IRSTI 28.23.02 https://doi.org/10.26577/jpcsit42202610

**Alisher Kaziz[1*], Richard Harrison[2]**

[1]Astana IT University, Astana, Kazakhstan
[2]University of Surrey, Guildford, UK
*e-mail: alisher.kaziz@gmail.com

# KNOWLEDGE FUSION THROUGH DATALESS BERT PARAMETER MERGING VIA A REGRESSION-BASED AVERAGING

**Abstract.** Recent progress in dataless knowledge fusion methods has focused on migrating knowledge from a teacher neural network to a student neural network without utilizing the teacher model's training data. As a result, fine-tuning pre-trained language models (PLM) has emerged as a simple method to improve the performance of Natural Language Processing (NLP) models in specific domains. These refined models are available as open source, but typically their training datasets are not, due to issues related to data privacy or intellectual property. This sets up an obstruction to combining knowledge from separate models to produce an enhanced single model. In this paper, we investigate the issue of combining separate models developed on distinct training datasets to create a unified model that performs on out-of-domain data for the student model. We suggest a dataless knowledge fusion technique that integrates models within their parameter space, directed by weights that reduce prediction discrepancies between the combined model and the separate models. Across a wide range of parameter configurations, our assessment indicates that the suggested approach substantially exceeds the performance of traditional baselines like Fisher-weighted averaging or model ensembling. Additionally, we observe that our approach serves as a viable alternative to multi-task learning, capable of maintaining and enhancing the individual models without requiring access to the training data. Our experiments confirm the dataless merging methods, integrating finely adjusted parameters to achieve robust multitask performance with negligible impact on class ratio degradation, approximately 2.1 %.

**Keywords:** NLP, RegMean, Knowledge Fusion, Dataless NLP, Kazakh NLP, BERT.

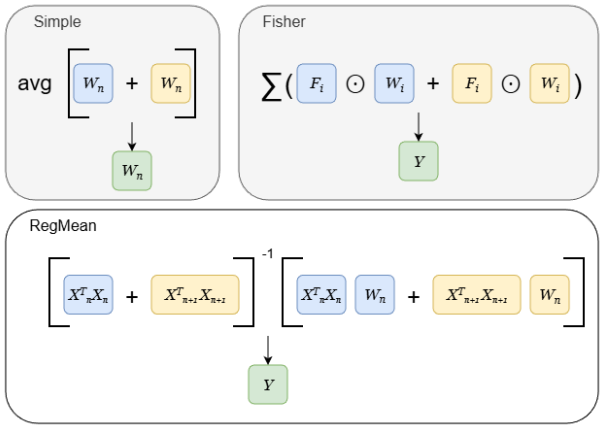
## 1. Introduction

Current trending paradigm for developing NLP model to solve tasks in specialized domains is to fine-tune a PLM, such as BERT, on task-specific labelled data [1]. This approach outperforms PLM, but it requires producing separate models for each language, domain, and task, which is excessive to maintain and difficult to deploy in real-world pipelines [2], [3]. These challenges tend to arise during the design of NLP models in low-resource settings, such as legal corpora, which are often protected by confidentiality constraints, where annotation is expensive, and cross-lingual transfer is limited by distribution changes of corpus and terminology mismatches [4], [5]. The use of multi-domain and multi-lingual data can partially solve these problems through shared representations across tasks and languages. However, several problems are associated with this approach [6]. First and foremost, centralised exposure to the underpinning labelled datasets is frequently essential for joint training, which may not be possible due to administrative, regulatory, or security constraints [7]. Following that, discovering which task combinations help, could become computationally prohibitive as the number of candidate sources grows [8]. These constraints motivate research in dataless approaches that are able to combine PLMs knowledge that are encoded in independently fine-tuned models without revisiting the original training data.

Model merging offers such a mechanism by combining multiple models directly in parameter space, producing a single model that can be used for inference across domains or languages without additional training [9]. In the context of cross-lingual knowledge fusion, merging is an appealing alternative to ensembling, it retains the development simplicity of a single model while enabling the integration of complementary expertise from multiple sources [10]. Moreover, merging is compatible with privacy-preserving workflows (such as federated or siloed fine-tuning), in which input models exchange parameters or updates rather than the underlying training data [10], [11].

Recent research has clarified both the promise and the limitations of dataless parameter merging for modern NLP systems [12]. Firstly, work on model soups and related weight-averaging approaches show that combining independently fine-tuned checkpoints can yield consistent gains when the models share the same architecture and originate from the same pre-trained initialization, suggesting that parameter-space fusion can work as a cheap alternative to retraining [13], [14]. Secondly, a line of work on task arithmetic and editing vectors indicates that certain task-specific changes can be approximated as linear directions in weight space, enabling controlled composition of capabilities under compatibility assumptions [15]. Thirdly, there are merging methods that incorporate curvature or dependency information [16]. For example, Fisher-weighted merging and regression-based merging such as Regression Mean (RegMean) and its extensions tend to be more robust-compared to naive averaging when source models are heterogeneous (Fig. 1). RegMean shows better account for interference between parameters in dynamics of weight changes to corpus ratio [2], [16].

**Figure 1**
*Comparison between Simple, Fisher, and RegMean parameters merging methods*




At the same time, large-scale evaluations on open-weight large language models highlight that merging heuristics are not universally reliable and can fail to transfer across model families, motivating careful validation in each target setting [17]. In parallel, multilingual evaluation efforts continue to emphasise persistent gaps for low-resource languages, even when using strong multilingual pre-trained models [18]. We also combine findings from semantically similar language (Turkish to Kazakh) to estimate the probability of label classes for Weighted-F1 [19], [20]. Together, these findings motivate our choice of a closed-form, interference-aware merging approach (RegMean) and our focus on cross-lingual and domain robustness in the domain [2].

The limitation can be attributed to secrecy, licensing limits, or organisational regulations, and it is typical in legislative materials, which may include delicate private or professional data [21], [22]. Our objective is to fuse NLP models by exploiting knowledge currently contained in present checkpoints, without the need for new labelled information or joint reconfiguration, to provide a single operational model that supports Kazakh legal NLP. Legal language makes cross-linguistic generalisation more difficult. Writing for legal purposes is characterised by extensive and complicated phrases, numerous allusions to legislation and case identification, and varying formulaic structures by jurisdiction [13]. The lack of high-quality annotated legal data sources and tooling in Kazakhstan makes it expensive to develop and upkeep distinct models for each activity and subsystem [18]. A dataless merging strategy combines tuned models into a single model while maintaining a straightforward implementation mechanism [23]. This study is guided by the following research questions (RQs):

- RQ1: How sensitive is the merged model to the choice and number of source checkpoints (e.g., Kazakh-only, multilingual, or legal vs. general-domain)?
- RQ2: What is the impact of merging on in-domain accuracy and out-of-domain robustness?

We only consider merging checkpoints for models that use the base architecture and start with the same pre-trained model. This makes it easy to align the model parameters and merge them directly. This choice keeps the parameters aligned and allows closed-form merging. Studies show engineering teams utilise PLMs [2], [18]. They take a pre-trained BERT model and create different versions for specific tasks or languages. A shared pre-trained BERT checkpoint is taken as a base, and multiple task- or language-specific variants are then produced. Within these conditions, the method suits teams that already run several fine-tuned models and want an inexpensive way to unify them

into a single model for use in various domain processing pipelines.

In this paper, we study Knowledge Fusion via Dataless BERT Parameter Merging with a focus on Kazakh legal modelling. We adopt RegMean as a practical, closed-form weight-merging method that can be applied to multiple fine-tuned checkpoints sharing the same architecture and initialization. We investigate how RegMean can fuse knowledge from models trained on different languages to yield a single BERT-based model that (i) maintains strong in-domain performance, (ii) improves robustness under cross-lingual and domain shift, and (iii) remains parameter-efficient compared to maintaining separate tuned models. Our contributions are three-fold:

1. we formulate a dataless cross-lingual model merging setup for legal NLP with an emphasis on Kazakh language;

2. we provide a systematic empirical evaluation of RegMean-based BERT parameter merging for cross-lingual and domain transfer in legal tasks;

3. we analyse the practical trade-offs of merged models;

## 2. Materials and methods

This section describes the data sources, model setup, and experimental protocol used to assess dataless parameter merging for Kazakh legal NLP. We summarize the source checkpoints, the RegMean merging procedure, and the downstream evaluation settings so that the experiments can be reproduced and interpreted in a consistent way. Our workflow follows three stages: (i) independently fine-tune multiple BERT checkpoints on language- and domain-specific datasets, (ii) merge the resulting parameters in a dataless manner using RegMean, and (iii) evaluate the merged model on Kazakh legal NLP tasks under both in-domain and cross-domain or cross-lingual conditions (Fig. 2). Importantly, the merging stage does not require any labelled data.

We use a BERT encoder as the shared backbone. All source checkpoints are required to have the same architecture (number of layers, hidden size, attention heads) and the same pre-trained initialization. This constraint ensures that corresponding tensors have identical shapes and can be merged directly in parameter space. We construct a pool of source models by fine-tuning the common backbone on datasets that differ in language (Kazakh, and multilingual) and domain (legal and general domain). Each source model is trained independently within its data silo, only the final model parameters are used for merging. We evaluate on Kazakh legal NLP tasks relevant to downstream document processing. Depending on availability, these include classification of legal documents or sentences, named entity recognition (NER) for legal entities, for which we use KazBERT and supplement it with our own manually labelled legal dataset to better cover domain-specific entities [24]. For each task, we define an in-domain test set and, when possible, an out-of-domain or cross-lingual test set to assess robustness.

Let set $\{\theta_i\}_{i=1}^{M}$ denote the parameters of $M$ fine-tuned source models sharing initialization $\theta_0$, and define $\Delta\theta_i = \theta_i - \theta_0$. RegMean computes a merged update $\Delta\theta$ by minimizing the functional discrepancy between intermediate representations.

$$\Delta\theta^* = \arg\min_{\Delta\theta} \sum_{i=1}^{M} \mathbb{E}_{x\sim\mathcal{D}} \left[ \left\| f_{\theta_0+\Delta\theta}(x) - f_{\theta_0+\Delta\theta_i}(x) \right\|_2^2 \right] \quad (1)$$

assume that for a fixed layer the representation is locally linear around $\theta_0$:

$$f_{\theta_0+\Delta\theta}(x) \approx f_{\theta_0}(x) + J(x)\Delta\theta \quad (1)$$

where $J(x)$ is the Jacobian of the layer output with respect to parameters. Then the objective in equation (1) is convex in $\Delta\theta$ and its unique minimiser is:

$$\Delta\theta^* = (\mathbb{E}_{x\sim\mathcal{D}}[J(x)^\top J(x)])^{-1} \left( \frac{1}{M} \sum_{i=1}^{M} \mathbb{E}_{x\sim\mathcal{D}} [J(x)^\top J(x)\Delta\theta_i] \right) \quad (2)$$

that $\mathbb{E}_{x\sim\mathcal{D}}[J(x)^\top J(x)]$ is positive definite in equation (2). Under the linear approximation,

$$f_{\theta_0+\Delta\theta}(x) - f_{\theta_0+\Delta\theta_i}(x) = \\ = J(x)(\Delta\theta - \Delta\theta_i) \tag{3}$$

then, substituting into equation (1) yields

$$\mathcal{L}(\Delta\theta) = \\ \sum_{i=1}^{M} \mathbb{E}_{x\sim\mathcal{D}}\left[(\Delta\theta - \Delta\theta_i)^\top J(x)^\top J(x)(\Delta\theta \right. \\ \left. - \Delta\theta_i)\right] \tag{4}$$

and were we define equation (4) as

$$H = \mathbb{E}_{x\sim\mathcal{D}}[J(x)^\top J(x)] \tag{5}$$

then,

$$\mathcal{L}(\Delta\theta) = \sum_{i=1}^{M} (\Delta\theta - \Delta\theta_i)^\top H(\Delta\theta - \Delta\theta_i) \tag{6}$$

Since $H$ is positive semidefinite, the objective is convex. If $H$ is positive definite, the objective is strictly convex and admits a unique minimiser.

Taking the gradient and setting it to zero:

$$\nabla_{\Delta\theta}\mathcal{L} = 2\sum_{i=1}^{M} H\,(\Delta\theta - \Delta\theta_i) = 0 \tag{7}$$

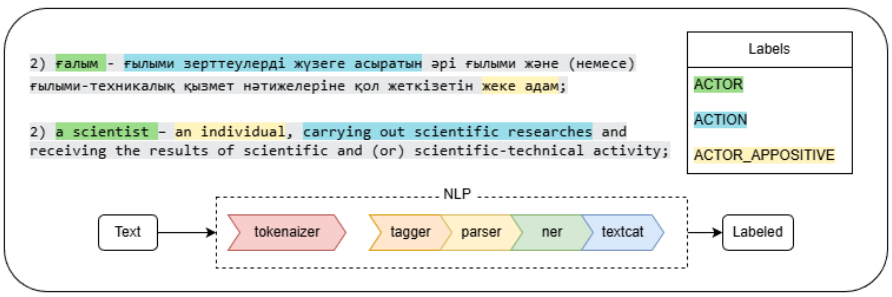
this implies,

$$MH\Delta\theta = H\sum_{i=1}^{M} \Delta\,\theta_i \tag{8}$$

Multiplying both sides by $H^{-1}$ yields the stated solution. In practice, expectations over $\mathcal{D}$ are approximated using a representative unlabelled corpus drawn from the target domain. Merging is performed once per source set and does not involve additional gradient-based training.

We compare RegMean to common alternatives: (i) selecting the best single source checkpoint per task, (ii) uniform weight averaging over source checkpoints (model soups), and (iii) prediction ensembling when applicable. These baselines separate gains from parameter-space fusion versus multi-model inference. Reported fine-tuning methods (optimizer, learning rate schedule, batch size, number of epochs, maximum sequence length) and model selection criterion ensured comparability, and a consistent fine-tuning procedure was implemented across source models within each experimental setting. We use standard task metrics (e.g., accuracy and macro-F1 for classification; entity-level weighted-F1 for NER). In addition to average performance, we analyse cross-lingual transfer to Kazakh language, robustness under domain shift. We apply a consistent NLP pipeline. This includes document segmentation, sentence splitting, and normalisation of common legal artifacts (e.g., article references, case numbers, dates, and monetary amounts). Where the downstream task benefits from it, we optionally anonymize personally identifiable information (PII) with placeholders to reduce leakage and to better simulate privacy-preserving deployment constraints. Since Kazakh language can be sensitive to subword segmentation quality, we specify the tokenizer used for all experiments and keep it fixed across source checkpoints and merged models. When source checkpoints use different tokenizers, we treat them as incompatible and do not merge them (Fig.2).

**Figure 2**
*Tokenizer for all evaluated models in cross lingual domain to keep minor bias on NER and entity relation tasks*



Study defines criteria for selecting source checkpoints for merging, including minimum validation performance, domain relevance, and language coverage. To mitigate negative transfer, we exclude checkpoints that are substantially misaligned with the target domain and report sensitivity to these selection decisions. All experiments are conducted with fixed random seeds. To analyse the conditions under which merging succeeds or fails, we examine (i) sensitivity to the number and composition of source checkpoints, (ii) robustness to variations in fine-tuning hyper-parameters, and (iii) qualitative error patterns on Kazakh legal data.

## 3. Results

Macro-F1 characterises class-balanced effectiveness by weighting each class equally. Weighted-F1 include observed class frequencies and therefore reflects performance under label imbalance. While we introduce MCC, which provides a correlation-based measure that remains informative when legal label distributions are skewed, validation across the metrics reported in Table 1. RegMean-merged model shows the strongest aggregate performance. When Macro-F1 assessment was conducted to measure specialised model performance with tuned RegMean model, the comparative analysis shows modest improvement in various scenarios. For instance, average improvements by 2.9 points with MCC by 0.05 was indicated. Even if prediction performance was enhanced, study reveals test scenarios where classes was not balanced specialised model outperform RegMean approach. Similar behaviour was revealed in Jin's research [2]. One reason why unbalanced classes scenario have declined, because it is RegMean method nature that merges parameters based on casual regression. The most likely causes of F1 poor gains is underestimation of parameters weights in RegMean merging process. The fact that Weighted-F1 is going up means that RegMean is still doing well with the majority class and it is also getting better at handling the minority labels.

In the Table 1 each model results are reported as the mean standard deviation across multiple random seeds. Higher Macro-F1 and MCC values correspond to better class-balanced effectiveness and greater overall consistency of predictions.

**Table 1**
*Performance comparison of specialised models and tuned models with RegMean approach for the Kazakh legal document classification task*

| Model | Macro F1 (± std) | Weighted F1 (± std) | MCC (± std) |
|---|---|---|---|
| KazBERT | 83.9 ± 0.6 | 85.2 ± 0.5 | 0.81 ± 0.01 |
| Multilingual Legal BERT | 81.5 ± 0.7 | 83.1 ± 0.6 | 0.78 ± 0.02 |
| Uniform Weight Averaging | 84.4 ± 0.5 | 85.9 ± 0.8 | 0.83 ± 0.01 |
| Prediction Ensemble | 85.7 ± 0.1 | 87.0 ± 0.3 | 0.85 ± 0.01 |
| RegMean (Merged Model) | **86.8 ± 0.3** | **87.3 ± 0.2** | **0.86 ± 0.01** |

The essence of RegMean approach is to combine models parameters to cover under-tuned space. Models fused together performed notable results in various test scenarios. While common methods of models tuning require holding each version and running many checks to breach the bottleneck of predictions thresholds, on other than RegMean method uses just one set of settings and gets results that are modest or outperform common methods in dataless environment. This shows that combining settings from models can bring together useful knowledge from models trained on different languages and areas without needing the original training data. RegMean approach is designed to perform in a specific area and keeps steady balanced across different domain categories by focusing on inter-domain accuracy and maintaining class-balanced results. Merged model keeps results steady on parent tuned models datasets and attains confident results on new task in a edge of the parent models domains. RegMean comes close to the effectiveness of prediction ensembling, while it retains the operational convenience of deploying a single model. Obtained results indicate that merging increases in-domain accuracy while preserving class-balanced performance across domain categories, with RegMean approach. As reported in Table 1, MCC shows an 8% relative gain over the Legal BERT baseline, alongside an approximately 5% relative improvement in Macro-F1. Earlier work on Fisher-weighted merging has typically

documented relative F1 increases of about 2 to 3% in heterogeneous scenarios, whereas the present cross-lingual legal setting exhibits a moderately larger effect size [16]. On the other hand, while gathered metrics reveal improvements model performance tuned with RegMean approach further imply that the benefits are not confined to majority classes. Instead, they point to broader gains in learned representations and in the quality of the induced decision boundaries. The relative improvement in minority-class F1 ($RI_c$)with respect to the baseline is summarised in Table 2:

**Table 2**
*Per-class F1 comparison between KazBERT and RegMean on the legal classification task*

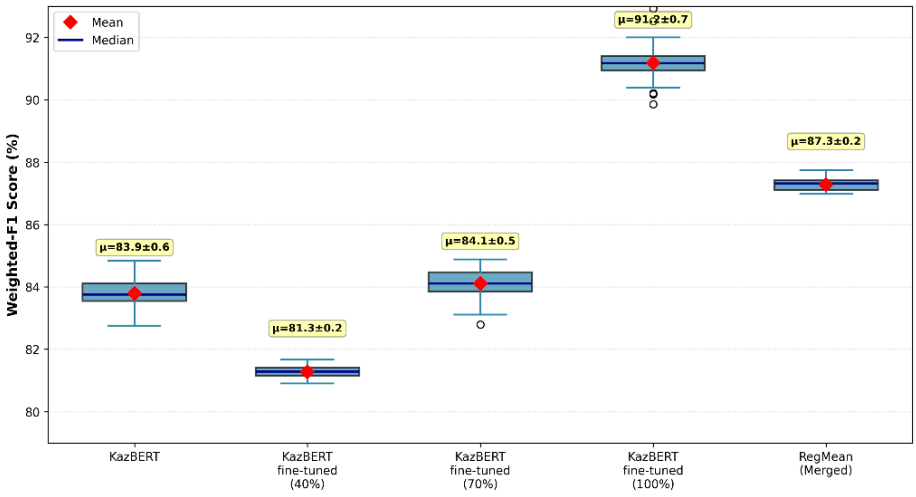
| Class | Sample size | KazBERT F1 | RegMean F1 | $RI_c$ % |
|---|---|---|---|---|
| PERSON | 13577 | 88.2 ± 0.1 | 88.7 ± 0.2 | +0.6 |
| LAW | 5693 | 84.2 ± 0.2 | 89.1 ± 0.3 | +5.9 |
| NORP | 3714 | 88.7 ± 0.5 | 87.3 ± 0.3 | -1.0 |
| PROJ | 2111 | 84.7 ± 0.2 | 86.4 ± 0.1 | +2.1 |

where,

$$RI_c = \frac{F1_{RegMean} - F1_{Baseline}}{F1_{Baseline}} \times 100\% \qquad (9)$$

In the Table 2, relative $\Delta$ reports the percentage change of RegMean over KazBERT for each label, highlighting where merging improves minor classes. Study shows minor improvements in couple classes, performance increasing caused from one of source model which specialised on law labelling. Our study also reports a comparison of Weighted-F1 scores for Kazakh legal classification using box plots as shown in Figure 3. We look at how the KazBERT model works when it is given amounts of labelled data to learn from. We compare the KazBERT model with versions that have been fine-tuned using 40%, 70% and 100% of the available training data and also the RegMean merged model. The box plots show how the results are spread out when we use random seeds, which helps us see both how well the model works on average and how much the results can vary. When we have labelled the data, the KazBERT model learns from it better. If Kazakh legal classification KazBERT model is trained using 40% of the dataset Kazakh legal classification KazBERT model gets a Weighted-F1 score of 81.3 give or take 0.2. In Figure 3 median line indicates typical performance, while boxes and whiskers show variability across random seeds.

**Figure 3**
*Box plots of Weighted-F1 under different data regimes (40%, 70%, 100%), the baseline KazBERT model, and the RegMean merged model*

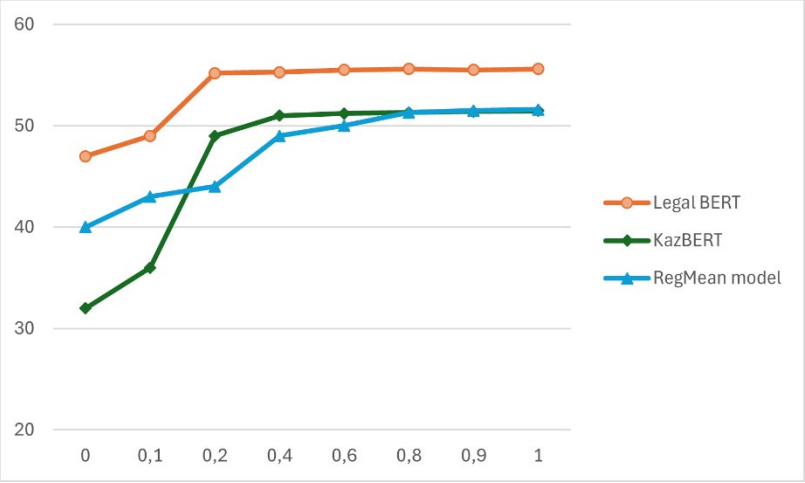


Box plot was introduced to show comparison of training data shortage and dataless RegMean approach. The box plot examine a range of values and short lines suggesting that the results are similar throughout the runs. However, this similarity means that the model is not very good at generalizing when it does not have data to learn from. The model does not learn well because it has not seen different types of legal issues and ways of writing. When we add data to 70% the models performance gets much better reaching 84.1 with a small variation of 0.5. The box plot for this data shows that the middle value is higher and the range of values is wider compared to when we had 40% of the data. However, even if variance across seeds does not mean the model explores more solutions, this wider range applicable in term of possible decision variance affected by low-resources settings. Even though the model does better with data it also becomes more sensitive to how it is set up and random changes.

Tuned KazBERT model on 100% of dataset outperforms RegMean tuned approach. Using all the data gives the results with a performance of 91.2 and a variation of 0.7. This supports our first research question. The box plot shows that the middle value of Weighted-F1 is higher than with data. However the range of values is also the widest among the models that were fine-tuned. The wider range of values and longer lines in the box plot show that the models performance varies across different runs. This is what happens when engineer tune a model, then they suggest that more data helps it to performance increasing on a task but it also makes the model more sensitive to changes [17]. The model could be tuned to bypass validation tests, however its performance varies more, between runs on various datasets.

KazBERT model shows modest scoring 83.9 with some room for improvement, common method to improve performance is to adjust dataset, this modest results was shown in between the 70% and 40% configurations in terms of dataset volume from complete one. The results for this model vary a bit. When we merge the model using RegMean, it scores 87.3. This model is stable with results close to the average, which indicate close or low variance among probes. It works well as fine-tuning 40% of the model but it shows underperformance of accuracy. RegMean does not outperform the model that's fully fine-tuned but it is a good balance, between accuracy and stability. Compared to a model that's 70% fine-tuned RegMean is more accurate and stable. RegMean perform notable improvements in term of performance and it has less variation by compared to the KazBERT model. RegMean is 3.4 points better and has a lower standard deviation. The 100% fine-tuned model achieves the highest peak performance but also shows the greatest variability (±0.7). In tested environments, such variability can translate into unpredictability during retraining or model updates. In contrast, RegMean demonstrates consistent performance across seeds (±0.2), indicating robustness to stochastic training effects.

**Figure 4**
*NER performance as a function of the RegMean mixing coefficient α, the curve highlights the stability region where merging yields consistent gains*



The narrow box distribution suggests that parameter-space merging aggregates complementary information from multiple checkpoints in a manner that smooths individual optimisation noise. This variance reduction is consistent with findings in large language models merging studies, where merged checkpoints demonstrate lower run-to-run dispersion compared to independently fine-tuned models [17]. In our case, standard deviation decreases from $\pm 0.7$ in full-data fine-tuning to $\pm 0.2$ in the merged model. This behaviour can be interpreted as a form of implicit regularization. To evaluate NEW performance we applied mixing coefficient approach for both our baseline models and RegMean tuned model, this approach fit to quantify sensitivity in sequence labelling (Fig. 4). Steady point and a plateau after $\alpha$ 0.7 indicate that the merged model is not overly sensitive to the exact parameters of baseline models but benefit from weighting elements. This suggests that RegMean provides a stable operating region in which practitioners can select $\alpha$ without extensive hyper-parameter tuning, while still retaining gains over individual baselines. Fine-tuning adapts parameters along a single optimisation trajectory, potentially converging to different local minima depending on initialisation. Parameter merging, by contrast, combines independently learned updates, effectively averaging task-relevant representations while attenuating idiosyncratic deviations. The resulting model occupies a more central region of the loss landscape, which may explain its reduced variance.

## 4. Discussion

The analysis of data efficiency shows a striking difference between RegMean and standard tuning on the reduced datasets. In particular, the training KazBERT on 40% of the entire data results in about 6 point drop in performance compared to the RegMean merged model. Even increasing the training set to 70% does not help, as the tuned model is still three points behind. This is somewhat ironic because the cross-lingual transfer in low-resource already usually yields around 2 to 5 points improvement.

Dataless parameter merging aside, this clearly demonstrates that the merge of parameters without data is also a way to get labelled data that is not available and it is the most effective way to take advantage of the knowledge from different models. While fine-tuning on a complete dataset ultimately yields the best performance, compiling and annotating extensive legal datasets is cost-prohibitive, especially for low-resource languages such as Kazakh. Therefore, unlike the common methods, RegMean provides solution that is both competitive and repeatable across runs.

The box-plot analysis reveals the following important points:

- First, the more training data we use, the better performance we can attain, but the greater the variability in the performance.
- Second, RegMean is effective in the trade-off between accuracy and stability. It surpasses the performance of fine-tuning with less data while also having significantly less variance.
- Third, the combination of the parameters makes the final training process less sensitive to the initial optimization conditions.

The implications of this observation are that not merely is dataless parameter merge an approach for acquiring knowledge of parameters, it is a method for increasing the robustness of a model. When data is extremely scarce in the labelled set, the justified power of consistent low-variance results is significant.

The NER $\alpha$-sweep (Fig. 4) is yet another indicator of robustness in results. We see a fairly broad performance plateau, which implies that the advantages of merging are not overly hyper-sensitive coefficient tuning. This is a significant advantage in deployment when applying merged models to new tasks or domains. This is naturally consistent with our results in classification tasks, confirming that sequence labelling tasks benefit by a similar level of robustness.

## 5. Conclusions

The consolidation of cross-lingual knowledge in low-resource legal NLP environments where training data is extremely limited is a key topic of this study. We designed a dataless merging framework that directly integrates independently trained BERT checkpoints in the parameter space through RegMean strategy. Notably, this approach does not require shared datasets or joint retraining which is often computationally costly.

Our experiments on Kazakh legal classification have demonstrated that the interference aware merging approach consistently outperforms individual specialist models and baseline parameter

averaging methods. By generating one single checkpoint for deployment, this method not only shows the advantages of model ensembles but also retains tunability. The resulting merged model displayed a significant rise in both Macro-F1 and MCC, as well as a decreased variance in different random seeds.

We indicate that the variance decline is because parameter-space fusion acts as a kind of implicit regularisation, which suppresses optimization noise and stops checkpoints trained on different data from interfering destructively. Our findings not only focus on the increase in raw accuracy but also highlight the fact that dataless merging serves as a stabilizer for limited supervision. On the contrary, it often surpasses traditional fine-tuning on reduced data sets and as such significantly closes the gap in performance with models trained on complete data. These features are essential for applications in low-resource languages or in fields that suffer from high costs due to data annotation. Moreover, the method presents a valid option for situations that face restrictions from either strict privacy regulations or organizational regulations thus excluding discreet sharing of training data.

This research confirms that even though dataless parameter merging cannot be regarded as a complementarity technique, it still is an excellent method for the transfer of knowledge between different models without any reduction in functionality or efficiency. The use of interference aware merging strategies has been shown to be a very efficient alternative to retraining on a central source, especially in cases when the access to data is limited or privacy is a matter of concern. The incorporation in legal domain NLP with RegMean parameter fusion in the early stages of the development is most likely to provide a multi-domain NLP tasks tuning with benefits from it. The statement is particularly relevant for low-resource domains that face the resource scarcity and data absence challenges, such as having under-labelled datasets, small validation sample sizes, or being completely dataless.

**Acknowledgments**

We would like to express our sincere gratitude to Dr Bolatzhan Kumalakov for his valuable guidance and support throughout this research.

**Funding**

This research received no external funding.

**Conflicts of Interest**

The authors declare no conflict of interest.

***Author Contributions***

*Conceptualization, A.K.and R.H.; methodology, A.K. and R.H.; software, A.K.; validation, A.K.; formal analysis, A.K.; resources, A.K. and R.H.; writing-original draft preparation, A.K.; writing-review and editing, A.K. and R.H.; visualization, A.K.; supervision, R.H.*

**References**

1. D. Khurana, A. Koli, K. Khatter, and S. Singh, “Natural language processing: state of the art, current trends and challenges,” *Multimedia Tools and Applications*, vol. 82, no. 3, pp. 3713–3744, Jul. 2022, doi: 10.1007/s11042-022-13428-4.

2. X. Jin, X. Ren, D. Preotiuc-Pietro, and P. Cheng, “Dataless knowledge fusion by merging weights of language models,” in *Proceedings of the 11th International Conference on Learning Representations*, pp. 1 – 19, Dec. 2023, doi: 10.48550/arxiv.2212.09849

3. H. Zheng, L. Shen, A. Tang, Y. Luo, H. Hu, B. Du, Y. Wen, and D. Tao, “Learning from models beyond fine-tuning,” *Nature Machine Intelligence*, vol. 7, no. 1, pp. 6–17, Jan. 2025, doi: 10.1038/s42256-024-00961-0.

4. X. Jiang, W. Wang, S. Tian, H. Wang, T. Lookman, and Y. Su, “Applications of natural language processing and large language models in materials discovery,” *Npj Computational Materials*, vol. 11, no. 1, Mar. 2025, doi: 10.1038/s41524-025-01554-0.

5. S. Joshi, M. S. Khan, A. Dafe, K. Singh, V. Zope, and T. Jhamtani, “Fine Tuning LLMs for Low Resource Languages,” in *5th International Conference on Image Processing and Capsule Networks (ICIPCN)*, pp. 511–519, Jul. 2024, doi: 10.1109/icipcn63822.2024.00090.

6. K. Liu, N. Uplavikar, W. Jiang, and Y. Fu, “Privacy-Preserving Multi-task Learning,” in *IEEE International Conference on Data Mining (ICDM)*, Nov. 2018, doi: 10.1109/icdm.2018.00147.

7. M. Tao, C. Zhang, Q. Huang, T. Ma, S. Huang, D. Zhao, and Y. Feng, “Unlocking the Potential of Model Merging for Low-Resource Languages,” in *Findings of the Association for Computational Linguistics*, Nov. 2024, pp. 8705–8720, doi: 10.18653/v1/2024.findings-emnlp.508.

8. Y. Hu, Y. Yao, N. Zhang, H. Chen, and S. Deng, “Exploring model kinship for merging large language models,” in *Findings of the Association for Computational Linguistics*, Nov. 2025, doi: 10.48550/arxiv.2410.12613.

9. C. Poth, J. Pfeiffer, A. Rücklé, and I. Gurevych, “What to Pre-Train on? Efficient intermediate task selection,” in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 10585–10605, Jan. 2021, doi: 10.18653/v1/2021.emnlp-main.827.

10. B. Y. Lin, C. He, Z. Ze, H. Wang, Y. Hua, C. Dupuy, R. Gupta, M. Soltanolkotabi, X. Ren, and S. Avestimehr, “FEDNLP: Benchmarking Federated Learning Methods for natural language processing tasks,” in *Findings of the Association for Computational Linguistics*, pp. 157–175, July 2022, doi: 10.18653/v1/2022.findings-naacl.13.

11. J. Zhang, S. Vahidian, M. Kuo, C. Li, R. Zhang, T. Yu, G. Wang, and Y. Chen, “Towards Building The Federatedgpt: Federated Instruction Tuning,” in *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 6915–6919, Mar. 2024, doi: 10.1109/icassp48485.2024.10447454.

12. J. Wang, M. Kulkarni, and D. Preotiuc-Pietro, “Multi-Domain Named Entity Recognition with Genre-Aware and Agnostic Inference,” in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 8476–8488, Jan. 2020, doi: 10.18653/v1/2020.acl-main.750.

13. Y. Wang, Y. Gu, Y. Zhang, Q. Zhou, Z. Yan, C. Xie, X. Wang, J. Yuan, and H. Yang, “Model merging scaling laws in large language models,” *arXiv (Cornell University)*, Sep. 2025, doi: 10.48550/arxiv.2509.24244.

14. G. Ortiz-Jimenez, A. Favero, and P. Frossard, “Task Arithmetic in the tangent Space: Improved editing of Pre-Trained models,” in *Proceedings of the 37th International Conference on Neural Information Processing System,* May 2023, doi: 10.48550/arxiv.2305.12827.

15. G. Ilharco, M. T. Ribeiro, M. Wortsman, S. Gururangan, L. Schmidt, H. Hajishirzi, and A. Farhadi, “Editing Models with Task Arithmetic,” in *International Conference on Learning Representations*, May. 2023, doi: 10.48550/arxiv.2212.04089.

16. M. Matena and C. Raffel, “Merging Models with Fisher-Weighted Averaging,,” in *Proceedings of the 36th International Conference on Neural Information Processing Systems*, pp. 17703–17716, Nov. 2022, doi: 10.48550/arXiv.2111.09832.

17. E. Yang, L. Shen, G. Guo, X. Wang, X. Cao, J. Zhang, and D. Tao, “Model Merging in LLMs, MLLMs, and beyond: methods, theories, applications, and opportunities,” *ACM Computing Surveys*, vol. 58, no. 8, pp. 1–41, Jan. 2026, doi: 10.1145/3787849.

18. N. Kadyrbek, Z. Tuimebayev, M. Mansurova, and V. Viegas, “The Development of Small-Scale Language Models for Low-Resource Languages, with a Focus on Kazakh and Direct Preference Optimization,” *Big Data and Cognitive Computing*, vol. 9, no. 5, p. 137, May 2025, doi: 10.3390/bdcc9050137.

19. C. Çetindağ, B. Yazıcıoğlu, and A. Koç, “Named-entity recognition in Turkish legal texts,” *Natural Language Engineering*, vol. 29, no. 3, pp. 615–642, Jul. 2022, doi: 10.1017/s1351324922000304.

20. M. Zampieri, P. Nakov, and Y. Scherrer, “Natural language processing for similar languages, varieties, and dialects: A survey,” *Natural Language Engineering*, vol. 26, no. 6, pp. 595–612, Nov. 2020, doi: 10.1017/s135132492000049.

21. F. Ariai, J. Mackenzie, and G. Demartini, “Natural Language Processing for the legal domain: A survey of tasks, datasets, models, and challenges,” *ACM Computing Surveys*, vol. 58, no. 6, pp. 1–37, Nov. 2025, doi: 10.1145/3777009.

22. P. Tulajiang, Y. Sun, Y. Zhang, Y. Le, K. Xiao, and H. Lin, “A Bilingual Legal NER Dataset and Semantics-Aware Cross-Lingual label transfer Method for Low-Resource languages,” *ACM Transactions on Asian and Low-Resource Language Information Processing*, vol. 24, no. 9, pp. 1–21, Jul. 2025, doi: 10.1145/3748325.

23. D. Shome and K. Yadav, “EXnet: Efficient In-context Learning for Data-less Text classification,” *arXiv (Cornell University)*, May 2023, doi: 10.48550/arxiv.2305.14622.

24. R. Yeshpanov, Y. Khassanov, and H. A. Varol, “KazNERD: Kazakh Named Entity Recognition Dataset,” in Proceedings of the Thirteenth Language Resources and Evaluation Conference, pp. 417 – 426, June 2022, doi: 10.48550/arXiv.2111.13419.

***Information about the authors:***

*Alisher Kaziz – PhD student (EP Computer Science) at the School of Software Engineering, Astana IT University. His research interests include multiagent systems, explainable artificial intelligence, and machine learning knowledge. He focuses on the development of interpretable AI models for machine learning knowledge fusion to improve decision making robustness (Astana, Kazakhstan, e-mail: alisher.kaziz@gmail.com. ORCID iD: 0009-0003-2554-6149).*

*Dr Richard Harrison – Director of Learning and Teaching for the Faculty of Engineering and Physical Sciences Foundation Year Programmes at the University of Surrey. He also Chairs the Faculty's Mathematics Education Working Group and is a Fellow of the Institute of Mathematics and its Applications. He has extensive experience as a Mathematician and Computer Scientist in industry and academia both in the UK and overseas (e-mail: r.harrison@surrey.ac.uk. ORCID iD: 0009-0005-5206-1117).*

*Submission received: 11 May, 2026*
*Revised: 15 June, 2026*
*Accepted: 15 June, 2026*