Quantifying Uncertainty in Transformer-Based Systems

STAAR Summer Workshop 2025

Shin Yoo | COINSE | KAIST

그 왕국의 지도술은 너무나 완벽한 수준에 이르러, 한 지방의 지도는 도시 하나를 덮을 만큼 컸고 왕국 전체의 지도는 지방 하나를 덮을 만했다. 이윽고 이 거대한 지도들도 만족감을 주지 못하게 되자, 지도 제작자 길드는 왕국과 크기가 같고 **1:1**로 대응하는 지도를 만들었다. 조상들만큼 지도 만들기에 관심이 없던 후세 사람들은 거대한 지도가 쓸모없다고 생각했고, 무심하게 지도를 태양과 겨울의 자비에 내맡겼다. 오늘날에도 서쪽 사막에는 너덜너덜해진 지도의 잔해 안에 동물과 걸인들이 살고 있다. 왕국 전체에 남은 지도술에 대한 유물은 이것 뿐이다.

“과학의 정확성에 대하여” - 호르헤 보르헤스, 1946

Paskaerte van't in comen
Van't CANAAL, hoemen
dat sal aen doen, als men uyt de West comt.
t'Amsteldam, by Pieter Goos

In that Empire, the Art of Cartography attained such Perfection that the map of a single Province occupied the entirety of a City, and the map of the Empire, the entirety of a Province. In time, those Unconscionable Maps no longer satisfied, and the Cartographers Guilds struck a Map of the Empire whose size was that of the Empire, and which coincided point for point with it. The following Generations, who were not so fond of the Study of Cartography as their Forebears had been, saw that that vast Map was Useless, and not without some Pitilessness was it, that they delivered it up to the Inclemencies of Sun and Winters. In the Deserts of the West, still today, there are Tattered Ruins of that Map, inhabited by Animals and Beggars; in all the Land there is no other Relic of the Disciplines of Geography.

“On Exactitude of Science” - Jorge Luis Borges, 1946

(OK I have my reasons)

(also, caveat: a very SW testing centric, ML ignorant, view)

Why do we care about uncertainty?

- ML systems are the most representative “non-testable programs” [1].
  - In that we use ML mainly because we do not know the answer ourselves (i.e., there is no easy and known computation that results in the answer).
- Without oracles in hand, testing becomes slightly… weird?:
  - Metamorphic Testing: essentially we go after relaxed oracles
  - Prioritization: we accept the high cost of oracles, but we only look at results that are likely to be incorrect

[1] Weyuker, E. J. On Testing Non-Testable Programs. The Computer Journal 25, 4 (November 1982), 465–470,

Uncertainty in Machine Learning 🌀

Two major sources

- Aleatoric Uncertainty (우연적 불확실성): randomness that cannot be avoided even under perfect knowledge
  - Results of coin flips, physical noise in sensor circuits, etc…
- Epistemic Uncertainty (인식론적 불확실성): uncertainty that **can** be reduced if you **learn harder** (typically by adding more training data, or changing the model)
  - Low accuracy due to too small a model or insufficient training data

Quantifying Uncertainty

- Exact, analytical quantification is only possible with mathematically grounded models that allows you to compute the confidence intervals.
- Training-based Approaches: MCMC, Bayesian Neural Networks, etc
  - Intuitively, try to capture and learn the variance in data during training
- Training-free Approaches: Bayesian Dropout, Gradient-based, Test-time Data Augmentation, etc.
  - Intuitively, try to inject small perturbation during inference and observe the variance (i.e., measure how “brittle” the model decision is “around” the original result)

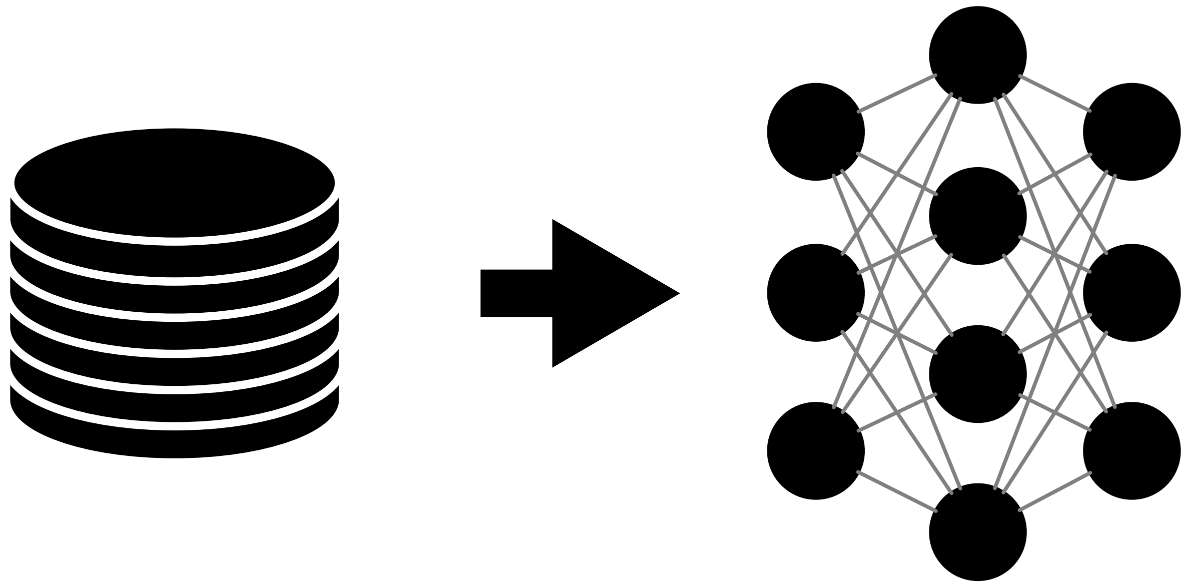
A Tangential Approach

(that has been successful… so far)

- Out-Of-Distribution-ness (OOD) is correlated with uncertainty:
  - If a new sample is close to the mean of training data distribution, the model will have low uncertainty (provided that the training has been effective)
  - If a new sample comes from an unseen distribution, the model is likely to show high epistemic uncertainty!
- This was the core idea behind Surprise Adequacy [2]

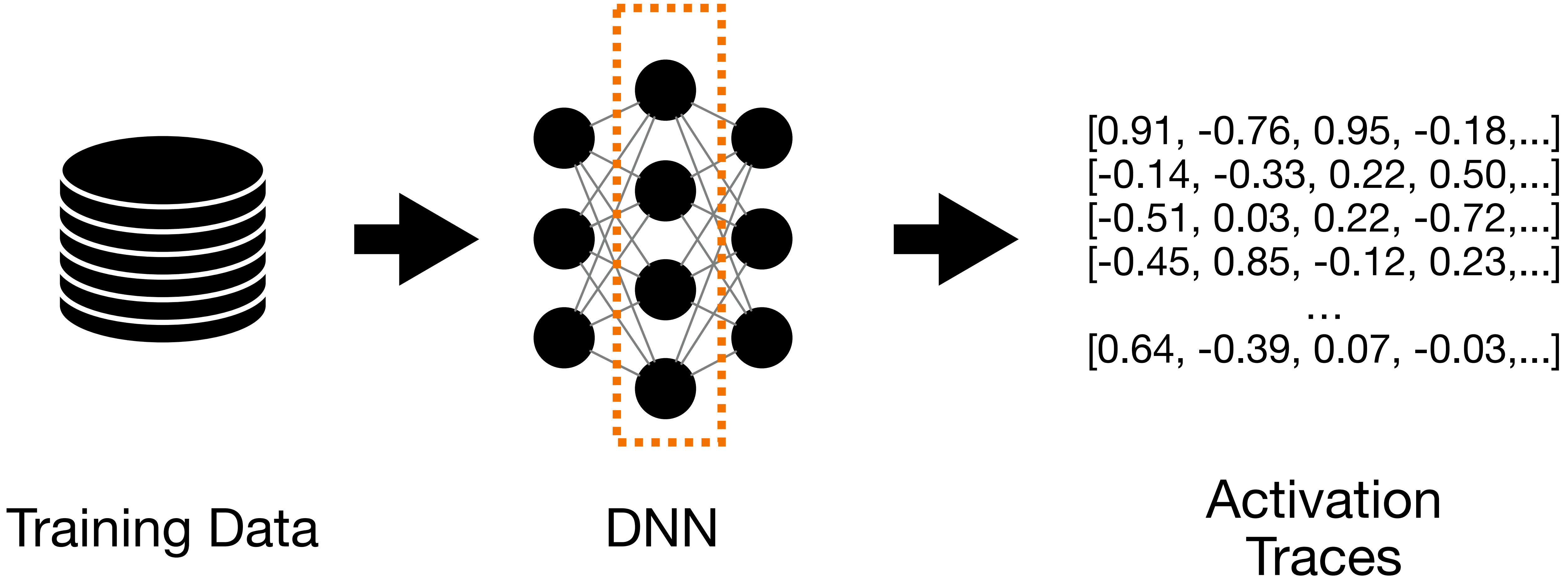
[2] Kim, J., Feldt, R., and Yoo, S. Guiding deep learning system testing using surprise adequacy. ICSE 2019, pp. 1039–1049.

Training Data



Training Data

DNN

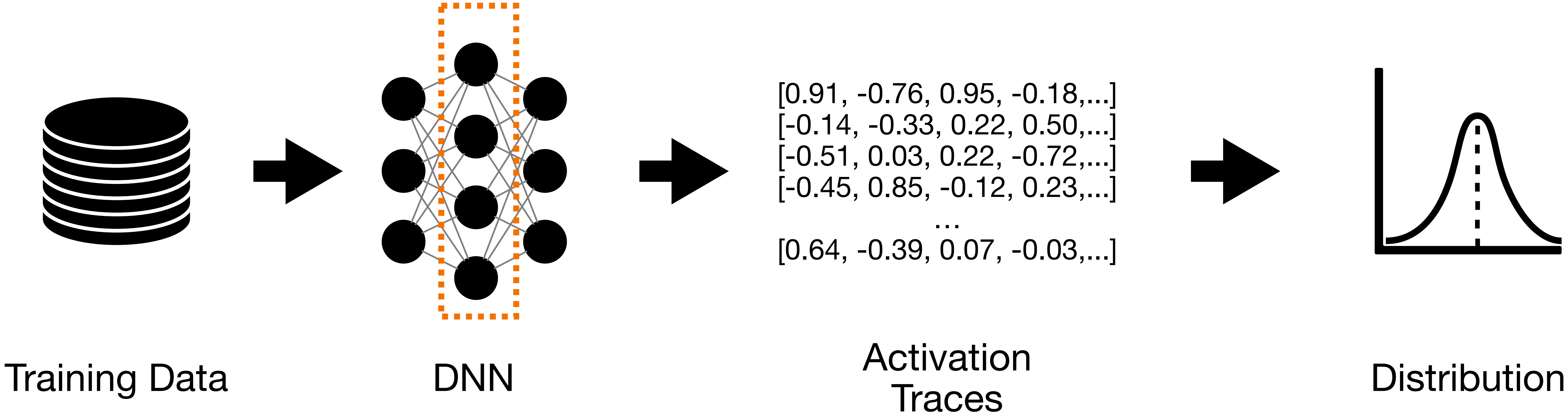


Training Data

DNN

[0.91, -0.76, 0.95, -0.18,...]
[-0.14, -0.33, 0.22, 0.50,...]
[-0.51, 0.03, 0.22, -0.72,...]
[-0.45, 0.85, -0.12, 0.23,...]
...
[0.64, -0.39, 0.07, -0.03,...]

Activation Traces



Training Data

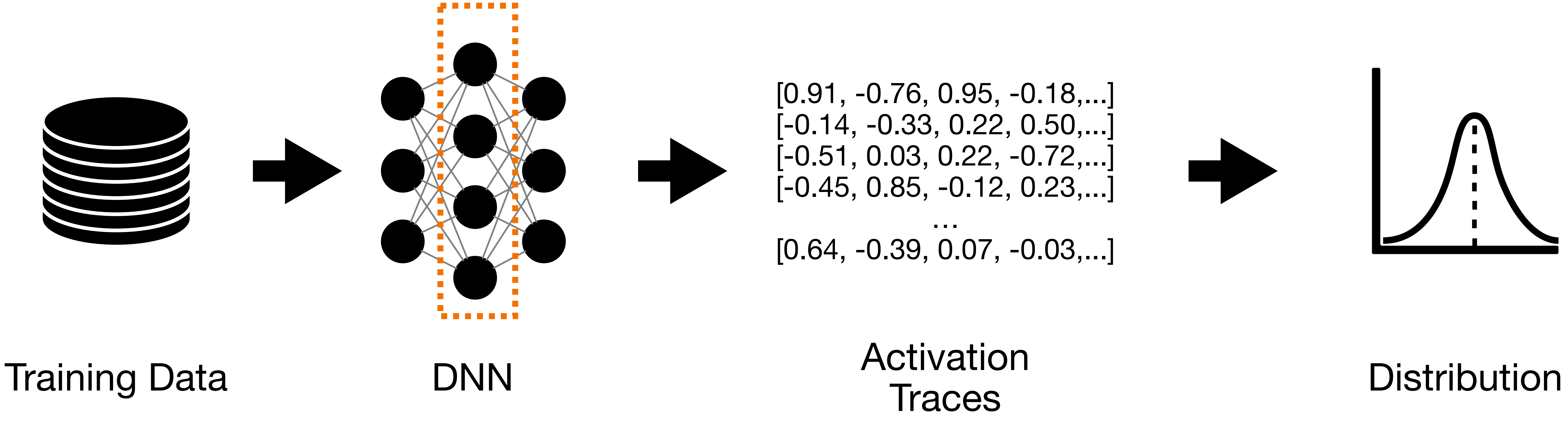
DNN

[0.91, -0.76, 0.95, -0.18,...]
[-0.14, -0.33, 0.22, 0.50,...]
[-0.51, 0.03, 0.22, -0.72,...]
[-0.45, 0.85, -0.12, 0.23,...]
...
[0.64, -0.39, 0.07, -0.03,...]

Activation
Traces

Distribution

New Sample



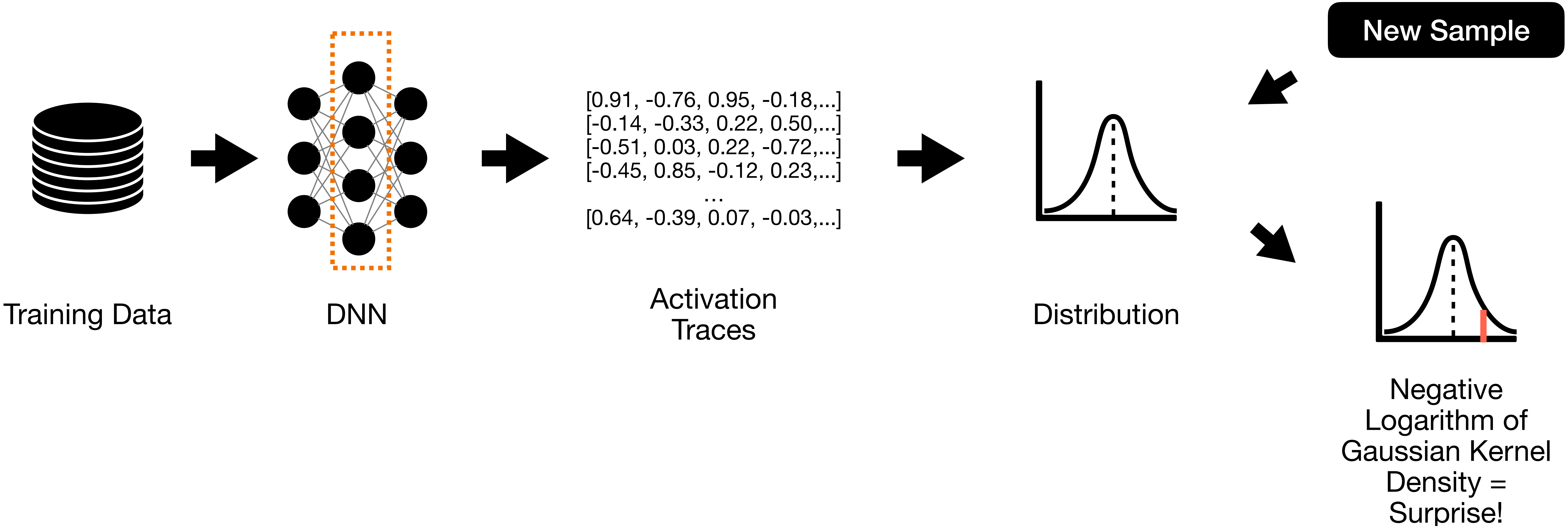
[0.91, -0.76, 0.95, -0.18,...]
[-0.14, -0.33, 0.22, 0.50,...]
[-0.51, 0.03, 0.22, -0.72,...]
[-0.45, 0.85, -0.12, 0.23,...]
...
[0.64, -0.39, 0.07, -0.03,...]

Training Data

DNN

Activation Traces

Distribution



New Sample

[0.91, -0.76, 0.95, -0.18,...]
[-0.14, -0.33, 0.22, 0.50,...]
[-0.51, 0.03, 0.22, -0.72,...]
[-0.45, 0.85, -0.12, 0.23,...]
...
[0.64, -0.39, 0.07, -0.03,...]

Training Data

DNN

Activation Traces

Distribution

Negative Logarithm of Gaussian Kernel Density = Surprise!

For large pre-trained LLMs, such as GPT4o, what is the training data?

(answer: the entire ~~human knowledge~~ internet)

For large pre-trained LLMs, such as GPT4o, what is the model accuracy?

(answer: autoregressive loss against… the entire ~~human knowledge~~ internet)

(or, even worse, some task-specific accuracy based on **emergent** inference capabilities…)

For large pre-trained LLMs, such as GPT4o, what is the model accuracy?

(answer: autoregressive loss against… the entire ~~human knowledge~~ internet)

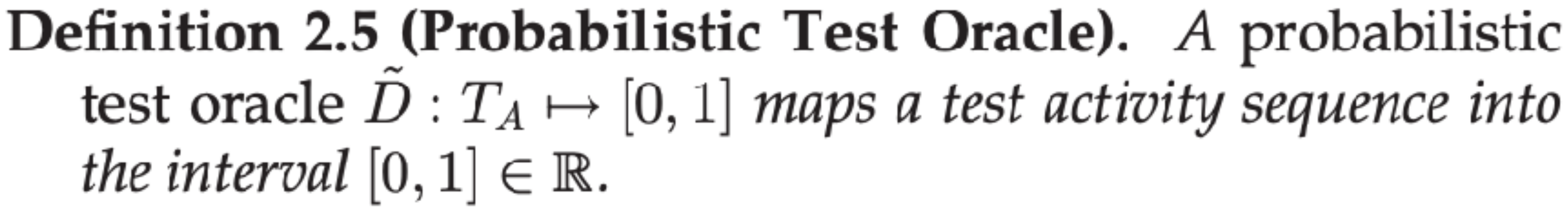
LLM Input Distribution $\simeq$ 1:1 map of the Empire

- Too big to handle, both conceptually and practically, i.e., too big to use as a reference to measure the OOD of a new sample.
- It would be nice to have a more manageable, but **sufficiently representative**, frame of reference.
  - Fine tuning dataset (if you have used fine-tuning)
  - Task-specific reference set (but how representative is it?)

Test Oracle

- For traditional programs, oracles are predicates that, given a stimulus to the system under test, tell you whether the observed behavior are as expected. They are discrete, Boolean.

- For ML systems, oracles are probability distributions that, given a new sample, tell you how likely it is that the resulting behavior are as expected. They are **continuous**, **approximate** [3].



Definition 2.5 (Probabilistic Test Oracle). *A probabilistic* test oracle $\tilde{D} : T_A \mapsto [0, 1]$ *maps a test activity sequence into the interval* $[0, 1] \in \mathbb{R}$.

[3] Barr, E., Harman, M., McMinn, P., Shahbaz, M., and Yoo, S. The oracle problem in software testing: A survey. IEEE Transactions on Software Engineering 41, 5 (May 2015), 507–525.

When and What to Measure

	OOD	**Variance**	**Token Probability**
Before Generation	• Last Hidden State of Input w.r.t. Reference Set • External Input Embeddings w.r.t. Reference Set		
After Generation (Internal)	• Last Hidden State of Output w.r.t. Reference Set • External Output Embedding w.r.t. Reference Set	• Variance of Last Hidden State of Outputs • Semantic Entropy • Test-time Augmentation	• (Max/Average) Token Probability • (Max/Average) Token Entropy
After Generation (External)		• Self-Consistency	

When and What to Measure

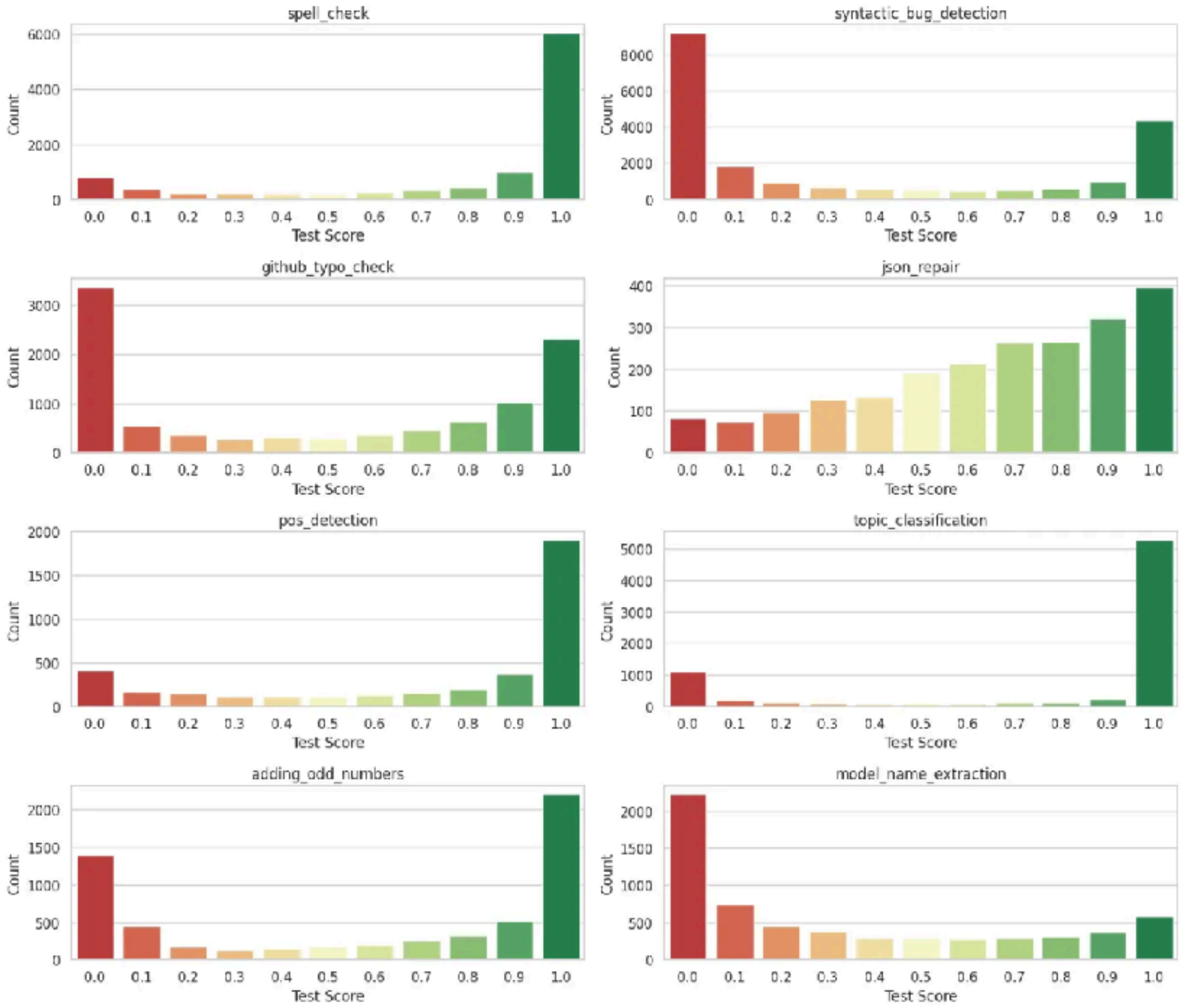
	OOD	Variance	Token Probability
Before Generation	Fit a Gaussian Mixture Model (GMM) to the reference set -> Estimate probability of the incoming sample using the GMM		
After Generation (Internal)		• Variance of Last Hidden State of Outputs • Semantic Entropy • Test-time Augmentation	• (Max/Average) Token Probability • (Max/Average) Token Entropy
After Generation (External)		• Self-Consistency	

When and What to Measure

	OOD	**Variance**	**Token Probability**
Before Generation	Fit a Gaussian Mixture Model (GMM) to the reference set -> Estimate probability of the incoming sample using the GMM	Perform k runs per input -> Compute covariance matrix -> Take the trace (i.e., sum of diagonals)	
After Generation (Internal)			• (Max/Average) Token Probability • (Max/Average) Token Entropy
After Generation (External)			

Dataset Statistics

Test Score = Ratio of Correct Responses out of 10 Queries / Llama 3.1 8B

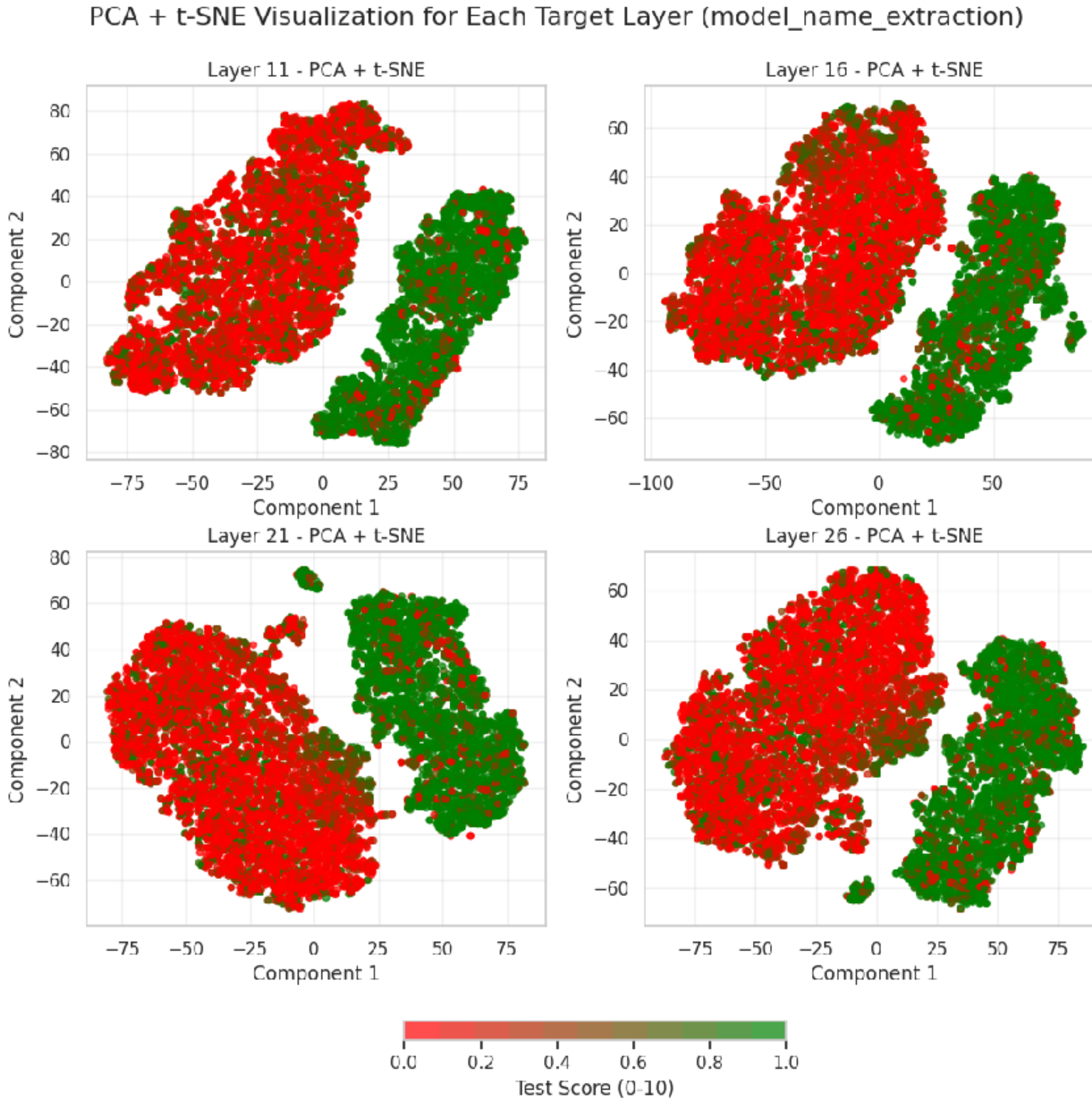


Task	# test cases	average test score	“all” correct ratio	“any” correct
syntactic_bug_detection	20518	0.36	0.21	0.55
spell_check	10000	0.80	0.60	0.92
github_typo_check	10000	0.48	0.23	0.66
json_repair	6563	0.66	0.19	0.96
pos_detection	15359	0.74	0.51	0.90
topic_classification	7600	0.78	0.70	0.85
adding_odd_numbers	6000	0.59	0.37	0.77
model_name_extraction	9810	0.54	0.32	0.76

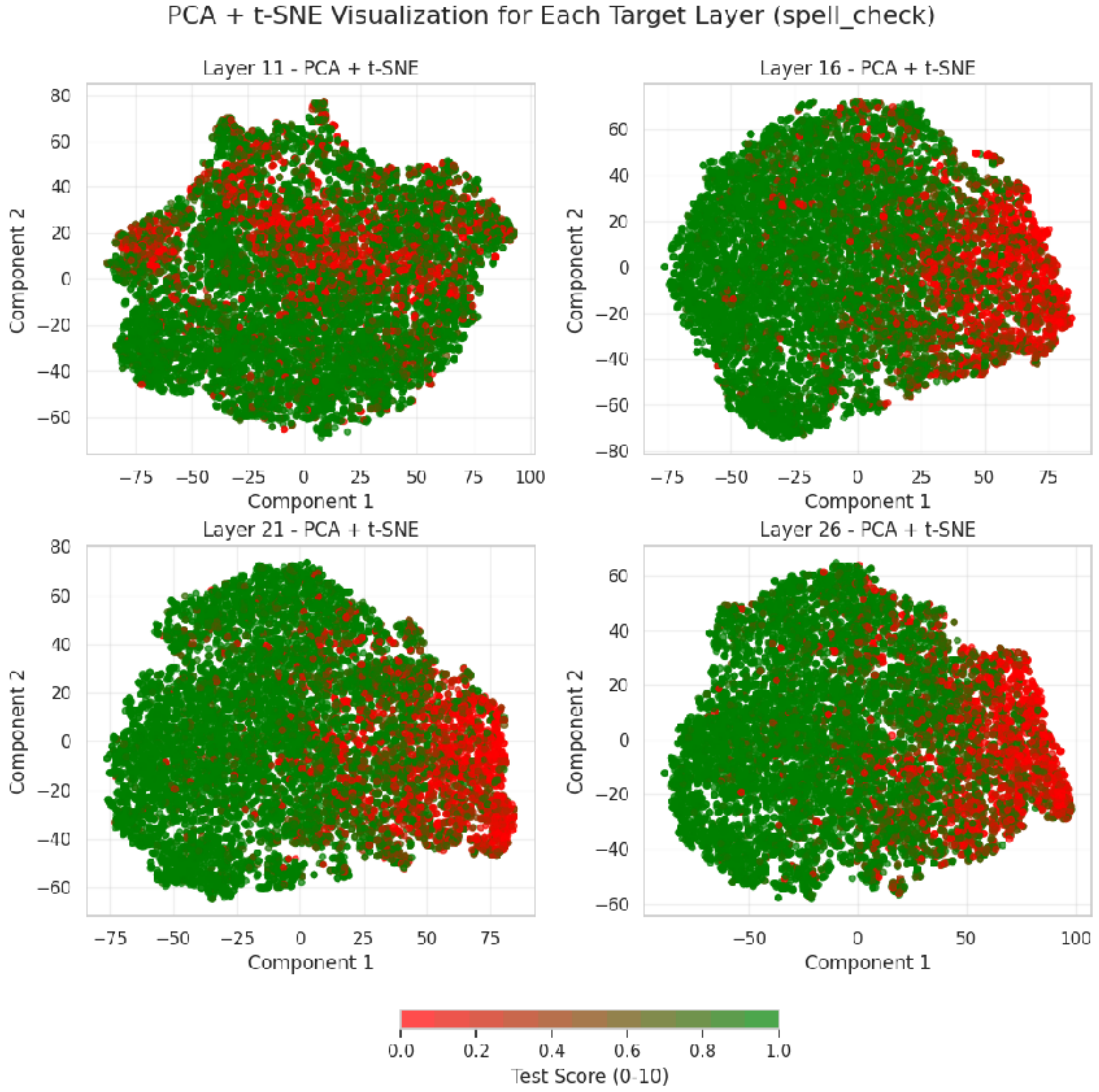
Metric Comparison

Correlation between Quantified Uncertainty and Test Scores

Task	LIH-OOD	LOH-VAR	Average Log Probability
syntactic_bug_detection	**0.3857**	0.2265	0.2184
spell_check	0.4267	**0.7086**	0.3922
github_typo_check	0.3680	**0.8049**	0.3737
json_repair	0.3922	**0.5990**	0.2636
pos_detection	0.1556	**0.6226**	0.2268
topic_classification	0.2918	**0.4329**	0.1199
adding_odd_numbers	**0.3721**	0.2439	0.3396
model_name_extraction	0.4011	**0.5651**	0.1620



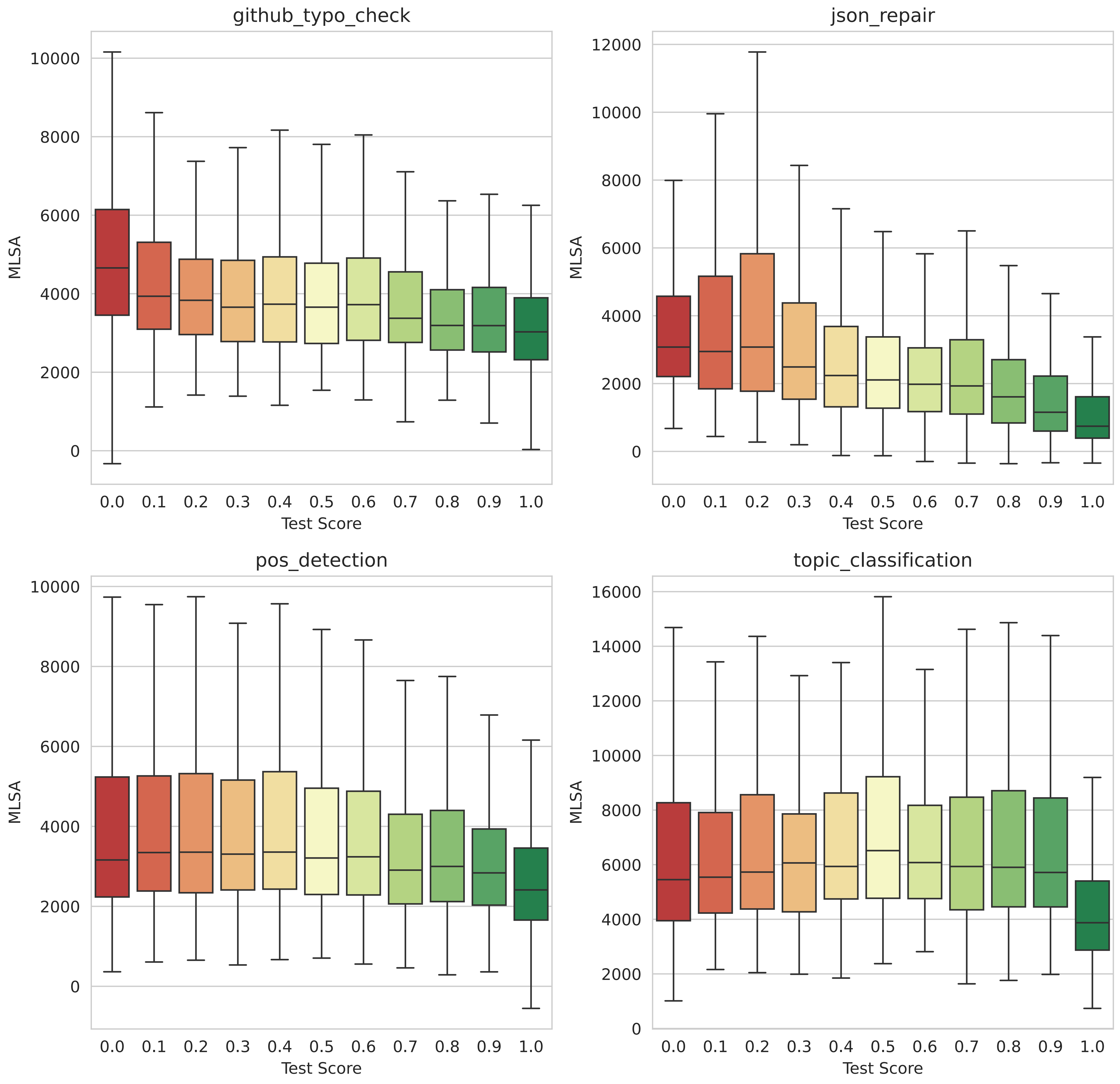
PCA + t-SNE Visualization for Each Target Layer (model_name_extraction)
Layer 11 - PCA + t-SNE
Layer 16 - PCA + t-SNE
Layer 21 - PCA + t-SNE
Layer 26 - PCA + t-SNE
Component 1
Component 2
80
60
40
20
0
−20
−40
−60
−80
−100
−75
−50
−25
0
25
50
75
0.0
0.2
0.4
0.6
0.8
1.0
Test Score (0-10)

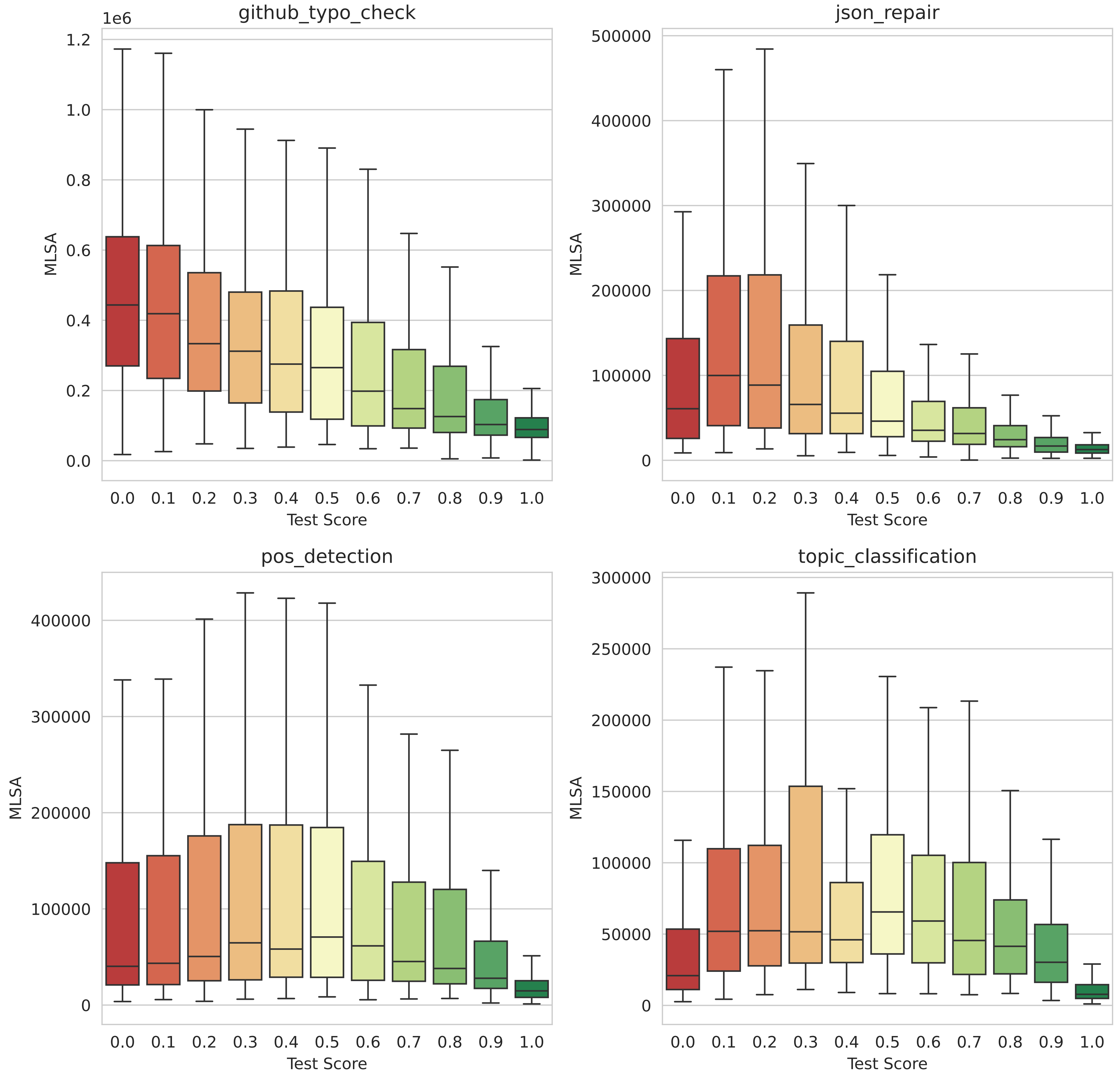


PCA + t-SNE Visualization for Each Target Layer (spell_check)
Layer 11 - PCA + t-SNE
Layer 16 - PCA + t-SNE
Layer 21 - PCA + t-SNE
Layer 26 - PCA + t-SNE
Component 1
Component 2
80
60
40
20
0
−20
−40
−60
−80
−75
−50
−25
0
25
50
75
100
0.0
0.2
0.4
0.6
0.8
1.0
Test Score (0-10)

LIH-OOD vs. LOH-VAR

(x-axis: test score, y-axis: OOD/VAR metric)





LIH-OOD vs. LOH-VAR

- OOD: If a new sample is out of distribution w.r.t. the known inputs (i.e., the reference set), then the performance goes down.
- VAR: However, an LLM can have low output variance and still be incorrect. That is, it can generate a wrong answer with high certainty - without anyone being adversarial!
- Open Issues
    - Test scores ≠ Original Model Performance (i.e., next token prediction): we are evaluating their emergent behavior, i.e., none of this is intended.
    - Reference set is in no way complete.

Going Forward

- How do we “refine” the probabilistic oracle? Essentially, we need to systemically augment the reference set so that the distribution sufficiently **covers** the **operational area** of the input space.
  - In the end, we have to rely on human inputs, but we want to minimise the human involvement as much as possible.
- Can we combine multiple oracles to have a more accurate one?