BEYOND FINE-TUNING: LORA MODULES BOOST NEAR-OOD DETECTION AND LLM SECURITY

Etienne Salimbeni[1,2], **Francesco Craighero**[1], **Renata Khasanova**[2], **Milos Vasic**[2], **Pierre Vandergheynst**[1]
[1] EPFL, Lausanne, Switzerland
{etienne.salimbeni, francesco.craighero, pierre.vandergheynst}@epfl.ch
[2] Oracle Labs, Zurich, Switzerland
{milos.vasic, renata.khasanova}@oracle.com

ABSTRACT

Under resource constraints, LLMs are usually fine-tuned with additional knowledge using Parameter Efficient Fine-Tuning (PEFT), using Low-Rank Adaptation (LoRA) modules. In fact, LoRA injects a new set of small trainable matrices to adapt an LLM to a new task, while keeping the latter frozen. At deployment, LoRA weights are subsequently merged with the LLM weights to speed up inference. In this work, we show how to exploit the unmerged LoRA's embedding to boost the performance of Out-Of-Distribution (OOD) detectors, especially in the more challenging near-OOD scenarios. Accordingly, we demonstrate how improving OOD detection also helps in characterizing wrong predictions in downstream tasks, a fundamental aspect to improve the reliability of LLMs. Moreover, we will present a use-case in which the sensitivity of LoRA modules and OOD detection are employed together to alert stakeholders about new model updates. This scenario is particularly important when LLMs are out-sourced. Indeed, test functions should be applied as soon as the model changes the version in order to adapt prompts in the downstream applications. In order to validate our method, we performed tests on Multiple Choice Question Answering datasets, by focusing on the medical domain as a fine-tuning task. Our results motivate the use of LoRA modules even after deployment, since they provide strong features for OOD detection for fine-tuning tasks and can be employed to improve the security of LLMs.

1 INTRODUCTION

Large Language Models (LLMs) are gaining popularity due to their general-purpose capabilities and are increasingly integrated into real-world applications, including medicine (Thirunavukarasu et al., 2023) and finance (Li et al., 2023). Their fast developing pace and ease of integration is alarming, since misconfiguration can be particularly damaging (Wang et al., 2024; Koessler & Schuett, 2023; Bickmore et al., 2018). New government regulations are focusing on LLM-based applications. The *EU AI Act* (EUA, 2023) and the *White House Executive Order* on AI systems (Whi, 2023) are setting plans for their safe deployment, including the "robust monitoring of AI systems" (EUA, 2023). Additionally, new *OWASP* (OWA, 2023) guidelines have been published, highlighting the security risks of integrating LLM into applications.

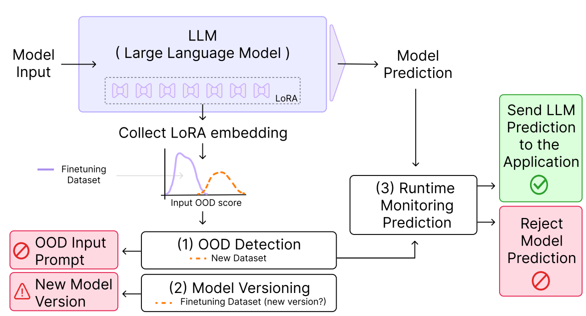

Figure 1: **Boosting LLM security with LoRA modules.** Given a fine-tuned LLM and the LoRA embeddings of the FT dataset, one can check: (1) if the LoRA embeddings of a new dataset are OOD, (2) if the model version has changed by detecting changes in LoRA embeddings, (3) if a prediction should be discarded due to an OOD input sample or low-confidence output.

One major challenge of Machine Learning is protecting against unexpected behaviours of the model. Indeed, real-world applications might involve data that differs from the training one due to distributional shifts (Yang et al., 2021). Coupled with random effects in the data, these shifts can make the model more uncertain about its predictions (Hüllermeier & Waegeman, 2021). Consequently, detecting such Out-Of-Distribution (OOD) instances (Yang et al., 2021) is crucial to allow users to discard untrustworthy predictions. While OOD detection has been a fast-growing field, especially on classification tasks, these approaches have been also been recently extended to LLMs and text generation (Ren et al., 2023). In this paper, we will study OOD detection in the context of fine-tuned LLMs employing Low-Rank Adaptation (LoRA) modules (Hu et al., 2022).

Fine-tuning is a common practice for adapting a model to a specific domain. However, recent results raise new concerns on the reliability of fine-tuned LLMs. Indeed, fine-tuning can deteriorate previous safety alignments enforced during pre-training (Qi et al., 2023). Moreover, it has been shown that fine-tuning worsens OOD robustness (Chen et al., 2023b). Low-Rank Adaptation (LoRA) modules (Hu et al., 2022) are commonly used to allow fine-tuning LLMs under resources constraints. Given a froze LLM, these small trainable modules are first injected for task adaptation and then merged with the model to reduce latency at inference time. Originally designed for fine-tuning purposes, LoRA modules are now being employed for greater control beyond their original purpose. Such applications include task arithmetic (add, combine or remove learned properties) (Zhang et al., 2023), scaling the influence of the fine-tuned task at inference(Shah et al., 2023), and switching tasks using dynamic LoRA module routing (Huang et al., 2024; Sheng et al., 2023).

In the following, we show how *unmerged* LoRA modules can also be exploited to improve the security and reliability of LLMs. First, we show that LoRA embeddings are more sensitive to near-OOD samples, allowing simpler OOD detectors such as the Mahalanobis Distance (Lee et al., 2018b) to perform well in most scenarios. Second, we will present a novel use-case of OOD detection for model inspection. Model updates might in fact require version checking (Hao et al., 2023), to prevent major security flaws such as backdoor attacks (Yang et al., 2023) as well as simple misconfiguration in the LLM service supply chain (Hao et al., 2023). With LoRA embeddings, one can easily detect model changes even under subtle updates. Last, we will test how LoRA embeddings improve runtime prediction monitoring, also known as selective prediction (Geifman & El-Yaniv, 2017; Lakshminarayanan et al., 2017; Tran et al., 2022), when an LLM is employed for downstream tasks such as question answering. While OOD detection accounts for unintended inputs, a prediction might be uncertainty also due to random effects in the data. These two sources of uncertainty are usually referred to as epistemic, due to lack of knowledge, and aleatoric uncertainty, due to the stochastic nature of the data-generating process (Hüllermeier & Waegeman, 2021). Similar to previous results on vision tasks (Kaur et al., 2021), we will show how aggregating OOD detection and the entropy of the model confidence can improve the reliability of LLMs. This combined approach, accounting for the two sources of uncertainty, improves the detection of incorrect predictions compared to taking each individual metric alone. We will focus on decoder-only LLMs and medical Multiple Choice Question Answering (MCQA). However, it's important to note that our methods are applicable to other large pre-trained models fine-tuned with LoRA and other generation tasks.

Contributions In Section 3.1 we show that LoRA Embedding boost the detection performance in near-OOD scenarios over the fine-tuning task. In Section 3.2 we present a novel use-case where OOD detection is employed to detect model updates. Last, in Section 3.3 we combine OOD detection and the output entropy to improve near-OOD runtime prediction monitoring in downstream tasks.

1.1 OOD Detection

Starting from the training data, a common approach to tackle OOD detection first builds a distribution of embeddings or outputs, such as the maximum softmax probability or the perplexity. Then, samples that significantly deviate from such distribution are rejected (Yang et al., 2021).

OOD Detection from embeddings Given the focus of this work on LoRA modules, we will consider distance- (Sun et al., 2022) and density-based approaches (Lee et al., 2018b; Ren et al., 2021) to detect OOD embeddings. The Deep Nearest Neighbors (Sun et al., 2022) method employs the distance to the K-th Nearest Neighbors (KNN) as a metric to measure the deviation of an embedding from the training distribution. The Mahalanobis Distance (MD) (Lee et al., 2018b) measures the OOD score of a test sample x as: $\text{MD}_{\text{train}}(x) := \text{MD}(x; \mu_{\text{train}}, \Sigma_{\text{train}}) := (x-\mu)^T\Sigma^{-1}(x-\mu)$.

where μ_{train} and Σ_{train} are obtained by fitting a multivariate Gaussian $\mathcal{N}(\mu, \Sigma), \mu \in \mathbb{R}^d, \Sigma \in \mathbb{R}^{d\times d}$ to the training data. Since MD might struggle on near-OOD samples, the Relative Mahalanobis Distance (RMD) (Ren et al., 2021) improves it by normalizing the training data likelihood $\text{MD}_{\text{train}}(x)$ with a background dataset $\text{MD}_{\text{bg}}(x)$: $\text{RMD}_{\text{train}}(x) := \text{MD}_{\text{train}}(x) - \text{MD}_{\text{bg}}(x)$.

Crucially, while the performance of RMD and KNN might be sensitive to the choice of the background dataset and the number of neighbors, respectively, this is not the case of MD, which has no additional requirements.

OOD Detection in LLMs Recently, OOD detection has been investigated in the context of conditional language models (Ren et al., 2023). More in detail, it has been shown that perplexity alone is unreliable for detecting OOD samples. On the other hand, combining perplexity with RMD on the last layer activation of both the encoder and decoder is a better performing alternative to discard low-quality outputs given OOD inputs.

2 METHODS

2.1 DATASETS AND MODEL

We integrated the previous results on OOD detection with RMD within the abstractive summarization and translation domains (Ren et al., 2023) by focusing on multiple question answering, which limits the number of generated token to 1. We selected three MCQA datasets and chose the medical domain as a fine-tuning task, by considering the MedMCQA (Pal et al., 2022) and the PubMedQA (Jin et al., 2019) datasets. Then, we employed the MMLU (Hendrycks et al., 2021) multi-domain dataset to define both near- and far-OOD samples (refer to Appendix A.1 to get the subtasks assigned to each category). In contrast to (Ren et al., 2023), we use a decoder only language model: Llama2-7B (Touvron et al., 2023). Llama2-7B has a vocabulary size of 32 000, an embedding size of 4 096 and has 32 layers. We fine-tuned the model with LoRA (Hu et al., 2022) on the MedMCQA training split, using a batch size of 32, the Adam optimizer and a learning rate of 2e-4. Moreover, we set LoRA to rank 16 and attached it to the query and value projections of each transformer layer. Concatenating all LoRA embeddings leads to a final embedding of size $32 \times 2 \times 16 = 2048$.

2.2 EMBEDDINGS

We compare two types of embeddings: last layer activations and LoRA embeddings. LoRA reparametrization of the i-th layer can be expressed as $l^i(x) = W_0^i x + B^i A^i x$, where W_0^i is the pretrained frozen weights and $B^i A^i$ are two matrices of the LoRA module. Now, given an input of N tokens, we define the last layer activation embedding as $E_{\text{LLA}}(x) := \frac{1}{N}\sum_{i=1}^{N} l_i^{\text{L}}(x)$ and LoRA embeddings as $E_{\text{LORA}}(x) = \frac{1}{N}\sum_{i=1}^{N} ||_{j=1}^{L} A^j x$. Where l_i^L is the final layer activation for token i, and $||_{j=1}^{L} A^j x$ denotes the concatenation of all LoRA modules intermediate activation $A^j x$ for the L layers (see Fig. 3). Both embeddings are scaled through division by the maximum value. Importantly, we considered multiple layer embeddings only with LoRA, due to its reduced dimensionality compared to the full-rank layers.

2.3 OOD DETECTION AND PREDICTION MONITORING

OOD Detection In order to perform OOD detection, we will compare all the three approaches mentioned in Section 1.1: MD (Lee et al., 2018a), RMD (Ren et al., 2021) like in Ren et al. (2023), using PubMedQA as the background dataset, and KNN (Sun et al., 2022), with $k = 100$ as number of neighbors. Both MD and RMD employed the embeddings on the fine-tuning dataset to compute μ_{train} and Σ_{train}.

Selective Prediction Similar to (Kaur et al., 2021), we will consider a combination of aleatoric and epistemic uncertainties (Hüllermeier & Waegeman, 2021) in order to define a stronger approach to monitor the predictions of our model, also called selective prediction (Geifman & El-Yaniv, 2017;

Lakshminarayanan et al., 2017; Tran et al., 2022). Likewise, in Ren et al. (2023) the perplexity of an LLM was combined with RMD for selective generation.

By detecting the embeddings that are far from the in-distribution ones, OOD detectors mostly capture the epistemic uncertainty of a model. For this experiment, we will consider the MD approach, that, thanks to the LoRA embeddings E_{LORA}, is comparable with RMD and KNN while having less requirements (see Table 1).

In order to estimate the aleatoric uncertainty, we simply compute the entropy of the token providing the answer to the question. Given a question x, let $f_i(x)$ be the output confidence of an LLM for the i-th answer. Then, the Shannon entropy of the output is defined as $H(x) = -\sum_i f_i(x) \log f_i(x)$.

While MD has no upper-bound, the entropy range is $[0, 1]$. Therefore, in order to combine them, it would be convenient to rescale the former. Since the squared Mahalanobis distance MD^2 follows a Chi-square ($\chi^{2,d}$) distribution with d degrees of freedom (Manly, 2014), where d is the number of dimensions of the data point, we can take the p-value of the $\chi^{2,d}$ instead of the MD to obtain a normalized value. The p-value associated with MD is calculated as $p_{MD}(x) = 1 - \text{CDF}_{\chi^{2,d}}(MD^2(x))$ where CDF stands for the cumulative distribution function. Then, given a question x with the associated LLM embeddings $E(x)$ (either E_{LLA} or E_{LoRA}), we can compute the p-value p_{MD} and the Shannon entropy $H(x)$ of the model prediction. The final combination is simply defined as $H(x) + p_{MD}(E(x))$.

3 Results

3.1 LoRA Modules Improve Near-OOD Detection

In Table 1 we compare the AUROC score for OOD detection of different embeddings (E_{LLA}, E_{LoRA}) on both near- and far-OOD datasets, as defined in Appendix A.1, against the test dataset of MedMCQA (our in-distribution fine-tuning domain). In accordance with the results reported in (Ren et al., 2023), the perplexity proves to be a poor choice as an OOD score, as it struggles to distinguish even far-OOD datasets. When employing the last layer embeddings, all the methods perfectly discriminate far-OOD datasets. However, in near-OOD scenarios only RMD demonstrates positive performance, while KNN and MD fail completely. On the other hand, LoRA embeddings allow KNN and MD to perform on par with RMD on the near-OOD datasets, while keeping the same high performance on the far-OOD ones. As clearly emerges from Fig. 4, LoRA embeddings boost the performance of the simpler MD approach, that neither requires hyperparameter tuning nor additional datasets like KNN and RMD, respectively. Indeed, RMD heavily depends on the goodness of the background dataset to perform well in the near-OOD dataset.

	Near OOD →				Far OOD
Method	clinical knowledge	anatomy	college biology	computer science	professional law
Perplexity	0.651	0.383	0.654	0.587	0.712
	Last Layer Activation (E_{LLA})				
KNN	0.387	0.296	0.786	**0.997**	**0.999**
MD	0.428	0.312	0.774	**0.997**	**0.999**
RMD* (baseline)	0.688	0.730	**0.998**	**0.997**	**0.999**
	LoRA (E_{LoRA})				
KNN	<u>0.819</u>	0.729	0.890	**0.997**	0.998
MD	0.814	<u>0.733</u>	0.890	0.996	0.994
RMD*	**0.828**	**0.762**	**0.998**	0.993	**0.999**

Table 1: **OOD detection AUROCs.** AUROCs distinguishing MMLU tasks from the MedMCQA dataset.
* RMD requires a background dataset.
(baseline) The approach of (Ren et al., 2023).

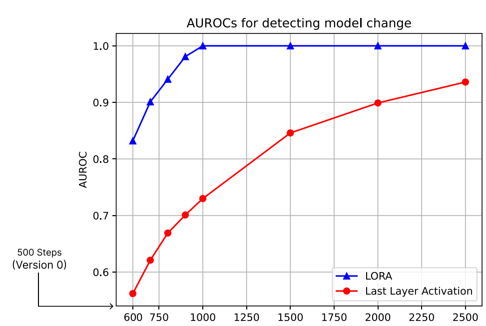

Figure 2: **AUROCs distinguishing the embeddings at different fine-tuning steps.** AUROCs of the Mahalanobis Distance distinguishing MedMCQA embeddings after 500 fine-tuning steps (model version 0) from the ones after > 500 steps (next model versions).

3.2 Detecting Model Updates

Given the good performance of the simple MD approach on LoRA embeddings, even in near-OOD scenarios, we investigate an interesting use-case to improve the security of fine-tuned LLMs: detecting the degree of change of a model version update. This time, instead of checking if an external dataset is OOD, we aim to detect whether the embeddings of the in-distribution data have changed due to a (possibly unexpected) model update. OpenAI's models endpoint degradation over time on some specific tasks (Chen et al., 2023a) underlines the practical significance of this issue. Existing methods, such as verifying model weights hashes (Hao et al., 2023) or using zero-knowledge proofs (South et al., 2024), offer only a binary indication of model change. Given a dataset of interest and a model version, our approach is instead able to quantify model change. Such a scenario is relevant when a stakeholder out-sources LLM for a specific fine-tuning task, where a model update might trigger a testing cascade on downstream tasks (Hao et al., 2023). Indeed, prompts may be invalidated on a different model version and malicious updates might inject backdoors in the model (Yang et al., 2023). In Fig. 2, we present the MD AUROCs for discriminating between the embeddings of our LLM fine-tuned for 500 steps on the MedMCQA training set (model version 0) and those obtained after fine-tuning for > 500 steps (next versions). Clearly, LoRA embeddings are much more sensitive to model updates than the last layer ones: while the latter has an AUROC $> .8$ 1000 fine-tuning steps after version 0 at 500 steps, the former after only 100.

3.3 Runtime Monitoring Predictions

In Table 2 we report the AUROCs when detecting incorrect model predictions, i.e., wrong answer choices. We tested the output entropy, MD on the two types of embeddings and a combination of the two. The results show again how LoRA helps to improve MD in the near-OOD scenario, even if the setting is different than Section 3.1. Moreover, aggregating MD and entropy achieves the best performance, due to the different sources of uncertainty captured by the two metrics, i.e. epistemic and aleatoric (Hüllermeier & Waegeman, 2021).

	MedMCQA	Near OOD	Far OOD
Entropy	0.554	0.541	0.547
MD LLA	0.528	0.509	0.510
MD LLA + Entropy	0.582	0.550	**0.549**
MD LORA	0.531	0.523	0.509
MD LORA + Entropy	**0.589**	**0.576**	0.543

Table 2: **AUROCs scores when differentiating correct and incorrect predictions**. We considered the MedMCQA validation dataset, and the near- and far-OOD datasets defined in Appendix A.1.

4 Conclusion

In our experiments, we found compelling evidence supporting the hypothesis that LoRA embeddings possess stronger near-OOD properties compared to last layer activations and perplexity in fine-tuning tasks, integrating previous research on OOD detection in LLMs (Ren et al., 2023). This enables LLM-based applications to better monitor whether the model is being used for the intended task, to quantify model version changes when the LLM is out-sourced, and to halt the model when the uncertainty about its predictions in a downstream task is too high. Importantly, LoRA modules allow us to employ simpler approaches for OOD detection, such as the Mahalanobis distance, that neither rely on additional data nor require hyperparameter tuning. Our findings suggest that LoRA weights should be kept also at deployment time to keep fine-grained control over the fine-tuning task for security purposes. Note that our work relies on the LoRA embedding being served from the LLM API endpoint, a technique not commonly employed on platforms such as HuggingFace which currently limits its adoption. While LoRA is a now widely adopted approach, future work could explore other PEFT methods and their sensitivity to OOD data. Moreover, our approach doesn't cover full fine-tuning, but studies on task vectors show promise for task-specific adaptation in large models (Ortiz-Jimenez et al., 2023; Ilharco et al., 2023). Preliminary results combining LoRA with task vectors (Zhang et al., 2023) hint at new ways to enhance OOD detection for full fine-tuning. Last, recently proposed uncertainty estimation approaches could be investigated as an alternative to the simple predictive entropy Lin et al. (2023); Kuhn et al. (2023) in our aggregated metric for runtime prediction monitoring.

4.1 FUNDING

F. Craighero is funded by the Swiss National Science Foundation (SNSF) Sinergia grant CRSII5_205884.

REFERENCES

Eu artificial intelligence act. https://www.europarl.europa.eu/doceo/document/TA-9-2023-0236_EN.pdf, 2023.

Owasp top 10 llm applications. https://owasp.org/www-project-top-10-for-large-language-model-applications/, 2023.

Us executive order on the safe, secure, and trustworthy development and use of artificial intelligence. https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/, 2023.

Timothy Bickmore, Ha Trinh, Stefán Ólafsson, Teresa O'Leary, Reza Asadi, Nathaniel Rickles, and Ricardo Cruz. Patient and consumer safety when using conversational assistants for medical information: Observational study (preprint). *Journal of Medical Internet Research*, 20, 07 2018. doi: 10.2196/11510.

Lingjiao Chen, Matei Zaharia, and James Zou. How is chatgpt's behavior changing over time? *arXiv preprint arXiv:2307.09009*, 2023a.

Sishuo Chen, Wenkai Yang, Xiaohan Bi, and Xu Sun. Fine-tuning deteriorates general textual out-of-distribution detection by distorting task-agnostic features. In Andreas Vlachos and Isabelle Augenstein (eds.), *Findings of the Association for Computational Linguistics: EACL 2023, Dubrovnik, Croatia, May 2-6, 2023*, pp. 552–567. Association for Computational Linguistics, 2023b. doi: 10.18653/V1/2023.FINDINGS-EACL.41. URL https://doi.org/10.18653/v1/2023.findings-eacl.41.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pp. 4885–4894, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

Wei Hao, Daniel Mendoza, Rafael da Silva, Deepak Narayanan, and Amar Phanishaye. Mgit: A model versioning and management system. 2023.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=d7KBjmI3GmQ.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Chengsong Huang, Qian Liu, Bill Yuchen Lin, Tianyu Pang, Chao Du, and Min Lin. Lorahub: Efficient cross-task generalization via dynamic lora composition. 2024.

Eyke Hüllermeier and Willem Waegeman. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine Learning*, 110:457–506, 2021.

Gabriel Ilharco, Marco Túlio Ribeiro, Mitchell Wortsman, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/pdf?id=6t0Kwf8-jrj.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. PubMedQA: A dataset for biomedical research question answering. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2567–2577, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1259. URL https://aclanthology.org/D19-1259.

Ramneet Kaur, Susmit Jha, Anirban Roy, Sangdon Park, Oleg Sokolsky, and Insup Lee. Detecting oods as datapoints with high uncertainty. *ICML 2021 Workshop on Uncertainty and Robustness in Deep Learning*, 2021.

Leonie Koessler and Jonas Schuett. Risk assessment at agi companies: A review of popular risk assessment techniques from other safety-critical industries. 2023.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=VD-AYtP0dve.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pp. 6402–6413, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/9ef2ed4b7fd2c810847ffa5fa85bce38-Abstract.html.

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pp. 7167–7177, 2018a. URL https://proceedings.neurips.cc/paper/2018/hash/abdeb6f575ac5c6676b747bca8d09cc2-Abstract.html.

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *Advances in neural information processing systems*, 31, 2018b.

Yinheng Li, Shaofei Wang, Han Ding, and Hang Chen. Large language models in finance: A survey. In *Proceedings of the Fourth ACM International Conference on AI in Finance*, pp. 374–382, 2023.

Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. Generating with confidence: Uncertainty quantification for black-box large language models. *arXiv preprint arXiv:2305.19187*, 2023.

Bryan F. J. Manly. *Multivariate Statistical Methods: A Primer, Third Edition*. Chapman and Hall/CRC, New York, 3 edition, May 2014. ISBN 978-0-429-17601-2. doi: 10.1201/b16974.

Guillermo Ortiz-Jimenez, Alessandro Favero, and Pascal Frossard. Task arithmetic in the tangent space: Improved editing of pre-trained models. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=0A9f2jZDGW.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In Gerardo Flores, George H Chen, Tom Pollard, Joyce C Ho, and Tristan Naumann (eds.), *Proceedings of the Conference on Health, Inference, and Learning*, volume 174 of *Proceedings of Machine Learning Research*, pp. 248–260. PMLR, 07–08 Apr 2022. URL https://proceedings.mlr.press/v174/pal22a.html.

Xiangyu Qi, Yi Zeng, Tinghao Xie, Pin-Yu Chen, Ruoxi Jia, Prateek Mittal, and Peter Henderson. Fine-tuning aligned language models compromises safety, even when users do not intend to! 2023.

Jie Ren, Stanislav Fort, Jeremiah Liu, Abhijit Guha Roy, Shreyas Padhy, and Balaji Lakshminarayanan. A simple fix to mahalanobis distance for improving near-ood detection. *arXiv preprint arXiv:2106.09022*, 2021.

Jie Ren, Jiaming Luo, Yao Zhao, Kundan Krishna, Mohammad Saleh, Balaji Lakshminarayanan, and Peter J. Liu. Out-of-distribution detection and selective generation for conditional language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/pdf?id=kJUS5nD0vPB.

Viraj Shah, Nataniel Ruiz, Forrester Cole, Erika Lu, Svetlana Lazebnik, Yuanzhen Li, and Varun Jampani. Ziplora: Any subject in any style by effectively merging loras. 2023.

Ying Sheng, Shiyi Cao, Dacheng Li, Coleman Hooper, Nicholas Lee, Shuo Yang, Christopher Chou, Banghua Zhu, Lianmin Zheng, Kurt Keutzer, Joseph E. Gonzalez, and Ion Stoica. S-lora: Serving thousands of concurrent lora adapters. 2023.

Tobin South, Alexander Camuto, Shrey Jain, Shayla Nguyen, Robert Mahari, Christian Paquin, Jason Morton, and Alex 'Sandy' Pentland. Verifiable evaluations of machine learning models using zksnarks, 2024.

Yiyou Sun, Yifei Ming, Xiaojin Zhu, and Yixuan Li. Out-of-distribution detection with deep nearest neighbors. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato (eds.), *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pp. 20827–20840. PMLR, 2022. URL https://proceedings.mlr.press/v162/sun22d.html.

Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. Large language models in medicine. *Nature medicine*, 29(8): 1930–1940, 2023.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models. 2023.

Dustin Tran, Jeremiah Zhe Liu, Michael W Dusenberry, Du Phan, Mark Collier, Jie Ren, Kehang Han, Zi Wang, Zelda E Mariet, Huiyi Hu, Neil Band, Tim G. J. Rudner, Zachary Nado, Joost van Amersfoort, Andreas Kirsch, Rodolphe Jenatton, Nithum Thain, E. Kelly Buchanan, Kevin Patrick Murphy, D. Sculley, Yarin Gal, Zoubin Ghahramani, Jasper Snoek, and Balaji Lakshminarayanan. Plex: Towards reliability using pretrained large model extensions. In *First Workshop on Pre-training: Perspectives, Pitfalls, and Paths Forward at ICML 2022*, 2022. URL https://openreview.net/forum?id=6x0gB9gOHFg.

Boxin Wang, Weixin Chen, Hengzhi Pei, Chulin Xie, Mintong Kang, Chenhui Zhang, Chejian Xu, Zidi Xiong, Ritik Dutta, Rylan Schaeffer, Sang T. Truong, Simran Arora, Mantas Mazeika, Dan Hendrycks, Zinan Lin, Yu Cheng, Sanmi Koyejo, Dawn Song, and Bo Li. Decodingtrust: A comprehensive assessment of trustworthiness in gpt models. 2024.

Haomiao Yang, Kunlan Xiang, Mengyu Ge, Hongwei Li, Rongxing Lu, and Shui Yu. A comprehensive overview of backdoor attacks in large language models within communication networks. 2023.

Jingkang Yang, Kaiyang Zhou, Yixuan Li, and Ziwei Liu. Generalized out-of-distribution detection: A survey. *arXiv preprint arXiv:2110.11334*, 2021.

Jinghan Zhang, Shiqi Chen, Junteng Liu, and Junxian He. Composing parameter-efficient modules with arithmetic operations. In *Advances in Neural Information Processing Systems*, 2023.

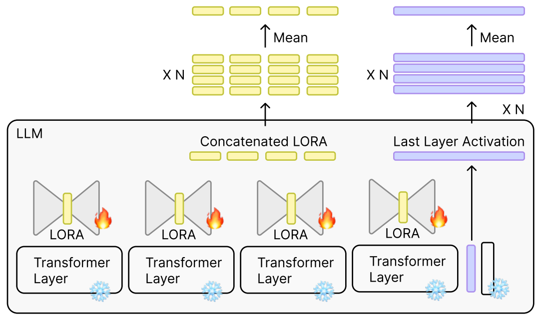

Figure 3: **Embedding generation.** For each N input token, we collect a concatenation of the LoRA embeddings and the last layer activation. The final LoRA embeddings are an average of all the concatenations. For Last Layer Activation embeddings, the same averaging process is applied.

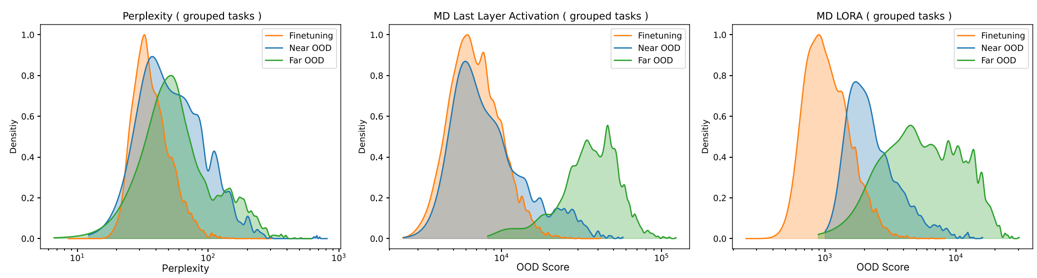

Figure 4: **OOD scores distributions.** Distribution of the OOD scores of MedMCQA validation set (in-distribution, in orange), compared to the MMLU medical subjects (near-OOD, in blue) and non-medical topics (far-OOD, in green) defined in Appendix A.1. From left to right: perplexity scores, Mahalanobis Distance (MD) on the last layer activation and on the LoRA embeddings.

A Appendix

A.1 Datasets (extended version)

We selected three Multiple Choice Question Answering (MCQA) datasets and chose the medical domain as a fine-tuning task, by considering the MedMCQA (Pal et al., 2022) and the PubMedQA (Jin et al., 2019) datasets. Then, we employed the MMLU (Hendrycks et al., 2021) multi-domain dataset to define both near- and far-OOD samples.

The MedMCQA dataset (Pal et al., 2022) contains around 194 000 multiple-choice questions, each with four options, derived from the Indian medical entrance exams (AIIMS and NEET). It includes 21 medical subjects and around 2 400 healthcare related topics.

The PubMedQA dataset (Jin et al., 2019) contains 1 000 expert-annotated and 211 300 artificially generated labelled Question Answering (QA) instances. The task involves generating a yes/no/maybe answers based on a context provided in the form of a PubMed abstract.

The MMLU dataset (Hendrycks et al., 2021) includes questions from 57 different domains. As near-OOD, we selected subtasks related to the medical domain such as "anatomy", "clinical knowledge", "college medicine", "medical genetics", "professional medicine", and "college biology". Conversely, as far-OOD we picked: "professional law", "international law", "business ethics", "computer security", "college computer science", "astronomy", "abstract algebra" and "college chemistry". These subtasks feature multiple-choice questions with four options and a known correct answer.