Multilingual Federated Low-Rank Adaptation for Collaborative Content Anomaly Detection across Multilingual Social Media Participants

Anonymous ACL submission

Abstract

Recently,the rapid development of multilingual social media platforms (SNS) exacerbates multilinguality challenges in SNS content anomaly detection due to data islands and linguistic imbalance. While federated learning (FL) and parameter-efficient fine-tuning (PEFT) offer potential solutions in most cases, when every client is multilingual, existing solutions struggle with multilingual heterogeneity: 1) entangled language-specific knowledge during aggregation, 2) noise from minority languages, and 3) unstable cross-platform collaboration. Based on the asymmetric nature of LoRA, we propose MuLA-F, a multilingual Federated LoRA introducing SVD-based language-specific disentanglement of LoRA blocks and a local orthogonal tuning strategy. Evaluations across 3 SNS content anomaly detection tasks demonstrate MuLA-F's superiority in multilingual performance while reducing multilingual knowledge conflicts and communication rounds.[1]

1 Introduction

As social media platforms (SNS) proliferate in recent years, coupled with escalating global unrest and instability, anomalous content (Geissler et al., 2023; Houston et al., 2015; Savage et al., 2014) spreads with alarming speed and magnitude across a vast network of vulnerable social media users (Chen et al., 2013; Mossie and Wang, 2020).

How can we safeguard the online ecosystem from the toxic contamination of fake news, hate speech, and other harmful content (Röttger et al., 2021; Wu et al., 2019)? How can we ensure that distant cries— e.g. those under crisis or depressions—are not drowned out amidst the noise (Zhang et al., 2019; Alam et al., 2021)? In response to these pressing concerns, academia has consistently pursued advancements in developing more effective content anomaly detectors (Aïmeur et al., 2023; Alam et al., 2021) for SNS online content. More recently, with the surge in applications of large language models (LLMs), numerous innovative works (Lei et al., 2025; Nan et al., 2024) based on Parameter-Efficient Fine-Tuning (PEFT) are proposed, achieving notable breakthroughs in SNS content anomaly detection.

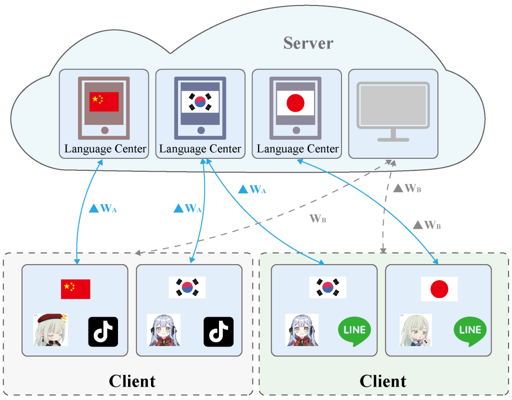

Figure 1: An illustration of addressing multilingual SNS content anomaly detection using Federated Learning.

However, as SNS continue to decentralize and show their inherent transcultural nature, and as user interest in cross-border communication grows, individuals speaking various native languages are flocking to popular or trending platforms (Kim et al., 2014). The influx of users speaking different native languages sparks a profound increase in linguistic diversity online. Consequently, SNS content anomaly detectors are now contending with the multilingual curse (Pfeiffer et al., 2022). Specifically, for a single data holder (e.g., an SNS operator's data storage center or an edged device), the dominant language among its users often prevails in usage proportion, while the data available in minority languages are insufficient to support the multilingual local training necessary for an effective detector against the abnormal content in these

[1]Our working code and data examples are available in supplementary materials (As our higher-ups do not allow us to release our code on Github before acceptance).

minor languages (Guo et al., 2024c; Weller et al., 2022). As a result, detectors trained locally (i.e. by a single SNS) could struggle to detect content anomaly in posts in minor languages—e.g. the APP “Little Red Note” fails to effectively filter out toxic remarks (even in English) posted by TikTok refugees (Press, 2025).

Confronted with this challenge and the heavy costs of acquiring multilingual in-domain annotated data (Wang et al., 2022), we argue that the growing multilingual user base gives rise to data islands in SNS content anomaly detection—the issue that lies at the heart of this paper's focus.

Empirically, previous studies tackle similar challenges by leveraging federated learning (FL) techniques, enabling the multilingual collaborative training of local models with multilingual or monolingual local datasets across diverse organizations and data sources. From a theoretical perspective, FL can effectively mitigate these challenges (Wang et al., 2022; Zhao et al., 2024). This is because, from a global standpoint, if data holders can collaborate, the completeness of multilingual data can be significantly enhanced (Yang et al., 2023), as each major language community gravitates toward its preferred SNS. Consequently, each widely-used language is dominantly prevalent on a certain number of SNS platforms (Khalil et al., 2024). Technically, assuming that local data is complete and FL is unnecessary, as mentioned above, applying LLMs as the backbone for detectors and introducing LoRA-PEFT (Hu et al.) emerges as a SOTA solution that effectively balances performance and cost (Yin et al., 2024; Wang et al., 2023a). Moreover, for FL-suited scenarios, recent strides in federated low-rank adaptation (FedLoRA) (Cho et al., 2024; Bai et al., 2024; Wu et al., 2024) make it possible to treat LLM-based detectors as local model backbones, i.e., only integrate additional modules for LoRA-PEFT in federated communication. As such, FedLoRA stands out as the most promising and compelling technical routine for us.

Nevertheless, despite recent studies (Guo et al., 2024c,b), significant technical challenges still persist for multilingual SNS content anomaly detection on social media—mainly regarding server-side operations and local weight uploading—which are outlined as follows:

(1-a): The nature of multilingual content detection, i.e. language gap based on multilingualism, is a kind of severe, threatening data heterogeneity (Huang et al., 2021; Tan et al., 2022).

(1-b): Since the language composition of each multilingual client is only a subset of the global language set, alleviating the multilingual curse and balancing the local detector's performance across the languages in the subset necessitates that the domain adaptation knowledge for each language be not only effective but also explicitly disentangled on the server-side (it can also be speculated that local knowledge should be also multilingually disentangled after local training). Moreover, the language-specific knowledge should not suffer from (catastrophic) forgetting and multilingual conflicts (Koohpayegani et al.; Xu et al., 2024) during aggregation and server-to-client distribution.

(2-a): Due to the severely imbalanced proportions of languages in local training data, after each local round, a local LoRA could contain mixed domain adaptation knowledge which shows varying degrees of effectiveness for various languages that appear in the corresponding client. Existing works (Khalil et al., 2024; Wang et al., 2022; Guo et al., 2024c) often treat it as a contribution to the client's primary language and then upload it. Unfortunately, it shows obvious drawbacks. First, the knowledge for local minority languages inevitably introduces noise into the primary language. Second, the contributions of minority languages are often overlooked. Specifically, if a local minority language never plays the role of primary language in a certain number of other clients, they will be excessively edged in federated collaboration. The current solution—aggregating the entire local weights into the part of global weight corresponding to the minority language—will introduce overwhelming noise. Thus, better countermeasures are needed

(2-b): In our task scenario, FL across SNS or users' data storage units always faces strict data-security legislation (Wen et al., 2023) and lower willingness to cooperate (compared to other fields, e.g. Medical) (Wu et al., 2022). Thus, the number of federated rounds also becomes a critical concern for participants in our task.

In light of this, to address these concerns, we propose MuLA-F. On the client-side of MuLA-F, we perform SVD on the local LoRA blocks and apply Diff-eRank (Wei et al., 2024) as the metric to identify the top-k most contributing feature subspaces for each language appeared in the local dataset. The selected feature subspaces are then reconstructed into the format of LoRA to achieve multilingual disentanglement of the local weights. Inspired by a previous theoretical work (Zhu et al.)

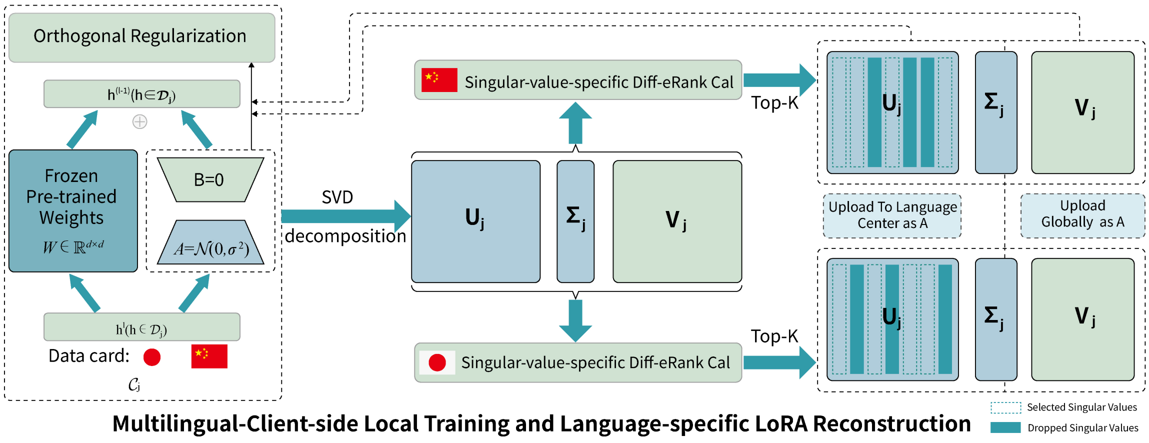

Figure 2: A client-side architecture overview of MuLA-F. We assume the local language composition is jp and zh.

which highlights the asymmetry between A and B in LoRA, i.e. A specializing in feature extraction while B specializes in feature transformation, we argue that multilingualism could influence client-side feature extraction by introducing data heterogeneity, while downstream SNS content anomaly detection is still a task with commonalities across languages (Yang et al., 2023; Dementieva and Panchenko, 2021; Xu et al., 2024). In light of this, we build multiple language centers on the server, perform language-specific federated aggregation regarding the uploaded A-matrices to achieve global disentanglement, and aggregate B-matrices globally for the downstream task. Meanwhile, in terms of local training, inspired by O-LoRA (Wang et al., 2023b), we leverage existing parameters in language centers to facilitate real-time orthogonalization of the locally reconstructed A-matrices to prevent catastrophic forgetting and knowledge conflicts between languages. Finally, after broadcasting the global B-matrices to the clients, based on each client's language composition, we customize the A-matrices with the language-specific A-matrices for this client, and then distribute them to it, ensuring the performance of the local detectors while economizing the number of federated rounds.

Experiments on three multilingual SNS content anomaly detection tasks demonstrate that MuLA-F significantly outperforms existing baselines.

2 Related Work

Since the proposal of LoRA (Hu et al.), several studies have explored incorporating LoRA into Federated Model Finetuning. For example, a study (Babakniya et al.) utilizes SVD combined with Federated Learning to initialize the LoRA blocks on local clients effectively. Additionally, (Zhang et al., 2024) integrates LoRA-based local updates with FedAvg for model aggregation. (Sun et al.) proposes a method to enhance LoRA's performance in Federated Learning settings; (Yan et al., 2024) addresses data heterogeneity by performing SVD on pretrained model weights, and (Qin et al.) reduces communication costs using zeroth-order optimization. FLoRA (Wang et al.) introduces stacking aggregation to alleviate data heterogeneity. FlexLoRA (Bai et al., 2024) introduces global SVD to allocate global knowledge across heterogeneous clients. The general problem formulation of FedLoRAs can be found in Appendix C.4.

However, there is limited research on Multilingual Federated PEFT (Parameter-Efficient Fine-Tuning). FedHLT (Guo et al., 2024b) and FedLFC (Guo et al., 2024c) have effectively utilized language family structures for federated LoRA aggregation. Existing multilingual federated finetuning methods (Khalil et al., 2024; Guo et al., 2024c) mostly focus on scenarios where each client is monolingual. The research on Federated LoRA for multilingual clients remains largely under-explored, which represents the primary technical challenge that MuLA-F addresses in our task scenario.

3 Methodology

Assuming there is a federated PEFT framework consisting of n client participants, which are denoted as $\{\mathcal{C}_1, \ldots, \mathcal{C}_n\}$. The local datasets of clients are denoted as $\{\mathcal{D}_1, \ldots, \mathcal{D}_n\}$. Before the start, the server investigates the languages that appear in at least one local dataset, to form the global language set $\mathcal{K}$, then assigns a "language center" to each language. The global language set is the union of all local language sets (the set of languages in each local dataset), expressed as $\mathcal{K} = \mathcal{K}_1 \cup \mathcal{K}_2 \cup \ldots \cup \mathcal{K}_n$. For each pair of clients, there is likely some overlap regarding the language composition of their corresponding local datasets. From a global perspective, for each language $k \in \mathcal{K}$, we identify all clients that incorporate the use of the language in their online activities, denoted as the set $\mathcal{S}^k$.

In the illustration of the client-side algorithm (see below), we focus on the client indexed by j. In detail, the local dataset $\mathcal{D}_j$ of client $\mathcal{C}_j$ consists of social media text data in multiple languages sourced from the platform or distributed storage unit, e.g. edge device, denoted by $\mathcal{D}_j = \bigoplus_{k \in \mathcal{K}_j} \mathcal{D}_j^k$, where $k_j \in \mathcal{K}_j$ is the primary language, and the languages other than the primary one are considered minority languages. In our setting, we suppose that most of the data heterogeneity among $\{\mathcal{D}_1, \ldots, \mathcal{D}_n\}$ can be attributed to the multilingual gap, as the nature of the downstream tasks across different languages are highly similar.

3.1 Multilingual Disentanglement for Client-side Language-specific Weights Uploading of LoRA Blocks

Considering the perspective put forward by Sutskever et al. and Wei et al. (Sutskever, 2023; Wei et al., 2024), the significance of the weight update in large language model training can be described as an operation specifically designed to eliminate redundant information within the training data. The process ensures that the representation of in-domain data for the given task scenario becomes more regularized and structured after undergoing additional transformations driven by the weight updates. Hence, we propose the first hypothesis: the Local LoRA blocks obtained from multilingual local training are linear combinations of rank-1 updates in multiple feature subspaces that are mutually independent. Each of these updates aids in removing redundant information and noise in data regarding one or more languages within the local dataset, then extracting more significant and structured patterns and features for it.

Furthermore, according to the insightful theoretical analysis regarding Asymmetry in LoRA by Zhu et al. (Zhu et al.), the following conclusion can be derived, which can be represented as:

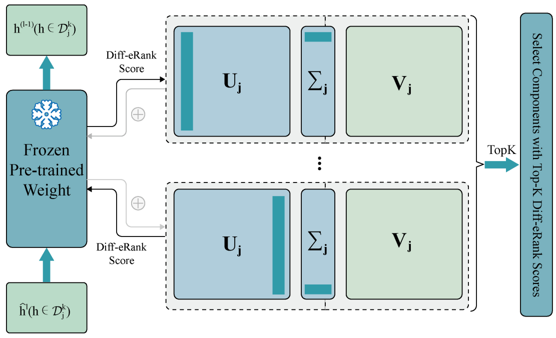

Figure 3: An illustration of multilingual disentanglement on local LoRA blocks performed by MuLA-F

$$\Delta W = BA = \phi_B \circ \varphi_A(\cdot), \tag{1}$$

where in the conclusion, A can be described as a feature extractor, while B acts more as task-oriented feature transformation, i.e. uses the extracted features to create the desired output. Building on this, we further refine our hypothesis: in our task scenario, the LoRA-based update of the backbone weight matrix can be reconstructed as a linear combination of multiple A-matrices, multiplied by a single B. In this context, the B serves the purpose of a general feature transformation for the received regularized and structured features (produced by A) towards the downstream content anomaly detection task. Each A, on the other hand, is bound with a specific language and represents a feature extractor formed by a linear combination of parts of the rank-1 updates (as described in the first hypothesis) which provide a certain contribution, e.g., removing redundant information and extracting effective patterns and features, to that language. In other words, these selected rank-1 updates jointly span the feature subspace for domain adaptation regarding the textual data composed in the language, in the given client.

Thus, specifically, given a LoRA block of $\mathcal{C}_j$ obtained by a local training round, we reconstruct it into the form of ΔW and then conduct an SVD on the matrix, which can be written as:

$$\mathrm{SVD}(B_j A_j) = U_j \Sigma_j V_j^T, \tag{2}$$

where U_j, Σ_j and V_j are the SVD components of $(B_j A_j)$, $U_j, V_j \in \mathbb{R}^{d \times r_0}$. Among them, each singular value and its corresponding singular vector can be reconstructed as a rank-1 weight matrix update. Subsequently, we introduce Diff-eRank (Wei et al., 2024), a simple yet effective metric that, from an information-theoretic perspective, measures the contribution of the rank-1 weight update in removing redundant information from features of the data and extracting more important ones, based on calculating the effective-rank (Schumacher, 1995) of the output hidden states (details can be found in Appendix A.1). For each language appeared in the local dataset, we compute the Diff-eRank contribution of each singular value. Taking the r-th singular value as an example, it can be written as:

$$e^k_{j,\ r} = \frac{1}{|\mathcal{D}^k_j|}\sum_{h\in\mathcal{D}^k_j}\big(e_{rank}(m^l_j(h^{(l-1)};\ W^{svd}_{j,\ r},\ \theta^l_j)) - e_{rank}(m^l_j(h^{(l-1)};\ \theta^l_j))\big), \tag{3}$$

where $m^l_j(\cdot)$ is the l-th transformer layer to which the LoRA block belongs, rather than the entire model. $W^{svd}_{j,\ r}\in\mathbb{R}^{d\times d}$ is a rank-1 update calculated by $(U_j[r,\ :]\Sigma_j[r,\ r]V_j[r,\ :]^T)$. $h^{(l-1)}$ denotes the hidden state fed into $m^l_j(\cdot)$. θ^l_j denotes the rest of the parameters in this layer (including other LoRA blocks in $m^l_j(\cdot)$). Here, the principle of controlling variables is strictly followed. Next, according to the Diff-eRank scores, we select the top-k singular values to retain, while masking out the other singular values (in their original positions).

$$M^k_j[r,\ r] = \begin{cases} 1, e^k_{j,\ r}\in topk\left(\left\{e^k_{j,\ t}\right\}_{t=1}^{r_0}\right) \\ 0, otherwise \end{cases}, \tag{4}$$

where M^k_j is the language-specific diagonal mask (Every local language respectively has one). Finally, the triplets after the masking operation are reconstructed into the $(B,\ A)$ format:

$$A^k_j = (\sqrt{\Sigma_j}M^k_j)V^T_j, \quad B^c_j = U_j\sqrt{\Sigma_j}. \tag{5}$$

Eventually, $\left\{\left\{A^k_j\right\}_{k\in\mathcal{K}_j}, B^c_j\right\}$ are uploaded to the server. Note that only $\left\{A^k_j\right\}_{k\in\mathcal{K}_j}$ are language-specific and uploaded as disentangled local weights specifically for the corresponding languages.

On the server-side, we perform global aggregation on all received B-matrices to obtain general feature transformation components for our multilingual SNS content anomaly detection task, which can be written as:

$$B^g = \frac{1}{\sum_{j=1}^n|\mathcal{D}_j|}\sum_{j=1}^n(|\mathcal{D}_j|B^c_j). \tag{6}$$

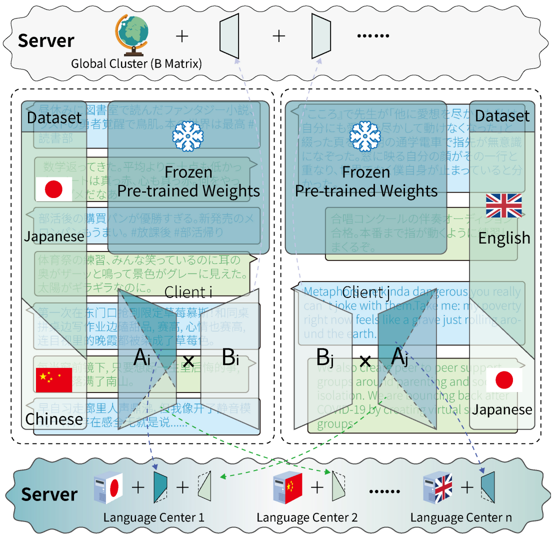

Figure 4: Architecture Overview of the proposed MuLA-F (Server-Side and Client-Server Communication)

Additionally, it is important to note that, since the entire LLM contains multiple LoRA blocks, we adopt a layer-wise (layer-by-layer) inference strategy when calculating the Diff-eRank scores

3.2 Server-side Language Centers

On the server side, the server initializes a center for each language that appears in the federated system. In each round of client-to-server federated communication, each client sends the language-specific reconstructed A-matrices to their corresponding language center for aggregation, which is:

$$A^{g,\ k} = \frac{1}{\sum_{j\in\mathcal{S}^k}|\mathcal{D}^k_j|}\sum_{j\in\mathcal{S}^k}(|\mathcal{D}^k_j|A^k_j), \tag{7}$$

where $A^{g,\ k}$ denotes the disentangled knowledge for language-specific feature extraction from a global perspective. In the server-to-client communication, the server, based on the language composition of each client's local data, selects the corresponding language centers to customize the A for that client, which can be written as:

$$A^u_j = \frac{1}{\sum_{k\in\mathcal{K}_j}|\mathcal{D}^k_j|}\sum_{k\in\mathcal{K}_j}(|\mathcal{D}^k_j|A^{g,\ k}), \tag{8}$$

where (A^u_j, B^g) is sent to $\mathcal{C}_j$, as the fruit of our proposed MuLA-F harvested by $\mathcal{C}_j$.

3.3 Orthogonal Tuning for Local Steps

Previous studies in multilingual PLMs and continual learning demonstrate that when feature subspaces of language centers overlap or conflict to some extent (Koohpayegani et al.), catastrophic forgetting can occur during the weight aggregation (in Eq.8). A recent study proposes a stacking strategy (Wang et al.) to tackle this challenge. However, it is not reliable enough for the selective aggregation in our method. Inspired by the finding of O-LoRA (Wang et al., 2023b) that constraining the feature subspaces of multiple A-matrices to be orthogonal can significantly avoid knowledge conflicts and mitigate catastrophic forgetting when aggregating them, in the local training phase of a given multilingual client $\mathcal{C}_j$, we introduce an orthogonal regularization term calculated from the real-time reconstructed A and other irrelevant language centers into the local objective function. Specifically, in each step, we perform a real-time low-rank approximate SVD on the in-training LoRA blocks, which can be written as:

$$\text{SVD}_{\text{low-rank}}(\hat{B}_j \hat{A}_j) = \hat{U}_j \hat{\Sigma}_j \hat{V}_j^T, \qquad (9)$$

where $(\hat{B}_j, \hat{A}_j)$ denotes the in-local-training LoRA block. Then, consider that the first singular value must be associated with the primary language, we compute the orthogonal loss between its corresponding singular vector and all other language centers, which is:

$$\mathcal{L}_{orth}^{a,\,1} = \sum_{i_2} ||(\hat{V}_j^T[1,\,:] \sum_{k \in \mathcal{K},\, k \neq k_j} A^{g,\,k})[1,\,i_2]||^2. \qquad (10)$$

For the other singular values, we measure the orthogonality of their corresponding singular vectors with the language centers that do not contribute at all to the current client, which is:

$$\mathcal{L}_{orth}^{a,\,ex} = \sum_{i_1,\,i_2} ||(\hat{V}_j^T[2:,\,:] \sum_{k \in \mathcal{K} \setminus \mathcal{K}_j} A^{g,\,k})[i_1,\,i_2]||^2. \qquad (11)$$

Meanwhile, to ensure that B focuses on serving as a common feature transformation, we try to unify the feature subspaces of B_j across clients. Specifically, starting from the second federated round, before each local round begins, we perform polar decomposition on B^g stored on the server to obtain the rotation matrix B_p^g, which is:

$$\text{SVD}(B^g) = U_b^g \Sigma_b^g V_b^{gT}, \quad B_p^g = U_b^g . V_b^{gT} \qquad (12)$$

Then, the orthogonality between the global rotation matrix and $\hat{B}_j$ is added into the regularization term to achieve our goal, which is:

$$\mathcal{L}_{orth}^{b} = -\sum_{i_1,\,i_2} ||(\hat{U}_j^T B_p^g)[i_1,\,i_2]||^2. \qquad (13)$$

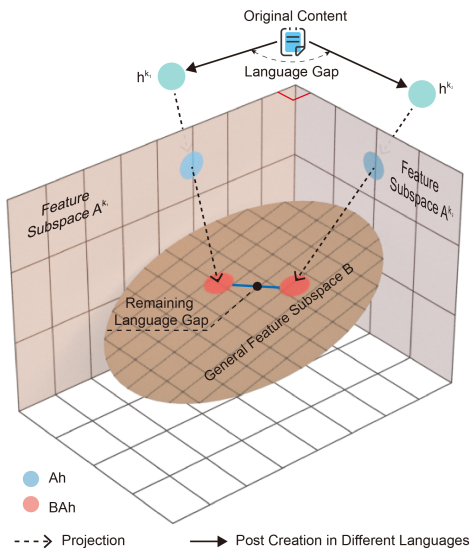

Figure 5: An illustration of our insights regarding LoRA Asymmetry in multilingual SNS content anomaly detection: language gap should be mitigated by A from language centers, while B is a general feature transformation towards detecting the content anomaly.

Finally, the modified local training objective which can mitigate the risk of forgetting when updating the selected language centers is:

$$\mathcal{L} = \mathcal{L}_{task} + \alpha(\mathcal{L}_{orth}^{a,\,1} + \mathcal{L}_{orth}^{a,\,ex}) + \beta \mathcal{L}_{orth}^{b}, \quad (14)$$

where $\mathcal{L}$ denotes the local training objective of the current client. Note that Eq. (14) is executed on the client side. It directly provides the input variables for Eq. (2).

See Appendix C.1, C.3, A.3 for further discussion of Section 3.1, 3.3, and a theoretical insight.

4 Experiments

4.1 Experimental Setups

4.1.1 Dataset

We collect publicly available social media text datasets, then filter and synthesize them into three datasets for distinct Multilingual SNS Content Anomaly Detection subtasks, which are: Fake News Detection (MM-COVID), Hate Speech Detection (CONAN), and Depression Detection (MD3D). Data statistics are provided in Table 2, and more details on global dataset construction are shown in Appendix B.1. In terms of language composition, MD3D mainly consists of commonly used East Asian languages, while the pre-processed versions of MM-COVID and CONAN are predominantly made up of Indo-European languages. All

Table 1: Comprehensive evaluations on multilingual SNS content anomaly detection datasets. Metric: federated averaged F1-Score (Fed-F1). We conduct at least 3 runs and report the averaged results (score $\pm$ std). The best results that pass $p \leq 0.005$ paired t-test are underlined (all baselines pass the paired t-test against LoRA w/o FL).

		MM-COVID		CONAN	MD3D	
		Sp-1	Sp-2	Sp-1	Sp-1	Sp-2
Qwen-2.5-7B	FedAVG	87.85 ± 0.16	86.69 ± 0.31	87.52 ± 0.19	87.14 ± 0.05	84.81 ± 0.14
	Vanilla	87.71 ± 0.45	86.37 ± 0.64	87.94 ± 0.27	86.92 ± 0.39	84.65 ± 0.58
	FFA-LoRA	88.28 ± 0.39	86.95 ± 0.33	87.69 ± 0.26	88.57 ± 0.16	85.41 ± 0.27
	FedSA	89.40 ± 0.24	87.18 ± 0.20	88.21 ± 0.15	89.86 ± 0.08	86.26 ± 0.13
	FLoRA	88.56 ± 0.54	87.03 ± 0.73	88.70 ± 0.41	91.20 ± 0.60	87.31 ± 0.72
	FlexLoRA	90.05 ± 0.31	87.91 ± 0.46	88.49 ± 0.31	92.61 ± 0.41	88.40 ± 0.65
	FedLFC	89.74 ± 0.13	88.33 ± 0.25	88.85 ± 0.10	92.03 ± 0.13	87.55 ± 0.19
	MuLA-F	<u>91.08 ± 0.22</u>	<u>89.46 ± 0.28</u>	<u>89.67 ± 0.17</u>	<u>93.35 ± 0.20</u>	<u>89.54 ± 0.24</u>
Qwen-2.5-14B	FedAVG	89.35 ± 0.13	87.62 ± 0.20	90.03 ± 0.16	90.57 ± 0.08	88.45 ± 0.13
	Vanilla	89.58 ± 0.38	87.17 ± 0.55	89.95 ± 0.29	90.11 ± 0.44	88.69 ± 0.56
	FFA-LoRA	91.22 ± 0.31	88.34 ± 0.42	90.34 ± 0.18	91.32 ± 0.22	88.90 ± 0.26
	FedSA	91.85 ± 0.26	89.59 ± 0.28	90.66 ± 0.11	91.66 ± 0.10	89.63 ± 0.18
	FLoRA	89.92 ± 0.39	87.99 ± 0.61	91.06 ± 0.35	93.59 ± 0.56	90.61 ± 0.75
	FlexLoRA	92.09 ± 0.21	89.60 ± 0.32	90.72 ± 0.20	94.02 ± 0.34	91.05 ± 0.47
	FedLFC	92.85 ± 0.09	90.11 ± 0.17	91.27 ± 0.09	92.88 ± 0.12	90.38 ± 0.14
	MuLA-F	92.66 ± 0.20	<u>91.24 ± 0.21</u>	<u>91.85 ± 0.12</u>	<u>95.25 ± 0.15</u>	<u>93.22 ± 0.19</u>

clients in the client settings reported in Table 1 are multilingual themselves. To address potential concerns regarding MuLA-F's performance on settings with monolingual clients, we additionally set up a client setting that includes both multilingual and monolingual clients. The corresponding additional results are reported in Appendix A.4. Details of client construction can be seen in in Appendix B.2.

4.1.2 Baselines

The selected competitive FedLoRAs as baselines except **FedAVG** (McMahan et al., 2017) include: **Vanilla**, **FFA-LoRA** (Sun et al.), **FedSA** (Guo et al., 2024a), **FLoRA** (Wang et al.), **FlexLoRA** (Bai et al., 2024), **FedLFC** (Guo et al., 2024c). Among them, FFA-LoRA and FedSA are simple but theoretically solid FedLoRA baselines. FLoRA and FlexLoRA are cutting-edge SOTA FedLoRAs. FedLFC is a dedicated SOTA Multilingual FedLoRA. Baseline introductions and implementation details can be found in Appendix B.3 and B.4.

Due to the inclusion of multiple East Asian languages in MD3D, we choose Qwen-2.5-7B and Qwen-2.5-14B as our base LLMs, as LLaMA-3.1-8B and Mistral-7B do not support these languages. Additional experimental results using LLaMA-3.1-8B on other datasets are shown in Appendix A.5.

4.2 Comprehensive Evaluations

The results of the comprehensive evaluations are reported in Table 1. Our findings are as follows:

(a) Across the three tasks of PEFT-based multilingual SNS anomaly detection, our proposed MuLA-F outperforms the best baseline methods by an average of approximately 1.2 percentage points. If we regard FedAVG as a reference point with no additional modules or modifications — effectively a relative zero — from this perspective, the advantage of MuLA-F will be further amplified.
(b) When the LoRA-rank is a normal value, FLoRA lags behind other methods, with almost no prominent local performance, especially on the CONAN dataset with less data. One reason is that while the stacking operation avoids introducing cross terms B_iA_j as noisy, the cost is that the global rank expands sharply, which can impair the validity of each singular value, causing feature subspace redundancy or multicollinearity.
(c) By effectively integrating the data resources of each language family, FedLFC performs better when the global data distribution is more equitable across language families. However, when more clients' local language composition spans multiple language families or some low-resource languages are consistently not primary across the clients, its

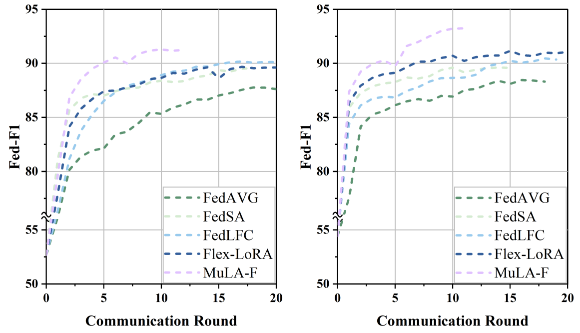

Figure 6: An analysis of federated communication rounds (Metric: Fed-F1; Base model: Qwen-2.5-14B; Dataset: (a) MD3D, (b) MM-COVID. For each method, We report the test results up to the corresponding checkpoint round in the section of comprehensive evaluation.

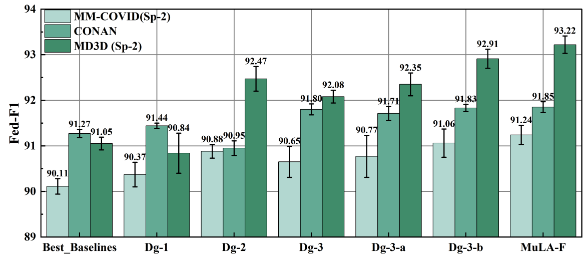

Figure 7: Ablation Study (Metric: Fed-F1; Base model: Qwen-2.5-14B)

Fed-F1 substantially decreases.
(d) The relatively low performance of FFA-LoRA and FedSA indicates that their interpretations of LoRA's asymmetry are successfully challenged by MuLA-F's in the context of multilingual SNS content anomaly detection.
(e) As the best-performing baseline on average, FlexLoRA also provides disentanglement of multilingual domain knowledge algorithmically. However, it occurs on the server-side, relatively late, which highlights that MuLA-F's early multilingual disentanglement (on the client-side) better leverages minor language data in local datasets, especially when there exist high data heterogeneity and low task heterogeneity.
(f) The language-specific evaluation results in Figure 8 further indicate that, even though MuLA-F does not show a significant advantage on primary Languages, it greatly balances the local model's performance for other locally minor languages.

Additional evaluations on special client setting including both multilingual and monolingual clients are reported in Appendix A.4.

4.3 Communication Rounds

We conduct a round-by-round analysis of MuLA-F and 4 critical baseline methods. The experimental results reported in Figure 6 show that, in terms of smoothness, FedLFC shows a more stable convergence per round, which may be because it does not perform complex decomposition operations. On the other hand, MuLA-F and FlexLoRA make larger strides toward convergence in the early rounds, although some fluctuation occurs. Despite MuLA-F having higher local round overheads, it surpasses the baselines within less than 40% of the total rounds. We emphasize that, in the context of our task, the number of federated communication rounds is a very sensitive parameter due to factors such as instability in multi-party cooperation intentions across social media platforms. Moreover, since each SNS participant always has sufficient resources for local model training, the sensitivity to local overhead is lower than in other scenarios.

4.4 Ablation Study

We create five degraded versions of MuLA-F. Specifically, we remove the following components: local disentanglement (directly submitting local weights; Dg-1), global disentanglement (aggregating language centers; Dg-2), orthogonal regularization (Dg-3), orthogonal regularization applied to A-matrices (Dg-3-a) and to B-matrices (Dg-3-b).

Experimental results reported in Figure 7 show that the effectiveness of language centers is significantly enhanced when local disentanglement is introduced. However, "language center" mechanism alone could not directly effectively utilize the weights submitted by multilingual clients. Moreover, the increase in the number of languages causes higher multilingual conflict, which manifests in the experimental results as a lift in the importance of the orthogonal term.

5 Conclusion

In this paper, to address the challenging issues faced by multilingual SNS content anomaly detection, we propose MuLA-F. MuLA-F leverages the asymmetry of LoRA, incorporating our proposed SVD-based multi-level multilingual knowledge disentangling and orthogonal regularization modules. These components significantly alleviate the multilingual curse and knowledge conflicts in our task scenario, enabling MuLA-F to outperform the cutting-edge FedLoRAs on multilingual clients.

Limitations

The limitations of MuLA-F are discussed as below.

Firstly, excessive constraints in feature decoupling may result in an overly rigid feature space, suppressing natural relationships between languages (such as the grammatical similarities between Japanese and Chinese), thereby affecting the model's adaptability to new language combinations or mixed languages. Over-decoupling may prevent the model from capturing shared features across languages, reducing generalization performance. Another drawback is the high computational cost of SVD decomposition and orthogonal constraints, particularly in scenarios involving large-scale language models or massive datasets, which could significantly slow down training speed and limit the scale of practical applications.

Nevertheless, as demonstrated by the experiments reported in Appendix A.6 and Appendix C.2, the two aforementioned drawbacks do not caused significant negative influence in our task setting. They are acceptable and do not undermine the significance of MuLA-F's advantages in comparison to baseline methods.

References

Esma Aïmeur, Sabrine Amri, and Gilles Brassard. 2023. Fake news, disinformation and misinformation in social media: a review. *Social Network Analysis and Mining*, 13(1):30.

Firoj Alam, Hassan Sajjad, Muhammad Imran, and Ferda Ofli. 2021. Crisisbench: Benchmarking crisis-related social media datasets for humanitarian information processing. In *Proceedings of the International AAAI conference on web and social media*, volume 15, pages 923–932.

Sara Babakniya, Ahmed Roushdy Elkordy, Yahya H Ezzeldin, Qingfeng Liu, Kee-Bong Song, MOSTAFA EL-Khamy, and Salman Avestimehr. Slora: Federated parameter efficient fine-tuning of language models. In *International Workshop on Federated Learning in the Age of Foundation Models in Conjunction with NeurIPS 2023*.

Jiamu Bai, Daoyuan Chen, Bingchen Qian, Liuyi Yao, and Yaliang Li. 2024. Federated fine-tuning of large language models under heterogeneous tasks and client resources. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Wei Chen, Carlos Castillo, and Laks VS Lakshmanan. 2013. *Information and influence propagation in social networks*. Morgan & Claypool Publishers.

Yae Jee Cho, Luyang Liu, Zheng Xu, Aldi Fahrezi, and Gauri Joshi. 2024. Heterogeneous lora for federated fine-tuning of on-device foundation models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12903–12913.

Yi-Ling Chung, Elizaveta Kuzmenko, Serra Sinem Tekiroglu, and Marco Guerini. 2019. Conan-counter narratives through nichesourcing: a multilingual dataset of responses to fight online hate speech. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2819–2829.

Daryna Dementieva and Alexander Panchenko. 2021. Cross-lingual evidence improves monolingual fake news detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: Student Research Workshop*, pages 310–320.

Dominique Geissler, Dominik Bär, Nicolas Pröllochs, and Stefan Feuerriegel. 2023. Russian propaganda on social media during the 2022 invasion of ukraine. *EPJ Data Science*, 12(1):35.

Pengxin Guo, Shuang Zeng, Yanran Wang, Huijie Fan, Feifei Wang, and Liangqiong Qu. 2024a. Selective aggregation for low-rank adaptation in federated learning. *arXiv preprint arXiv:2410.01463*.

Zhihan Guo, Yifei Zhang, Zhuo Zhang, Zenglin Xu, and Irwin King. 2024b. Fedhlt: Efficient federated low-rank adaption with hierarchical language tree for multilingual modeling. In *Companion Proceedings of the ACM on Web Conference 2024*, pages 1558–1567.

Zhihan Guo, Yifei Zhang, Zhuo Zhang, Zenglin Xu, and Irwin King. 2024c. Fedlfc: Towards efficient federated multilingual modeling with lora-based language family clustering. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 1519–1528.

J Brian Houston, Joshua Hawthorne, Mildred F Perreault, Eun Hae Park, Marlo Goldstein Hode, Michael R Halliwell, Sarah E Turner McGowen, Rachel Davis, Shivani Vaid, Jonathan A McElderry, et al. 2015. Social media and disasters: a functional framework for social media use in disaster planning, response, and research. *Disasters*, 39(1):1–22.

Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Yutao Huang, Lingyang Chu, Zirui Zhou, Lanjun Wang, Jiangchuan Liu, Jian Pei, and Yong Zhang. 2021. Personalized cross-silo federated learning on non-iid data. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 7865–7873.

Samar Samir Khalil, Noha S Tawfik, and Marco Spruit. 2024. Federated learning for privacy-preserving depression detection with multilingual language models in social media posts. *Patterns*.

Suin Kim, Ingmar Weber, Li Wei, and Alice Oh. 2014. Sociolinguistic analysis of twitter in multilingual societies. In *Proceedings of the 25th ACM conference on Hypertext and social media*, pages 243–248.

Soroush Abbasi Koohpayegani, KL Navaneet, Parsa Nooralinejad, Soheil Kolouri, and Hamed Pirsiavash. Nola: Compressing lora using linear combination of random basis. In *The Twelfth International Conference on Learning Representations*.

Zhenyu Lei, Yushun Dong, Weiyu Li, Rong Ding, Qi Wang, and Jundong Li. 2025. Harnessing large language models for disaster management: A survey. *arXiv preprint arXiv:2501.06932*.

Yichuan Li, Bohan Jiang, Kai Shu, and Huan Liu. 2020. Mm-covid: A multilingual and multimodal data repository for combating covid-19 disinformation. *arXiv preprint arXiv:2011.04088*.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. 2017. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR.

Zewdie Mossie and Jenq-Haur Wang. 2020. Vulnerable community identification using hate speech detection on social media. *Information Processing & Management*, 57(3):102087.

Qiong Nan, Qiang Sheng, Juan Cao, Beizhe Hu, Danding Wang, and Jintao Li. 2024. Let silence speak: Enhancing fake news detection with generated comments from large language models. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 1732–1742.

Jonas Pfeiffer, Naman Goyal, Xi Lin, Xian Li, James Cross, Sebastian Riedel, and Mikel Artetxe. 2022. Lifting the curse of multilinguality by pre-training modular transformers. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3479–3495.

Associated Press. 2025. Tiktok refugees are pouring to xiaohongshu. what to know about the rednote app.

Zhen Qin, Daoyuan Chen, Bingchen Qian, Bolin Ding, Yaliang Li, and Shuiguang Deng. Federated full-parameter tuning of billion-sized language models with communication cost under 18 kilobytes. In *Forty-first International Conference on Machine Learning*.

Paul Röttger, Bertie Vidgen, Dong Nguyen, Zeerak Waseem, Helen Margetts, and Janet Pierrehumbert. 2021. Hatecheck: Functional tests for hate speech detection models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 41–58.

David Savage, Xiuzhen Zhang, Xinghuo Yu, Pauline Chou, and Qingmai Wang. 2014. Anomaly detection in online social networks. *Social networks*, 39:62–70.

Benjamin Schumacher. 1995. Quantum coding. *Physical Review A*, 51(4):2738.

Youbang Sun, Zitao Li, Yaliang Li, and Bolin Ding. Improving lora in privacy-preserving federated learning. In *The Twelfth International Conference on Learning Representations*.

Ilya Sutskever. 2023. Stronger compressors find more shared structure. The Ilya's Talk. Accessed: 2023-10-15.

Alysa Ziying Tan, Han Yu, Lizhen Cui, and Qiang Yang. 2022. Towards personalized federated learning. *IEEE transactions on neural networks and learning systems*, 34(12):9587–9603.

Haoyu Wang, Handong Zhao, Yaqing Wang, Tong Yu, Jiuxiang Gu, and Jing Gao. 2022. Fedkc: Federated knowledge composition for multilingual natural language understanding. In *Proceedings of the ACM Web Conference 2022*, pages 1839–1850.

Neng Wang, Hongyang Yang, and Christina Dan Wang. 2023a. Fingpt: Instruction tuning benchmark for open-source large language models in financial datasets. *arXiv preprint arXiv:2310.04793*.

Xiao Wang, Tianze Chen, Qiming Ge, Han Xia, Rong Bao, Rui Zheng, Qi Zhang, Tao Gui, and Xuan-Jing Huang. 2023b. Orthogonal subspace learning for language model continual learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10658–10671.

Ziyao Wang, Zheyu Shen, Yexiao He, Guoheng Sun, Hongyi Wang, Lingjuan Lyu, and Ang Li. Flora: Federated fine-tuning large language models with heterogeneous low-rank adaptations. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Lai Wei, Zhiquan Tan, Chenghai Li, Jindong Wang, and Weiran Huang. 2024. Diff-erank: A novel rank-based metric for evaluating large language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Orion Weller, Marc Marone, Vladimir Braverman, Dawn Lawrie, and Benjamin Van Durme. 2022. Pre-trained models for multilingual federated learning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1413–1421.

Jie Wen, Zhixia Zhang, Yang Lan, Zhihua Cui, Jianghui Cai, and Wensheng Zhang. 2023. A survey on federated learning: challenges and applications. *International Journal of Machine Learning and Cybernetics*, 14(2):513–535.

Chuhan Wu, Fangzhao Wu, Lingjuan Lyu, Yongfeng Huang, and Xing Xie. 2022. Fedctr: Federated native ad ctr prediction with cross-platform user behavior data. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 13(4):1–19.

Feijie Wu, Zitao Li, Yaliang Li, Bolin Ding, and Jing Gao. 2024. Fedbiot: Llm local fine-tuning in federated learning without full model. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 3345–3355.

Liang Wu, Fred Morstatter, Kathleen M Carley, and Huan Liu. 2019. Misinformation in social media: definition, manipulation, and detection. *ACM SIGKDD explorations newsletter*, 21(2):80–90.

Rongwu Xu, Zehan Qi, Zhijiang Guo, Cunxiang Wang, Hongru Wang, Yue Zhang, and Wei Xu. 2024. Knowledge conflicts for llms: A survey. *arXiv preprint arXiv:2403.08319*.

Yuxuan Yan, Qianqian Yang, Shunpu Tang, and Zhiguo Shi. 2024. Federa: Efficient fine-tuning of language models in federated learning leveraging weight decomposition. *arXiv preprint arXiv:2404.18848*.

Yingguang Yang, Renyu Yang, Hao Peng, Yangyang Li, Tong Li, Yong Liao, and Pengyuan Zhou. 2023. Fedack: Federated adversarial contrastive knowledge distillation for cross-lingual and cross-model social bot detection. In *Proceedings of the ACM Web Conference 2023*, pages 1314–1323.

Kai Yin, Chengkai Liu, Ali Mostafavi, and Xia Hu. 2024. Crisissense-llm: Instruction fine-tuned large language model for multi-label social media text classification in disaster informatics. *arXiv preprint arXiv:2406.15477*.

Cheng Zhang, Chao Fan, Wenlin Yao, Xia Hu, and Ali Mostafavi. 2019. Social media for intelligent public information and warning in disasters: An interdisciplinary review. *International Journal of Information Management*, 49:190–207.

Jianyi Zhang, Saeed Vahidian, Martin Kuo, Chunyuan Li, Ruiyi Zhang, Tong Yu, Guoyin Wang, and Yiran Chen. 2024. Towards building the federatedgpt: Federated instruction tuning. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6915–6919. IEEE.

Wanru Zhao, Yihong Chen, Royson Lee, Xinchi Qiu, Yan Gao, Hongxiang Fan, and Nicholas Donald Lane. 2024. Breaking physical and linguistic borders: Multilingual federated prompt tuning for low-resource languages. In *The Twelfth International Conference on Learning Representations*.

Jiacheng Zhu, Kristjan Greenewald, Kimia Nadjahi, Haitz Sáez de Ocáriz Borde, Rickard Brüel Gabrielsson, Leshem Choshen, Marzyeh Ghassemi, Mikhail Yurochkin, and Justin Solomon. Asymmetry in low-rank adapters of foundation models. In *Forty-first International Conference on Machine Learning*.

A Extensive Analysis

A.1 Sensitivity Analysis

In MuLA-F, two critical configuration parameters are: the number of singular value components selected for each local language after Diff-eRank evaluation, and the coefficient α of the orthogonal regularization term.

The experimental results shown in Figure 9 indicate that for $\mathcal{C}_j$, the optimal number of selected singular values should be slightly greater than $r/||K_j||$. α depends on the complexity of the local language composition. Furthermore, the experimental results suggest that the optimal value of the orthogonal coefficient is partially influenced by the level of heterogeneity across languages within the current dataset.

In this part, we also evaluate several of the most competitive FedLoRAs with respect to their sensitivity to the rank of LoRA. The experimental results shown in Figure 10 indicate that an increase in the global number of languages or the complexity of local datasets indicates a higher rank required. In contrast, FLoRA is more suitable for low-rank local LoRA, while MuLA-F and FlexLoRA are better suited for higher ranks.

A.2 Introduction to Diff-eRank

Diff-eRank is a novel evaluation metric for large language models (LLMs) based on information theory and Effective Ranks. Diff-eRank assesses model performance by analyzing the effective rank of hidden representations. This approach quantifies how LLMs eliminate redundant information during training and how LLMs make the data representations more structural for feature transformations, offering evaluation insights regarding their internal information processing.

Specifically, regarding the algorithm, given an arbitrary input x, Diff-eRank calculates the hidden representation respectively with the model before (M_0) and after (M_1) the training:

$$h_0 = M_0(x), h_1 = M_1(x),$$

where h_0, h_1 are two-dimensional sequential hidden representations with the shape [seq-len, d]. Furthermore, it respectively calculates the covariance

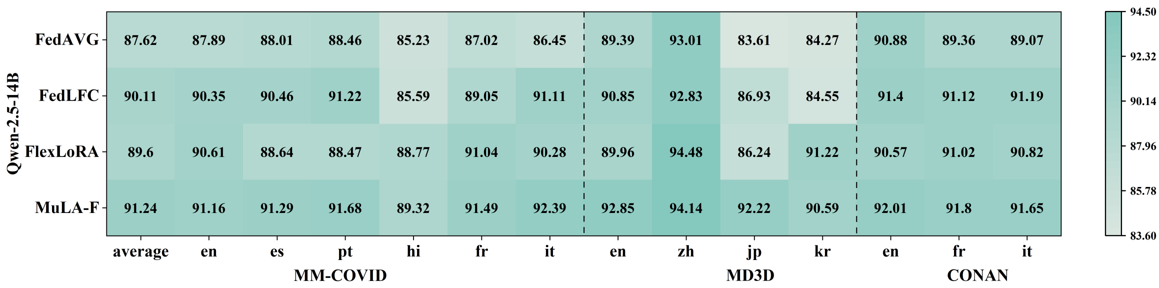

Figure 8: Multilingual comprehensive evaluation results (Base model: Qwen-2.5-14B; Metric: Fed-F1)

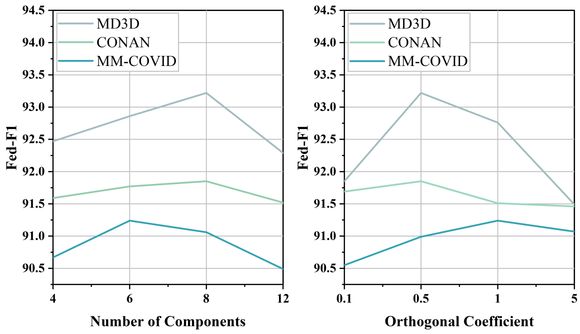

Figure 9: Sensitivity analysis of Diff-eRank-selected singulars and orthogonal tuning coefficient (Dataset: MD3D (Sp-2), CONAN, MM-COVID (Sp-2); Base model: Qwen-2.5-14B; Metric: Fed-F1)

matrix of h_0, h_1, as A_0, A_1. Finally, the effective rank of each covariance matrix can be calculated as:

$$e_{rank}(A) = \exp\left(-\frac{\sum_{i=1}^{Q} \sigma_i}{\sum_{i=1}^{Q} \sigma_i \log \sigma_i}\right),$$

where σ denote the singular values of A In MuLA-F, as for all local languages of a given client, (M_0) is consistent. Hence, we only need to rank the values of $e_{rank}(A_1)$

A.3 A Theoretical Insight

Although all the authors of this paper come from a team that primarily focuses on empirical work, we still provide an interesting theoretical insight to enhance the soundness of MuLA-F.

In our setting, the gradient of the orthogonal loss function forces the column vectors of v_j to align with the orthogonal space, thereby correcting the update direction of the client's parameters. When this constraint aligns with the objective function (for example, separating noise features), the convergence speed will be accelerated. When the gradient descent applies the orthogonal constraint, it restricts the parameters within the set of the Stiefel manifold (the space of orthogonal matrices), which is written as:

$$V(m,n) = \{W \in \mathbb{R}^{m \times n} | W^T W = I_n\},$$

so that it can be regarded as a non-convex optimization problem, which reduces the risk of catastrophic forgetting. Note that at this point, $V(m,n)$ is defined as the set of $m \times n$ matrices that satisfy the column orthogonality condition, closely approximating the set of decomposed A-matrices in MuLA-F. It allows the optimization problem with orthogonal constraints to be framed within Riemannian optimization, making its convergence less questionable. Specifically, the gradient in the embedding space (in Euclidean space) is calculated as:

$$G = \nabla_W f(W),$$

where $f(W)$ is the objective function being minimized, and is then projected onto the tangent space of the Stiefel manifold at the point W:

$$gradf(W) = Proj_w(G) = G - W \cdot sym(W^T G),$$

After that, during the optimization process, the tangent vector p can be further projected back to the manifold (similar to the Cayley transform) to maintain its orthogonality:

$$R_w(p) = (I - \frac{s}{2}p)^{-1}(I + \frac{s}{2}p)W,$$

where s represents the step size. If the actual step size used during updates satisfies the Wolfe conditions, then this gradient descent can converge to a stable point within the framework of Riemannian gradient descent. This means that the various language centers are sufficiently modularized and orthogonalized. At the same time, in addition to reducing catastrophic forgetting, this also avoids irrational update directions and updates on redundant parameters, enhancing numerical stability.

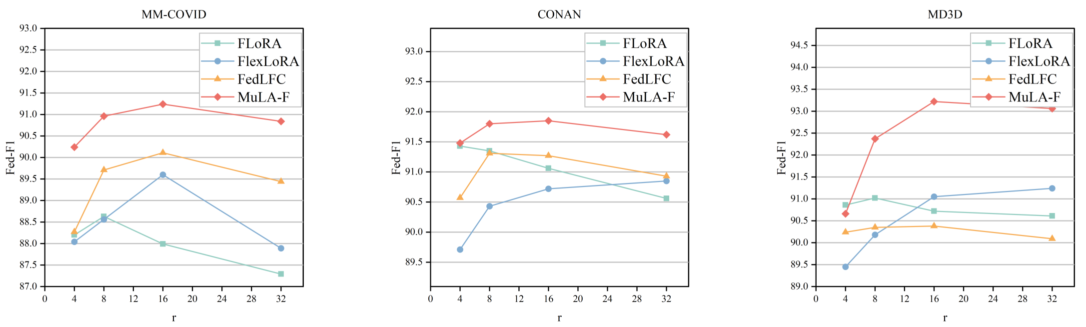

Figure 10: Impact of LoRA Rank across MuLA-F and three important baselines (Dataset: MD3D (Sp-2), CONAN, MM-COVID (Sp-2); Base model: Qwen-2.5-14B; Metric: Fed-F1)

Table 2: Comprehensive evaluations on multilingual settings where part of clients are monolingual (Base Model: Qwen-2.5-7B, Metrics: Fed-F1). The best results that pass $p \leq 0.005$ paired t-test are shaded

Model	MM-COVID Sp-3	MD3D Sp-3
FedAVG	89.15 ± 0.19	89.34 ± 0.32
Vanilla	89.12 ± 0.57	89.22 ± 0.83
FFA-LoRA	90.01 ± 0.36	91.67 ± 0.33
FedSA	90.27 ± 0.24	92.23 ± 0.30
FLoRA	90.12 ± 0.66	92.45 ± 0.28
FlexLoRA	90.55 ± 0.61	92.81 ± 0.50
FedLFC	90.72 ± 0.19	92.96 ± 0.24
MuLA-F	91.43 ± 0.41	93.50 ± 0.35

A.4 Evaluations on Client Settings with Monolingual Clients.

Although the proposed MuLA-F dedicatedly targets scenarios where the clients are multilingual, in the context of SNS content anomaly detection, some clients might still be considered monolingual (e.g., Yahoo, which has a highly localized user profile). At the same time, there are also potential concerns about whether MuLA-F still performs outstandingly in settings where monolingual clients are present. Therefore, it would be meaningful to compare MuLA-F with baseline methods in multilingual scenarios where part of clients are monolingual. In consideration of this, by respectively creating an additional monolingual client for each language involved (based on Sp-1), we create a special client setting on MM-COVID and MD3D, named as Sp-3 (details see in Appendix B.2). We use the Qwen-2.5-7B model as the base model and report the experimental results in Table 2.

The results show that there is a noticeable reduction in the performance advantage of MuLA-F. The main reason for this phenomenon is, when a client is monolingual, it implies that the client-side multilingual disentanglement module (Para 3.1) of MuLA-F methodologically doesn't work. However, overall, the advantage of MuLA-F still remains statistically significant.

A.5 Evaluations on Extra LLM Base Model

When selecting base LLMs for the main experiments, we encounter a minor challenge — as the language composition of our data is quite rich, most of the widely-used small LLMs are not suitable for the language composition of our multilingual SNS anomaly detection tasks (e.g., LLaMA 3.1-8B is only applicable to English, Spanish, German, French, Hindi, Thai, Italian, and Portuguese; Mistral is mainly suitable for English, French, German, and Spanish). Therefore, we could only choose qwen-2.5-7B and qwen-2.5-14B as base LLM models, as their training corpus covers all the languages appeared in the datasets of our experiments. However, to alleviate potential concerns regarding the singularity of base LLM selection, we conduct additional evaluations using LLaMA-3.1-8B as the base LLM model only on MM-COVID and CONAN. The results of MuLA-F and four most competitive baseline methods are reported in Table 3.

The experimental results show that the performance advantage of MuLA-F compared to the baselines is still sufficiently significant. Additionally, FlexLoRA and FedLFC remain the most competitive baselines. The findings indicate that base model selection does not affect the overall experimental conclusions.

Table 3: Comprehensive evaluations using LLaMA-3.1-8B as base model (Metrics: Fed-F1, Dataset: MM-COVID, CONAN (LLaMA-3.1-8B does not support Chinese, Japanese and Korean appeared in MD3D)).The best results that pass $p \leq 0.005$ paired t-test are shaded

Method	MM-COVID Sp-1	MM-COVID Sp-2	CONAN Sp-1
FFA-LoRA	87.72 ± 0.20	86.65 ± 0.49	86.85 ± 0.72
FedSA	88.69 ± 0.31	87.48 ± 0.65	87.21 ± 0.44
FLexLoRA	90.01 ± 0.28	87.60 ± 0.53	87.74 ± 0.46
FedLFC	89.16 ± 0.27	88.11 ± 0.61	88.09 ± 0.21
MuLA-F	**90.78** ± 0.37	**88.87** ± 0.48	**88.93** ± 0.25

Table 4: Time overhead statistics (Base Model: Qwen-2.5-7B; Metrics: GPU-Hour).

Model	MM-COVID	MD3D
FedSA	18.6	11.8
FLexLoRA	21.9	14.2
FedLFC	20.3	13.5
MuLA-F	24.4	15.7

A.6 Time Overhead Analysis

The most obvious limitation of MuLA-F is that, due to the need to perform inference on each individual singular value as described in Section 3.1—although in a layer-by-layer manner—its time overhead will be higher than that of other FedLoRAs. To evaluate this, we take Qwen-2.5-7B base model as an example and record the time overhead (average of Sp-1 and Sp-2 for MM-COVID and MD3D) incurred by each method up to the checkpoint round. The results are reported in Table 4.

Experimental results show that although MuLA-F has slightly higher time overhead, the difference compared to baseline methods is not substantial. This is because the dominant source of the time cost in FedLoRA still lies in the LoRA PEFT training across multiple epochs in each round. Nevertheless, as shown in Figure 6, MuLA-F requires only about 70% of the federated communication rounds on average, compared to the baseline methods. This indicates that, in practical SNS scenarios, compared to others, MuLA-F's participants can share parameters for fewer times, thereby lowering the collaboration threshold, which can also be regarded as a compensation for the higher time overhead.

Table 5: Global Statistics of Datasets

Dataset	Train + Val	Test
MM-COVID	48268	8519
CoNAN	8027	1417
MD3D	20262	3576

B Experiment Details

B.1 Datasets

The detailed description of the datasets is provided below. The global language composition statistics and data statistics are respectively shown in Figure 11 and Table 5.

MM-COVID (Li et al., 2020): This dataset consists of English, Spanish, Portuguese, Hindi, French, and Italian. Among these, English, Spanish, and Portuguese are high-resource languages, while Italian, French, Hindi are considered a low-resource language. Given the extreme distribution of the original dataset, we perform 50% downsampling on the following categories: en-real, en-fake, fr-fake, pt-fake, and es-fake.

CONAN (Chung et al., 2019): The high-quality, manually constructed dataset includes three languages—English, French, and Italian. We retain all original pairs and augmented pairs. However, for each pair, we only keep one of the positive or negative samples. Additionally, we discard all English-translated pairs, as they might introduce information leakage into the samples from other languages.

MD3D: We leverage three publicly available datasets to construct MD3D (Note that the dataset can be also referred to as MU3D. All data consist of social media posts):

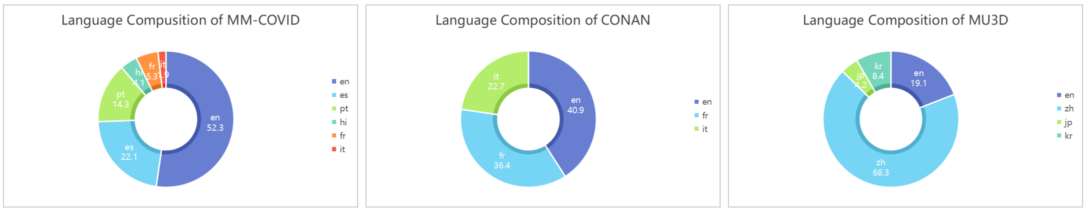

Figure 11: Global language composition statistics of full datasets (%).

(1) Depression detection data [2] collected from a Korean daily-learning app, as well as from Twitter in English, Korean, and Japanese-speaking regions.
(2) Depression detection data collected from Weibo (a Chinese alternative to Twitter) [3]. Specifically, we perform 50% downsampling on the user set. Since each user has multiple posts, we concatenate the longest and most recent posts from each user to form a representative post for that user.
(3) Posts from suspected depression patients on Reddit [4].

B.2 Client Construction

Considering the generally low applicability of the LLaMA-3 series to East Asian languages, we select Qwen-2.5-7B and Qwen-2.5-14B as the local LLM backbones. Our hypothesis suggests that, while there are shared commonalities, different languages have distinct characteristics, which is what introduces data heterogeneity among clients in our scenario.

Thus, MuLA-F differs from other existing FedLoRAs that sample global datasets using a Dirichlet distribution to create clients with heterogeneous data. In MuLA-F's data partitioning, for each local dataset, the language set is a subset of the global language set, and the elements within this subset exhibit some degree of relatedness (in terms of linguistic or socio-cultural background).

Overall, each client consists of 2-4 languages. "Sp-1" refers to a dataset split with data from 5 clients, while "Sp-2" refers to a dataset split with data from 10 clients. Note that, due to the multilingualism nature, only MM-COVID and MD3D have an Sp-2 split setting. Additionally, the data splitting strategy varies across each dataset.

The details of client construction are listed below:

[2] https://github.com/dxlabskku/Mental-Health/tree/main/data
[3] https://github.com/aidenwang9867/Weibo-User-Depression-Detection-Dataset
[4] https://github.com/usmaann/Depression_Severity_Dataset

MM-COVID: During the data splitting process, we make every effort to ensure that languages from the Romance language family, which are closely related, appear together on certain clients. For bilingual, trilingual, and quadrilingual clients, the proportion of the primary language is set to be greater than 60%, greater than 50%, and greater than 40%, respectively. Among these, languages within the Romance language family exhibit a high degree of affinity. The language splits are as follows: Sp-1: (1) en-fr-it (2) pt-es-en (3) hi-en; (4) en-es-fr-it; (5) es-pt-it-fr Sp-2: (1); (2); (3); (4); (5); (6) en-fr-it; (7) es-pt-fr; (8) en-hi; (9) en-es-it-fr (10) pt-es-fr-it.

CONAN: The local clients' language composition is as follows: (1) en-fr; (2) fr-en; (3) en-it; (4) it-en; (5) fr-it-en. The proportion of the primary language for each client is between 60% and 80%. We prioritize sampling data from clients where English is the primary language.

MD3D: We provide two client and data split strategies, Sp-1 and Sp-2. Overall, 60% of the clients are bilingual, while 40% are trilingual. To simulate a realistic industry ecosystem, the language composition of each client is as follows (with the primary language listed first): Sp-1: (1) jp-en; (2) kr-en; (3) zh-en; (4) en-kr-jp; (5) zh-jp-kr. Sp-2: (1); (2); (3); (4); (5); (6) en-jp; (7) en-kr; (8) zh-en; (9) zh-en-jp; (10) zh-en-kr. For each bilingual client, the primary language accounts for 60%-95% of the data; for each trilingual client, the primary language accounts for 50%-95%. Since the amount of data for Japanese and Korean is relatively small, prior to the data split, we first designate 50% of the data for these two languages to construct clients where either Japanese or Korean is the primary language. The remaining data is then allocated to other clients involving these two languages according to a Dirichlet distribution. Specifically, for clients

Table 6: An Example of multilingual local instruction-tuning datasets for MuLA-F and FedLoRA baselines (using "client (1)" and "client (5)" in MD3D-Sp-1 as the clients). Other multilingual examples in Chinese, Japanese and Korean are respectively shown in Figure 12, Figure 13 and Figure 14.

Task Type	Multilingual Depression Detection
Post Content Language	English
Client ID	1 (MD3D-Sp-1)
Local Language Composition	Japanese, English
Task Instruction	You will receive a social media post written by an English user who is at risk of depression. You must analyze whether the post clearly shows depression or subtly suggests depressive tendencies through word choice, phrasing, or viewpoints. Based on your analysis, assess whether the user has depression.
Input (Post Content)	When I felt the coldness from water on the skin of my temple. I thought I would feel fear but all I felt was relief and how easy it would have been to end my overthinking, torturing anxiety brain. I think about everything I've said and done and it feels like fight of flight all the time.
Output (Label)	["Depressed"]
Explanation for Readers	The author suffers from severe anxiety and suicidal tendencies.

where Japanese or Korean is not the primary language, we prioritize constructing clients (6) and (7), followed by (4) and (5), then (9) and (10), and finally others.

A similar strategy is also applied to the data splitting for MM-COVID and CONAN.

Overall, each local dataset after data-split shows data characteristics that can be mapped to a real-world social media platform. For each local dataset, we divide the data into training, valid and test sets at a 75%/10%/15% ratio. Since each round of local LoRA-PEFT only involves two epochs (unchanged), we do not use the validation set to schedule the local epoch.

It's important to note that, to intuitively demonstrate the effects of locally trained instruction-tuning data and prompts, we provide one example per language using the MD3D dataset. These examples can be found in Table 6, Figure 12, Figure 13, and Figure 14.

B.3 Baselines

Details of the baseline methods in this paper are listed below.

FedAVG: The most classic baseline method, used to demonstrate that MuLA-F indeed makes a positive contribution.

Vanilla: Almost all FedLoRA researchers have considered performing federated aggregation with dual centers on the A and B matrices, which can be written as: However, unfortunately, terms like B_iA_j, without special conditions, would introduce significant noise, making its performance unstable compared to FedAVG. Nevertheless, this method still needs to be mentioned and compared in experiments.

FFA-LoRA: A simple yet SOTA FedLoRA baseline method, with a conflicting perspective against ours. It ignores and freezes A , only performing FedAVG on B.

FedSA: A simple yet SOTA general FedLoRA, which is a compromise between the previous method and MuLA-F in terms of core ideas. It acknowledges the importance of federated aggregation for A but downplays the significance of A on local heterogeneous datasets. The core insight of FedSA is to use the asymmetry of LoRA to globally aggregate A and locally personalize B. The underlying logic for utilizing asymmetry conflicts with the perspective of MuLA-F.

FLORA: The authors of FLoRA argue that additive aggregation operation is the root cause of the problem. In light of this, they modify it to stacking A and B.

Task Type	Multilingual Depression Detection
Post Content Language	Chinese
Client ID	5 (in MU3D-Sp-1)
Local Language Composition	Chinese, Korean, Japanese
Task Instruction	You will receive a social media post written by a Chinese user who is at risk of depression. You must analyze whether the post clearly shows depression or subtly suggests depressive tendencies through word choice, phrasing, or viewpoints. Base on your analysis, assess whether the user has depression.
Input (Post Content)	我讨厌钢琴，我讨厌给学生上课，我讨厌所有东西，我好累。我感觉身上手上像挂了好多东西。我发现我对我一年多从启蒙带起来的十几个孩子都一点不留恋，我感觉很无所谓——想要我教就教 不要我 ok 我无所谓。星期五我的一个5岁很内向平时完全不喜欢说话但是特别聪明的学生，自己生病了2周好不容易来了知道我也生病请假了，给我做了一个贺卡，上面有一句很简单的祝福注意身体，天天开心还配了一个画，一个女孩子放风筝。对小朋友来说，放风筝就是开心。小朋友的想法很简单，但是这个简单但是很真诚的祝福我百感交集。我手上很多很内向，我当时接触花了特别多心思和办法一点点和他们熟络，然后让他们接受这个其实是很差的我。现在我感觉 我播下的种子开始发芽长大，移到别人的花盆里也无所谓。.
Output (Label)	"Depressed"
Explanation for Readers	A summary of the post content is: "As a junior student, exhausted and detached, the author reflects on teaching, emotional numbness, and a child' s sincere gesture triggering mixed feelings." Clearly, the post includes chronic fatigue, emotional detachment, lack of joy, apathy toward students, and disproportionate emotional response signal underlying depressive symptoms

Figure 12: A Chinese example of multilingual local instruction-tuning data for MuLA-F and FedLoRAs.

FlexLoRA: A novel FedLoRA, SOTA for clients with heterogeneous data or resources. FlexLoRA performs global SVD on the server side and, based on the characteristics of local data in terms of statistical distribution and resources, assigns different low-rank reconstruction matrices to each client.

FedLFC: A recent SOTA multilingual FedLoRA based on language clustering. In the original task scenario of FedLFC, multilinguality only exists from the server's perspective, i.e., each local dataset is monolingual. FedLFC performs multi-center aggregation on the low-rank reconstruction matrix of each local LoRA block based on its language family. Note that when selecting baselines, we skip FedHLT (Guo et al., 2024b) because FedLFC and FedHLT have a strong theoretical relationship, with the latter being a lower-level alternative to the former.

B.4 Implementation Details

In our experiment, for each local client, we set the rank of LoRA to 16 and the LoRA- α to 32. In each federated round, the local client performs two LoRA tuning epochs, followed by disentanglement and upload. We set the number of selected singular values in MuLA-F as 8. For the two orthogonalization coefficients, we assume that the absolute values of α and β are equal, and then conduct a grid search for the optimal setting from the set 0.1, 0.5, 1, 5. We also perform a grid search for the learning rate in the range 1e-4, 5e-4, 1e-3, 5e-3. We set the maximum federated communication round as 20, with an early-stopping patience as 5. All experiments are carried out using two NVIDIA A800 80GB GPUs.

Task Type	Multilingual Depression Detection
Post Content Language	Japanese
Client ID	1 (in MU3D-Sp-1)
Local Language Composition	Japanese, English
Task Instruction	You will receive a social media post written by a Japanese user who is at risk of depression. You must analyze whether the post clearly shows depression or subtly suggests depressive tendencies through word choice, phrasing, or viewpoints. Base on your analysis, assess whether the user has depression.
Input (Post Content)	がんばったんだけどなあ。 皆に楽しんで欲しくて、言われるままに飲んだし、出来る限りのことを尽くしたんだけど、あれじゃあダル絡みだから、ダメだって。
Output (Label)	"Depressed"
Explanation for Readers	The English translation is: "I really tried my best.I just wanted everyone to have fun, so I drank as they told me to, and did everything I could. But then they said it was just annoying drunken rambling and that it was no good." Clearly, she feels unappreciated despite trying hard to please others and being criticized harshly.

Figure 13: A Japanese example of multilingual local instruction-tuning data for MuLA-F and FedLoRAs.

C Additional Discussions

We provide further clarification and discussion on certain statements in the Methodology and Limitation sections that may cause confusion.

C.1 Why Diff-eRANK + SVD ?

In this part, we discuss our theoretical motivation regarding why we combine Diff-eRank and SVD for client-side multilingual knowledge disentanglement.

According to the theoretical analysis provided by the original Diff-eRANK paper (Wei et al., 2024), if after model weight updates (Fine-tuning), a post's token representations become highly structured or compressed, we can conclude that this update reduces the uncertainty in the representation space (from an information-theoretic perspective) and removes redundant information irrelevant to general tasks (from an empirical perspective). It also implies that the model can more effectively extract patterns and regularities from the data. Moreover, it is well acknowledged that a weight matrix can be reconstructed through SVD decomposition into a linearly independent combination of several low-rank matrix components (as does $B \times A$), where each low-rank matrix can be regarded as a feature which represents a direction/semantic

Task Type	Multilingual Depression Detection
Post Content Language	Korean
Client ID	2 (in MU3D-Sp-1)
Local Language Composition	Korean, English
Task Instruction	You will receive a social media post written by a Korean user who is at risk of depression. You must analyze whether the post clearly shows depression or subtly suggests depressive tendencies through word choice, phrasing, or viewpoints. Base on your analysis, assess whether the user has depression.
Input (Post Content)	자존감은 일상의 성실로부터 온다 하세요. 걍 하라고. 생각하지 말고 그냥 하세요. 어디 불 지르고 사람 찌를 일 아니면 그냥 하라고. 일단 하세요. 아침에 일어나면 그냥 세수하세요. 세수하고 나서 뭐라도 입에 집어넣으세요. 창문 열고 환기하고 커튼 쳐서 햇빛 들어오게 하세요. 바깥 풍경 잠깐 보다가 바로 책상에 앉고 공부 시작하세요. 눈에 안 들어온다고 다시 내려오지 말고, 일단 계속 책상에 앉아 있으세요. 그것만 해도 잘한 거임.
Output (Label)	"Depressed"
Explanation for Readers	A summary of the post content is: "Build self-esteem through small daily actions—don't overthink, just start, keep going, and recognize effort as meaningful progress." Clearly, the post includes repeated self-persuasion, low motivation, reliance on minimal tasks indicate depressive coping and internal struggle.

Figure 14: A Korean example of multilingual local instruction-tuning data for MuLA-F and FedLoRAs.

(which can also be understood as neurons). Hence, we discover an interesting collaboration between Diff-eRANK and SVD: Suppose a local client's data comprises three languages, and given the asymmetry function of LoRA as demonstrated in the paper "Asymmetry in low rank adapters of foundation models" (Zhu et al.) (also demonstrated in Figure 5 in our paper) the expected role of A-Matrices in Fed-LoRA is inherently "more effective extraction of patterns and regularities from data, reducing uncertainty in the representation space, and isolating general features relevant to specific tasks", which is highly similar to the focus of Diff-eRank. The reconstructed low-rank matrices are orthogonal to each other, naturally leading one to consider, "how much each low-rank matrix contributes from this perspective to the local data of each language."

Additionally, previous works on lifting the multilingual curse, such as the paper "Lifting the curse of multilinguality by pre-training modular transformers", have already provided clear empirical conclusions: Despite the overlap and conflict between domain adaptation knowledge across various languages, they can be disentangled during the PEFT process through modularization (and the reconstructed low-rank matrices are themselves in an overly disentangled state). Thus, in this client, a logically sound reasoning is that the knowledge associated with each language can be approximated as a combination of several selected reconstructed low-rank matrices, to simulate an appropriate level of disentanglement. This selection, as mentioned, is aptly handled by Diff-eRank in our task scenario. For each local language, low Diff-eRank score matrices can be seen as a concrete representation of the multilingual curse.

Table 7: Quantitative evaluations for "over-decoupling". (Metrics: Fed-F1.)

Method	MM-COVID Sp-1	MM-COVID Sp-2
FedSA	92.47	90.44
FlexLoRA	92.09	89.60
FedLFC	92.85	90.11
MuLA-F	92.66	91.24
MuLA-F-C	93.01	91.69

C.2 An Illustration of "Over-Decoupling"

In the Limitation section, we express a potential concern that MuLA-F might lead to over-decoupling across languages. In this part, we aim to quantitatively evaluate the possible impact of the concern. First, we'd like to give a more detailed explanation of the concern. In multilingual experimental settings, due to linguistic features and other reasons, the affinity/differences between languages could vary. Some languages may share part of vocabularies, grammatical structures, or exhibit a high degree of similarity in expression patterns (especially in a specific task scenario), thus having many shared features. In such cases, they are more suited to share a single language center, rather than having separate ones.

In light of this, we conduct an additional experiment to evaluate its potential impact. For the MM-COVID dataset, we keep data in only three languages: French, Portuguese, and Spanish, to construct a degraded version. On this degraded MM-COVID, we created a variant of MuLA-F, namely MuLA-F-C, where considering the strong affinity between Portuguese and Spanish, we build only two language centers: one for French, and one shared by Spanish and Portuguese. Using Qwen-2.5-14B as the base model, our experimental results are shown in Table 6 (FedSA uses the same language division as MuLA-F-C, while FedLFC and FlexLoRA are unaffected).

The experimental results reported in Table 7 show that, due to the excessive disentangling and decoupling of the A-Matrices related to Spanish and Portuguese, the performance of MuLA-F is not as good as that of MuLA-F-C, which shares this domain adaptation knowledge. Moreover, this phenomenon is not unique to MuLA-F (e.g., also appeared in FedSA). However, the impact is still acceptable in our settings.

C.3 Motivations of Our Orthogonal PEFT Strategy

In Para 3.3, we convert the process of "sequentially aggregating weights for each language center" into an approximated continual learning process. Furthermore, supported by the theoretical analysis provided by O-LoRA (Wang et al., 2023b), MuLA-F ensures that the feature subspaces occupied by each global language center are more orthogonal (less overlapping) to each other. Consequently, the orthogonality further extends to the language-specific reconstructed A-matrices in each local client, thereby reducing the catastrophic forgetting that might be caused by multilingual conflicts. The theoretical robustness of Orthogonal LoRA PEFT has also been demonstrated in the original paper of O-LoRA. For B-matrices, however, we utilize the opposite insight (as shown in Figure 5 in our paper, a local B matrix should generally be responsible for feature transformation for the downstream task).

C.4 Problem Formulation of FedLoRA

We consider a federated learning setting with n clients collaboratively finetuning a LLM base model for a classification task (i.e. multilingual SNS content anomaly detection in our paper). Each client $j \in \{1, 2, \ldots, n\}$ holds a private local dataset $\mathcal{D}_j$ of size $|\mathcal{D}_|$|, which may be non-iid across clients. To reduce communication and memory costs, FedLoRAs adopt Low-Rank Adaptation (LoRA) for fine-tuning a shared pretrained model f_θ. Rather than updating the full model parameters, in each federated round, each client learns a pair of low-rank matrices (A_j, B_j) on the local-side, and the local effective adaptation is given by $\Delta_j = B_j A_j$.

The overall goal is to collaboratively learn a global LoRA update across clients. The most common pipeline is, following the FedAvg paradigm, in each communication round, clients locally compute Δ_j based on their data and send it to the server. The server then performs a weighted aggregation of these updates:

$$\Delta_{\text{global}} = \frac{1}{\sum_{j=1}^{n} D_j} \sum_{j=1}^{n} D_j \cdot \Delta_j,$$

and broadcasts Δ_{global} back to all clients. Each client then updates its local model using this aggregated low-rank adaptation on top of the fixed pretrained weights θ.

The objective is to minimize the average empirical loss over all clients:

$$\min_{\Delta_{\text{global}}} \frac{1}{n} \sum_{j=1}^{n} \mathbb{E}_{(x,y)\sim\mathcal{D}_j} \left[\ell(f_{\theta+\Delta_{\text{global}}}(x), y) \right],$$

while ensuring collaborative domain adaption and preserving data privacy. In our paper, the metrics is set as federated F1-score, which is written as:

$$\text{Fed-F1} = \frac{2 \cdot \sum_{j=1}^{n} |\mathcal{D}_j| \cdot \text{Precision}_j \cdot \text{Recall}_j}{\sum_{j=1}^{n} |\mathcal{D}_j| \cdot (\text{Precision}_j + \text{Recall}_j)}.$$

Nevertheless, in many cutting-edge FedLoRA variants, B_j and A_j are separately processed, shaped, transformed and exchanged either on the client-side or on the server side.

C.5 Ethics Statement

In this section, we provide a detailed discussion of the ethical considerations involved in our work, with a particular focus on two main aspects: the use of A.I. assistants in the writing process and the handling of data ethics in our experimental design. We believe that addressing these issues explicitly is essential to ensure transparency, uphold academic integrity, and align with the ethical guidelines of the research community.

With respect to the A.I. assistant, all innovations and arguments presented in this paper are entirely authored by the researchers. GPT-4o is only employed for limited proofreading and grammar checking during the writing process, which is fully compliant with the ARR submission guidelines.

Regarding data ethics, both the MM-COVID and CONAN datasets undergo thorough desensitization by their original authors prior to release. For the MD3D dataset, we carefully remove all potentially sensitive information—such as IP addresses, usernames, and user profiles—from the portion collected from open-source platforms. We are confident that all our experiments strictly adhere to ethical policies.