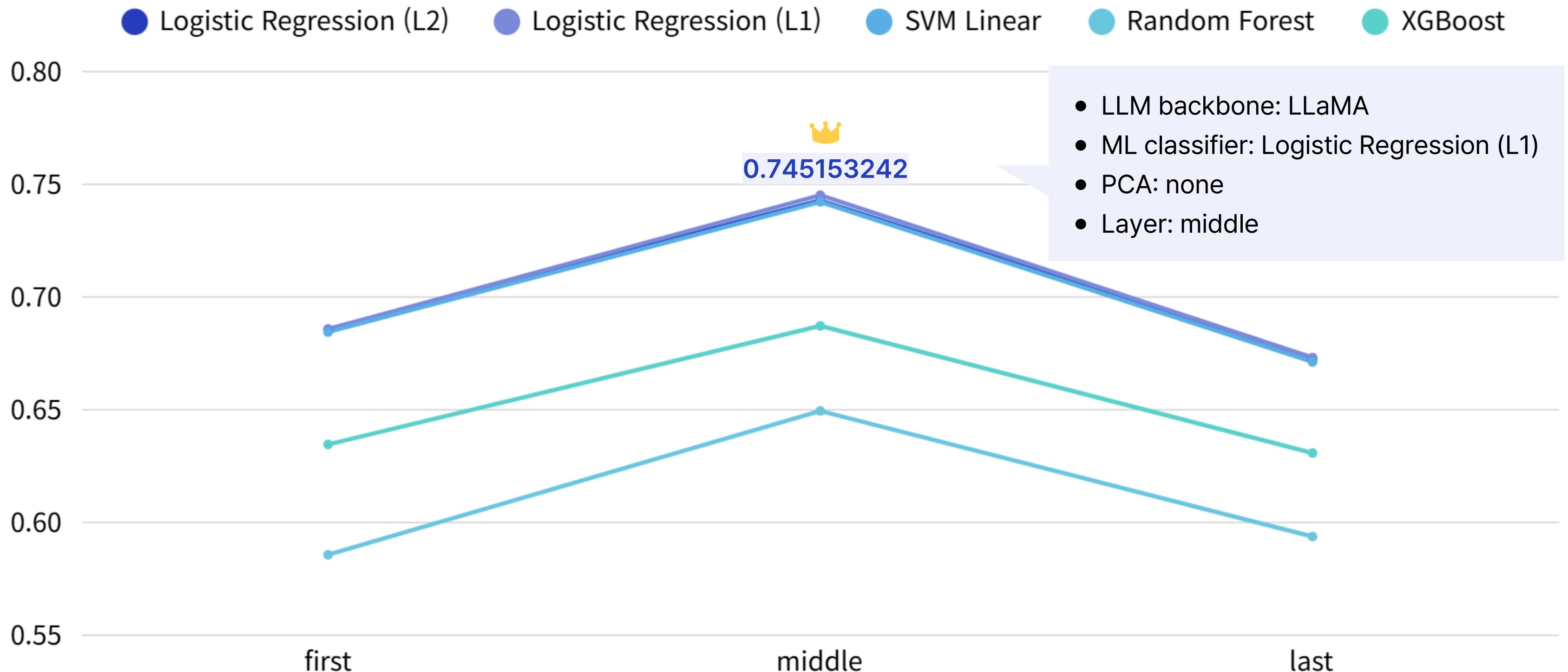
실험 결과: Falcon AUROC



실험 결과: LLaMA 모델별 AUROC (PCA none)



실험 결과: LLaMA 모델별 Accuracy (PCA none)



실험 결과 해석 및 인사이트

Layer별 결과

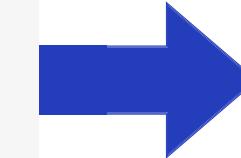
- Middle layer > First / Last
- 대부분의 LLM·모델에서 일관적
- AUROC: middle ≈ 0.80 , first/last $\approx 0.72\text{--}0.76$

모델별 경향

- 선형 모델(Logistic, SVM) 안정적
- 트리 기반 모델은 고차원 공간에서 불리

PCA 영향

- PCA(10D) 적용 시 성능 급락
- hallucination 신호는 저차원 선형 구조가 아님



- LLM 중간 레이어는 hallucination 판별 정보를 이미 포함
- 추가 LLM inference 없이도 경량 ML classifier로 유의미한 탐지 가능

왜 ML인가

- 비용·속도·해석 가능성
- 딥러닝을 다시 쓰지 않아도 representation은 충분히 강력

한계 및 향후 방향

한계

환각 정의의 데이터셋 의존성

- 본 연구는 HaluEval 벤치마크에 기반
- HaluEval의 환각 정의는 context 기반 factual consistency에 초점
- 다른 유형의 환각 (추론 오류, 수치 오류, 창작형 환각 등)에 그대로 일반화되기 어려움

정적 평가에 그침

- 레이어 표현을 사후적으로 추출하여 분류
- 실시간 제어 또는 개입까지는 다루지 않음

성능 최적화보다는 가능성 검증에 초점을 둔 설계

- 하이퍼파라미터 튜닝, 양상블 등 성능 향상을 위한 추가적인 최적화에 집중하지 않음
- 보고된 성능은 상한선이 아니라 보수적인 추정치

향후 방향

1. 성능 최적화를 위한 추가 실험

- 적극적인 하이퍼파라미터 탐색, 양상블 등
- 비선형 모델 및 간단한 신경망 분류기 도입

2. 학습 기반 차원 축소 기법 도입

- PCA 대신 Autoencoder, Supervised bottleneck
- 환각 판별에 유효한 정보만 압축

3. 레이어 결합 및 정보 흐름 반영

- 단일 레이어 → 다중 레이어 결합
- 레이어별 가중치 학습 등으로 확장

4. 실제 추론 단계로의 확장

- LLM 추론 중간에 auxiliary hallucination evaluator로 사용



감사합니다.

D조 | 박제우, 박윤서, 허지원, 서정원