Biomedical machine translation for low-resource Arabic-script languages via cross-lingual transfer and LoRA adapter merging

Alabdullah, A.; Eslamighayour, A.; Harbalioglu, S.; Han, L.

Biomedical Machine Translation for Low-Resource Arabic-Script Languages via Cross-Lingual Transfer and LoRA Adapter Merging

Abdullah Alabdullah [1], Arash Eslamighayour [1], Sarp Harbalioglu [1], Lifeng Han[2],

[1]School of Informatics, University of Edinburgh, The United Kingdom
[2]Leids Universitair Medisch Centrum & LIACS, Universiteit Leiden, NL

Correspondence: a.alabdullah@sms.ed.ac.uk & l.han@lumc.nl,liacs.leidenuniv.nl

Abstract

We present a systematic study of healthcare-domain cross-lingual transfer to address the scarcity of biomedical NMT resources for Arabic-script languages. We use Arabic and Persian as higher-resource pivots to improve translation for **four severely low-resource** targets: Dari (Afghan Persian, a standardised variety of Persian), Pashto, Sorani Kurdish (Central Kurdish, a major standardized variety of Kurdish), and Urdu (closely related to Hindi). Using LoRA fine-tuning on small decoder-only LLMs, we train *domain-specific pivot adapters* and evaluate **three transfer strategies**: few-shot in-context learning, minimal supervised adaptation, and, to the best of our knowledge, for the first time in this setting, zero-data LoRA adapter merging. Supervised adaptation with just 500 sentences achieves near pivot-language quality for Dari (CHrF++ 41.01) and meaningful gains for Urdu (28.88), while adapter merging reaches within 3.5 CHrF++ of supervised adaptation for Dari at zero additional cost. Pashto and Sorani Kurdish remain insufficient for high-stakes clinical deployment exposing the limits of cross-lingual transfer when structural distance from the pivots is too great. LoRA adapter merging works surprisingly well for closely related languages, even without target-language biomedical data.

1 Introduction

Neural machine translation (NMT) has achieved remarkable progress for high-resource language pairs in general-domain settings (Johnson et al., 2017). However, translating biomedical and clinical text remains a fundamentally harder challenge (Bawden et al., 2019). Unlike general-domain translation, where occasional inaccuracies cause minor miscommunication, errors in medical translation carry direct, life-threatening consequences (Mehandru et al., 2023). Biomedical text is characterised by highly specialised terminology, context-sensitive acronyms (e.g., "BP" may refer to blood pressure or bipolar disorder), complex multi-word expressions, and a register that demands strict precision over fluency (Han et al., 2024). These characteristics make biomedical NMT qualitatively distinct from general-domain MT and require dedicated system development, domain adaptation, and safety-aware evaluation.

Despite growing reliance on AI-based MT in healthcare for translating discharge instructions, facilitating doctor–patient communication, and supporting clinical decision-making (Zappatore and Ruggieri, 2024), the vast majority of biomedical MT research has concentrated on well-represented European languages (Neves et al., 2024; Han et al., 2024). Arabic-script languages spoken across the Middle East and Central Asia remain critically under-studied in this domain (Abouzahir et al., 2026), despite collectively serving hundreds of millions of speakers in contexts where interpreters are often unavailable and mistranslation carries direct clinical consequences (Al Shamsi et al., 2020).

A central obstacle is the lack of domain-specific parallel data for these languages (Nigatu et al., 2025; Merx et al., 2025). While multilingual models such as NLLB-200 provide broad language coverage, they are trained on general-domain data and require domain-specific adaptation to perform reliably in healthcare settings (Team et al., 2022; Han et al., 2023). The large parallel corpora needed for such adaptation are typically unavailable for low-resource languages, motivating transfer-based approaches that leverage knowledge from related higher-resource languages — an approach supported by evidence of positive cross-lingual transfer under joint training (Lakew et al., 2018; Aharoni et al., 2019).

In this work, we investigate transfer from two higher-resource pivot languages (Talwar and Laasri, 2025) (Arabic (ar) and Persian (fa)) to four severely low-resource targets: 1) Dari (prs), a standard va-

riety of Persian; 2) Pashto (ps), an Eastern Iranian language ; 3) Sorani Kurdish (ckb), a major dialect of Kurdish; and 4) Urdu (ur), closely related to Hindi language. All four languages have their scripts written right-to-left with some comparisons listed in Table 1.[1] Dari shares substantial lexical, grammatical, and orthographic overlap with Persian. Pashto and Sorani show greater morphological and orthographic divergence, while Urdu benefits primarily from script overlap and extensive Persian/Arabic lexical borrowing. Both pivots have sufficient publicly-available biomedical parallel data for meaningful fine-tuning, and together provide complementary transfer signals: Arabic shares script, vocabulary, and loanwords with all four target languages, while Persian is closely related to Dari (a mutually intelligible variety), shares script and vocabulary with Pashto and Sorani Kurdish, and has substantially influenced Urdu at the lexical level. The four target languages are selected for their extreme data scarcity, limited MT research coverage, and linguistic proximity to the pivots.

Our **research questions** are: RQ1) How effective is biomedical domain transfer from Arabic and Persian to related low-resource Arabic-script languages? RQ2) Can zero-shot LoRA adapter merging approximate supervised adaptation? RQ3) How does linguistic distance influence transfer effectiveness? To investigate these, we first fine-tune LoRA adapters (Hu et al., 2022) on healthcare-domain data for English-to-Arabic (en→ar) and English-to-Persian (en→fa) translation, then examine whether the domain knowledge encoded in these pivot adapters transfers to four low-resource target languages through three complementary cross-lingual transfer strategies. This paper makes the following contributions and interesting findings:

- We present the first systematic evaluation of LoRA adapter merging via tensor arithmetic for biomedical NMT across Arabic-script low-resource languages.
- We show that supervised adaptation with just 500 sentences achieves pivot-level quality for Dari (CHrF++ 41.01), demonstrating the viability of cross-lingual transfer under severe data constraints.
- We demonstrate that LoRA adapter merging enables zero-resource biomedical domain adaptation for closely related low-resource Arabic-script languages, achieving performance surprisingly close to supervised fine-tuning while requiring no target-language biomedical parallel data.
- We have an interesting finding of a model inversion effect: a weaker pivot model (Llama 3.2 3B) transfers more effectively to low-resource targets than a stronger one (Gemma 2 2B), suggesting a possible trade-off between pivot specialisation and cross-lingual generalisation.

The rest of the paper is organised as below: Section 2 reviews related work; Section 4 describes the data and task; Section 3 presents the methodology; Section C reports experiments and results, and Section 6 concludes.

Language	Family	Script	Persian Sim.	Difficulty
Dari	W. Iranian	Shared	Very High	Low
Pashto	E. Iranian	Shared	Low–Mod.	High
Sorani	NW. Iranian	Partial	Moderate	Mod.–High
Urdu	Indo-Aryan	Shared	Low	Moderate

Table 1: Expected transfer difficulty based on linguistic relationship to the pivot languages.

2 Related Work

This work sits at the intersection of three bodies of work: LLMs for low-resource translation, cross-lingual transfer from high-resource pivot languages, and parameter-efficient adapter composition. We review each in turn.

2.1 LLMs for low-resource and biomedical translation

NMT has traditionally relied on encoder–decoder architectures (Bahdanau et al., 2014; Sutskever et al., 2014), but decoder-only LLMs have recently demonstrated competitive performance with less parallel data (Xu et al., 2024). However, their effectiveness remains limited for low-resource languages: Zhu et al. (2024) evaluated LLMs across 102 languages and found consistently poor performance on low-resource directions. Mao and Yu (2024) address this with contrastive alignment instructions for unseen languages, while Liang et al. (2025) show that language-specific LoRA fine-tuning substantially improves LLM-based MT for low-resource settings. In the medical domain, conventional MT systems still outperform LLMs on complex clinical texts (Li et al., 2026), motivating our strategy of fine-tuning domain-specific

[1] Pashto: written in a modified Perso-Arabic script, as one of the official languages of Afghanistan and widely spoken in Pakistan, overall 40-60 million speakers https://www.britannica.com/topic/Pashto-language.

LoRA adapters rather than relying on zero-shot LLM capabilities.

2.2 Cross-lingual transfer

Cross-lingual transfer leverages knowledge from high-resource languages to improve low-resource performance, with effectiveness depending on linguistic similarity (Conneau et al., 2020; Aharoni et al., 2019). Chronopoulou et al. (2023) show that language-family adapters outperform language-agnostic ones, suggesting typologically proximate pivot languages yield stronger transfer. For Arabic-script languages, shared script aids subword-level transfer, though character-level variation weakens this advantage (Touileb and Barnes, 2021). Garcia et al. (2023) show that few-shot in-context learning can match fine-tuned baselines when the base model is sufficiently capable. In our setting, Arabic and Persian serve as pivot languages for transfer to Dari, Pashto, Sorani Kurdish, and Urdu.

2.3 Adapter merging for cross-lingual transfer

Parameters of independently fine-tuned models can be combined through tensor arithmetic on model weights (Wortsman et al., 2022; Ilharco et al., 2023). These approaches have been used primarily in multi-task settings, where models trained on different tasks are merged to produce a single model with combined capabilities.

More recently, work has extended this idea to adapter-level merging and cross-lingual transfer. Zhao et al. (2025) decompose cross-lingual ability into task and language adapters and merge them for NLU. Bandarkar and Peng (2025) show that merging language and task experts via layer-swapping is effective for cross-lingual reasoning. Tao et al. (2024) demonstrate that merging continually pre-trained models with English task models outperforms standard fine-tuning when target-language data is scarce. However, none of these works merge *domain-specific* LoRA adapters trained on *different languages* to construct a single multilingual adapter for translation in specialised domains. This is the gap we address in this work.

3 Methodology

3.1 Fine-tuning with Low-Rank Adaptation

All models used in our study are autoregressive decoder-only transformers, which generate each token conditioned on all previously generated tokens and the source input. All six models share this architecture but differ in scale (1.5B–3B parameters), vocabulary size, and multilingual pretraining data.

We apply LoRA to fine-tune domain-specific adapters for English-to-Arabic and English-to-Persian translation, forming the foundation for all subsequent transfer experiments. Large pre-trained language models contain billions of parameters, making full fine-tuning computationally expensive. Low-Rank Adaptation (LoRA) (Hu et al., 2022) addresses this by freezing the original weights and injecting small trainable matrices into selected layers. For a weight matrix $\mathbf{W}_0 \in \mathbb{R}^{d\times k}$, LoRA defines the update as:

$$\mathbf{W} = \mathbf{W}_0 + \mathbf{BA}, \tag{1}$$

where $\mathbf{B} \in \mathbb{R}^{d\times r}$ and $\mathbf{A} \in \mathbb{R}^{r\times k}$ are the only trainable parameters, with rank $r \ll \min(d, k)$. The update is scaled by α/r, where α controls its magnitude. Only $\mathbf{A}$ and $\mathbf{B}$ are trained, allowing trainable parameter reduction by up to 99% while preserving pre-trained knowledge. The resulting adapter can be stored and applied independently of the base model, with each adapted layer adding approximately $2rd$ parameters.

3.2 Multilingual Adapter Training

We use joint multilingual training to produce a single Arabic–Persian adapter, serving as an upper bound reference for our adapter merging approach. Rather than training separate adapters per language, joint training learns a single adapter from a shuffled multilingual dataset, with a language-conditioned prompt specifying the target language per input. Since all languages share the same adapter parameters, they compete for limited representational capacity; a rank that is too small risks interference, where gains in one language degrade the other.

3.3 Zero-Shot and Few-Shot Transfer Learning

We apply the pivot-language adapters directly at inference time to evaluate how much domain knowledge transfers to low-resource target languages without any additional training.

Transfer learning leverages knowledge from a source language to improve performance on a target language with limited labelled data (Pan and Yang, 2010). In our setting, we explore two inference-time strategies that require no additional training on target-language data.

In zero-shot transfer, the Arabic and Persian adapters are applied directly to Dari, Sorani Kurdish, Pashto, and Urdu without any target-language examples. This is most effective when languages share script, morphology, or vocabulary (Aharoni et al., 2019; Conneau et al., 2020), all of which hold to varying degrees across our target languages.

In few-shot in-context learning (ICL) (Brown et al., 2020), k target-language translation examples are prepended to the input at inference time. This steers the model toward the target language without updating any parameters, and has been shown to rival fine-tuned models when the base model is sufficiently capable (Garcia et al., 2023).

3.4 Model Merging for Parameter-Efficient Adapters

We merge the Arabic and Persian LoRA adapters via tensor arithmetic to produce a zero-data multilingual adapter for transfer to low-resource target languages.

Model merging combines the capabilities of multiple independently fine-tuned models into a single model without additional training (Wortsman et al., 2022; Ilharco et al., 2023). Rather than retraining on a joint dataset, it operates directly in weight space by combining model parameters through tensor arithmetic. When the models are sufficiently aligned, the resulting model can retain the competencies of each constituent (Ilharco et al., 2023).

This approach has negligible computational cost and works best when models share the same base architecture and initialisation, ensuring aligned parameter spaces for meaningful element-wise operations (Yadav et al., 2023). Merging can be applied to full models or to adapters (e.g. LoRA matrices). In the latter case, adapters trained from a shared frozen backbone lie in a common parameter space, enabling their weights to be combined directly (Hu et al., 2022).

Formally, let each LoRA adapter be indexed by $i \in \{1, 2\}$. Let $\boldsymbol{\theta}_0$ denote the full parameter vector of the frozen base model, and $\boldsymbol{\theta}_i$ denote the effective parameters after applying adapter i. Since the base weights are frozen during LoRA training, the only learned parameters are the adapter matrices, and the difference $\boldsymbol{\tau}_i = \boldsymbol{\theta}_i - \boldsymbol{\theta}_0$, referred to as the *task vector* (Ilharco et al., 2023), is entirely determined by the adapter. Because both adapters are trained from the same frozen base with identical LoRA configurations, they share the same tensor keys and shapes, making element-wise operations on their parameters directly applicable. The merged adapter is then obtained by combining the two task vectors and adding them back to the base:

$$\boldsymbol{\theta}_{\text{merged}} = \boldsymbol{\theta}_0 + f(\boldsymbol{\tau}_1, \boldsymbol{\tau}_2), \tag{2}$$

where f denotes a merging function applied to the task vectors. The specific choices of f are described in the following subsections. We describe more details of weighted averaging, TIES-Merging, and DARE (Drop and rescale) in Section B.

4 Dataset and Task Formalisation

4.1 Task Formulation

We address sentence-level, domain-specific NMT from English (en) into six target languages: Arabic (ar), Persian (fa), Dari (prs), Sorani Kurdish (ckb), Pashto (ps), and Urdu (ur). Arabic and Persian serve as *pivot languages* with available in-domain training data; Dari, Sorani Kurdish, Pashto, and Urdu are *low-resource target languages* for which no large-scale healthcare-domain parallel data exists. This data constraint motivates a transfer learning framework in which domain knowledge encoded by pivot-language adapters is redirected to typologically related target languages. Concretely, our task decomposes into three sub-problems: (i) fine-tuning domain-specific LoRA adapters for ar and fa, (ii) transferring these adapters to the four low-resource target languages via zero-shot or few-shot inference, and (iii) investigating whether minimal target-language supervision or zero-data adapter merging can further improve transfer quality.

4.2 Data

4.2.1 Training Data

As no single healthcare-domain parallel dataset covers both en→ar and en→fa, we use two separate corpora. For en→ar, we use PEACH (Al-Sabbagh, 2024), a professionally translated, sentence-aligned corpus of patient information leaflets and patient education materials. PEACH's patient-facing content (e.g., instructional sentences) aligns closely with our goal of translating public health communication. The dataset is manually aligned, ensuring high quality.

For en→fa, we use the Medical Science subset of Esposito (Esalati et al., 2024), consisting of sentence-aligned pairs from bilingual scientific journal abstracts. While its content is more scientific, its medical vocabulary remains relevant to

healthcare text translation. Unlike PEACH, Esposito is automatically aligned; however, human evaluation of 1,500 samples shows that approximately 82% of sentence pairs are semantically well aligned ("Good" or "Very Good" labelled). It has also been validated for MT, with models trained on it outperforming general-domain baselines (Esalati et al., 2024).

Both datasets were preprocessed using sentence-length filtering, retaining pairs with source lengths between 5 and 80 words. This removes short fragments with limited translation value (e.g., headers) and long outliers likely caused by alignment errors — a standard filtering criterion in NMT preprocessing (Koehn et al., 2007). After filtering, only 25,000 sentence pairs were sampled from each corpus for computational feasibility and shuffled to remove document-ordering effects. All 25,000 pairs from each corpus are used for training; no held-out split is taken from the training corpora.

For our four low-resource target languages, no healthcare-domain parallel data exists. We instead sample 500 sentence pairs per language from FLORES-200 (Team et al., 2022) — a professionally translated, general-domain multilingual corpus derived from Wikipedia and Wikinews — covering Dari (prs), Sorani Kurdish (ckb), Pashto (ps), and Urdu (ur). These small general-domain sets are used exclusively for two purposes: as source demonstrations in few-shot in-context learning (§C.3), and as fine-tuning data in our adaptation experiment (§C.4).

4.2.2 Evaluation and Test Data

For evaluation, we use the TICO-19 (Translation Initiative for COvid-19) benchmark (Anastasopoulos et al., 2020), a professionally translated multilingual dataset comprising healthcare and public health sentences covering all six languages in our study. We use the provided split: 971 sentences for intermediate evaluation and model selection, and 2,100 for final testing. The dataset spans patient-facing guidance, scientific text, and general public health content, making it both domain-relevant and a realistic benchmark relative to our training data. No preprocessing was required for either sets.

TICO-19's high-quality translation process, widespread use in low-resource MT research, and cross-lingual comparability (due to shared English source sentences) make it well-suited to this study. We use TICO-19 exclusively for evaluation and testing, ensuring no data contamination with training corpora.

4.3 Evaluation Setup

Following the recent Arabic NMT work from Alabdullah et al. (2025), we use CHrF++ (Popović, 2017) as our primary evaluation metric. CHrF++ computes an F-score over character n-grams (up to length 6) augmented with word unigram and bigram overlap, weighted toward recall by default. This design makes it particularly well-suited to our setting for three reasons. First, character-level overlap awards partial credit to morphological variants that differ only in prefix, suffix, or inflection — a critical advantage for Arabic and Persian, where a single lemma can surface in dozens of inflected forms that exact word matching would treat as entirely incorrect (Chauhan et al., 2023). Second, its recall orientation penalises translations that omit content, not only those that produce incorrect words — important in a medical domain where missing information can be as dangerous as mistranslation. Third, CHrF++ has been adopted as the primary metric in recent low-resource MT shared tasks precisely because of its robustness to morphological richness (Popović, 2015), making our results directly comparable to the broader low-resource MT literature. BLEU (Papineni et al., 2002) is reported in Appendix B but not used as primary metric: its exact word-level matching systematically penalises valid morphological variants, a well-documented limitation for inflectionally rich languages (Chauhan et al., 2023; Popović, 2017). COMET (Rei et al., 2020) is similarly relegated to the Appendix, as it is trained predominantly on European-language judgements and offers reduced reliability for our Arabic-script targets (Rei et al., 2020). All CHrF++ scores are computed using `sacrebleu` on the 0–100 scale.

Text Normalisation Arabic-script languages often encode the same character via multiple Unicode representations (Doctor et al., 2022), which can cause metrics to penalise valid translations. We therefore apply normalisation to both hypothesis and reference at evaluation time: CAMeL Tools (Obeid et al., 2020) for Arabic, and Hazm (Roshan) for Persian and Dari. No normalisation is applied to Pashto, Sorani Kurdish, or Urdu, as no standardised tools exist for these languages.

We investigate our research question (§1) using a six-stage experimental pipeline that progressively builds toward our cross-lingual transfer goal. We first describe the configurations used across

experiments (§C.1), followed by model selection, and monolingual and joint multilingual fine-tuning (§C.2). We then introduce three complementary transfer strategies: few-shot in-context learning (§C.3), adaptation fine-tuning (§C.4), and zero-data adapter merging (§C.5). For each experiment, we explain its motivation and relevance to the research objective, followed by a technical description of the experimental procedure. We describe such details on experimental setup, pilot language adapter training, few-shot cross-lingual transfer, adaptation fine-tuning on low-resource target languages, and LoRA adapter merging for zero-data cross-lingual transfer in Section C for reproducibility.

5 Results and Discussion

Configuration	PRS	CKB	PS	UR
Zero-shot (Gemma)	5.44	3.23	2.77	13.40
Zero-shot (Llama)	12.64	3.10	2.56	4.08
Few-shot (AR)	25.94	8.93	6.10	12.25
Few-shot (FA)	14.64	8.83	6.57	8.16
Few-shot (AR+FA) (Llama)	34.84	15.42	16.70	22.26
Adaptation (AR)	**41.01**	12.34	9.94	**28.88**
Adaptation (FA)	40.25	**12.97**	10.31	28.74
Merge: Simple Avg	37.56	10.03	12.23	15.67
Merge: Wtd. (α=0.3)	36.50	11.34	**12.87**	13.94
Merge: Wtd. (α=0.7)	36.12	7.04	10.53	16.69
Merge: TIES	36.49	7.67	9.03	15.05
Merge: DARE	37.43	9.91	11.99	15.26

Table 2: CHrF++ on four low-resource target languages. Rows group into zero-shot, few-shot transfer, adaptation ($\sim$500 sentences), and adapter merging. **Bold**: best per column. Unless otherwise specified the model used is Gemma

5.1 Model Selection

Gemma 2 2B and Llama 3.2 3B produce meaningful zero-shot outputs on Arabic and Persian and improve with fine-tuning, whereas BLOOM 1.7B, XGLM 1.7B, and mGPT 1.3B generate near-random outputs with negligible gains (e.g., BLOOM Arabic CHrF++ 5.86→5.94). Despite their multilingual design, the latter models fail entirely, indicating that at this scale, the quantity and quality of Arabic-script pretraining data matter more than multilinguality alone.

Gemma 2 2B achieves the highest zero-shot scores on Arabic (29.91) and Persian (28.49), rising to 33.97 and 40.09 after fine-tuning on 500 examples, suggesting strong Arabic and Persian pretraining. Llama 3.2 3B shows similar behaviour, improving from 13.97→26.97 (Arabic) and 10.66→30.58 (Persian) with minimal data. All models achieve near-zero CHrF++ on low-resource languages (Dari, Sorani Kurdish, Pashto, & Urdu), confirming the need for dedicated adaptation and motivating transfer-learning experiments. Accordingly, Gemma 2 2B and Llama 3.2 3B are selected for subsequent experiments.

5.2 Pivot Language Baselines

Scaling pivot-language training from 500 to 25,000 sentences yields strong monolingual adapters for Gemma 2 2B, with asymmetric gains: Arabic improves substantially (33.97→40.42), while Persian remains unchanged (40.09→40.06), saturating after 500 examples. This likely reflects strong Persian pretraining (requiring minimal domain adaptation) and limited lexical diversity in the Esposito corpus, reducing returns from additional data. Arabic, by contrast, continues to benefit from scaling, suggesting our Arabic fine-tuning data is more distant from Gemma's priors. Llama 3.2 3B scales more uniformly (AR: 26.97→30.48; FA: 30.58→35.15) but reaches lower ceilings on both languages.

Joint AR+FA fine-tuning shows that both languages can be encoded in a single adapter without interference: Gemma improves substantially above its monolingual baselines (AR: 40.42→42.45; FA: 40.06→42.20), establishing the strongest pivot baselines. Llama, by contrast, exhibits cross-lingual interference under joint training, with both languages degrading below monolingual performance (AR: 30.48→28.87; FA: 35.15→31.44), indicating that the capacity for positive joint transfer is model-dependent.

5.3 Few-Shot Transfer Learning

Pivot-language adapters allow transfer to all four unseen target languages, with quality closely tracking linguistic proximity. Three in-context examples raise Dari well above zero-shot (Gemma AR: 16.33→25.94), demonstrating that inference-time adaptation yields large gains at zero cost. Performance varies sharply: Dari achieves the highest scores (Gemma AR adapter: 25.94; Llama AR+FA adapter: 34.84), while Sorani Kurdish, Pashto, and Urdu remain much lower (6–22), consistent with greater structural distance.

A key finding is a model inversion: although Gemma outperforms Llama on pivot languages (40–42 vs. 28–35), Llama transfers better than Gemma across all low-resource languages, espe-

cially Pashto (16.70 vs. 6.28) and Urdu (22.26 vs. 9.96). One possible explanation is a trade-off between pivot specialisation and cross-lingual generalisation: intensive LoRA fine-tuning on Arabic and Persian overwrites Gemma's broader multilingual representations, reducing transfer capacity, whereas Llama retains more generalisable cross-lingual priors.

Adapter source reveals a further asymmetry: under Gemma, the Arabic adapter outperforms the Persian adapter for Dari (25.94 vs. 14.64) despite closer linguistic similarity. Under Llama, the expected pattern holds (FA: 32.87 $>$ AR: 31.33), indicating model-specific interactions between pivot-language adapter training and transfer. Across both models, the combined AR+FA adapter consistently matches or exceeds the best single adapter.

5.4 Adaptation with Minimal Supervision

Minimal target-language fine-tuning (with AR and FA adapters) is highly effective for Dari and Urdu but exposes a performance ceiling for Pashto and Sorani Kurdish that $\sim$500 sentences cannot overcome. Dari reaches 41.01 (AR adapter) and 40.25 (FA adapter), a +15.1 gain over the best few-shot result (25.94), matching Arabic and Persian pivot baselines (40.42 and 40.06). The few-shot advantage of the AR adapter (25.94 vs. 14.64) disappears after adaptation, with both adapters converging, consistent with Dari's status as a Persian variety. This suggests that without target-language data, corpus quality drives transfer, whereas with supervision, linguistic proximity dominates. Urdu shows even larger gains (+16.6), reaching 28.88 (AR) and 28.74 (FA), indicating that shared script and lexical borrowing enable rapid adaptation despite its Indo-Aryan typology.

In contrast, Pashto and Sorani Kurdish improve only +3.4–4.2 CHrF++ and remain below 13, far from usable performance. Pashto (FA: 10.31; AR: 9.94) diverges in morphosyntax and script, while Sorani Kurdish (FA: 12.97; AR: 12.34) shares vocabulary but differs grammatically, gaps that 500 general-domain sentences cannot bridge. Addressing this likely requires more data, in-domain supervision, or a closer pivot language. Across all targets, AR and FA adapters converge within 1.0 CHrF++, indicating that pivot choice becomes irrelevant once target-language supervision is available.

5.5 Adapter Merging as a Zero-Data Strategy

Merging AR and FA adapters yields asymmetric effects across pivot languages. On the Persian dev set, weighted averaging ($\alpha = 0.7$) achieves 42.54, the best Persian result in the study, surpassing both the AR+FA jointly-trained adapter (42.20) and the monolingual FA adapter (40.06) at zero additional cost. This suggests the AR adapter contributes complementary Arabic-script features (e.g., shared vocabulary and biomedical terminology) that enhance Persian performance. The reverse does not hold: on Arabic dev set, all merging methods degrade performance, with the best result ($\alpha = 0.7$, 39.95) falling 0.47 below the monolingual AR adapter and 2.5 below AR+FA jointly-trained adapter. Practically, merging provides a cost-free improvement for Persian outperforming even joint training where strong single-language adapters exist.

Transfer to low-resource targets shows strong language-dependent variation. Dari reaches 37.56 CHrF++ (Simple Average), within 3.5 points of adaptation (41.01) despite using no Dari data, and far exceeding few-shot (25.94), making it a strong zero-data model. For Pashto, weighted averaging ($\alpha = 0.3$) achieves 12.87, outperforming adaptation (10.31) by +2.56 — a rare case where zero-data merging surpasses supervised fine-tuning. This may reflect domain mismatch (where our domain-agnostic adaptation data is shifting adapters away from biomedical domain) or insufficient data to capture Pashto's structural diversity, though confirming this requires further ablation. Urdu shows the opposite pattern: the best merging result (16.69, $\alpha = 0.7$) lags far behind adaptation (28.88; -12.2), indicating that weight-space interpolation cannot replace in-language supervision under high linguistic distance. Sorani Kurdish lies between these extremes, with merging (11.34, $\alpha = 0.3$) trailing adaptation (12.97) by only 1.6, though neither reaches usable performance.

Across target languages, optimal merging tracks linguistic proximity to the pivots: Persian-biased weighting ($\alpha = 0.3$) performs best for Iranian-family targets (Dari, Pashto, Sorani Kurdish), while Arabic-biased weighting ($\alpha = 0.7$) is preferred for Urdu and the pivot languages. TIES ($k = 0.2$) consistently underperforms, as trimming 80% of LoRA parameters removes shared signal rather than noise. Table 1 shows the full comparison; overall, merging is the most cost-effective starting point for Iranian-family targets without parallel data, whereas Urdu

Configuration	AR	FA	PRS	CKB	PS	UR
Few-shot (AR)	—	—	29.80	13.65	15.20	20.27
Few-shot (FA)	—	—	31.66	14.55	15.65	20.40
Few-shot (AR+FA)	—	—	33.33	15.16	16.72	21.91
Adaptation (AR)	—	—	36.95	12.60	10.73	28.32
Adaptation (FA)	—	—	**38.52**	**13.14**	**11.45**	**28.79**
Merge: Simple Avg	38.66	42.58	37.16	11.08	14.21	17.29
Merge: Wtd. (α=0.3)	36.61	41.83	36.00	12.14	**14.28**	15.30
Merge: Wtd. (α=0.7)	**39.99**	42.85	35.77	7.81	13.56	19.19
Merge: TIES	38.67	**43.34**	36.98	9.98	13.38	18.35
Merge: DARE	38.52	42.82	36.96	11.05	14.25	17.75

Table 3: CHrF++ on all target languages for few-shot transfer, adaptation (∼500 sentences), and adapter merging. Stage 4 uses Llama 3.2 3B; all other stages use Gemma 2 2B. Normalised scores used where available (CAMeL for AR; Hazm for FA/PRS). **Bold**: best per column. '—' indicates the configuration was not evaluated on that target.

requires in-language supervision.

5.6 Test Set Evaluation

Evaluation on the held-out TICO-19 test set (2,100 sentences) closely mirrors development-set rankings, confirming that observed trends reflect genuine generalisation rather than overfitting. TIES-Merging achieves the highest score in the study on Persian (43.34 CHrF++), while adapter merging remains the strongest strategy for Pashto, with weighted averaging (α=0.3) outperforming supervised adaptation (14.28 vs. 10.73–11.45) — confirming this as a robust effect. Dari approaches pivot quality under Simple Averaging (37.16), within 1.36 points of supervised adaptation at zero data cost, while Urdu continues to require in-language supervision, with the best merge result lagging adaptation by nearly 10 points (19.19 vs. 28.79). Sorani Kurdish remains insufficient for high-stakes clinical deployment.

5.7 Analysis of Linguistic Similarity and Transfer

We present in Table 4 on the transfer performance compared with linguistic relationship to the pivot languages, where it shows that the transfer quality decreases as lexical and grammatical distance increase (RQ3).

6 Conclusion and Future Work

This study investigated whether healthcare-domain adaptation on Arabic and Persian can support neural machine translation for four underrepresented Arabic-script languages — Dari, Pashto, Sorani Kurdish, and Urdu — through cross-lingual transfer. We successfully fine-tuned LoRA adapters for en→ar and en→fa, achieving CHrF++ scores of 40.42 and 40.06 respectively under monolingual training, and up to 42.45 and 42.20 under joint training, establishing strong pivot-language baselines from which transfer could be attempted.

Language	Best CHrF++	Family Relation to Persian	Script
Dari	41.01	Same language continuum	Shared
Urdu	28.88	Different family	Shared
Sorani	12.97	Iranian (distant)	Partial
Pashto	10.31	Iranian (distant)	Shared

Table 4: Transfer performance compared with linguistic relationship to the pivot languages. Languages with closer genealogical and lexical relationships generally benefit more from cross-lingual biomedical transfer.

Cross-lingual transfer to all four target languages was then evaluated across three complementary strategies. For Dari, the closest target language to our pivots, supervised adaptation with just 500 sentences reached near-pivot quality (CHrF++ ≈ 38–41), confirming that cross-lingual transfer is a practical strategy for Iranian-family languages even under severe data constraints. Zero-data adapter merging alone achieved 37.16 on the test set (within 1.36 points of supervised adaptation) making it a viable option when no target-language data is available. For Urdu, adaptation reached 28.32–28.79, demonstrating that shared script and lexical borrowing enable meaningful transfer despite its Indo-Aryan typology, though merging lagged by over 9 points, indicating that in-language supervision is necessary at this level of structural distance. For Pashto and Sorani Kurdish, no strategy exceeded 16 CHrF++, demonstrating the limits of cross-lingual transfer when structural distance from the pivot languages is too great.

The most important finding is that biomedical knowledge encoded in higher-resource Arabic and Persian adapters can be transferred to related low-resource languages through weight-space operations alone. For closely related languages such as Dari, zero-data adapter merging approaches the effectiveness of supervised adaptation, suggesting a practical path toward biomedical MT development when target-language resources are unavailable.

Limitations

The main limitation is that Pashto and Sorani Kurdish remain insufficient for high-stakes clinical deployment, likely due to their greater structural dis-

tance from the pivots and the domain mismatch between our general-domain adaptation data and the biomedical evaluation set. Future work should therefore focus on collecting in-domain biomedical adaptation data for these languages, identifying structurally closer pivot languages, and designing training objectives that explicitly preserve cross-lingual representations during domain specialisation.

Limited target-language resources. The target languages considered in this study differ substantially in the quantity and quality of available biomedical parallel data. Although all datasets were processed using the same pipeline, differences in corpus size, annotation quality, and domain coverage may influence downstream translation performance. Consequently, some performance differences attributed to cross-lingual transfer may also partially reflect underlying dataset characteristics. Future work should evaluate transfer methods on larger and more diverse biomedical benchmarks to better isolate the effect of linguistic similarity from data availability.

Limitations of instruction-tuned LLMs for machine translation. The models evaluated in this work are general-purpose instruction-tuned language models rather than systems specifically optimized for machine translation. Their outputs may therefore be affected by generation variability, hallucination, and prompt sensitivity. Furthermore, automatic metrics such as BLEU and CHRF++ may not fully capture semantic adequacy when multiple medically correct translations exist. Future work should incorporate human evaluation and semantic metrics such as COMET to provide a more comprehensive assessment of translation quality.

Coverage of biomedical terminology. The biomedical corpora used in this study represent only a subset of the terminology encountered in real-world healthcare communication. As a result, the reported performance may not generalize equally across all medical subdomains, including specialized clinical, pharmaceutical, or diagnostic terminology. Although our results demonstrate effective transfer within the evaluated datasets, further validation on broader biomedical benchmarks and terminology-focused test sets is required before drawing conclusions about comprehensive biomedical translation capability.

Confounding factors in cross-lingual transfer. A central hypothesis of this work is that linguistic proximity facilitates cross-lingual biomedical transfer. However, linguistic similarity is correlated with several other factors, including shared scripts, lexical overlap, corpus availability, and potential differences in pretraining exposure. Consequently, the observed transfer patterns cannot be attributed solely to language-family relationships. Future work should employ controlled comparisons and quantitative measures of linguistic distance to better disentangle these interacting effects.

Generalizability of adapter merging. The effectiveness of LoRA adapter merging was evaluated using two base models and a specific biomedical adaptation setting. While the results suggest that merged adapters can transfer domain knowledge to related low-resource languages, it remains unclear whether the same behavior would generalize to larger language models, alternative adapter architectures, or non-biomedical domains. Additional experiments across broader model families and domains are needed to establish the robustness of this approach.

Ethical Considerations

This work investigates biomedical machine translation for under-represented Arabic-script languages. While improved translation systems may increase access to healthcare information for speakers of low-resource languages, translation errors in clinical settings can also lead to misunderstanding, delayed treatment, or patient harm. Consequently, the models presented in this paper are intended solely for research purposes and should not be deployed for clinical decision-making without rigorous human evaluation and domain-specific validation. Our experiments use publicly available datasets and pretrained language models. We do not collect personal health records or other sensitive patient information. Nevertheless, the biomedical corpora used in this work may inherit biases present in the original source materials, including uneven coverage of medical conditions, terminology, and language varieties. Future work should investigate fairness across dialects and demographic groups and include evaluation by professional medical translators before deployment in real-world healthcare settings.

Large language models may generate fluent but factually incorrect translations or hallucinated medical terminology. Such behaviour is particularly concerning in healthcare applications where seemingly minor translation errors can have significant

consequences. Therefore, human oversight remains essential whenever these systems are used in safety-critical environments. At the same time, improving translation quality for severely under-represented languages has the potential to increase equitable access to health information in multilingual communities that currently lack high-quality translation technology.

References

Chaimae Abouzahir, Congbo Ma, Nizar Habash, and Farah E. Shamout. 2026. Cross-lingual empirical evaluation of large language models for Arabic medical tasks. In *Proceedings of the 1st Workshop on Linguistic Analysis for Health (HeaLing 2026)*, pages 158–171, Rabat, Morocco. Association for Computational Linguistics.

Roee Aharoni, Melvin Johnson, and Orhan Firat. 2019. Massively multilingual neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3874–3884, Minneapolis, Minnesota. Association for Computational Linguistics.

Rania Al-Sabbagh. 2024. Peach: a sentence-aligned parallel english–arabic corpus for healthcare. *Corpora*, 19(3):395–410.

Hilal Al Shamsi, Abdullah G Almutairi, Sulaiman Al Mashrafi, and Talib Al Kalbani. 2020. Implications of language barriers for healthcare: a systematic review. *Oman medical journal*, 35(2):e122.

Abdullah Alabdullah, Lifeng Han, and Chenghua Lin. 2025. Advancing dialectal arabic to modern standard arabic machine translation. *arXiv preprint arXiv:2507.20301*.

Antonios Anastasopoulos, Alessandro Cattelan, Zi-Yi Dou, Marcello Federico, Christian Federmann, Dmitriy Genzel, Francisco Guzmán, Junjie Hu, Macduff Hughes, Philipp Koehn, Rosie Lazar, Will Lewis, Graham Neubig, Mengmeng Niu, Alp Öktem, Eric Paquin, Grace Tang, and Sylwia Tur. 2020. TICO-19: the translation initiative for COvid-19. In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*, Online. Association for Computational Linguistics.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Lucas Bandarkar and Nanyun Peng. 2025. The unreasonable effectiveness of model merging for cross-lingual transfer in LLMs. In *Proceedings of the 5th Workshop on Multilingual Representation Learning (MRL 2025)*, pages 131–148, Suzhuo, China. Association for Computational Linguistics.

Rachel Bawden, Kevin Bretonnel Cohen, Cristian Grozea, Antonio Jimeno Yepes, Madeleine Kittner, Martin Krallinger, Nancy Mah, Aurelie Neveol, Mariana Neves, Felipe Soares, Amy Siu, Karin Verspoor, and Maika Vicente Navarro. 2019. Findings of the WMT 2019 biomedical translation shared task: Evaluation for MEDLINE abstracts and biomedical terminologies. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 29–53, Florence, Italy. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Shweta Chauhan, Philemon Daniel, Archita Mishra, and Abhay Kumar. 2023. Adableu: A modified bleu score for morphologically rich languages. *IETE Journal of Research*, 69(8):5112–5123.

Alexandra Chronopoulou, Dario Stojanovski, and Alexander Fraser. 2023. Language-family adapters for low-resource multilingual neural machine translation. In *Proceedings of the Sixth Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2023)*, pages 59–72, Dubrovnik, Croatia. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Michael Han Daniel Han and Unsloth team. 2023. Unsloth.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient finetuning of quantized LLMs. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Raiomond Doctor, Alexander Gutkin, Cibu Johny, Brian Roark, and Richard Sproat. 2022. Graphemic Normalization of the Perso-Arabic Script. In *Proceedings of Grapholinguistics in the 21st Century, 2022*, volume 9 of *Grapholinguistics and Its Applications*, pages 315–376, Brest. Fluxus Editions.

Mersad Esalati, Mohammad Javad Dousti, and Heshaam Faili. 2024. Esposito: An English-Persian scientific parallel corpus for machine translation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 6299–6308, Torino, Italia. ELRA and ICCL.

Xavier Garcia, Yamini Bansal, Colin Cherry, George Foster, Maxim Krikun, Melvin Johnson, and Orhan Firat. 2023. The unreasonable effectiveness of few-shot learning for machine translation. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 10867–10878. PMLR.

Lifeng Han, Gleb Erofeev, Irina Sorokina, Serge Gladkoff, and Goran Nenadic. 2023. Investigating massive multilingual pre-trained machine translation models for clinical domain via transfer learning. In *Proceedings of the 5th Clinical Natural Language Processing Workshop*, pages 31–40, Toronto, Canada. Association for Computational Linguistics.

Lifeng Han, Serge Gladkoff, Gleb Erofeev, Irina Sorokina, Betty Galiano, and Goran Nenadic. 2024. Neural machine translation of clinical text: an empirical investigation into multilingual pre-trained language models and transfer-learning. *Frontiers in Digital Health*, Volume 6 - 2024.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2790–2799. PMLR.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. 2023. Editing models with task arithmetic. In *The Eleventh International Conference on Learning Representations*.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

Surafel Melaku Lakew, Mauro Cettolo, and Marcello Federico. 2018. A comparison of transformer and recurrent neural networks on multilingual neural machine translation. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 641–652, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Andy Li, Wei Zhou, Rashina Hoda, Chris Bain, and Peter Poon. 2026. Comparing large language models and traditional machine translation tools for translating medical consultation summaries: Quantitative pilot feasibility study. *JMIR Form Res*, 10:e85169.

Xiao Liang, Yen-Min Jasmina Khaw, Soung-Yue Liew, Tien-Ping Tan, and Donghong Qin. 2025. Toward low-resource languages machine translation: A language-specific fine-tuning with lora for specialized large language models. *IEEE Access*, 13:46616–46626.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Zhuoyuan Mao and Yen Yu. 2024. Tuning LLMs with contrastive alignment instructions for machine translation in unseen, low-resource languages. In *Proceedings of the Seventh Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2024)*, pages 1–25, Bangkok, Thailand. Association for Computational Linguistics.

Nikita Mehandru, Sweta Agrawal, Yimin Xiao, Ge Gao, Elaine Khoong, Marine Carpuat, and Niloufar Salehi. 2023. Physician detection of clinical harm in machine translation: Quality estimation aids in reliance and backtranslation identifies critical errors. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11633–11647, Singapore. Association for Computational Linguistics.

Raphael Merx, Hanna Suominen, Trevor Cohn, and Ekaterina Vylomova. 2025. OpenWHO: A document-level parallel corpus for health translation in low-resource languages. In *Proceedings of the Tenth Conference on Machine Translation*, pages 142–160, Suzhou, China. Association for Computational Linguistics.

Mariana Neves, Cristian Grozea, Philippe Thomas, Roland Roller, Rachel Bawden, Aurélie Névéol, Steffen Castle, Vanessa Bonato, Giorgio Maria Di Nunzio, Federica Vezzani, Maika Vicente Navarro, Lana Yeganova, and Antonio Jimeno Yepes. 2024. Findings of the WMT 2024 biomedical translation shared task: Test sets on abstract level. In *Proceedings of*

the Ninth Conference on Machine Translation, pages 124–138, Miami, Florida, USA. Association for Computational Linguistics.

Hellina Hailu Nigatu, Nikita Mehandru, Negasi Haile Abadi, Blen Gebremeskel, Ahmed Alaa, and Monojit Choudhury. 2025. Viability of machine translation for healthcare in low-resourced languages. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 10584–10598, Suzhou, China. Association for Computational Linguistics.

Ossama Obeid, Nasser Zalmout, Salam Khalifa, Dima Taji, Mai Oudah, Bashar Alhafni, Go Inoue, Fadhl Eryani, Alexander Erdmann, and Nizar Habash. 2020. CAMeL tools: An open source python toolkit for Arabic natural language processing. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 7022–7032, Marseille, France. European Language Resources Association.

Sinno Jialin Pan and Qiang Yang. 2010. A survey on transfer learning. *IEEE Trans. on Knowl. and Data Eng.*, 22(10):1345–1359.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 311–318, USA. Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Roshan. Hazm - persian nlp toolkit.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems*, volume 27, pages 3104–3112.

Abhimanyu Talwar and Julien Laasri. 2025. Pivot language for low-resource machine translation. *Preprint*, arXiv:2505.14553.

Mingxu Tao, Chen Zhang, Quzhe Huang, Tianyao Ma, Songfang Huang, Dongyan Zhao, and Yansong Feng. 2024. Unlocking the potential of model merging for low-resource languages. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 8705–8720, Miami, Florida, USA. Association for Computational Linguistics.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, and 20 others. 2022. No language left behind: Scaling human-centered machine translation. *Preprint*, arXiv:2207.04672.

Samia Touileb and Jeremy Barnes. 2021. The interplay between language similarity and script on a novel multi-layer Algerian dialect corpus. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3700–3712, Online. Association for Computational Linguistics.

Leandro von Werra, Younes Belkada, Lewis Tunstall, Edward Beeching, Tristan Thrush, Nathan Lambert, Shengyi Huang, Kashif Rasul, and Quentin Gallouédec. 2020. TRL: Transformers Reinforcement Learning.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Huggingface's transformers: State-of-the-art natural language processing. *Preprint*, arXiv:1910.03771.

Mitchell Wortsman, Gabriel Ilharco, Samir Ya Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, and Ludwig Schmidt. 2022. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 23965–23998. PMLR.

Di Wu, Shaomu Tan, Yan Meng, David Stap, and Christof Monz. 2024. How far can 100 samples go? unlocking zero-shot translation with tiny multi-parallel data. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 15092–15108, Bangkok, Thailand. Association for Computational Linguistics.

Haoran Xu, Young Jin Kim, Amr Sharaf, and Hany Hassan Awadalla. 2024. A paradigm shift in machine translation: Boosting translation performance of large language models. In *The Twelfth International Conference on Learning Representations*.

Prateek Yadav, Derek Tam, Leshem Choshen, Colin Raffel, and Mohit Bansal. 2023. Ties-merging: Resolving interference when merging models. In *Advances in Neural Information Processing Systems*,

volume 36, pages 7093–7115. Curran Associates, Inc.

Le Yu, Bowen Yu, Haiyang Yu, Fei Huang, and Yongbin Li. 2024. Language models are super mario: Absorbing abilities from homologous models as a free lunch. In *Forty-first International Conference on Machine Learning*.

Marco Zappatore and Gilda Ruggieri. 2024. Adopting machine translation in the healthcare sector: A methodological multi-criteria review. *Computer Speech & Language*, 84:101582.

Yiran Zhao, Wenxuan Zhang, Huiming Wang, Kenji Kawaguchi, and Lidong Bing. 2025. AdaMergeX: Cross-lingual transfer with large language models via adaptive adapter merging. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 9785–9800, Albuquerque, New Mexico. Association for Computational Linguistics.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2024. Multilingual machine translation with large language models: Empirical results and analysis. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2765–2781, Mexico City, Mexico. Association for Computational Linguistics.

A Model Illustrations

We illustrate LoRA and LoRA Adapter Merging in Figures 1 and 2 respectively.

B Model Merging Details

B.1 Weighted Averaging

The simplest merging strategy is weighted averaging, which interpolates between two task vectors via a blending coefficient $\alpha \in [0, 1]$:

$$\theta_{\text{merged}} = \theta_0 + \alpha\tau_1 + (1 - \alpha)\tau_2. \quad (3)$$

Simple averaging is the special case $\alpha = 0.5$, treating both adapters as equally informative. The general weighted form is more appropriate for our cross-lingual setting, where the two pivot languages are not equally related to each target language. By biasing toward the Persian adapter ($\alpha \to 0$) or the Arabic adapter ($\alpha \to 1$), we can exploit known linguistic proximity without any additional training. For Iranian-family targets such as Dari and Sorani Kurdish, Persian-biased weighting is a principled choice; for Urdu, where Arabic script influence is stronger, Arabic-biased weighting is preferred.

This approach is well-motivated when both adapters share the same frozen base model and LoRA initialisation, as their task vectors lie in a comparable parameter space. Arabic and Persian share script, morphological patterns, and biomedical vocabulary, making it plausible that their adapter updates encode compatible representations. When this compatibility holds, the interpolated midpoint remains a low-loss solution for both languages (Wortsman et al., 2022).

B.2 TIES-Merging

Weighted averaging can fail when adapters assign opposite signs to the same parameter, causing destructive cancellation — a concrete risk given Arabic–Persian morphological divergence. TIES-Merging (Yadav et al., 2023) mitigates this via three operations on each task vector $\tau_i = \theta_i - \theta_0$: *trim*, which zeros all parameters outside the top-k absolute-magnitude mass, discarding low-signal noise; *elect*, which determines a dominant sign per position from the aggregate signed mass $\sum_i \tilde{\tau}_i[j]$; and *disjoint merge*, which averages only the adapter values consistent with the elected sign, fully excluding conflicting updates. The result preserves the dominant task-specific direction at each parameter

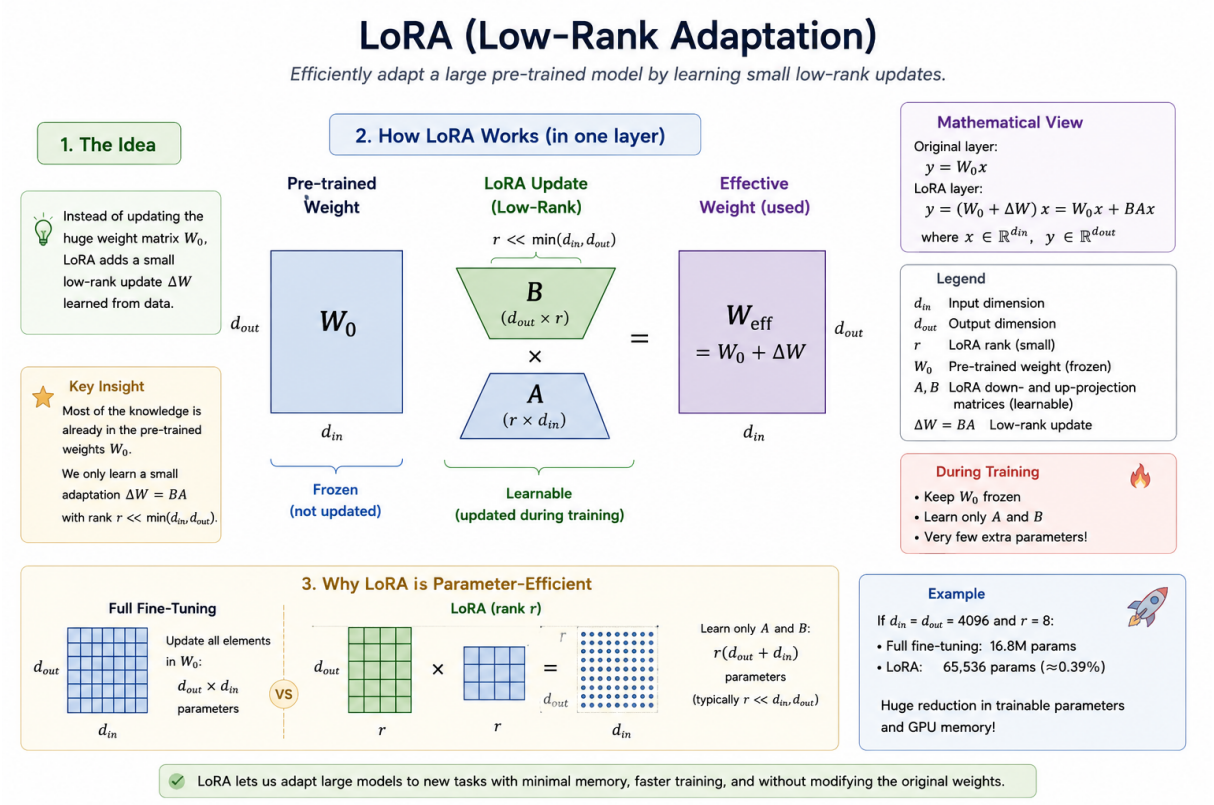

Figure 1: LoRA Illustration

position rather than collapsing opposing updates into a attenuated compromise.

B.3 DARE

Drop And REscale (DARE) (Yu et al., 2024) addresses interference from a complementary angle. Rather than resolving conflicts deterministically, it reduces the probability that conflicting updates co-occur at the same position at all. This is motivated by the observation that many fine-tuned parameter updates are redundant: in our setting, where Arabic and Persian share substantial biomedical vocabulary, both adapters may reinforce the same domain signal through overlapping parameters. By independently subsampling each adapter before merging, DARE reduces unnecessary overlap without requiring prior knowledge of which parameters conflict.

Formally, for each adapter i and parameter j, a Bernoulli mask $m_{ij} \sim \text{Bernoulli}(1-p)$ is sampled, where p is the drop rate. The masked adapter is rescaled to preserve the expected magnitude of the update:

$$\tilde{\tau}_i[j] = \frac{m_{ij}}{1-p} \cdot \tau_i[j]. \tag{4}$$

The rescaling factor $1/(1-p)$ ensures $\mathbb{E}[\tilde{\tau}_i[j]] = \tau_i[j]$. The merged adapter is then:

$$\theta_{\text{merged}} = \frac{1}{2}(\tilde{\tau}_1 + \tilde{\tau}_2) + \theta_0. \tag{5}$$

At $p = 0.3$, the probability that both adapters retain the same parameter is $(1-p)^2 = 0.49$, halving the expected number of conflicting positions relative to standard averaging. Unlike TIES, which excludes conflicting parameters entirely, DARE reduces interference stochastically — a softer form of conflict resolution that retains more of each adapter's signal at the cost of some residual noise.

C Experimental Investigation Details

C.1 Experimental Setup

All models are loaded with 4-bit NormalFloat (Dettmers et al., 2023) quantisation via `bitsandbytes`. Experiments are run on NVIDIA L4 and A40 GPUs, enabling bfloat16 mixed precision for both training and inference. Models supported by Unsloth (Gemma 2 2B, Llama 3.2 3B, and Qwen 2.5 1.5B) are loaded via Unsloth's `FastLanguageModel`, while incompatible architectures (BLOOM 1.7B, mGPT 1.3B, and XGLM 1.7B) are loaded via HuggingFace

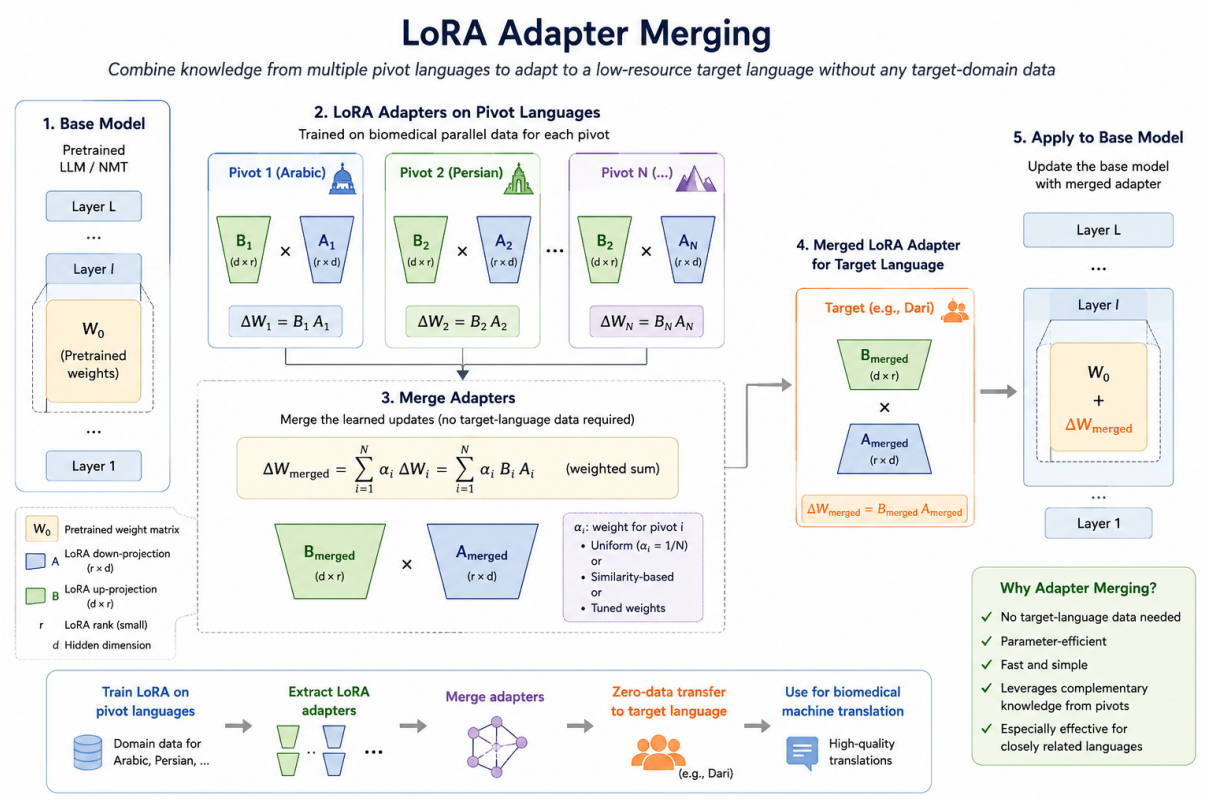

Figure 2: LoRA Adapter Merging Illustration

Transformers and PEFT with the same quantisation setup.

Parameter-efficient fine-tuning is applied using LoRA (Hu et al., 2022), with rank $r = 16$, scaling factor $\alpha = 16$ (i.e. $\alpha/r = 1.0$), and no dropout, applied to all seven projection matrices: `q_proj`, `k_proj`, `v_proj`, `o_proj`, `gate_proj`, `up_proj`, and `down_proj`.

Optimisation is performed using AdamW (8-bit) (Loshchilov and Hutter, 2019) with a learning rate of 1×10^{-4}, linear decay, 50 warmup steps, and weight decay of 0.01. The effective batch size is 32, and the maximum sequence length is 256 tokens. Sequence packing is disabled to prevent cross-contamination between training examples. All runs use 1 epoch to avoid memorisation. A fixed random seed of 3407 is used throughout. The final checkpoint is used for all evaluations.

All models use the same Alpaca-style instruction template (see Appendix), ensuring differences in translation quality arise from model or training configuration rather than prompting. We use greedy decoding throughout to ensure that performance differences across experiments reflect model and training configuration only, not decoding hyperparameters. All fine-tuning is implemented using HuggingFace Transformers (Wolf et al., 2020), PEFT (Houlsby et al., 2019), and TRL's `SFTTrainer` (von Werra et al., 2020). Unless otherwise specified, all experiments in this section follow the above-mentioned configurations. More information about our experimental setup can be found in Appendix A.

C.2 Pivot Language Adapter Training

Before fine-tuning, we identify the most suitable base model by evaluating six small (1B–3B) open-source decoder-only LLMs: Gemma 2 2B, Llama 3.2 3B, Qwen 2.5 1.5B, BLOOM 1.7B, mGPT 1.3B, and XGLM 1.7B. Each is evaluated zero-shot across all seven translation directions and fine-tuned on 500 sentence pairs from our en→ar and en→fa training sets (§4.2.1), using a reduced learning rate (5×10^{-5}) and warmup (5 steps). This identifies which models exhibit strong multilingual capability and adapt efficiently with minimal data — both prerequisites for cross-lingual transfer.

We then fine-tune LoRA adapters for the two best-performing models on en→ar and en→fa using the full 25,000-pair training sets. These monolingual adapters serve two roles: establishing strong pivot-language baselines, and acting

as inputs to all subsequent transfer experiments (§§ C.3–C.5).

Finally, we train a joint AR+FA adapter by concatenating, shuffling, and formatting both training sets with a language-conditioned prompt. Only Gemma 2 2B and Llama 3.2 3B are used. The LoRA rank is increased to 32 ($\alpha = 64$) to provide sufficient capacity for multilingual encoding. This joint adapter serves as an upper bound for our zero-data adapter merging approach (§C.5), and tests whether Arabic and Persian representations are compatible within a shared parameter space — a prerequisite for merging to succeed.

C.3 Few-Shot Cross-Lingual Transfer

Having established monolingual dapters, we now evaluate few-shot in-context learning (ICL) as the **first cross-lingual transfer strategy**. The source-language adapter is applied at inference time with target-language demonstrations prepended to the prompt, requiring no additional training. This establishes a lower bound on transfer-learning performance at zero training cost, against which more involved strategies (§C.4 and §C.5) can be compared, while also revealing which source adapter provides the strongest transfer signal for each target language.

Three adapter configurations are evaluated: the monolingual Arabic adapter, the monolingual Persian adapter, and the joint AR+FA adapter. Comparing all three reveals whether a single pivot language provides sufficient transfer signal, or whether combining both — either through joint training or merging — is necessary. Both Gemma 2 2B and Llama 3.2 3B are included, since the model inversion hypothesis (that stronger pivot performance does not imply stronger transfer) can only be tested by holding the adapter fixed and varying the base model. Each prompt prepends three target-language examples from FLORES-200, a number small enough to avoid context-window saturation while providing sufficient steering signal (Garcia et al., 2023).

C.4 Adaptation Fine-Tuning on Low-Resource Target Languages

Few-shot ICL (§C.3) is limited by the number of demonstrations that fit within the context window. We therefore investigate whether directly updating a source adapter on a small target-language dataset yields stronger transfer. Rather than training from scratch, we continue training from our monolingual Arabic and Persian adapters §C.2. This warm-start approach is motivated by evidence that minimal parallel data can enable effective multilingual transfer in pre-trained models (Wu et al., 2024).

We hypothesise that the source adapter already encodes biomedical domain knowledge from its 25k training pairs, and that a small target-language dataset is sufficient to redirect this knowledge without re-learning the domain. This experiment tests whether minimal adaptation data enables effective cross-lingual transfer, and compares Arabic- versus Persian-initialised adaptation to assess how source–target linguistic relatedness affects transfer.

For each target language, we run two warm-start experiments: one initialised from the Arabic adapter and one from the Persian adapter. This comparison is informative because Arabic and Persian differ in their structural proximity to each target language — if linguistic relatedness drives adaptation efficiency, we expect Persian initialisation to yield stronger results for Dari and Sorani Kurdish, and Arabic initialisation to be more effective for Urdu. The adaptation data is general-domain (FLORES-200), which is a deliberate choice: we are not trying to teach the model new biomedical knowledge, but to redirect already-encoded domain knowledge toward the target language's surface forms. Whether 500 general-domain sentences are sufficient for this redirection — or whether in-domain adaptation data is necessary — is itself a finding this experiment is designed to reveal.

C.5 LoRA Adapter Merging for Zero-Data Cross-Lingual Transfer

The transfer strategies explored so far require target-language data. We now investigate whether high-quality translation can be achieved by merging the weight vectors of the source-language adapters (§C.2) via post-hoc tensor arithmetic, requiring no additional GPU training. Comparing its performance with joint-training and adaptation quantifies the cost of eliminating target-language data entirely.

We apply four merging configurations (formalised in §3.4): weighted averaging with $\alpha \in \{0.3, 0.7\}$, where $\alpha = 0.3$ favours Persian and $\alpha = 0.7$ favours Arabic; TIES-Merging with density $k = 0.2$; and DARE with drop rate $p = 0.3$. All four are evaluated against the joint-training upper bound (§C.2) and supervised adaptation (§C.4).

D Experimental Setup (further details)

This section adds a more detailed explanation of the experimental setup described in C.1.

We use 4-bit nf4 quantisation to reduce GPU memory consumption sufficiently to fit training within a single GPU's VRAM budget. For our Unsloth-supported models (Gemma 2 2B, Llama 3.2 3B, Qwen 2.5 1.5B), Unsloth's optimised training kernels and custom gradient checkpointing are used, reducing VRAM usage by approximately 30% (Daniel Han and team, 2023) relative to standard PyTorch and increasing training throughput. These optimisations are mathematically equivalent to standard HuggingFace training and do not affect model weights or outputs. The per-device batch size is 2 with gradient accumulation over 16 steps, yielding the effective batch size of 32 reported in the main text.

During training, the model's EOS token is appended after `{target}`. During inference, `{target}` is left empty and the model generates after the `### {Language} translation:` marker; translations are extracted by splitting the decoded output on this marker and retaining the final segment. Inference uses `max_new_tokens = 128` and input truncation at 256 tokens. Inference is run in batches of 64 sentences — a throughput choice only, as greedy decoding is deterministic per sample regardless of batch size. The prompt template used across all experiments is:

```
Translate the below text from English
    to {Language}. Only output the
final translation in {Language}; do not
    include any additional text.

### English text:
{source}

### {Language} translation:
{target}
```

To monitor training progress, a `CheckpointMetricsCallback` evaluates BLEU and CHrF++ on the development set at every checkpoint save (approximately every $\lfloor \text{total steps}/3 \rfloor$ steps). This produces mid-epoch learning curves used for analysis.