Leveraging the Regularizing Effect of Mixing Industrial and Open Source Data to Prevent Overfitting of LLM Fine Tuning

Mohamed Salah Jebali,[1] , **Anna Valanzano,**[1] , **Giacomo Veneri**[1] ,
Malathi Murugesan [1] , **Giovanni De Magistris**[1]

[1]Baker Hughes
mohamedsalah.jebali@bakerhughes.com, anna.valanzano@bakerhughes.com

Abstract

Language models have demonstrated important advancements across various natural language processing (NLP) tasks. However, the availability of high-quality and domain-specific data remains a challenge for training these models, particularly in industry-specific applications. In this paper, we propose a methodology to fine-tune a large language model (LLM) using a mixture of private company data and open-source data.

Our empirical investigation reveals that combining private and open-source data during the fine-tuning process leads to superior performance, mitigating the risk of overfitting that can occur when training solely on narrow, domain-specific datasets. We observed that incorporating open-source data alongside the private data helps to reduce the distribution shift between the source and target data, effectively acting as a regularizer and enhancing the model's ability to generalize.

Furthermore, we compare the divergence between the private and open-source datasets with the test loss of the fine-tuned model. Our results suggest a correlation between reduced data divergence and improved model performance, indicating that carefully selecting and curating the dataset mixture can be a crucial step in preventing overfitting and ensuring the model's effective adaptation to industry-specific use cases.

This study provides a practical solution for industry-specific adaptation of LLMs, demonstrating how the strategic blending of private and open-source data can unlock the full potential of these models while addressing critical concerns around data privacy and model reliability in real-world applications.

1 Introduction

The integration of large language models (LLMs) into industry-specific applications has the potential to transform operations across various sectors, notably in the energy industry. They can automate and enhance tasks such as predictive maintenance, regulatory compliance, and customer service. The process of fine-tuning on specific industrial data offers several key advantages:

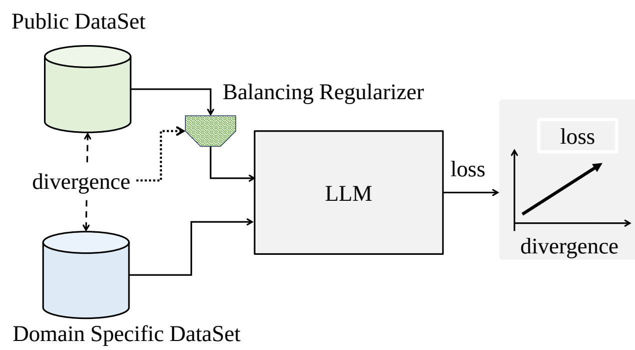


Figure 1: Strategy to use mixture of public dataset and domain specific dataset improves fine tuning of LLM only if divergence between the two dataset is small.

- Cost Savings: Utilizing pre-trained models and fine-tuning them for specific purposes significantly reduces the resources needed for training from scratch. This efficient use of computational power results in considerable cost savings by eliminating the necessity for extensive data collection and expensive hardware procurement for training large models.
- Privacy and Security: Fine-tuning enables organizations to customize pre-trained models using their own datasets, keeping sensitive information within the organization and minimizing exposure to external risks. This localized training approach ensures that data remains under the organization's control, protecting privacy and complying with relevant regulations.
- Tailored Applications: Fine-tuning opens up possibilities for various specialized applications. For example, chatbots trained on customer service data capable of addressing specific product inquiries, or research assistants fine-tuned on scientific literature to assist researchers in their work.

Moreover, supervised fine-tuning on domain-specific data is advantageous for Retrieval Augmented Generation (RAG) systems [Gao *et al.*, 2023] and becomes a fundamental and crucial step for improved alignment techniques such as Reinforcement Learning from Human Feedback (RLFH) [Ouyang *et al.*, 2022] and Direct Preference Optimization (DPO) [Rafailov *et al.*, 2024]. These cutting-edge techniques leverage human feedback and oversight to steer LLMs towards desired behaviors and mitigate potential risks, but their effectiveness heavily relies on the initial domain-specific fine-tuning step.

Deploying LLMs, including RAG systems, in sectors like energy, where data are not only domain-specific but also highly confidential, presents significant challenges [Cuconasu *et al.*, 2024]. Standard fine-tuning practices on such narrow datasets, especially when the data is substantially out-of-distribution compared to the training corpus used in the initial model training, can lead to severe overfitting. This overfitting hinders the model's ability to generalize effectively, undermining the potential benefits of subsequent alignment techniques and ultimately limiting the model's practical utility in industrial settings.

This study explores an innovative approach to fine-tuning LLMs by using a mixed dataset of open-source and proprietary domain-specific data. Our empirical results indicate that this method significantly mitigates overfitting, enabling the models to generalize better across a broader range of tasks without compromising data confidentiality. Thus, we provide a practical solution for industry-specific adaptation of LLMs, ensuring enhanced performance while adhering to stringent privacy standards. Crucially, this robust domain-specific fine-tuning sets the stage for effective application of advanced alignment techniques, unlocking the full potential of these models in industrial contexts while addressing critical concerns around safety and reliability.

Additionally, our approach enriches the understanding of domain adaption in machine learning. In fact, fine-tuning LLMs on a mixed dataset effectively addresses a domain adaptation problem where the goal is to adapt the model to fit a mixture of distributions. The integration of open-source data with the domain specific data reduces the divergence between the target and source distributions. This strategic blending facilitates a more robust training process, helping the model to better adapt and thus enhancing resilience against overfitting.

2 Related works

The integration of LLMs into domain-specific applications is a crucial research challenge. For example, the RAG system is a common strategy to bridge the gap between pre-trained LLMs and industry-specific data [Cuconasu *et al.*, 2024; Lewis *et al.*, 2020]. The RAG approach combines the generative capabilities of LLMs with information retrieval, allowing the model to access relevant documents and incorporate that contextual information into its outputs.

Recent studies, such as [Cuconasu *et al.*, 2024], demonstrate the promising results of RAG systems. However, their performance may not be optimal, especially when applied to domain-specific data. In this case, fine-tuning can improve the performance of RAG systems in these tasks, or even outperform it.

Beyond the RAG system, researchers have explored other techniques to fine-tune and adapt LLMs to specialized domains.[Arora *et al.*, 2023] proposed a simplified prompting strategy that generates multiple questions and aggregates the most reliable responses, aiming to improve the LLM's performance on domain-specific tasks.

Recognizing the limitations of pre-trained LLMs in capturing domain-specific knowledge, researchers have explored hybrid approaches that combine the flexibility of LLMs with enterprise-specific knowledge graphs. [Baldazzi *et al.*, 2023] demonstrated how this integration of LLMs and ontological reasoning can effectively capture and augment domain knowledge, enhancing the model's capabilities in industry-specific applications.

Fine-tuning LLMs on large additional text corpora has also been shown to be effective in improving performance on various NLP tasks. For example, the FinBERT model [Liu *et al.*, 2021] was fine-tuned on a financial domain-specific dataset, leading to improved results on finance-related tasks. Similarly, [Xia *et al.*, 2024] fine-tuned LLMs using a manufacturing-domain corpus to better adapt the models to the nuances of the manufacturing field.

However, the lack of available domain-specific training data remains a significant challenge in many specialized industries. [Saxena *et al.*, 2024] reported difficulties in finding appropriate datasets for their domain-specific applications, highlighting the need for novel approaches to address this data scarcity.

To tackle the data scarcity issue, researchers have explored parameter-efficient fine-tuning techniques, where only a few external parameters are fine-tuned instead of the entire LLM [Hu *et al.*, 2023]. Additionally, data cleaning and curation approaches, such as the one proposed by Lin et al. [Lin *et al.*, 2024], have shown promise in improving the fine-tuning performance of LLMs.

Finally, a comprehensive study by [Zhang *et al.*, 2024] explored the impact of different scaling factors on the fine-tuning performance of LLMs, emphasizing the data- and task-dependent nature of these fine-tuning methods.

Overall, the existing literature underscores the importance of adapting LLMs to domain-specific applications, while also highlighting the challenges posed by data availability and distribution shifts. Our work builds upon these insights and proposes a novel approach to fine-tune LLMs using a strategic mixture of domain-specific and open-source data, with the aim of mitigating the risk of overfitting and enhancing the model's generalization capabilities.

3 Proposed Method

The key challenge we aim to address is the tendency of large language models (LLMs) to overfit when fine-tuned on limited, domain-specific data. This phenomenon can lead to poor generalization, especially on out-of-distribution samples. To mitigate this issue, we propose a novel fine-tuning approach that leverages a mixture of data distributions.

Our proposed method is motivated by the principles of domain adaptation. When fine-tuning an LLM, the target data distribution (e.g., a company's private data) often differs sig-

nificantly from the original distribution the model was trained on (e.g., general web data). This domain shift can exacerbate the overfitting problem, as the model struggles to generalize from the source distribution to the target distribution.
To address this, we fine-tune the LLM on a mixture of data sources, combining the company's private data with publicly available, related data (e.g., open-source documents). By exposing the model to a more diverse set of data during fine-tuning, we aim to reduce the discrepancy between the source and target distributions, thereby improving the model's ability to generalize.
Specifically, our fine-tuning approach involves training the LLM on a balanced mixture of private company data and relevant open-source data. The intuition is that the open-source data, while not identical to the target domain, can help the model learn more robust representations that are less sensitive to the idiosyncrasies of the private data alone. By reducing the domain shift between the fine-tuning data and the original model training data, we expect to enhance the model's performance on a wide range of samples from the target domain.
We evaluate the effectiveness of our approach on text generation tasks, such as question-answering, and find that fine-tuning on a mixture of private and open-source data indeed helps reduce overfitting compared to fine-tuning on private data alone. This is a promising result, as direct access to the original data used to train the LLM is often unavailable, making our approach a practical solution for domain-specific fine-tuning.

3.1 Problem Statement

We consider a supervised learning task where the datasets are defined as Cartesian products between features spaces $\mathcal{X}$ and label spaces $\mathcal{Y}$. We consider different datasets, each as a collection of points generated by an underlying distribution, as follows:

- Source distribution $\mathcal{S}$ over $\mathcal{X} \times \mathcal{Y}$, where the collection of points $S = \{(x_i, y_i) \in (\mathcal{X} \times \mathcal{Y})\}_{i=1}^{m}$ are drawn independently and identical distributed (i.i.d.) from $\mathcal{S}$. It indicates the set used to pre-train the foundation LLM.
- Domain-specific distribution $\mathcal{D}$, such that the points in the dataset $D = \{(x_i, y_i) \in (\mathcal{X} \times \mathcal{Y})\}_{i=1}^{n}$, drawn i.i.d. from $\mathcal{D}$, represent the proprietary data.
- Open-source distribution $\mathcal{O}$, assumed similar to $\mathcal{S}$, such that the points in $O = \{(x_i, y_i) \in (\mathcal{X} \times \mathcal{Y})\}_{i=1}^{n}$ are drawn i.i.d. from $\mathcal{O}$.
- Target distribution $\mathcal{T}$ over $\mathcal{X} \times \mathcal{Y}$, a mixture of $\mathcal{D}$ and $\mathcal{O}$:
$$\mathcal{T} = \alpha\mathcal{D} + (1 - \alpha)\mathcal{O}, \tag{1}$$
where $\alpha \in [0, 1]$ is the mixing ratio. In this case, the collection of points in $T = \{(x_i, y_i) \in (\mathcal{X} \times \mathcal{Y})\}_{i=1}^{n}$, are drawn from $\mathcal{T}$.

Let $\mathcal{H}$ be an hypothesis space such that the hypothesis $h \in \mathcal{H}$ is a function $h : \mathcal{X} \to \mathcal{Y}$. Consider $h_{\mathcal{S}}$ the hypothesis learnt by minimizing the expected loss over $\mathcal{S}$ (corresponding to the LLM to fine tune in our case). The hypothesis $h_{\mathcal{S}}$ is the solution of the empirical risk minimization (ERM) problem over the source distribution, given by:
$$h_{\mathcal{S}} = \arg\min_{h\in\mathcal{H}} \mathbb{E}_{(x,y)\sim\mathcal{S}}[\ell(h(x), y)]. \tag{2}$$
The ℓ term is a loss function which in our case is the cross-entropy loss, defined as:
$$\ell(h(x), y) = -\sum_{c=1}^{C} y_c \log h_c(x), \tag{3}$$
where y is one-hot encoded label vector, C is the number of classes, and $h_c(x)$ is the predicted probability of class c for input x. In our case, the domain adaptation problem involves adapting the h_S function to minimize the loss over the target distribution $\mathcal{T}$. The learnt hypothesis $h_{\mathcal{T}}$ is then the solution of the following ERM problem:
$$h_{\mathcal{T}} = \arg\min_{h\in\mathcal{H}|_{h_{\mathcal{S}}}} \mathbb{E}_{(x,y)\sim\mathcal{T}}[\ell(h(x), y)], \tag{4}$$
where $\mathcal{H} \; |_{h_{\mathcal{S}}}$ means that the exploration of the hypothesis space is initialized at the the point $h_{\mathcal{S}}$.

Adapting the hypothesis $h_{\mathcal{S}}$ to the distribution $\mathcal{T}$ involves addressing the discrepancy between the distributions $\mathcal{S}$ and $\mathcal{D}$. In fact, even though we are interested in applying the learnt hypothesis $h_{\mathcal{T}}$ on the points $(\mathbf{x}, \mathbf{y}) \sim \mathcal{D}$, fitting $h_{\mathcal{S}}$ directly to $\mathcal{D}$ can lead to overfitting due to the divergence between $\mathcal{S}$ and $\mathcal{D}$. By performing fine-tuning on the target distribution $\mathcal{T}$, which is a mixture of distributions $\mathcal{O}$ and $\mathcal{D}$, we aim to leverage the similarity between $\mathcal{S}$ and $\mathcal{O}$ to mitigate the outlying nature of $\mathcal{D}$ more effectively than $\mathcal{D}$ alone would do. Based on empirical observations, we find that:
$$\text{JS-Divergence}(\mathcal{D}, \mathcal{O}) > \text{JS-Divergence}(\mathcal{D}, \mathcal{T}), \tag{5}$$
and under the assumption that $\mathcal{S}$ and $\mathcal{O}$ are similar, we claim that:
$$\text{JS-Divergence}(\mathcal{D}, \mathcal{S}) \geq \text{JS-Divergence}(\mathcal{D}, \mathcal{T}). \tag{6}$$
This inequality explains the effectiveness of fine tuning over a mixture of distribution between the domain specific data and opensource data which are assumed similar to the source one. By reducing the overall divergence between the training distribution $\mathcal{T}$ and the original source distribution $\mathcal{S}$, it facilitates a smoother adaptation of the model $h_{\mathcal{S}}$, enhancing its ability to generalize from training to real-world data.
The divergence used is the *Jensen-Shannon* divergence (JS-Divergence) [Lin, 1991], which is a popular measure of distance between two probability distributions. It is defined as the average of the Kullback-Leibler divergences [Kullback and Leibler, 1951] of each distribution to the mean of both distributions, providing a symmetric and bounded measure. Mathematically, it is given by:
$$\text{JS-Divergence}(P, Q) = \frac{1}{2}\text{KL}(P \parallel M) + \frac{1}{2}\text{KL}(Q \parallel M), \tag{7}$$
where P and Q are the two distributions, $M = \frac{1}{2}(P + Q)$ is their mean, and KL denotes the Kullback-Leibler divergence. This measure is particularly useful in scenarios where the distributions may not overlap completely, as it remains finite under such conditions. The properties of being symmetric and bounded between 0 and 1 make JS-Divergence a robust tool for quantifying distributional discrepancies, especially in domain adaptation scenarios.

3.2 Generalization and Theoretical Bounds

The generalization performance of $h_{\mathcal{T}}$ is influenced not only by the number of samples used during training but also by the divergence between the mixed training distribution $\mathcal{T}$ and the target distribution $\mathcal{D}$. Building upon the foundational work by [Mansour *et al.*, 2009] we can provide a generalization error of a hypothesis bounded by the Rademacher Complexity ($\mathcal{R}$) of a hypothesis space $\mathcal{H}$ [Shalev-Shwartz and Ben-David, 2014; Mohri *et al.*, 2018], and the *Discrepancy* introduced in [Mansour *et al.*, 2009; Mohri and Muñoz Medina, 2012], which is a type of $\mathcal{H}\Delta\mathcal{H} - divergence$ between two distributions. With probability at least $1-\delta$ we have:

$$\mathcal{L}_{\mathcal{D}}(\hat{h}^*_{\mathcal{T}}) - \mathcal{L}_{\mathcal{D}}(h^*_{\mathcal{D}}) \leq 4\mathcal{R}_{\mathcal{T},n}(\mathcal{H}\mid_{h_{\mathcal{S}}}) + d_{\mathcal{H}|_{h_{\mathcal{S}}}\Delta\mathcal{H}|_{h_{\mathcal{S}}}}(\mathcal{T},\mathcal{D}) + d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{T},\mathcal{D}), \quad (8)$$

which becomes

$$\mathcal{L}_{\mathcal{D}}(\hat{h}^*_{\mathcal{T}}) - \mathcal{L}_{\mathcal{D}}(h^*_{\mathcal{D}}) \leq \mathcal{O}\left(\frac{\sqrt{d+\log 1/\delta}}{\sqrt{n}}\right) + d_{\mathcal{H}|_{h_{\mathcal{S}}}\Delta\mathcal{H}|_{h_{\mathcal{S}}}}(\mathcal{T},\mathcal{D}) + d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{T},\mathcal{D}), \quad (9)$$

where $\mathcal{L}_{\mathcal{D}(h)} = \mathbb{E}_{(x,y)\sim\mathcal{D}}[\ell(h(x),y)]$, $\hat{h}^*_{\mathcal{T}}$ is the empirical risk minimizer over the mixture distribution $\mathcal{T}$, $h^*_{\mathcal{D}}$ is the true minimizer over $\mathcal{D}$, $\mathcal{H} \mid h_{\mathcal{S}}$ represents the hypothesis space constrained by source domain knowledge, $d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{T},\mathcal{D})$ is the *Discrepancy* and n is the number of training points. This formulation supports our methodology by highlight how mixed-data fine-tuning serves as an effective domain adaptation strategy. By blending domain-specific and open-source data, we aim to craft a hypothesis that not only fits well to the training data but also exhibits robust generalization across diverse real-world applications. In practice, while the theoretical model uses $\mathcal{H}\Delta\mathcal{H} - divergence$ for its rigorous properties, our empirical evaluation employs JS-Divergence due to its computational efficiency and its practical effectiveness in capturing the essential aspect of distributions shifts.

4 Experiments

4.1 Overview

In this study, we fine-tuned an open-source LLM, specifically the mistral-7b-Instruct-v02 [Jiang *et al.*, 2023], utilizing a combination of domain-specific data and open-source data. The fine-tuning process employed QLORA (Quantized Low-Rank Adaptation) [Dettmers *et al.*, 2024], which is designed to reduce the model complexity and size, thereby enabling efficient fine-tuning. We explored various mixtures of private company data and open-source data from public domain to construct our mixed training set to use during the fine-tuning phase. This approach allowed us to examine the impact of different data proportions on the model's performance. The evaluation was consistently carried out on a held-out test set comprised exclusively of private company data, corresponding the domain specific data. We performed experiments focusing on a text-generation tasks, specifically for question-and-answer scenarios.

Moreover, to assess the distribution shift between the datasets, we computed the Jensen-Shannon divergence, providing a quantifiable measure of dataset similarity.

4.2 Dataset

The domain-specific dataset was derived from technical manuals and documentation related to the manufacturing, testing, and assembly of industrial assets within the energy sector. This dataset is highly specialized, containing information that has likely never been exposed to the public domain, thus representing a set of highly out-of-distribution samples with respect the source distribution the original LLM was trained on. Some examples of question and answering reported below:

- "Q": *"What are the traditional monitored parameters for oil?"*
- "A": *"The traditional monitored parameters for oils are viscosity and oxidation."*
- "Q": *"Why should the traditional monitored parameters be used for turbine ?"*
- "A": *"These parameters are used to trend and predict the remaining useful life of the asset, helping to prevent operational problems from developing due to the condition of external environment."*

The dataset comprises approximately 3,000 samples of specific question-and-answer texts, meticulously curated to reflect the unique context of the energy industry.
For the open-source dataset, we utilized the Alpaca dataset [Taori *et al.*, 2023], publicly available on Hugging Face and comprising 52,000 instructions and demonstrations generated by OpenAI's text-davinci-003 engine. This dataset is commonly used for conducting instructional tuning on language model to enhance their ability to follow directions more precisely [Jiang *et al.*, 2023]. For the purposes of this experiments, we selected 3,000 samples from the Alpaca dataset to match the number of points used in the domain-specific dataset.
The two datasets were then blended in varying proportions, with subsequent testing conducted solely on the domain-specific dataset to isolate the effects of the mixed training data.

4.3 QLORA Technique

The QLORA technique is a variant of the Low-Rank Adaptation (LORA) methodology [Hu *et al.*, 2021] which involves modifying the parameterization of the neural network by introducing low-rank matrices that approximate the update to the weights during training. This method significantly reduces the number of trainable parameters, which minimizes memory usage and computational demands, making it suitable for fine-tuning large models on specialized datasets. By applying QLORA, we aim to maintain or even enhance the model's performance while mitigating the risk of overfitting to the highly specialized domain data.

4.4 Experimental Setup

The experiments were conducted on high-performance computing environment equipped with DGX NVIDIA 8xA100

GPUs, with with 40 GB of memory. This setup ensured efficient handling of the large models and extensive data involved. We implemented the experiments using Python, leveraging libraries such as Transformers [Wolf *et al.*, 2019] and PyTorch [Paszke *et al.*, 2019], which provide a robust frameworks for training and manipulating large-scale language models.

For our experiments, QLORA was applied to all linear layers of the model to efficiently adapt the pre-trained weights with minimal computational overhead. We set the low-rank adaption factor α equal to 8 and the rank r equal to 8 as well. We employed Adam optimizer [Kingma and Ba, 2014] and a cosine learning rate scheduler starting from an initial learning rate of 2×10^{-4}, which adapts the learning rate cyclically based on epoch count. A weight decay of 0.1 was applied to prevent overfitting, alongside a dropout rate of 0.05. Models were trained for up to 150 epochs, with early stopping employed to halt the training if the validation loss ceased to improve, ensuring efficient use of the computational resources.

Each fine-tuning experiment varied the ratio of domain-specific to open-source data to identify the optimal conditions for model performance. We meticulously tracked the model's behavior under each configuration to asses how variations in data mixture affect the learning outcomes. The performance metrics and detailed analysis of these experiments will be presented in the results section.

5 Results and Analysis

5.1 Jensen-Shannon Divergence Calculation

To evaluate the effectiveness of our mixed-data fine-tuning approach, we quantified the distribution shifts between different dataset configurations using the Jensen-Shannon divergence (JS-Divergence). This metric was instrumental in assessing how well the mixed dataset aligns with both the domain-specific and open-source datasets, providing a basis for understanding the impact of our data blending strategy on model generalization.

Dataset 1	Dataset 2	Discrepancy
Domain Specific	Open Source	0.62
Domain Specific	Mixed Dataset	0.37

Table 1: Jensen-Shannon divergence values quantifying discrepancies between different dataset configurations. Lower values indicate greater similarity between datasets.

The results, as summarized in Table 1, reveal significant insights into the dynamics of dataset integration. The divergence between the domain-specific and open-source dataset was notably high 0.62, indicating substantial differences in their distributions. By introducing a mixed dataset, the divergence from the domain-specific dataset decrease to 0.37. This reduction in divergence suggests that mixing the data has effectively made the distribution of the training data more representative.

These findings support the hypothesis that a well-blended training dataset can bridge the gap between diverse data sources, thus mitigating potential overfitting issues when the model is applied exclusively to domain-specific data. The reduced Jensen-Shannon divergence value indicates that the mixed dataset shares more characteristics with both parent sets, potentially leading to improved generalization across varied data domains. Given the impact of dataset mixing on reducing discrepancy we aim to observe a mitigation of the overfitting effect.

5.2 Results of Fine-Tuning

We conducted several experiments where we modified the composition of the training set while keeping the held-out test set unchanged (solely composed of proprietary domain specific data). Initially, we focused on fine-tuning the model using the proprietary data of our company. Figure 2a illustrates that when the model is trained solely on domain-specific data, there is a significant deviation between the test loss and the training loss. This discrepancy indicates that the model is overfitting to the domain data. As a result, we proceeded with an additional experiment, fine-tuning the model exclusively on open-source data and evaluating its performance on the same domain-specific test set. The plots in Figure 2b once again reveal a divergence between the test and training losses, confirming the occurrence of overfitting in this case as well. These findings are consistent with the results obtained from the JS-Divergence analysis, which highlights the substantial dissimilarity between the open-source data and the domain-specific data.

We then proceeded on performing fine-tuning of the model over a mixed dataset comprising both open-source and domain-specific data, varying the mixing weight α as described in Equation 1. We tested different levels of mixture, ranging from $\alpha = 1$ (using only domain-specific data) to $\alpha \in [0.2, 0.5, 0.8]$ (mixing different proportions of domain and open-source data), and finally $\alpha = 0.0$ (fine-tuning solely on open-source data). For each training run, we evaluated the model on the same held-out test set composed exclusively of domain-specific data, as our primary interest was to improve the model's performance on the specific data domain.

Figure 3b illustrates the results obtained on the test set. Surprisingly, we observe that fine-tuning over a mixture dataset effectively mitigated the overfitting phenomenon. Indeed, all the test loss curves obtained by the models trained on the mixture dataset were lower than those obtained differently, and they tended to converge within the training loss, while the others diverged, as we can observe by comparing the test loss curves and the training ones in Figure 3a. Moreover, the best result was achieved using $\alpha = 0.5$, indicating that a balanced mix between open and private data led to a better mitigation effect.

We hypothesize that the regularization effect was achieved by reducing the distribution shift between the mixed data and the domain-specific data. This is supported by the observation that the Jensen-Shannon divergence between the open-source data and the domain-specific data is higher compared to the divergence between the mixed data and the domain-specific data. It is worth noting that there exists a positive correlation between reducing the divergence among the datasets and the mitigating effect observed. This implies that supervised fine-tuning on a mixed dataset serves as a regularization technique. Our empirical findings demonstrate that, particularly during fine-tuning on out-of-distribution data such as domain-specific data, the integration of datasets with reduced discrepancy values can help prevent the occurrence of overfitting.

Effect of Increasing Dataset Size

To further investigate the benefits of fine-tuning on a mixed dataset, we conducted an additional experiment where we doubled the size of the training data while maintaining the optimal mixing ratio of $\alpha = 0.5$ between open-source and domain-specific data. We then evaluated the fine-tuned model on the held-out domain-specific test set, and the results are depicted in Figure 4.

Interestingly, the model trained on the larger mixed dataset did not exhibit overfitting, and more importantly, it achieved better convergence compared to the previous experiments with a smaller dataset. This observation not only reinforces the notion that mixing datasets can effectively prevent overfitting, but it also suggests that increasing the number of training samples can lead to further performance improvements.

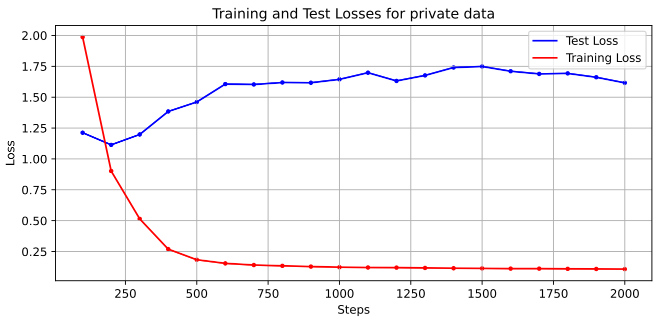
The training loss curve and the test loss curve in Figure 4 demonstrate a desirable trend: the test loss closely follows the training loss, indicating that the model generalizes well to the unseen domain-specific data. This behavior contrasts with the overfitting observed when training solely on the domain-specific dataset or on the open-source dataset alone, where the test loss diverged from the training loss.

Our findings suggest that the regularization effect introduced by mixing datasets can be amplified by increasing the overall size of the training data. The larger and more diverse the combined dataset, the more effective the regularization becomes, leading to better generalization and prevention of overfitting. This phenomenon aligns with the well-established principle in machine learning that larger and more diverse training datasets can help models capture broader patterns and improve their ability to generalize to unseen data distributions.

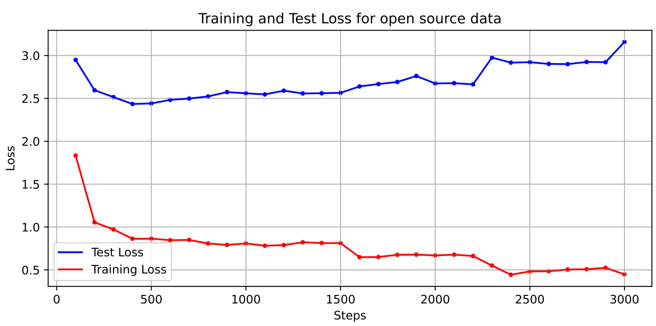
In summary, this experiment not only corroborates the effectiveness of mixing datasets in mitigating overfitting but also highlights the potential benefits of increasing the overall training data size when working with domain-specific or out-of-distribution data. By combining these two strategies, we can leverage the regularizing effect of data mixing while also benefiting from the increased diversity and information provided by larger datasets.

6 Conclusions

This study presents a novel approach to fine-tuning LLMs for industry-specific applications by leveraging a strategic combination of private company data and open-source data. Our findings demonstrate that this mixed-dataset fine-tuning methodology can effectively mitigate the risk of overfitting that often arises when training LLMs solely on narrow, domain-specific datasets. The key contributions of this work include empirical evidence that incorporating open-source data alongside private company data during the fine-tuning process leads to superior performance compared to using private data alone. This blended approach helps to reduce the distribution shift between the training and test data, acting

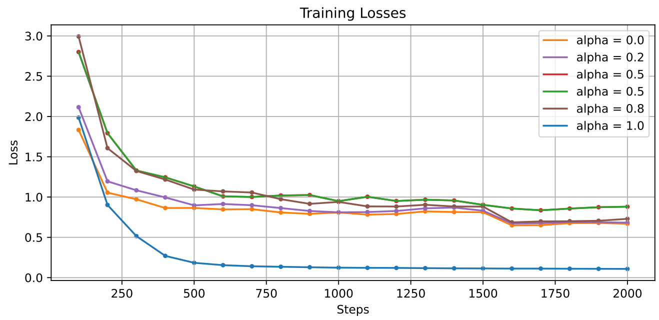


(a) Training loss curve during fine-tuning on domain-specific (private) data, and the corresponding test loss curve evaluated on a held-out domain-specific dataset.

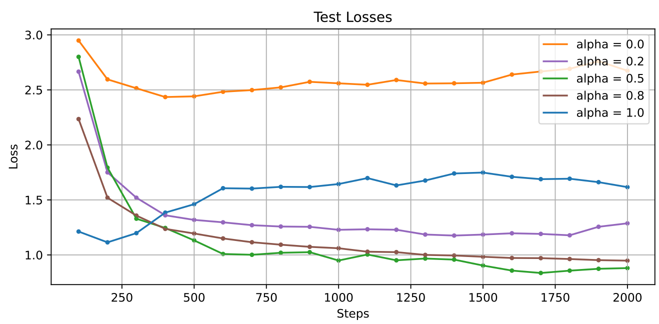


(b) An overfitting case: training loss curve during fine-tuning on open-source data, and the test loss curve evaluated on a domain-specific dataset, demonstrating a divergence between the two losses.

Figure 2: Comparison of fine-tuning performance on domain-specific (private) data and open-source data. **(a)** Fine-tuning on domain-specific data only, showing the training loss curve on the domain-specific data and the corresponding test loss curve evaluated on a held-out domain-specific dataset. **(b)** Fine-tuning on open-source data only evaluated on a held-out domain-specific dataset, illustrating an overfitting case where the test loss on the domain-specific dataset diverges from the training loss on the open-source data. The results highlight the challenges of domain shift and the need for regularization techniques when fine-tuning on out-of-distribution data.



(a) The figure depicts the training losses of the model during the fine-tuning process using datasets that combine open-source and domain-specific data. The legend denotes the different mixing weights used: $\alpha = 1$ corresponds to the utilization of exclusively domain-specific data, $\alpha = 0$ represents the exclusive use of open-source data, and $\alpha \in [0.2, 0.5, 0.8]$ indicates varying proportions of domain and open-source data.



(b) The figure illustrates the test losses of the model throughout the fine-tuning process, utilizing training datasets that combine open-source and domain-specific data. The alpha values represent the proportions of private and open-source data used for training, as explained in the caption of Figure 3a.

Figure 3: Comparison of training and test losses during fine-tuning on mixed datasets comprising open-source and domain-specific data and tested. **(a)** Training losses for models fine-tuned with varying mixing weights (α) between the two data sources. **(b)** Test losses evaluated on a held-out domain-specific dataset for the same fine-tuned models. The results demonstrate the effectiveness of fine-tuning on a mixed dataset in mitigating overfitting, with the optimal performance achieved at $\alpha = 0.5$, indicating a balanced mix between open-source and domain-specific data.

Figure 4: Fine-tuning performance on a larger mixed dataset with mixing ratio $\alpha = 0.5$. The training and test loss curves on the held-out domain-specific dataset exhibit desirable convergence, with the test loss closely following the training loss, contrasting the overfitting observed when fine-tuning on single datasets. The results suggest that increasing the overall training data size while maintaining an optimal mixing ratio can amplify the regularization effect of dataset mixing.

as a regularizer and enhancing the LLM's ability to generalize, by mitigating the risk of overfitting. Additionally, our analysis of the Jensen-Shannon divergence between the private and open-source datasets reveals a correlation between reduced data divergence and improved model performance, underscoring the importance of carefully curating the dataset mixture to bridge the gap between diverse data sources and prevent overfitting. This work also provides a practical solution for industry-specific adaptation of LLMs, demonstrating how the strategic combination of private and open-source data can unlock the full potential of these models while addressing critical concerns around data privacy and model reliability in real-world applications. By addressing the challenges of data scarcity and domain shift in industry-specific settings, this study paves the way for more effective deployment of LLMs in sectors such as energy, where data confidentiality and model performance are of paramount importance. Overall, the principles and insights gleaned from this work can also inform the broader field of domain adaptation in machine learning, enriching our understanding of how to leverage diverse data sources to enhance model generalization.

6.1 Future works

In future research, our investigation will focus on assessing the effectiveness of the proposed fine-tuning methodology on other domain-specific datasets. Given the limited availability of additional data, it becomes imperative to expand the scope of analysis in order to evaluate the generalizability and robustness of the approach across various domains. It is important to emphasize the significance of supervised fine-tuning on domain-specific data, as it serves as a fundamental and crucial step for improved alignment techniques such as RLFH and DPO. By incorporating these advanced techniques, we can further enhance the performance and alignment of the model. Therefore, in our ongoing fine-tuning process, exploration DPO emerges as a valuable next step.

7 Acknowledgement

Thanks also to valuable contribution of Athira Puthanveetil Madathil, Sachin Shetty, Luca Cuomo, Andrea Panizza e Alessandro Bettini. Experiments have been conducted thanks to internal and external funding Regione Calabria Contribution and R&T Program.

References

[Arora *et al.*, 2023] Simran Arora, Avanika Narayan, Mayee F Chen, Laurel Orr, Neel Guha, Kush Bhatia, Ines Chami, and Christopher Re. Ask me anything: A simple strategy for prompting language models. In *The Eleventh International Conference on Learning Representations*, 2023.

[Baldazzi *et al.*, 2023] Teodoro Baldazzi, Luigi Bellomarini, Stefano Ceri, Andrea Colombo, Andrea Gentili, and Emanuel Sallinger. Fine-tuning large enterprise language models via ontological reasoning. In Anna Fensel, Ana Ozaki, Dumitru Roman, and Ahmet Soylu, editors, *Rules and Reasoning*, pages 86–94, Cham, 2023. Springer Nature Switzerland.

[Cuconasu *et al.*, 2024] Florin Cuconasu, Giovanni Trappolini, Federico Siciliano, Simone Filice, Cesare Campagnano, Yoelle Maarek, Nicola Tonellotto, and Fabrizio Silvestri. The power of noise: Redefining retrieval for rag systems, 2024.

[Dettmers *et al.*, 2024] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms. *Advances in Neural Information Processing Systems*, 36, 2024.

[Gao *et al.*, 2023] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2023.

[Hu *et al.*, 2021] 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. *arXiv preprint arXiv:2106.09685*, 2021.

[Hu *et al.*, 2023] Zhiqiang Hu, Lei Wang, Yihuai Lan, Wanyu Xu, Ee-Peng Lim, Lidong Bing, Xing Xu, Soujanya Poria, and Roy Ka-Wei Lee. Llm-adapters: An adapter family for parameter-efficient fine-tuning of large language models, 2023.

[Jiang *et al.*, 2023] Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.

[Kingma and Ba, 2014] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[Kullback and Leibler, 1951] Solomon Kullback and Richard A Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951.

[Lewis *et al.*, 2020] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474, 2020.

[Lin *et al.*, 2024] Xinyu Lin, Wenjie Wang, Yongqi Li, Shuo Yang, Fuli Feng, Yinwei Wei, and Tat seng Chua. Data-efficient fine-tuning for llm-based recommendation. *ArXiv*, abs/2401.17197, 2024.

[Lin, 1991] Jianhua Lin. Divergence measures based on the shannon entropy. *IEEE Transactions on Information theory*, 37(1):145–151, 1991.

[Liu *et al.*, 2021] Zhuang Liu, Degen Huang, Kaiyu Huang, Zhuang Li, and Jun Zhao. Finbert: A pre-trained financial language representation model for financial text mining. In *Proceedings of the twenty-ninth international conference on international joint conferences on artificial intelligence*, pages 4513–4519, 2021.

[Mansour *et al.*, 2009] Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. Domain adaptation: Learning bounds and algorithms. *arXiv preprint arXiv:0902.3430*, 2009.

[Mohri and Muñoz Medina, 2012] Mehryar Mohri and Andres Muñoz Medina. New analysis and algorithm for learning with drifting distributions. In *Algorithmic Learning Theory: 23rd International Conference, ALT 2012, Lyon, France, October 29-31, 2012. Proceedings 23*, pages 124–138. Springer, 2012.

[Mohri *et al.*, 2018] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. MIT press, 2018.

[Ouyang *et al.*, 2022] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

[Paszke *et al.*, 2019] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

[Rafailov *et al.*, 2024] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36, 2024.

[Saxena *et al.*, 2024] Shreya Saxena, Siva Prasad, Muneeswaran I, Advaith Shankar, Varun V, Saisubramaniam Gopalakrishnan, and Vishal Vaddina. Automated tailoring of large language models for industry-specific downstream tasks. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, WSDM '24, page 1184–1185, New York, NY, USA, 2024. Association for Computing Machinery.

[Shalev-Shwartz and Ben-David, 2014] Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

[Taori *et al.*, 2023] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca, 2023.

[Wolf *et al.*, 2019] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*, 2019.

[Xia *et al.*, 2024] Liqiao Xia, Chengxi Li, Canbin Zhang, Shimin Liu, and Pai Zheng. Leveraging error-assisted fine-tuning large language models for manufacturing excellence. *Robotics and Computer-Integrated Manufacturing*, 88:102728, 2024.

[Zhang *et al.*, 2024] Biao Zhang, Zhongtao Liu, Colin Cherry, and Orhan Firat. When scaling meets LLM finetuning: The effect of data, model and finetuning method. In *The Twelfth International Conference on Learning Representations*, 2024.