Your Finetuned Large Language Model is Already a Powerful Out-of-Distribution Detector

Andi Zhang [1,3] Tim Z. Xiao [2,4,5] Weiyang Liu [3,5] Robert Bamler[2] Damon Wischik[3]

[1]University of Manchester [2]University of Tübingen [3]University of Cambridge
[4]IMPRS-IS [5]Max Planck Institute for Intelligent Systems, Tübingen

Takeaway

Your finetuned large language model (LLM) already functions as a powerful out-of-distribution (OOD) detector. After finetuning an LLM, you obtain a new distribution $p_{\theta'}$ while still having access to the pretrained distribution p_θ. For any input sentence x, you can easily calculate the likelihood ratio $p_\theta(x)/p_{\theta'}(x)$, as an effective criterion for OOD detection.

Currently, it is straightforward to access both a finetuned LLM and its pre-trained version from online platforms such as Huggingface. Calculating our proposed OOD criterion requires only feeding the input through each model once, with no additional computational cost. Implementing the method requires only three lines of code: calculate the log likelihood for each model separately, then subtract them to obtain the criterion.

(Unsupervised) OOD Detection

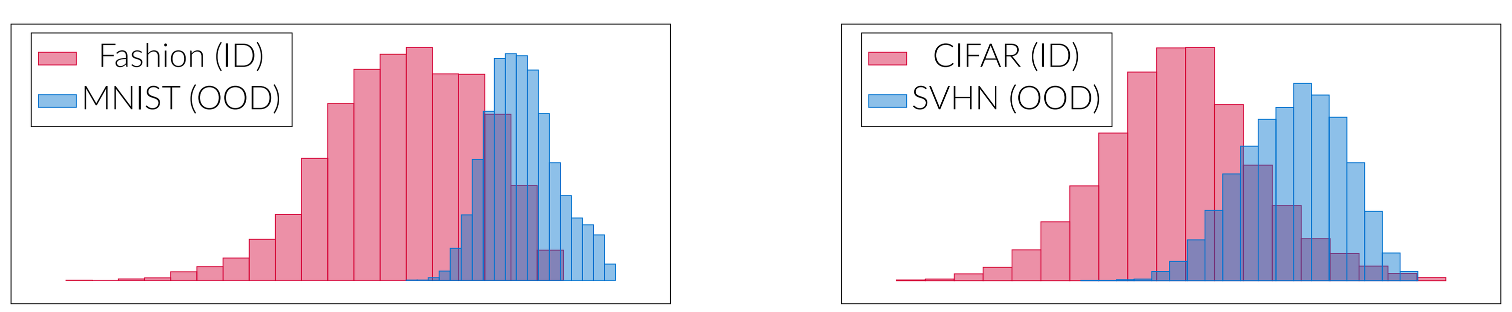
Hendrycks and Gimple [4] established a baseline for deep learning OOD detection where a model trained on $\mathcal{D}_{\text{in}}^{\text{train}}$ provides a detection criterion S. Performance is evaluated by applying S to samples from $\mathcal{D}_{\text{in}}^{\text{test}} \cup \mathcal{D}_{\text{out}}^{\text{test}}$ and measuring AUROC, AUPR, and FPR95 [11].

The term unsupervised refers to the setting where labels for in-distribution data are not accessible.

Nalisnick's Paradox

Given an input x, using the likelihood of in-distribution $p(x)$ as an OOD criterion seems straightforward, since in-distribution data should have higher density within the in-distribution region. In practice, researchers fit a probabilistic generative model (PGM) p_θ on $\mathcal{D}_{\text{in}}^{\text{train}}$ and use $p_\theta(x)$ as the criterion to detect OOD samples [1].

However, Nalisnick et al. [6] find that for high-dimensional data such as images, sometimes $p_\theta(x)$ is higher for samples from $\mathcal{D}_{\text{out}}^{\text{test}}$ than for samples from $\mathcal{D}_{\text{in}}^{\text{test}}$. As illustrated in the figure below, OOD data obtain higher likelihood, making OOD detection ineffective — worse than random guessing.

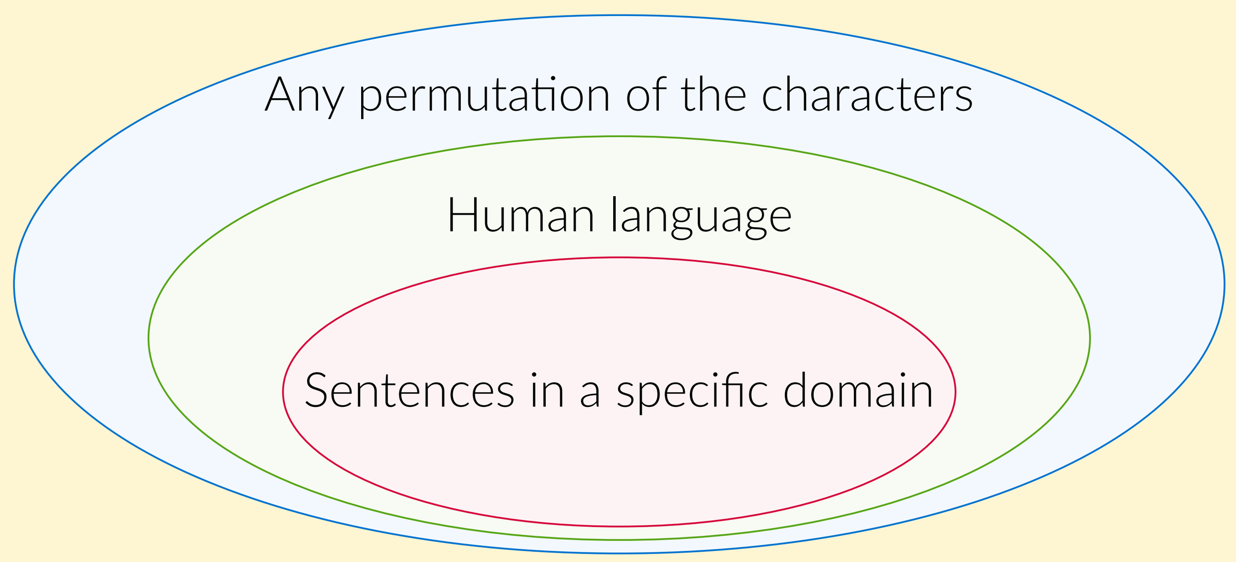


Likelihood Ratio and OOD Proxy

To address this paradox, several studies [8, 10, 9, 13, 2] have proposed using likelihood ratios as the criterion for identifying OOD data. [12] integrates these techniques into a comprehensive structure called the OOD proxy framework. In this framework, we assume that OOD data follow a distribution p_{out} which we cannot directly access. A tractable solution is to build a proxy distribution $p_{\text{out}}^{\text{proxy}}$ to represent p_{out}. The construction of these proxies incorporates empirically-based subjective understanding of OOD data: for example, [8] found that 'background statistics' are shared between in-distribution and OOD data, while [13] discovered that local features are shared between in-distribution and OOD data, making these effective OOD proxies. The criterion: $S(x) = p_{\text{out}}^{\text{proxy}}(x)/p_{\text{in}}(x)$.

Pretrained LLM as an OOD Proxy

Given the assumption that pretrained LLMs are trained on a comprehensive corpus of natural languages, we can assert that these models capture the shared features across the entire spectrum of natural languages. This characteristic naturally positions pretrained LLMs as effective OOD proxies.



The figure illustrates how pretrained LLM OOD proxies more accurately distinguish between in-distribution sentences and natural language compared to uniform OOD proxies.

Likelihood Ratio OOD Detection For QA Systems

For QA systems, detecting OOD questions is vital but difficult due to their brevity. Our approach leverages the observation that finetuned LLMs produce reasonable answers for in-distribution questions but unreasonable ones for OOD questions. We propose generating an answer for each question, then applying OOD detection to the question-answer pair rather than the question alone.

Experiments

Due to space limitations, only the results for Near OOD detection are presented here.

Dataset	In-D Label	Model	AUROC ↑	AUPR ↑	FPR95 ↓
ROSTD	No	Gangal et al. [3]	0.981	0.958	0.077
		Jin et al. [5]	0.990	0.973	0.041
		Llama-7B LH	0.960	0.890	0.168
		Llama-7B LR	**0.994**	0.984	0.023
		Mistral-7B LH	0.964	0.901	0.158
		Mistral-7B LR	0.992	0.978	0.033
		Llama-13B LH	0.965	0.905	0.166
		Llama-13B LR	**0.994**	**0.988**	**0.018**
	Yes	Podolskiy et al. [7]	0.998	0.994	0.008
SNIPS	No	Gangal et al. [3]	0.955	0.903	0.192
		Jin et al. [5]	0.963	0.910	0.145
		Llama-7B LH	0.912	0.829	0.391
		Llama-7B LR	0.993	0.986	0.029
		Mistral-7B LH	0.912	0.819	0.417
		Mistral-7B LR	0.987	0.968	0.087
		Llama-13B LH	0.942	0.872	0.280
		Llama-13B LR	**0.995**	**0.988**	**0.028**
	Yes	Podolskiy et al. [7]	0.978	0.933	0.120
CLINC150	No	Gangal et al. [3]	0.883	0.677	0.463
		Jin et al. [5]	0.902	0.703	0.417
		Llama-7B LH	0.821	0.456	0.538
		Llama-7B LR	**0.917**	**0.766**	**0.384**
		Mistral-7B LH	0.823	0.454	0.540
		Mistral-7B LR	0.913	0.730	0.399
		Llama-13B LH	0.820	0.450	0.546
		Llama-13B LR	0.915	0.742	0.386
	Yes	Podolskiy et al. [7]	0.982	0.939	0.092

From the table, we can observe that using the likelihood ratio between the finetuned model and the pretrained model yields the best performance in the unsupervised OOD detection.

References

[1] C. M. Bishop. Novelty detection and neural network validation. *IEE Proceedings-Vision, Image and Signal processing*, 141(4):217–222, 1994.

[2] A. L. Caterini and G. Loaiza-Ganem. Entropic issues in likelihood-based ood detection. In *I (Still) Can't Believe It's Not Better! Workshop at NeurIPS 2021*, pages 21–26. PMLR, 2022.

[3] V. Gangal, A. Arora, A. Einolghozati, and S. Gupta. Likelihood ratios and generative classifiers for unsupervised out-of-domain detection in task oriented dialog. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7764–7771, 2020.

[4] D. Hendrycks and K. Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*, 2016.

[5] D. Jin, S. Gao, S. Kim, Y. Liu, and D. Hakkani-Tür. Towards textual out-of-domain detection without in-domain labels. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:1386–1395, 2022.

[6] E. Nalisnick, A. Matsukawa, Y. W. Teh, D. Gorur, and B. Lakshminarayanan. Do deep generative models know what they don't know? *International Conference on Learning Representations*, 2019.

[7] A. Podolskiy, D. Lipin, A. Bout, E. Artemova, and I. Piontkovskaya. Revisiting mahalanobis distance for transformer-based out-of-domain detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13675–13682, 2021.

[8] J. Ren, P. J. Liu, E. Fertig, J. Snoek, R. Poplin, M. Depristo, J. Dillon, and B. Lakshminarayanan. Likelihood ratios for out-of-distribution detection. In *Advances in Neural Information Processing Systems*, pages 14707–14718, 2019.

[9] R. Schirrmeister, Y. Zhou, T. Ball, and D. Zhang. Understanding anomaly detection with deep invertible networks through hierarchies of distributions and features. *Advances in Neural Information Processing Systems*, 33:21038–21049, 2020.

[10] J. Serrà, D. Álvarez, V. Gómez, O. Slizovskaia, J. F. Núñez, and J. Luque. Input complexity and out-of-distribution detection with likelihood-based generative models. *arXiv preprint arXiv:1909.11480*, 2019.

[11] J. Yang, P. Wang, D. Zou, Z. Zhou, K. Ding, W. Peng, H. Wang, G. Chen, B. Li, Y. Sun, et al. Openood: Benchmarking generalized out-of-distribution detection. *Advances in Neural Information Processing Systems*, 35:32598–32611, 2022.

[12] A. Zhang and D. Wischik. Falsehoods that ml researchers believe about ood detection. *arXiv preprint arXiv:2210.12767*, 2022.

[13] M. Zhang, A. Zhang, and S. McDonagh. On the out-of-distribution generalization of probabilistic image modelling. *Advances in Neural Information Processing Systems*, 34, 2021.