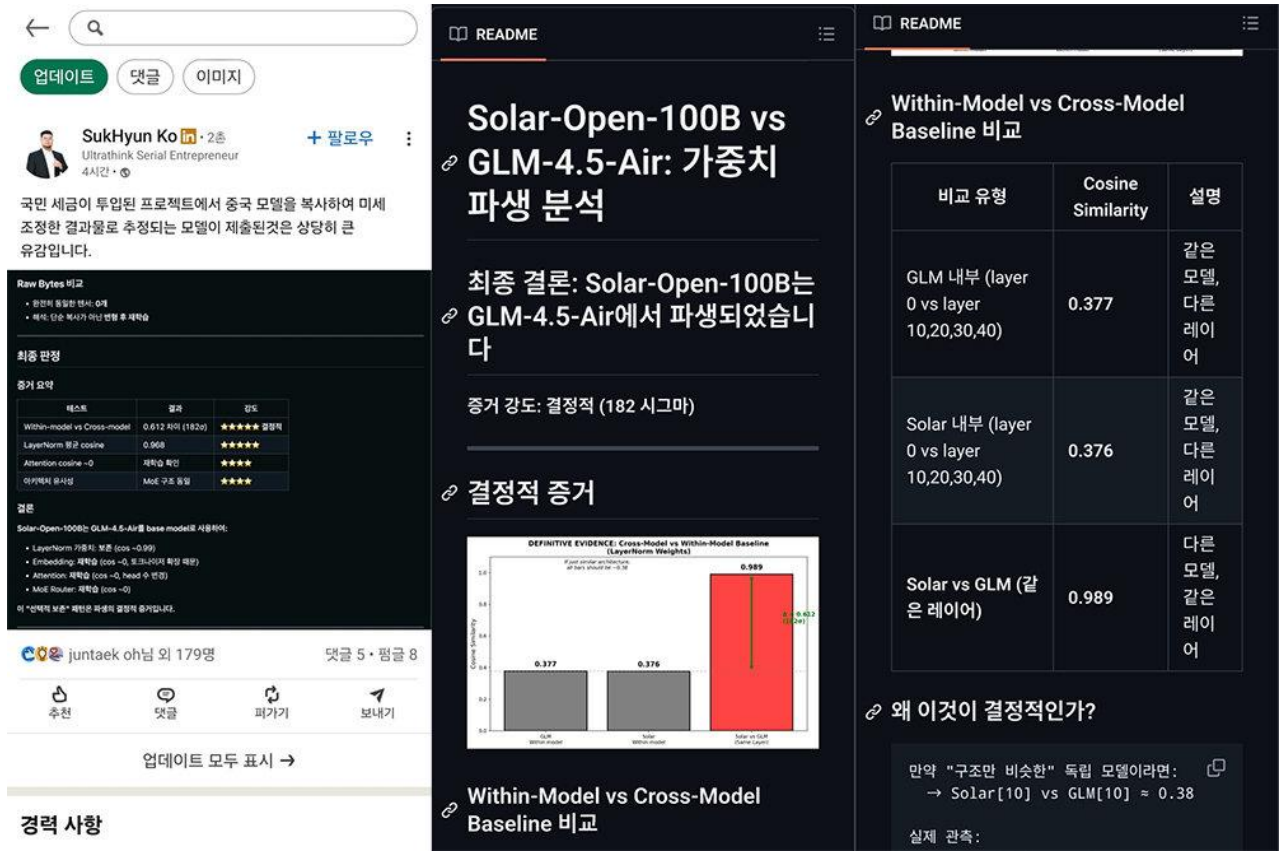
Fingerprint of LLMs

발표자: 김민혁

Korea University

Natural Language Processing
& Artificial Intelligence

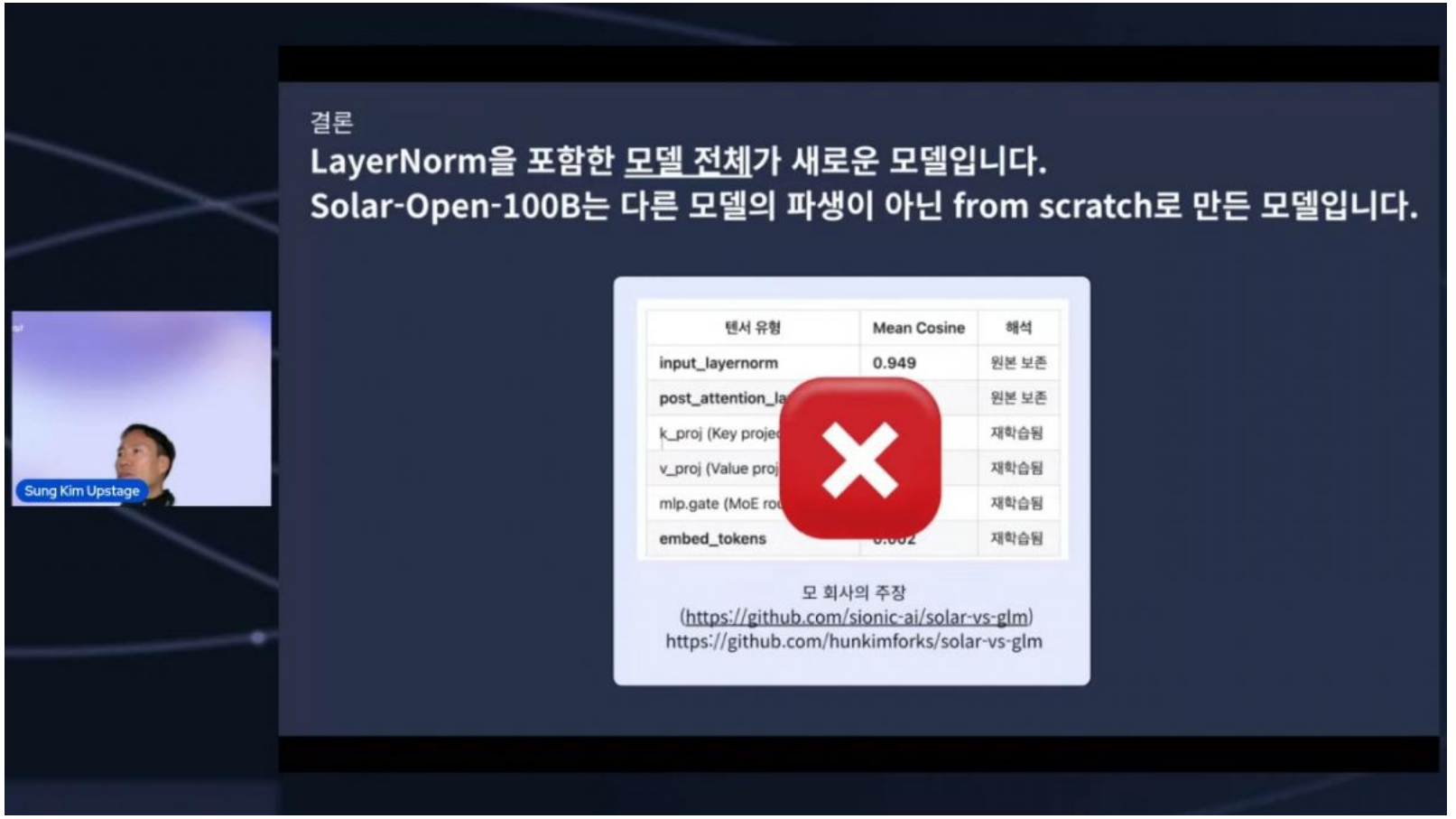
Overview



Solar-Open-100B(Upstage) - GLM 복제 이슈

- Layernorm 간의 cosine similarity가 매우 높음
- 라이선스 및 기타 코드에 GLM의 흔적이 남아있음

Overview



Solar는 From scratch가 맞습니다

- Layernorm간의 코사인 유사도는 모델 복제의 근거가 될 수 없음
- Solar ↔ Phi, Phi ↔ GLM, Solar ↔ GLM 모두 높은 코사인 유사도

Layernorm은 가진 정보량이 적다

- Attention, MLP에 비해 매우 적은 가중치만을 지니고 있음

Layernorm의 weight은 1.0으로 초기화된다

- Xavier, He 등의 초기화 대신 1.0으로 초기화하여 상대적으로 이미 방향성이 정해진 상태에서 학습됨
  - Xavier, He: 이전 층의 노드 개수 n을 표준편차에 반영하여 정규분포로 초기화하는 방법

의혹 제기 실험의 논리적 결함

- 0번째 레이어 vs 타 레이어의 layernorm간의 비교만 수행
  - 주장: 같은 모델의 레이어간 layernorm의 코사인 유사도는 낮게 측정되는데, 다른 모델의 layernorm의 코사인 유사도가 높다면 복제된 모델이다

```
from transformers import AutoModel

model = AutoModel.from_pretrained("Qwen/Qwen3-8B")
```

✓ 6.7s

```
Loading checkpoint shards: 100%|██████████| 5/5 [00:05<00:00,  1.19s/it]
```

```
model.layers[0]
```

✓ 0.0s

```
Qwen3DecoderLayer(
  (self_attn): Qwen3Attention(
    (q_proj): Linear(in_features=4096, out_features=4096, bias=False)
    (k_proj): Linear(in_features=4096, out_features=1024, bias=False)
    (v_proj): Linear(in_features=4096, out_features=1024, bias=False)
    (o_proj): Linear(in_features=4096, out_features=4096, bias=False)
    (q_norm): Qwen3RMSNorm((128,), eps=1e-06)
    (k_norm): Qwen3RMSNorm((128,), eps=1e-06)
  )
  (mlp): Qwen3MLP(
    (gate_proj): Linear(in_features=4096, out_features=12288, bias=False)
    (up_proj): Linear(in_features=4096, out_features=12288, bias=False)
    (down_proj): Linear(in_features=12288, out_features=4096, bias=False)
    (act_fn): SiLUActivation()
  )
  (input_layernorm): Qwen3RMSNorm((4096,), eps=1e-06)
  (post_attention_layernorm): Qwen3RMSNorm((4096,), eps=1e-06)
)
```

생각해봅시다

모델의 복제 여부를 파악할 수 있는 다른 합리적인 방법은 없을까?

Intrinsic Fingerprint of LLMs: Continue Training is NOT All You Need to Steal A Model!

Do-hyeon Yoon, Minsoo Chun, Thomas Allen, Hans Müller, Min Wang, and Rajesh Sharma
Honest AGI Community
honestagi@outlook.com

Black-Box Identification

- 모델 가중치 사용 X, API-only 모델에 적합
- Behavioral Fingerprinting
  - 모델의 생성문에서 통계적 패턴을 분석하거나 스타일을 분석
  - Decoding randomness에 취약함
- Watermarking
  - 학습 또는 토큰 단위 perturbation을 통해 모델이 detectable한 signal을 삽입하도록 함
  - Output editing으로 인해 무효화 가능

White-Box Identification

- Weight, activation 등 모델의 내부 수치를 활용
- Representation-based Fingerprinting
  - Hidden representation을 분석 (Gradient statistics)
  - 데이터 의존적이며, 훈련 데이터와의 potential correlation 존재 가능
- Weight-based Fingerprinting
  - 모델 weights에 대한 data-free analysis

Methodology

1. 모델의 모든 attention matrix에 대해 각각의 표준 편차를 구함
2. 이를 attention matrix의 구성요소에 따라 분류하여 sequence를 만듦
3. 각 sequence를 정규화한 sequence S를 구함
4. 모델 간의 유사성은 각 sequence 간의 상관 계수를 계산하여 평가
   - Pearson 상관계수를 사용

- Layer 개수가 다르다면?
  - 선형 보간 사용

$$\mathbf{S}^Q = [\sigma_1^Q, \sigma_2^Q, \ldots, \sigma_L^Q] \quad (2)$$

$$\mathbf{S}^K = [\sigma_1^K, \sigma_2^K, \ldots, \sigma_L^K] \quad (3)$$

$$\mathbf{S}^V = [\sigma_1^V, \sigma_2^V, \ldots, \sigma_L^V] \quad (4)$$

$$\mathbf{S}^O = [\sigma_1^O, \sigma_2^O, \ldots, \sigma_L^O] \quad (5)$$

$$\hat{\mathbf{S}}^M_{\text{interp}} = \text{interp1d}(\mathbf{i}_{\text{short}}, \hat{\mathbf{S}}^M_{\text{short}}, \mathbf{i}_{\text{target}})$$

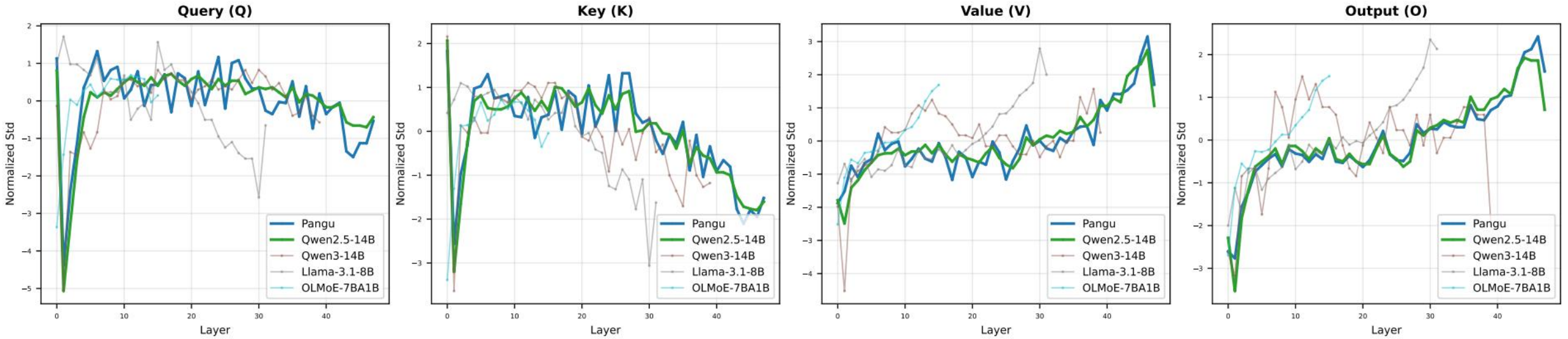
$$\rho^M = \text{corr}(\hat{\mathbf{S}}^M_{\text{interp}}, \hat{\mathbf{S}}^M_{\text{long}})$$

Intrinsic Fingerprint of LLMs: Continue Training is NOT All You Need to Steal A Model

Experiments

- Model
  - Qwen 계열 모델
  - Llama 계열 모델
  - Moe 모델 (OLMoE-7BA1B, Qwen1.5-MoE-A2.7B, **Pangu Pro MoE**)
- 분석 종류
  - Cross-Family Model
  - Quantitative Correlation
  - Validation Through Known Model Lineages
  - Feed-Forward Network Analysis

Intrinsic Fingerprint of LLMs: Continue Training is NOT All You Need to Steal A Model



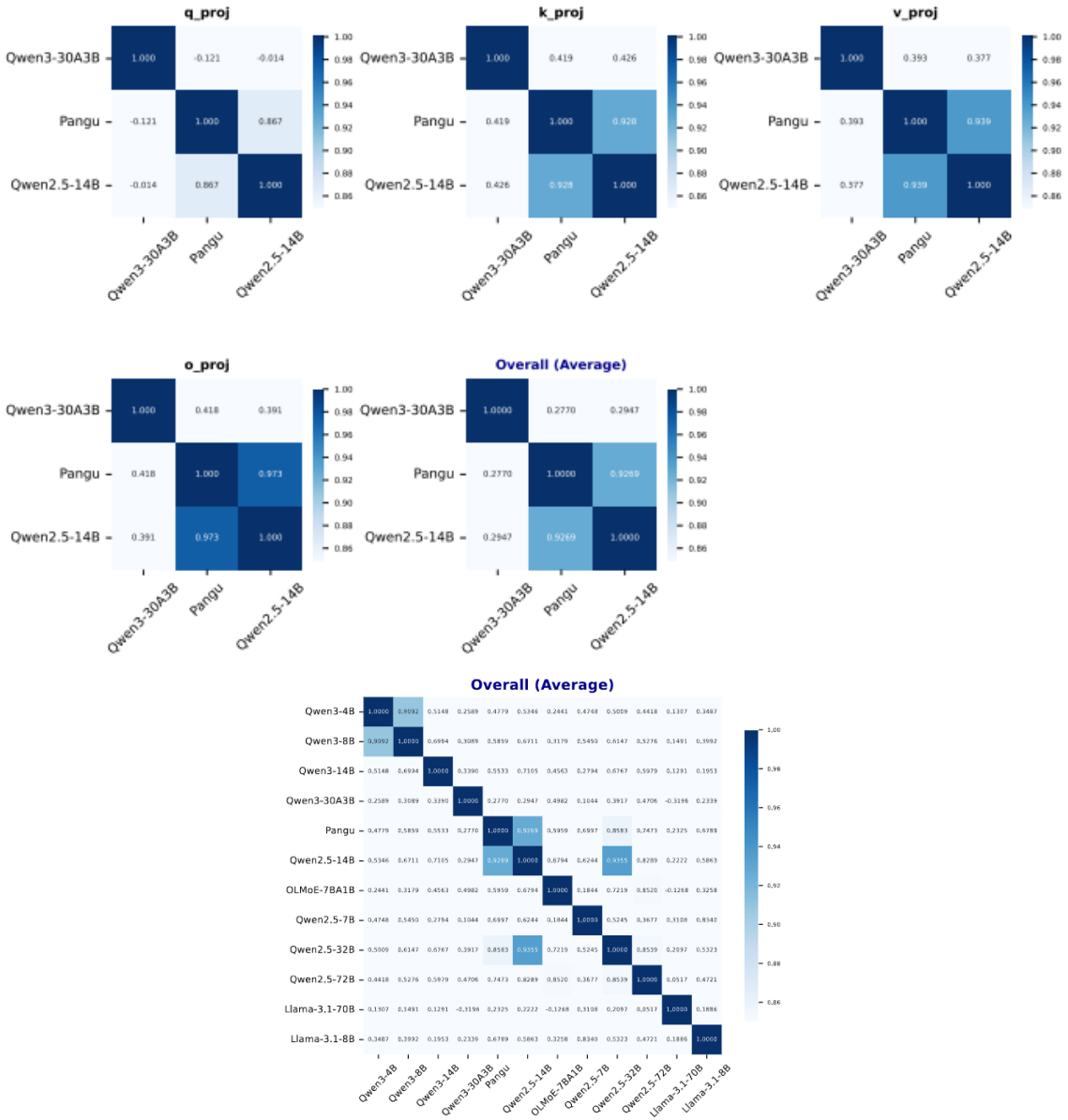
Cross-Family Model Analysis

- Model Family 중 대표 모델간의 비교
- 각각의 모델은 고유한 패턴이 존재함
- Pangu와 Qwen2.5-14B는 네 가지 attention matrix 유형에서 모두 거의 동일한 패턴을 보이고 있음

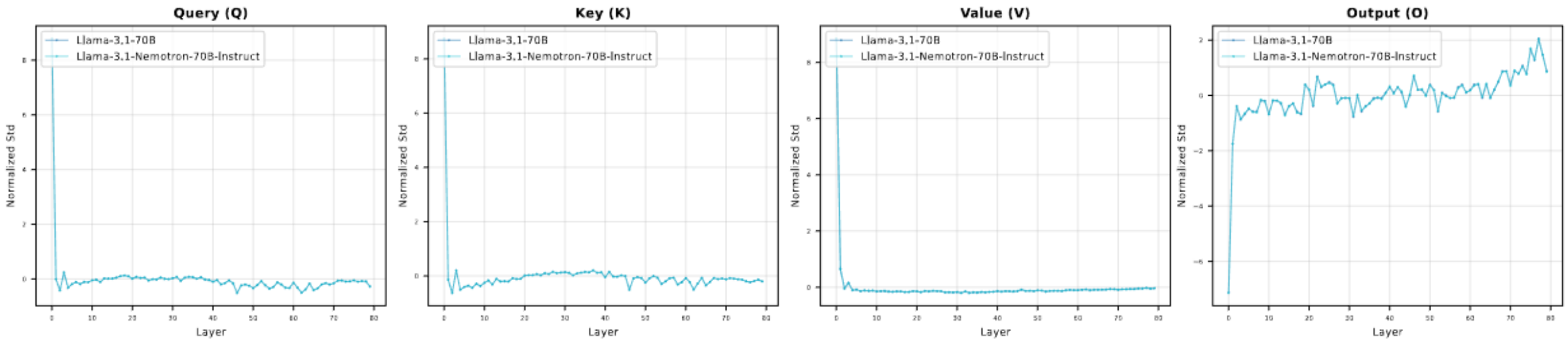
Intrinsic Fingerprint of LLMs: Continue Training is NOT All You Need to Steal A Model

Quantitative Correlation Analysis

- Pangu와 Qwen2.5-14B
  - Q: 0.867
  - K: 0.928
  - V: 0.939
  - O: 0.973
- 이 정도의 수치는 같은 Family 계열 모델에서도 찾아보기 힘든 수치



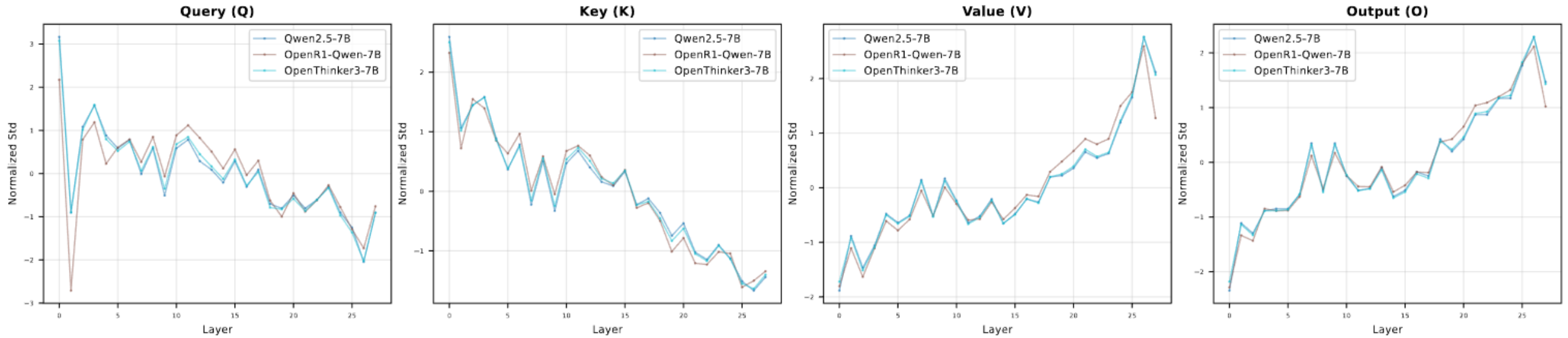
Intrinsic Fingerprint of LLMs: Continue Training is NOT All You Need to Steal A Model



Validation Through Known Model Lineages

- 방법론이 유효한가에 대해 model derivation 관계가 공식화된 모델 간의 비교를 진행
- Llama-3.1-70B vs. Llama-3.1-Nemotron-70B-Instruct
- Instruction tuning 및 safety alignment를 거치더라도, attention parameter의 distribution은 거의 동일하게 유지

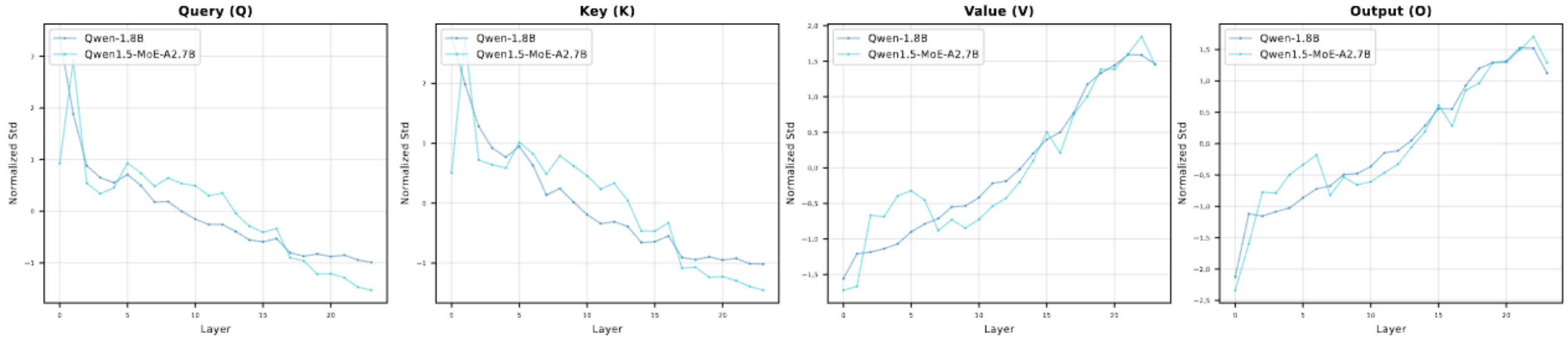
Intrinsic Fingerprint of LLMs: Continue Training is NOT All You Need to Steal A Model



Validation Through Known Model Lineages

- Qwen2.5-7B vs. OpenR1-Qwen-7B vs. OpenThinker3-7B
- Instruction-tuning의 목표가 다르더라도 상당한 일관성을 보임

Intrinsic Fingerprint of LLMs: Continue Training is NOT All You Need to Steal A Model



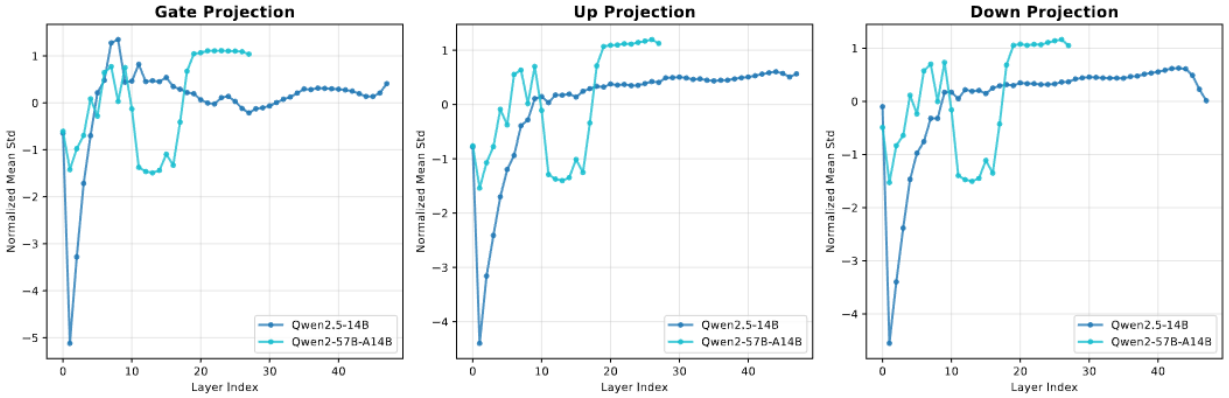
Validation Through Known Model Lineages

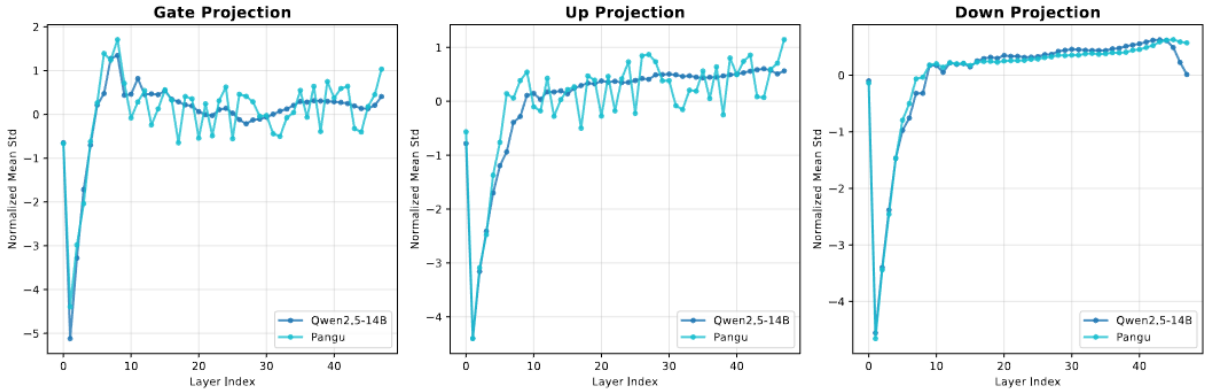
- Qwen-1.8B vs Qwen1.5-MoE-A2.7B
- Qwen1.5-MoE-A2.7B는 Qwen-1.8를 upcycle해서 만들었는데, 상당한 아키텍처 수정에도 불구하고 기존 모델과 매우 유사한 흐름을 보임
- Pangu(MoE)가 Qwen2.5-14B(Dense)를 upcycle해서 만들었다는 주장을 뒷받침

Intrinsic Fingerprint of LLMs: Continue Training is NOT All You Need to Steal A Model

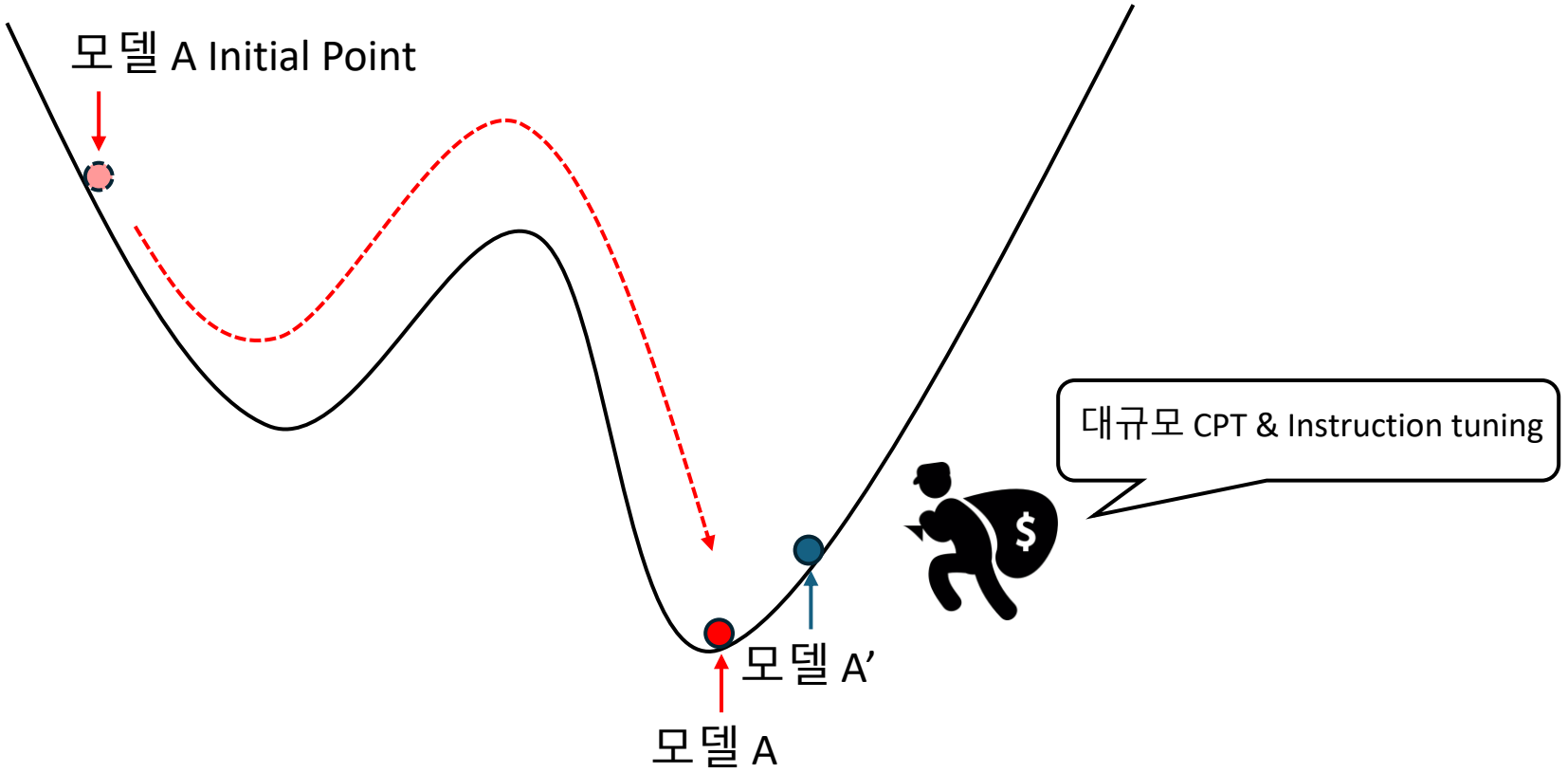
Feed-Forward Network Analysis

- Qwen2.5-14B vs Qwen2-57B-A14B
- MoE 모델과 Dense 모델은 Feed-forward 층에서 그 아키텍처가 상이하기 때문에 동일한 패턴을 지니기가 힘듦
- 그러나 Pangu와 Qwen2.5-14B는 아키텍처의 차이에도 불구하고 강한 유사성을 띰
- Pangu(MoE)가 Qwen2.5-14B(Dense)를 upcycle해서 만들었다는 주장을 뒷받침





Intrinsic Fingerprint of LLMs: Continue Training is NOT All You Need to Steal A Model



Ghost in the Transformer: Detecting Model Reuse with Invariant Spectral Signatures

Suqing Wang[1*], **Ziyang Ma**[2*], **Li Xinyi**[2], **Zuchao Li**[1†]

[1]School of Artificial Intelligence, Wuhan University
[2]School of Computer Science, Wuhan University
{wangsuqing, maziyang, xyli-lucia, zcli-charlie}@whu.edu.cn

Ghost in the Transformer: Detecting Model Reuse with Invariant Spectral Signatures

Preliminaries

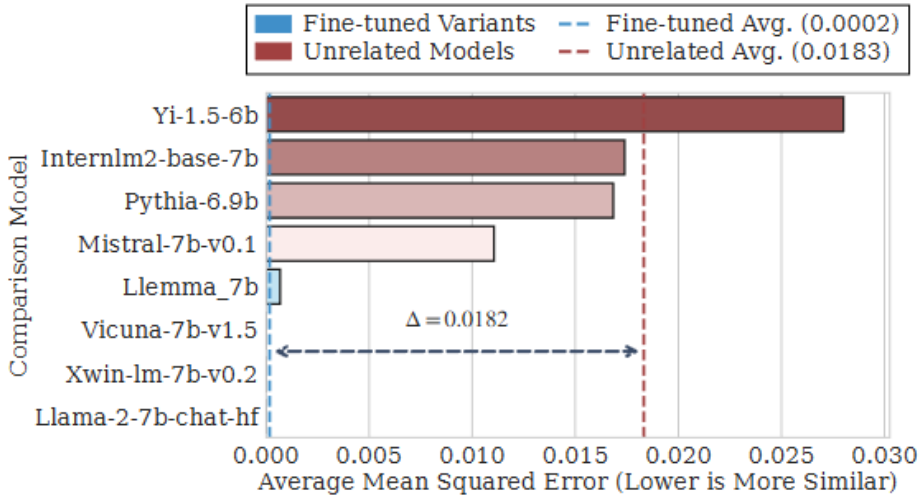
- Random Matrix Theory
  - 원소가 확률적으로 분표하는 행렬의 성질을 이해하는 연구
  - Marchenko–Pastur 법칙: bulk(Marchenko–Pastur에 해당하는 연속분포)와 그 바깥의 spikes로 구분 가능
    - Spikes는 보통 모델이 학습한 유의미한 방향을 나타냄. 이를 제거하면 perplexity가 상당히 증가함
    - SVD에서 상위 특이값이 spike로 간주됨
  - PT 시에 큰 특이값은 안정적으로 유지, 모델의 전반적인 동작을 고정
  - 사후 학습 시 큰 특이값보다는 작은 특이값과 관련된 방향에 영향을 미침

→ 사후 학습을 해도 original 모델과 상위 특이값이 크게 변하지 않는다!

Ghost in the Transformer: Detecting Model Reuse with Invariant Spectral Signatures

Experiments Setup

- Fine-tuned Variants
  - Llama-2-7b & variants
  - Llama-2-7b & 관련 없는 모델 (Mistral-7B-v0.1, Internlm2-base-7b ..)
- Quantifying Spectral Similarity
  - 비교대상 모델 A, B의 각 레이어의 유형별 attention matrix에 대해서 특이값 벡터를 추출함
  - 얻어진 벡터쌍을 각 쌍의 최소 유효랭크만큼의 길이로 통일해서 자르고, 정규화
    - 유효랭크 = exp(-∑(행렬 각 원소의 값 * log (행렬 각 원소의 값)))
  - 최종 벡터쌍의 MSE를 둘 간의 스펙트럼 거리로 정의



Ghost in the Transformer: Detecting Model Reuse with Invariant Spectral Signatures

Experiments Setup

- 비교군
  - Fine-Tuning
  - Model Pruning
  - Model Merging
  - Model Upcycling: Dense LLM을 MoE로 변환하는 경우를 포함
  - Permutation and Scaling Transformations
  - Unrelated Models
- Baselines
  - QueRE: 특정 query에 대한 응답 패턴을 통해 모델 특성을 식별
  - Logits: 모델의 출력 분포를 통해 모델 특성을 식별
  - REEF: activation / embedding간의 유사도를 비교
  - PCS: weight의 통계적 특성을 분석해 모델의 유사성을 식별

Methodology

- 개별 attention matrix의 특이값을 직접 비교하는 것은 여러 기법에 의해 무의미해질 수 있음
  - 변환 공격에 비교적 robust하게 동작하는 matrix product에서 파생된 스펙트럼을 분석

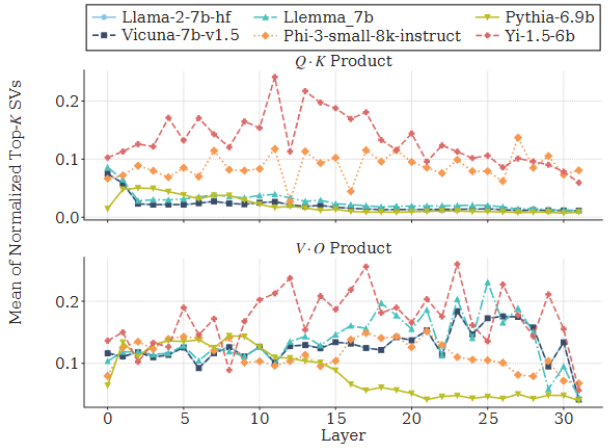
$$M_{qk}^{(i)} = W_q^{(i)}(W_k^{(i)})^T \quad \text{and} \quad M_{vo}^{(i)} = W_v^{(i)}W_o^{(i)}$$

- Similarity Metrics
  - GhostSpec-mse: 레이어별 비교 (1.0에 가까울수록 유사)
    - 구성 요소별 **특이값 벡터 간의 스펙트럼 거리**의 MSE 평균
    - POSA 알고리즘을 통해 레이어 개수가 다른 비교군에서도 방법론 적용 가능

$$\mathrm{Sim}_{\mathrm{MSE}}(A,B) = 1 - \frac{1}{1 + e^{-k(d_{\mathrm{path}} - \tau)}}$$

Avg MSE

  - GhostSpec-corr: 전반적인 추세를 비교
    - 각 레이어별 matrix product (q,k), (v,o)의 **특이값 평균**을 측정하여 sequence 구성 후 이 **sequence간의 거리를 비교**
    - 마찬가지로 레이어 수가 달라 sequence의 길이가 다를 경우 POSA 알고리즘과 유사하게 정렬



Ghost in the Transformer: Detecting Model Reuse with Invariant Spectral Signatures

Primary Model: Llama-2-7b							
Method	Data Dep.	Model Fine-tuning (↑)		Adversarial Transforms (↑)		Unstructured Pruning (↑)	
		Vicuna-7b-v1.5	Llemma_7b	Llama-2-7b -scaled	Llama-2-7b -permuted	Pruned-50% -Retrained	Pruned-70% -Retrained
QueRE	Data-Aware	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
Logits	Data-Aware	0.9767	0.8400	1.0000	1.0000	0.8567	0.8533
REEF	Data-Aware	0.9992	0.9979	1.0000	1.0000	0.9968	0.9948
PCS	Data-Free	0.9986	0.5052	0.5970	0.3863	0.9061	0.7829
GhostSpec-corr	Data-Free	0.9992	0.7595	1.0000	1.0000	0.8967	0.7045
GhostSpec-mse	Data-Free	0.9760	0.9532	0.9761	0.9761	0.9727	0.9653
Method	Data Dep.	Structured Pruning (↑)		Merging & Expansion (↑)		Unrelated Models (↓)	
		Sheared-Llama 1.3B	Sheared-Llama 2.7B	Llama2-7b-func -call-slerp	Camelidae-8x7B	Qwen2.5-7B	OPT-6.7b
QueRE	Data-Aware	1.0000	1.0000	0.0910	1.0000	0.3410	1.0000
Logits	Data-Aware	1.0000	1.0000	1.0000	0.9500	0.9967	0.2200
REEF	Data-Aware	0.9315	0.9487	0.9996	0.9991	0.2513	0.2692
PCS	Data-Free	0.0000	0.0000	0.9993	0.0204	0.0000	0.0000
GhostSpec-corr	Data-Free	0.9398	0.9414	0.9998	0.9999	0.2940	0.3423
GhostSpec-mse	Data-Free	0.8886	0.9045	0.9760	0.9761	0.0000	0.5025
Primary Model: Mistral-7B							
Method	Data Dep.	Fine-tuning (↑)	Merging (↑)	Expansion (↑)	Pruning (↑)	Unrelated Models (↓)	
		OpenHermes-2.5 -Mistral-7B	Triunvirato-7b	Chunky-Lemon -Cookie-11B	OpenHermes-2.5 -Mistral-7B-pruned50	Qwen2.5-7B	Yi-1.5-6B
QueRE	Data-Aware	1.0000	1.0000	1.0000	1.0000	0.3410	0.0819
Logits	Data-Aware	0.9933	0.9967	1.0000	0.9867	0.9567	0.2067
REEF	Data-Aware	0.8949	0.8538	0.8495	0.8596	0.7473	0.8301
PCS	Data-Free	0.9999	0.9997	0.8987	0.9979	0.0000	0.0000
GhostSpec-corr	Data-Free	0.9999	0.9997	0.9981	0.9896	0.2708	0.4304
GhostSpec-mse	Data-Free	0.9760	0.9759	0.9758	0.9753	0.0083	0.0581

Results

- Data-aware, Data-free 방법론을 막론하고 모든 케이스에 대해서 정확하게 모델 복제 여부를 파악해냄
- 기존 방법론들이 판단에 어려움을 겪었던 Structured Pruning, Merging&Expansion, False Positive에서도 강건한 성능을 보임

Ghost in the Transformer: Detecting Model Reuse with Invariant Spectral Signatures

Primary Model: Llama-2-7b							
Method	Data Dep.	Model Fine-tuning (↑)		Adversarial Transforms (↑)		Unstructured Pruning (↑)	
		Vicuna-7b-v1.5	Llemma_7b	Llama-2-7b -scaled	Llama-2-7b -permuted	Pruned-50% -Retrained	Pruned-70% -Retrained
QueRE	Data-Aware	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
Logits	Data-Aware	0.9767	0.8400	1.0000	1.0000	0.8567	0.8533
REEF	Data-Aware	0.9992	0.9979	1.0000	1.0000	0.9968	0.9948
PCS	Data-Free	0.9986	0.5052	0.5970	0.3863	0.9061	0.7829
GhostSpec-corr	Data-Free	0.9992	0.7595	1.0000	1.0000	0.8967	0.7045
GhostSpec-mse	Data-Free	0.9760	0.9532	0.9761	0.9761	0.9727	0.9653
Method	Data Dep.	Structured Pruning (↑)		Merging & Expansion (↑)		Unrelated Models (↓)	
		Sheared-Llama 1.3B	Sheared-Llama 2.7B	Llama2-7b-func -call-slerp	Camelidae-8x7B	Qwen2.5-7B	OPT-6.7b
QueRE	Data-Aware	1.0000	1.0000	0.0910	1.0000	0.3410	1.0000
Logits	Data-Aware	1.0000	1.0000	1.0000	0.9500	0.9967	0.2200
REEF	Data-Aware	0.9315	0.9487	0.9996	0.9991	0.2513	0.2692
PCS	Data-Free	0.0000	0.0000	0.9993	0.0204	0.0000	0.0000
GhostSpec-corr	Data-Free	0.9398	0.9414	0.9998	0.9999	0.2940	0.3423
GhostSpec-mse	Data-Free	0.8886	0.9045	0.9760	0.9761	0.0000	0.5025
Primary Model: Mistral-7B							
Method	Data Dep.	Fine-tuning (↑)	Merging (↑)	Expansion (↑)	Pruning (↑)	Unrelated Models (↓)	
		OpenHermes-2.5 -Mistral-7B	Triunvirato-7b	Chunky-Lemon -Cookie-11B	OpenHermes-2.5 -Mistral-7B-pruned50	Qwen2.5-7B	Yi-1.5-6B
QueRE	Data-Aware	1.0000	1.0000	1.0000	1.0000	0.3410	0.0819
Logits	Data-Aware	0.9933	0.9967	1.0000	0.9867	0.9567	0.2067
REEF	Data-Aware	0.8949	0.8538	0.8495	0.8596	0.7473	0.8301
PCS	Data-Free	0.9999	0.9997	0.8987	0.9979	0.0000	0.0000
GhostSpec-corr	Data-Free	0.9999	0.9997	0.9981	0.9896	0.2708	0.4304
GhostSpec-mse	Data-Free	0.9760	0.9759	0.9758	0.9753	0.0083	0.0581

Results

- Fine-tuning, Adversarial transform(성능을 유지하며 내부 가중치 분포를 바꿈), Unstructured pruning(개별 weight 을 0으로 만듦)을 통한 변형의 경우, 모든 baselines가 쉽게 이를 구분할 수 있음

Ghost in the Transformer: Detecting Model Reuse with Invariant Spectral Signatures

Primary Model: Llama-2-7b							
Method	Data Dep.	Model Fine-tuning (↑)		Adversarial Transforms (↑)		Unstructured Pruning (↑)	
		Vicuna-7b-v1.5	Llemma_7b	Llama-2-7b -scaled	Llama-2-7b -permuted	Pruned-50% -Retrained	Pruned-70% -Retrained
QueRE	Data-Aware	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
Logits	Data-Aware	0.9767	0.8400	1.0000	1.0000	0.8567	0.8533
REEF	Data-Aware	0.9992	0.9979	1.0000	1.0000	0.9968	0.9948
PCS	Data-Free	0.9986	0.5052	0.5970	0.3863	0.9061	0.7829
GhostSpec-corr	Data-Free	0.9992	0.7595	1.0000	1.0000	0.8967	0.7045
GhostSpec-mse	Data-Free	0.9760	0.9532	0.9761	0.9761	0.9727	0.9653
Method	Data Dep.	Structured Pruning (↑)		Merging & Expansion (↑)		Unrelated Models (↓)	
		Sheared-Llama 1.3B	Sheared-Llama 2.7B	Llama2-7b-func -call-slerp	Camelidae-8x7B	Qwen2.5-7B	OPT-6.7b
QueRE	Data-Aware	1.0000	1.0000	0.0910	1.0000	0.3410	1.0000
Logits	Data-Aware	1.0000	1.0000	1.0000	0.9500	0.9967	0.2200
REEF	Data-Aware	0.9315	0.9487	0.9996	0.9991	0.2513	0.2692
PCS	Data-Free	0.0000	0.0000	0.9993	0.0204	0.0000	0.0000
GhostSpec-corr	Data-Free	0.9398	0.9414	0.9998	0.9999	0.2940	0.3423
GhostSpec-mse	Data-Free	0.8886	0.9045	0.9760	0.9761	0.0000	0.5025
Primary Model: Mistral-7B							
Method	Data Dep.	Fine-tuning (↑)	Merging (↑)	Expansion (↑)	Pruning (↑)	Unrelated Models (↓)	
		OpenHermes-2.5 -Mistral-7B	Triunvirato-7b	Chunky-Lemon -Cookie-11B	OpenHermes-2.5 -Mistral-7B-pruned50	Qwen2.5-7B	Yi-1.5-6B
QueRE	Data-Aware	1.0000	1.0000	1.0000	1.0000	0.3410	0.0819
Logits	Data-Aware	0.9933	0.9967	1.0000	0.9867	0.9567	0.2067
REEF	Data-Aware	0.8949	0.8538	0.8495	0.8596	0.7473	0.8301
PCS	Data-Free	0.9999	0.9997	0.8987	0.9979	0.0000	0.0000
GhostSpec-corr	Data-Free	0.9999	0.9997	0.9981	0.9896	0.2708	0.4304
GhostSpec-mse	Data-Free	0.9760	0.9759	0.9758	0.9753	0.0083	0.0581

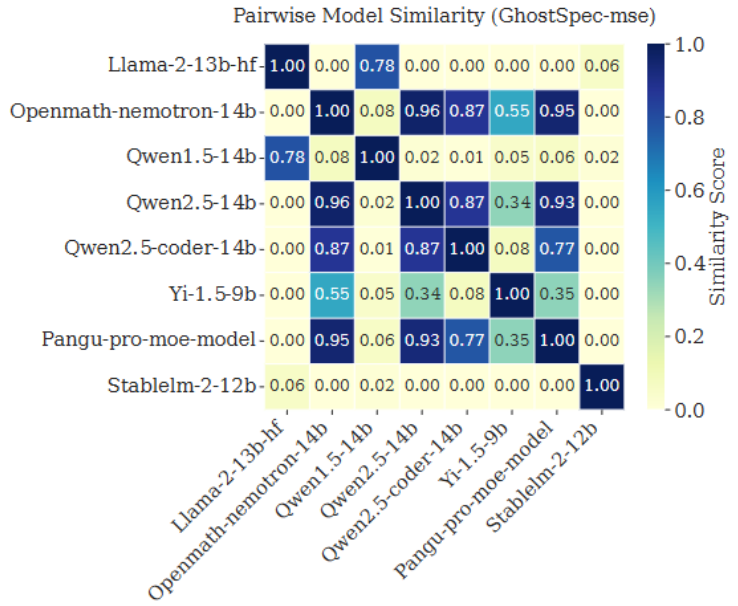
Results

- PCS의 경우 Structured Pruning(블록 단위로 Pruning), Merging 과 같은 기법을 적용했을 때 이를 구분하지 못하는 모습을 보임
- QueRE(응답 분석)의 경우 Unrelated Model를 구분하지 못하고 유사도를 높게 부여함

Ghost in the Transformer: Detecting Model Reuse with Invariant Spectral Signatures

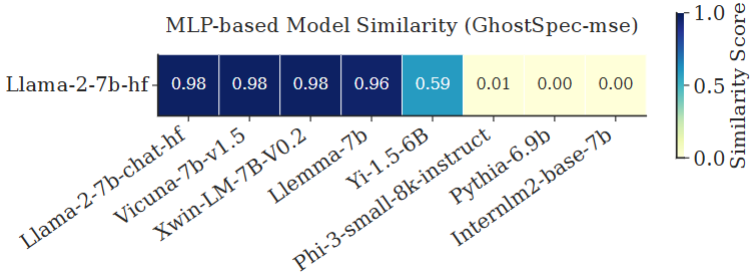
Case Study

- Pangu-pro-moe vs. 다른 모델
  - OpenMath-Nemotron14B (Qwen2.5-14B fine-tuning), Qwen2.5-14B와 가장 높은 유사성을 보임
  - Yi-1.5-9B, Llama-2-13b-hf와는 유사성을 보이지 않음



Analysis of MLP Module Spectra

- Llama-2-7b-hf 기준으로 MLP 레이어에 대해 GhostSpec-mse를 통한 분석을 진행
  - 파생된 모델과 그렇지 않은 모델 간의 유의미한 유사도 차이가 나타남
- 그러나 Dense-to-MoE 확장에는 MLP 레이어를 보는 것이 robust 하지 않을 수 있음
- 따라서 attention-based fingerprints가 더 효율적이고 안정적임



Thank you

Q&A