

# From Out-of-Distribution Detection to Hallucination Detection: A geometric view.

Litian Liu<sup>1</sup>, Reza Pourreza<sup>1</sup>, Yubing Jian<sup>1</sup>, Yao Qin<sup>2</sup>, Roland Memisevic<sup>1</sup>

1. Qualcomm AI Research 2. UC Santa Babara

# Overview

- Uncertainty scores from Out-of-Distribution (OOD) Detectors
  - fDBD: Fast Decision Boundary based OOD Detector (*Liu and Qin, ICML2024*)
  - NCI: Neural-Collapse Inspired OOD Detector (*Liu and Qin, CVPR2025*)
- LLM Hallucination Detection
  - Connection to OOD detection.
  - Bridging the gap between OOD and Hallucination detection.
- Experiments

# Out-of-Distribution Detection in Classification

- Out-of-Distribution Detection - when test samples belong to unseen classes during training.

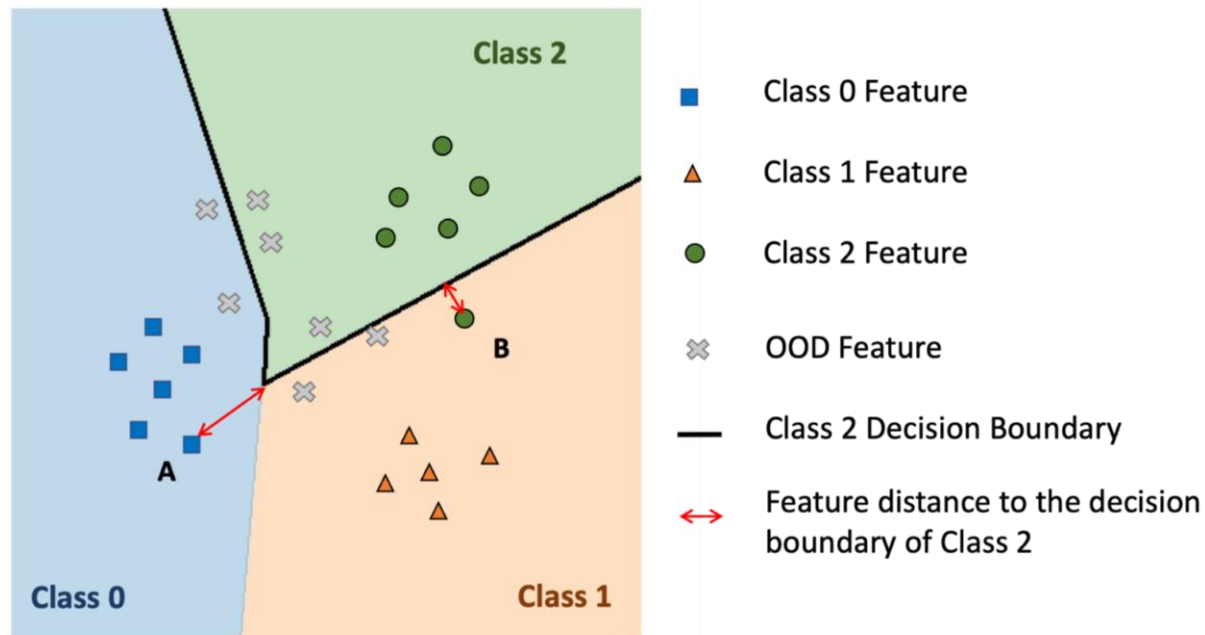


Dog Classifier



# fDBD: Feature Distance to Decision Boundary

- Definition: Minimum perturbation magnitude to change the prediction to a target class  $c$
- Intuition: Larger distance  $\leftrightarrow$  Higher model confidence  $\leftrightarrow$  ID.
- Closed-form Estimation: 
$$\widetilde{D}_f(z_x, c) := \frac{|(w_{f(x)} - w_c)^T z_x + (b_{f(x)} - b_c)|}{|w_{f(x)} - w_c|_2}$$
- OOD detection score: Average distance to each alternative class.



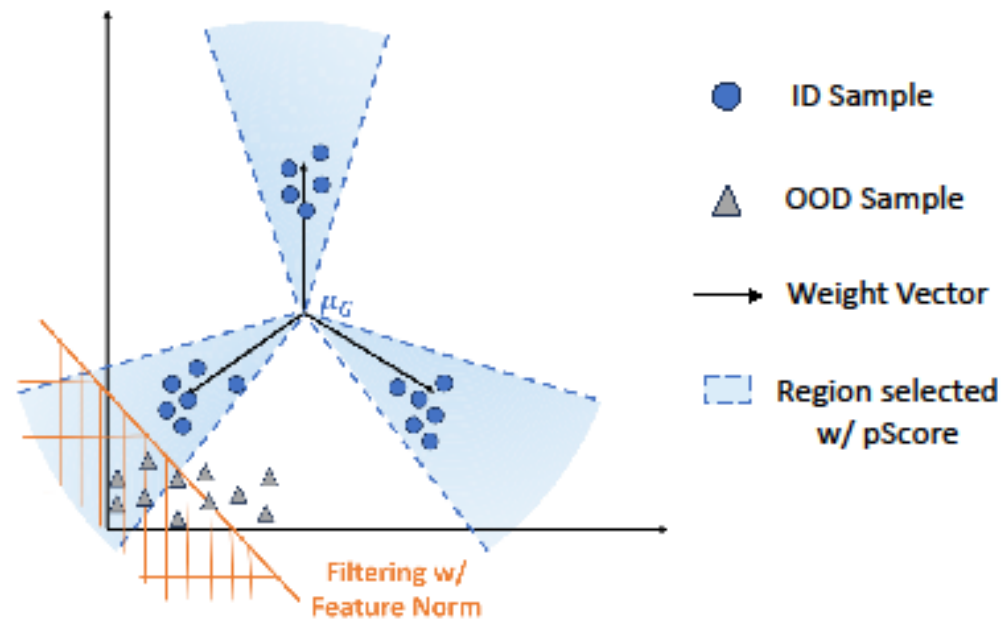
# NCI: Feature Proximity to Weight Vectors

- Intuition from Neural Collapse:

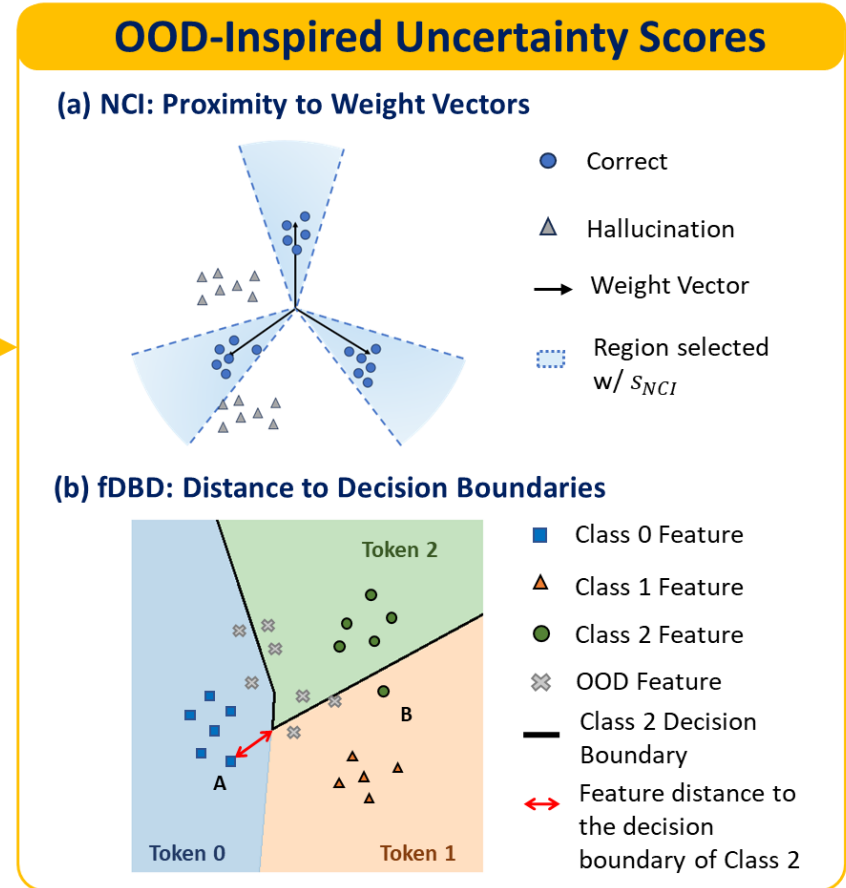
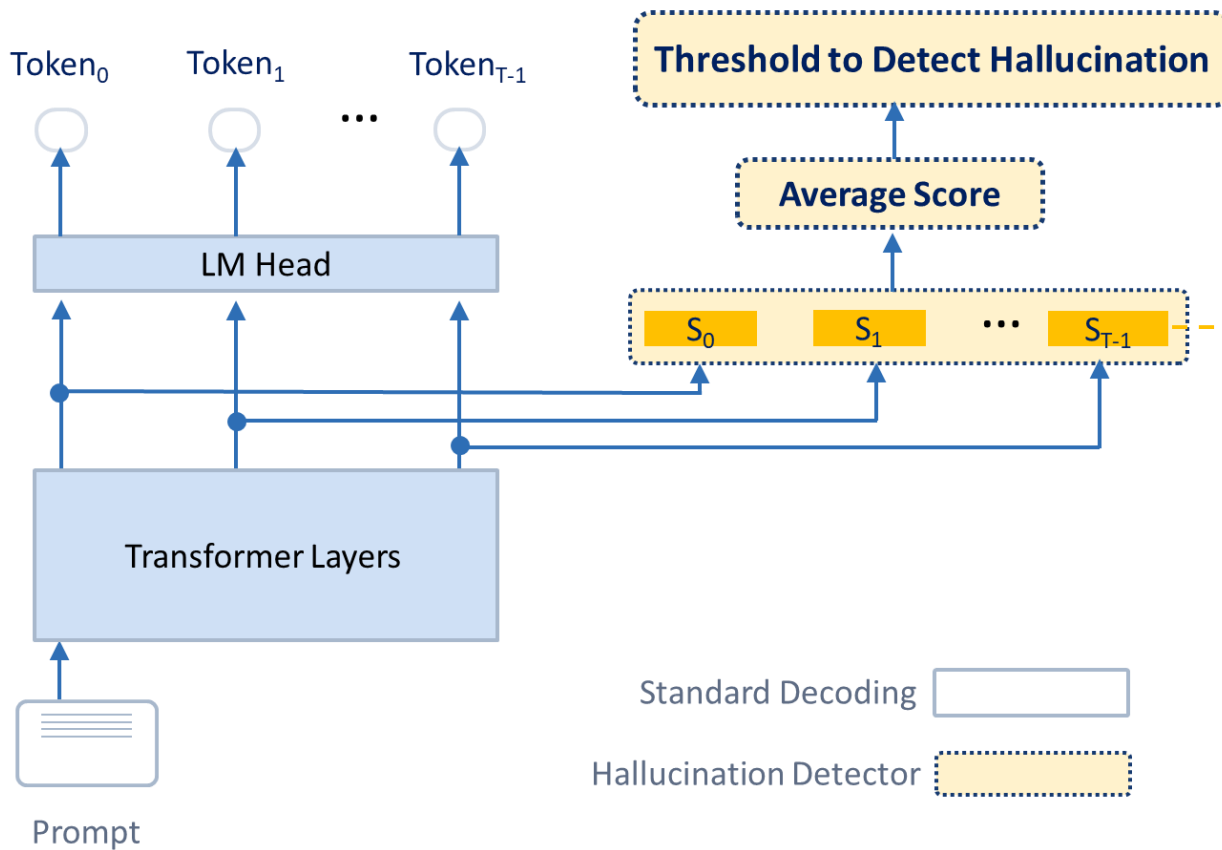
Centered ID features tend to cluster near the weight vector of predicted class.

- Definition:

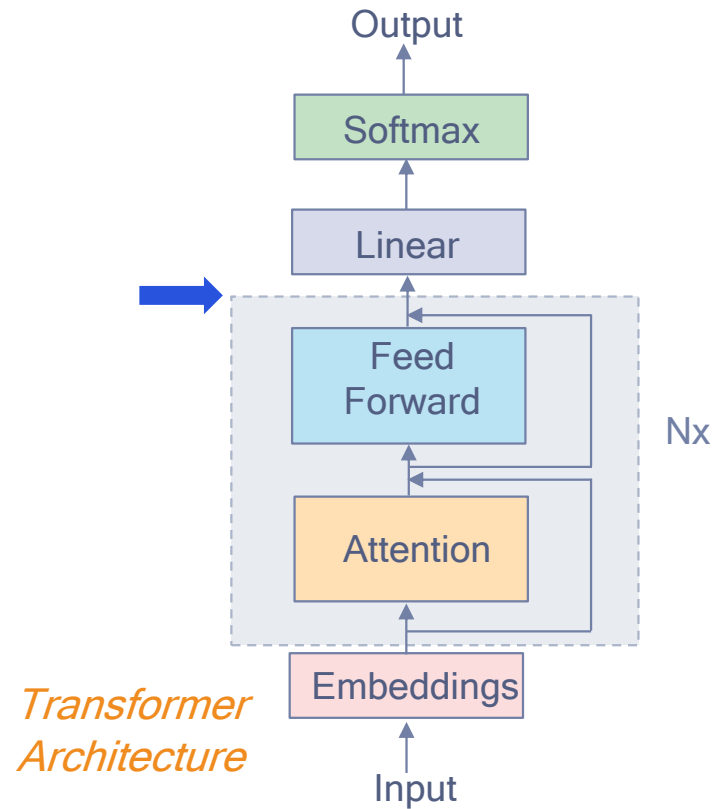
$$\text{pScore} = \cos(\omega_c, h - \mu_G) \|\omega_c\|_2, \text{ where } \cos(\omega_c, h - \mu_G) = \frac{(h - \mu_G)\omega_c}{\|h - \mu_G\|_2 \|\omega_c\|_2}.$$



# From OOD Detection to Hallucination Detection



# Bridging the Gap from OOD Detection



## Challenge I: Estimating Training Statistics at Scale

- fDBD: 
$$\frac{1}{|C| - 1} \sum_{c \in C, c \neq f(x)} \frac{\widetilde{D}_f(z_x, c)}{\|z_x - \mu_G\|_2}$$
- NCI: 
$$\cos(\omega_c, h - \mu_G) \|\omega_c\|_2$$

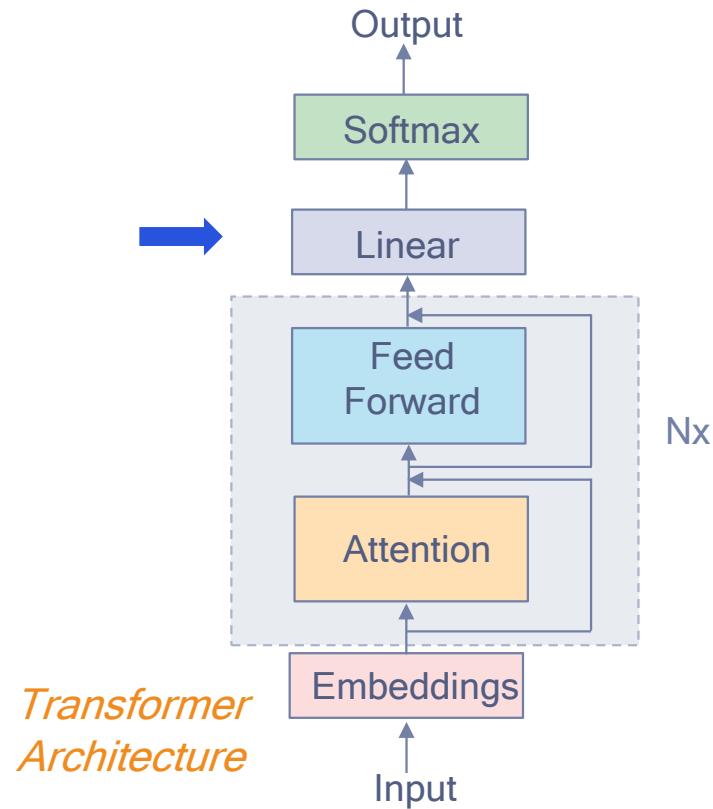
→ Analytical Proxy:

Mean of features  $\leftrightarrow$  center of language head

Decision-Neutral Closest Point

$$\hat{z}_* = -(W^T P W)^\dagger W^T P b$$

# Bridging the Gap from OOD Detection



## Challenge II: Massive Vocabulary Space

- Classification labels: ~1000 (ImageNet)
- LLM Vocabulary: 128,256 (Llama-3.2-3B)

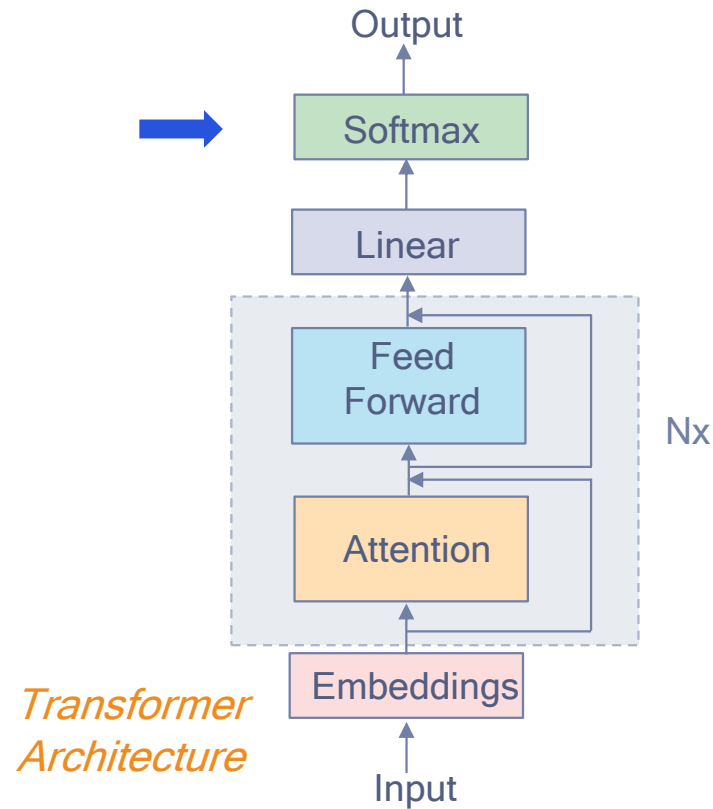
Aggregating signal from all labels: High noise and cost.

Selective Aggregation of fDBD:

- Step 1: select top k most likely tokens per step
- Step 2: aggregate per step across selected tokens

$$S_{fDBD}^k = \frac{1}{k} \sum_{c \in K_t} \frac{\widetilde{D}_f(z^t, c)}{\|z^t - \mu_G\|_2}$$

# Bridging the Gap from OOD Detection



## Challenge III: Stochastic Decoding

- Both fDBD and NCI can be imprecise.
- During reasoning, the metrics “catch up”:

Method	T = 0.2	T = 0.5	T = 0.8	T = 1.0
Perplexity	63.49 ± 0.75	62.08 ± 2.34	63.45 ± 0.71	62.68 ± 1.08
NCI	67.07 ± 0.53	66.04 ± 1.59	67.53 ± 0.88	67.93 ± 1.65
fDBD	69.30 ± 0.56	68.19 ± 1.47	69.12 ± 0.56	69.19 ± 1.98

# Results

## Superior Detection Effectiveness

### Minimal Latency

	Latency (ms/token)
Standard	31.94
Perplexity	32.88
NCI	32.54
fDBD	32.71

Model	Methods	Single Sample	CSQA	GSM8K	AQuA
Llama-3.2-3B-Instruct	Perplexity	✓	63.23	69.63	72.85
	Predictive Probability	✓	61.63	70.88	69.07
	LN Predictive Probability	✓	61.51	70.68	68.98
	Max P	✓	66.01	73.90	66.02
	P(True)	✓	47.73	51.02	39.38
	CoE-R	✓	47.06	50.12	45.55
	CoE-C	✓	58.82	60.69	62.56
	Lexical Similarity	✗	62.94	73.66	71.48
	SelfCheckGPT NLI	✗	64.18	74.29	66.01
	Semantic Entropy	✗	60.61	64.40	64.71
	<b>NCI</b>	✓	66.07	<u>76.32</u>	74.41
	<b>fDBD</b>	✓	<u>68.15</u>	75.59	<u>75.80</u>
	<b>fDBD (selected k)</b>	✓	<b>69.24</b>	<b>76.36</b>	<b>76.20</b>
Qwen-2.5-7B-Instruct	Perplexity	✓	61.94	71.54	71.66
	Predictive Probability	✓	64.91	73.29	73.37
	LN Predictive Probability	✓	65.19	73.01	74.17
	Max P	✓	49.90	50.00	50.83
	P(True)	✓	68.01	70.31	72.86
	CoE-R	✓	62.75	75.13	72.13
	CoE-C	✓	66.89	75.50	72.04
	Lexical Similarity	✗	60.57	72.02	72.62
	SelfCheckGPT NLI	✗	60.18	76.22	70.90
	Semantic Entropy	✗	59.10	66.83	69.62
	<b>NCI</b>	✓	<u>71.60</u>	75.83	<u>78.19</u>
	<b>fDBD</b>	✓	71.50	<u>77.19</u>	<u>77.08</u>
	<b>fDBD (selected k)</b>	✓	<b>72.47</b>	<b>77.19</b>	<b>78.22</b>

# Thank you



Follow us on: [in](#) [twitter](#) [instagram](#) [youtube](#) [facebook](#)

For more information, visit us at:

[qualcomm.com](http://qualcomm.com) & [qualcomm.com/blog](http://qualcomm.com/blog)

All data and information contained in or disclosed by this document is confidential and proprietary information of Qualcomm Technologies, Inc. and/or its affiliated companies and all rights therein are expressly reserved. By accepting this material the recipient agrees that this material and the information contained therein will not be used, copied, reproduced in whole or in part, nor its contents revealed in any manner to others without the express written permission of Qualcomm Technologies, Inc. Nothing in these materials is an offer to sell any of the components or devices referenced herein.

©2018-2023 Qualcomm Technologies, Inc. and/or its affiliated companies.  
All Rights Reserved.

Qualcomm is a trademark or registered trademark of Qualcomm Incorporated. Other products and brand names may be trademarks or registered trademarks of their respective owners.

References in this presentation to “Qualcomm” may mean Qualcomm Incorporated, Qualcomm Technologies, Inc., and/or other subsidiaries or business units within the Qualcomm corporate structure, as applicable. Qualcomm Incorporated includes our licensing business, QTL, and the vast majority of our patent portfolio. Qualcomm Technologies, Inc., a subsidiary of Qualcomm Incorporated, operates, along with its subsidiaries, substantially all of our engineering, research and development functions, and substantially all of our products and services businesses, including our QCT semiconductor business.

Snapdragon and Qualcomm branded products are products of Qualcomm Technologies, Inc. and/or its subsidiaries. Qualcomm patented technologies are licensed by Qualcomm Incorporated.