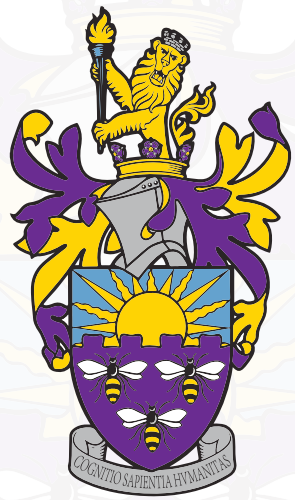
Your Finetuned Large Langugae Model is Already a Powerful OOD Detector



Andi Zhang, Tim Z. Xiao, Weiyang Liu,
Robert Bamler, Damon Wischik

Takeaway

- θ: parameters of the pre-trained LLM
- θ': parameters of the fine-tuned LLM on a **specific domain**

$$\frac{p_\theta(x)}{p_{\theta'}(x)} \quad : \quad \textbf{powerful OOD detector}$$

Takeaway

- θ: parameters of the pre-trained LLM
- θ': parameters of the fine-tuned LLM on a **specific domain**

$$\frac{p_\theta(x)}{p_{\theta'}(x)} \quad : \quad \textbf{powerful OOD detector}$$

Implementation in 3 lines:

```
pre_log_likelihood  = -pretrained_model(x).loss * len(x)
fine_log_likelihood = -finetuned_model(x).loss * len(x)
log_ratio = pre_log_likelihood - fine_log_likelihood
```

Note: loss of autoregressive LLMs = average NLL per token (perplexity).

Why?

Assumption of OOD Distribution

- Data come from a **true distribution** p_{true}

Assumption of OOD Distribution

- Data come from a **true distribution** p_{true}
- We only observe **in-distribution data** $\Rightarrow$ assume they follow p_{in}

Assumption of OOD Distribution

- Data come from a **true distribution** p_{true}
- We only observe **in-distribution data** $\Rightarrow$ assume they follow p_{in}
- **Extra assumption (not typical):** OOD data also follow some distribution p_{out}

Assumption of OOD Distribution

- Data come from a **true distribution** p_{true}
- We only observe **in-distribution data** $\Rightarrow$ assume they follow p_{in}
- **Extra assumption (not typical):** OOD data also follow some distribution p_{out}

Neyman–Pearson lemma: Given x, the optimal test is

$$\frac{p_{\text{out}}(x)}{p_{\text{in}}(x)}.$$

OOD Proxy

- The true p_{OOD} is **unknown**

[1] Andi Zhang and Damon Wischik. "Falsehoods that ML researchers believe about OOD detection". In: *arXiv preprint arXiv:2210.12767* (2022).

[2] Christopher M Bishop. "Novelty detection and neural network validation". In: *IEE Proceedings-Vision, Image and Signal processing* 141.4 (1994), pp. 217–222.

OOD Proxy

- The true p_{OOD} is **unknown**
- Zhang & Wischik[1]: use a **proxy distribution** $p_{\text{OOD}}^{\text{proxy}}$ to approximate the OOD distribution

$$\frac{p_{\text{OOD}}^{\text{proxy}}(x)}{p_{\text{in}}(x)}$$

[1] Andi Zhang and Damon Wischik. "Falsehoods that ML researchers believe about OOD detection". In: *arXiv preprint arXiv:2210.12767* (2022).

[2] Christopher M Bishop. "Novelty detection and neural network validation". In: *IEE Proceedings-Vision, Image and Signal processing* 141.4 (1994), pp. 217–222.

OOD Proxy

- The true p_{OOD} is **unknown**
- Zhang & Wischik[1]: use a **proxy distribution** $p_{\text{OOD}}^{\text{proxy}}$ to approximate the OOD distribution

$$\frac{p_{\text{OOD}}^{\text{proxy}}(x)}{p_{\text{in}}(x)}$$

Special case: If $p_{\text{OOD}}^{\text{proxy}}$ is uniform $\Rightarrow$ criterion reduces to in-distribution likelihood[2]

[1] Andi Zhang and Damon Wischik. "Falsehoods that ML researchers believe about OOD detection". In: *arXiv preprint arXiv:2210.12767* (2022).

[2] Christopher M Bishop. "Novelty detection and neural network validation". In: *IEE Proceedings-Vision, Image and Signal processing* 141.4 (1994), pp. 217–222.

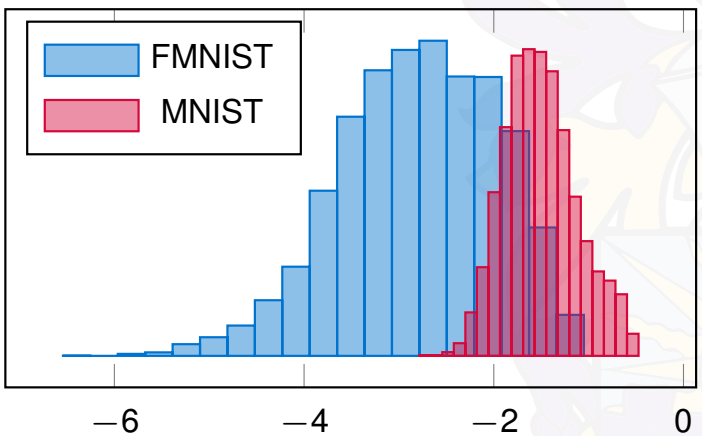
Failure of the Uniform OOD Proxy

- Nalisnick et al.[3] find: using uniform OOD proxy ($p_{\text{in}}(x)$ as the criterion) **can fail**.

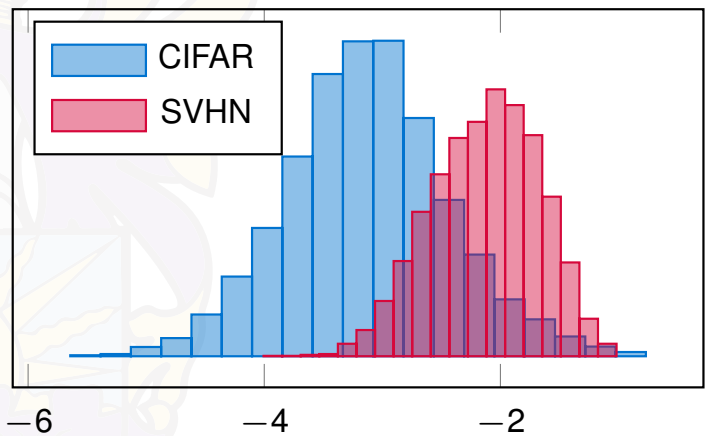
[3] Eric Nalisnick et al. "Do deep generative models know what they don't know?" In: *International Conference on Learning Representations* (2019).

Failure of the Uniform OOD Proxy

- Nalisnick et al.[3] find: using uniform OOD proxy ($p_{\text{in}}(x)$ as the criterion) **can fail**.
- In some cases, OOD samples get **higher likelihood** than in-distribution samples.



Trained on Fashion-MNIST



Trained on CIFAR-10

Minus BPD $\propto$ likelihood. OOD likelihood $>$ in-distribution likelihood $\Rightarrow$ uniform OOD proxy is **misleading**.

[3] Eric Nalisnick et al. "Do deep generative models know what they don't know?" In: *International Conference on Learning Representations* (2019).

Pretrained LLM as OOD Proxy

The University of Manchester

- We need some **smarter** OOD proxy

Pretrained LLM as OOD Proxy

- We need some **smarter** OOD proxy
- In practice, OOD $\neq$ random characters from uniform distribution which can be **filted by low level methods**

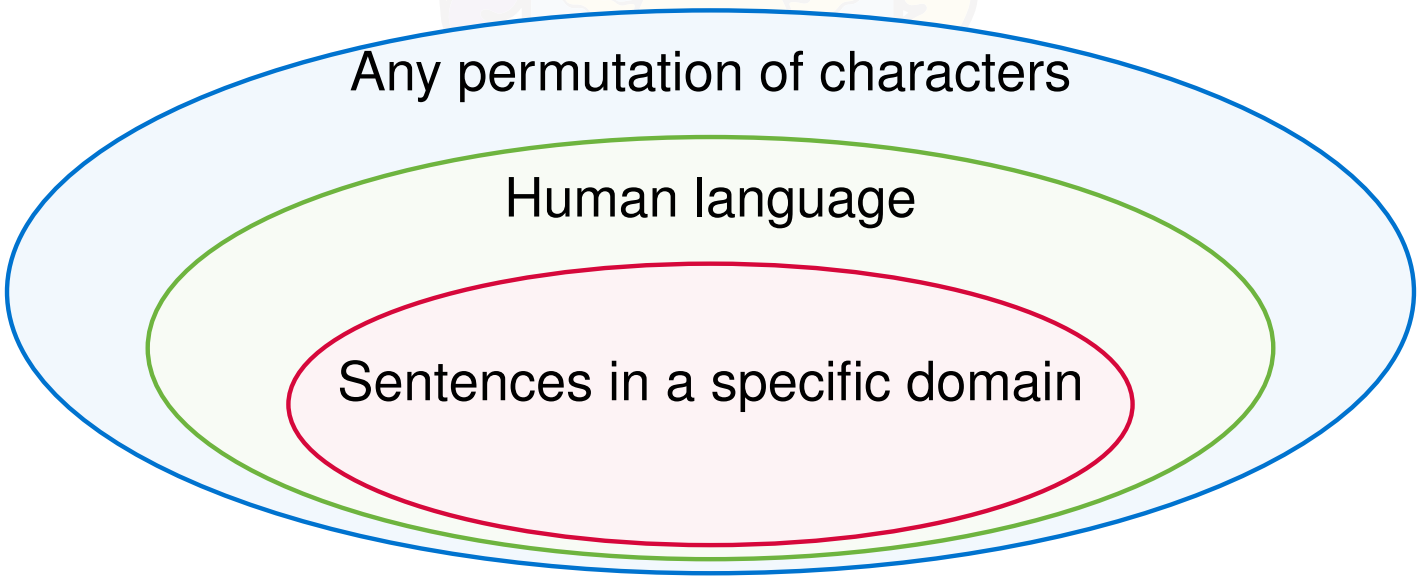
Pretrained LLM as OOD Proxy

- We need some **smarter** OOD proxy
- In practice, OOD $\neq$ random characters from uniform distribution which can be **filted by low level methods**
- OOD of interest: **human language outside the domain**

Pretrained LLM as OOD Proxy

- We need some **smarter** OOD proxy
- In practice, OOD $\neq$ random characters from uniform distribution which can be **filted by low level methods**
- OOD of interest: **human language outside the domain**
- Modern LLMs are trained on massive human-language corpora $\Rightarrow$ **can serve as an OOD proxy**:

$$\frac{p_{\text{OOD}}^{\text{proxy}}(x)}{p_{\text{in}}(x)} = \frac{p_\theta(x)}{p_{\theta'}(x)} = \frac{\text{pretrained}}{\text{finetuned}}$$



Final Takeaway

The University of Manchester

- A finetuned LLM is already a **powerful** OOD detector

Final Takeaway

The University of Manchester

- A finetuned LLM is already a **powerful** OOD detector
- Just calculate the **likelihood ratio**:

$$\frac{p_\theta(x)}{p_{\theta'}(x)}$$

Final Takeaway

- A finetuned LLM is already a **powerful** OOD detector
- Just calculate the **likelihood ratio**:

$$\frac{p_\theta(x)}{p_{\theta'}(x)}$$

- Can be implemented by **3 lines of code**

Final Takeaway

- A finetuned LLM is already a **powerful** OOD detector
- Just calculate the **likelihood ratio**:

$$\frac{p_\theta(x)}{p_{\theta'}(x)}$$

- Can be implemented by **3 lines of code**
- Gives you an OOD criterion **for free** if you already have a finetuned LLM.

Thanks!