Repairing Fused Language Models for Low-Resource Chinese–Tibetan Translation and Tibetan Medical QA via Sub-module Fine-tuning

Dejie Wang **Congfei Luo** **Chuanjie Wen** **Ning Ma***
Northwest Minzu University
{906543141, 2827573062, 3026893194}@qq.com, maning@xbmu.edu.cn

Abstract

Chinese–Tibetan medical information applications motivate combining translation from Chinese into Tibetan with medical answer selection in Tibetan within one model. Static parameter fusion offers a compact way to merge a translation expert (TB) and a Tibetan medical answer-selection expert (ACC-TB), but the task pair examined here shows severe translation degradation after merging. This paper asks whether performance can be recovered through low-rank updates to selected sub-modules while keeping the merged backbone frozen. With 500 additional mixed examples, updating the feed-forward gating matrix (GATE) raises translation BLEU from 0.25 to 48.91 and chrF++ from 0.28 to 42.10; updating the attention value matrix (V) gives a four-choice answer-selection accuracy of 34.99%, against 34.18% for GATE. Under the evaluated settings, update locations exhibit different task trade-offs; local updates improve out-of-domain translation over the unrepaired merge, although losses remain relative to the translation expert; and the tested cross-module combinations give inconsistent gains. Selective sub-module fine-tuning is therefore a low-cost recovery option under the evaluated backbone and configuration.

Keywords: Chinese–Tibetan MT; Low-Resource MT; Model Merging and Repair; Parameter-Efficient Fine-Tuning; Tibetan Medical Answer Selection

1 Introduction

Machine translation between Chinese and Tibetan is a long-standing low-resource task: parallel corpora are orders of magnitude smaller than for Chinese–English, and the two languages differ in word order, morphology, and script. Deployment adds a second constraint: a Tibetan medical expert system is expected both to translate Chinese medical text into Tibetan and to answer questions in Tibetan, and serving both with separate 7B-scale models requires two sets of weights and can increase the memory footprint and maintenance burden — which motivates studying compact multi-task models.

Parameter fusion combines expert models directly in weight space: it requires no joint training and yields a single dense model of unchanged size (Li et al., 2025; Stoica et al., 2024). Here the two experts differ radically in task structure — one maps a Chinese source to Tibetan text, the other selects from a closed set — so their output formats and sequence-level demands differ despite sharing a next-token prediction objective. Merging them statically collapses translation: COMET drops to −195.53, N-gram repetition reaches 80.07%, and EOS success falls to 8.75%, while answer-selection accuracy stays near 27%. The accompanying BLEU of 0.25 and chrF++ of 0.28 also indicate poor overlap with reference translations; these output-level observations motivate a recovery study but do not by themselves establish the cause of the degradation.

Three lines of work are relevant: methods that merge better (Ilharco et al., 2023; Yadav et al., 2023; Yu et al., 2024; Wortsman et al., 2022; Shoemake, 1985; Jang et al., 2024); the CCMT line on this pair, where Wang et al. (2022) improved Tibetan–Chinese NMT with mRASP-based pre-training and Zhuang et al. (2025) report the CCMT2024 evaluation, in which 7B systems with secondary pre-training and SFT set the reference results; and MoE routing (Kudugunta et al., 2021), which shifts the difficulty to learning a router (Chi et al., 2022; Zoph et al., 2022) and, under a few hundred mixed examples, leaves its behavior an empirical question while retaining multiple experts increases the parameter storage requirement (Section 5.2). Yang et al. (2024) instead reduce representation bias with a lightweight plugin; this study asks a narrower question: how do low-rank updates at different internal projections affect a fixed merged model?

This paper reports three contributions within the evaluated task pair. **(1) Chinese–Tibetan MT.** A

*Corresponding author.

failed static merge of a translation expert (TB) and a Tibetan medical answer-selection expert (ACC-TB) expert is substantially repaired by low-rank updates to a single sub-module: with only 500 mixed examples, updating the gating matrix restores translation COMET from -195.53 to 64.30, against 63.48 for full-parameter fine-tuning under the same budget; Merged-FullFT remains higher on BLEU and chrF++. **(2) Low-resource cross-lingual technology.** Beyond showing that low-rank repair helps a failed merge, we localize *where* task recovery occurs: on the evaluated backbone, updating the FFN gating matrix recovers translation while the attention value matrix improves answer selection, and jointly updating attention+FFN projections can renew degradation rather than help. The effective update location is therefore task-dependent — GATE is effective for translation recovery, whereas V gives the highest single-matrix answer-selection accuracy (34.99%); these observations concern the tested configurations and do not establish a stable ranking across random seeds or languages. **(3) MT system development and deployment.** The tested combinations of joint updates to self-attention and FFN projections do not yield consistent gains and can show renewed degradation: O+GATE and V+O+GATE+UP degrade translation COMET to -204.09 and -203.74. Local updates also improve several out-of-distribution scores over the unrepaired merge, with residual losses relative to the translation expert. GATE and V use approximately 10.08 M and 1.83 M trainable parameters respectively, with a peak training VRAM of 18 GB per GPU; these costs refer to the additional recovery stage.

2 Background and Motivation

2.1 Parameter Fusion and Its Limits in Low-Resource MT

The fusion strategies below were studied under different task collections, architectures, and initialization assumptions; low-resource Chinese–Tibetan MT departs from that profile, the repair budget is limited, and the partners perform structurally different tasks, so reported gains should not be assumed to transfer. Static merging integrates experts in weight space: Model Soups (Wortsman et al., 2022), TIES (Yadav et al., 2023), DARE (Yu et al., 2024), Model Stock (Jang et al., 2024) and Franken Merging (Goddard et al., 2024) average, sparsify, or splice fine-tuned weights, while SLERP (Shoemake, 1985) interpolates along an angular path. Their behavior can depend on partners and settings; here the tested merges exhibit repetition and low task scores. The six algorithms are compared before repair; later experiments use SLERP as one fixed starting point, without ranking the repairability of all merging methods.

2.2 Dynamic MoE and Parameter-Efficient Repair

Dynamic expert architectures such as the Switch Transformer (Fedus et al., 2022) address conflicts through routing (Zoph et al., 2022; Chi et al., 2022). Neither result determines whether the 500-example router used here has converged; under the present budget a dense merged model with a small local update is one option, so Dynamic MoE is retained as a higher-cost baseline. LoRA (Hu et al., 2022) is usually applied to the original pretrained model; this study instead considers a post-fusion setting in which the merged model may retain knowledge but exhibits degraded task control, so local updates test whether limited supervised adaptation can improve it — only the gating matrix or selected attention projections are updated. The experiments do not separate reactivation of retained knowledge from learning induced by the additional supervision, and do not constitute a validation of a deployed medical service.

3 Methodology

3.1 Task Definition and Parameter Fusion

Two tasks are considered throughout. **Chinese–Tibetan medical translation** takes a Chinese medical sentence or departmental query and generates its Tibetan counterpart, evaluated with COMET, BLEU, chrF++, and ROUGE-L. **Tibetan medical answer selection** takes a Tibetan medical question with four candidates and selects the correct option; because free generation in low-resource Tibetan is unstable, it is cast as four-way multiple choice and evaluated with accuracy. The two represent bilingual access to medical information and a constrained answer-selection task. Four-choice accuracy does not measure the reliability of free-form medical advice. The two single-task experts are TB (Chinese$\rightarrow$Tibetan translation) and ACC-TB (Tibetan medical answer selection, four-choice); throughout, references to a "QA expert" denote ACC-TB. They are trained separately and then merged.

Given a pre-trained base Θ_{base}, the two experts are obtained by independent training on two structurally distinct task sets, with weights Θ_{TB} and Θ_{ACC-TB}, and are merged with spherical linear interpolation (Shoemake, 1985). SLERP interpolates along an angular path between parameter vectors; this choice does not guarantee preservation of either expert's task performance or of arbitrary vector norms. Equation (1) gives one pair of corresponding tensors, with the ratio t possibly hierarchical by depth; let ω be their angle and $t \in [0, 1]$ the interpolation coefficient:

$$\Theta_{merged} = \frac{\sin((1-t)\omega)}{\sin\omega}\Theta_{TB} + \frac{\sin(t\omega)}{\sin\omega}\Theta_{ACC-TB} \tag{1}$$

Coefficients for the 28-layer implementation are given in Section 4. The observed degradation is compatible with task interference, but task scores alone do not establish the geometry or location of a representation conflict. The SLERP model retains usable answer-selection outputs but degrades markedly on translation, and is used as the base for the repair experiments.

3.2 Construction of Sub-module Fine-tuning Objectives

Degradation after static fusion is not assumed to imply complete loss of relevant knowledge; localized component mismatch is one possible explanation, rather than an established cause. The study therefore intervenes only in a few critical components, leaving the fused base otherwise intact, and applies the same repair objective to a 1:1 mixture of translation and answer-selection examples so that each resulting model is evaluated on both tasks (Figure 1).

Let S denote the selected weight matrices and Θ_{frozen} the remaining backbone weights, with $\Theta_{merged} = \Theta_{frozen} \cup S$; all original weights, including those in S, remain frozen and only the added low-rank factors are trainable. For any target $W \in \mathbb{R}^{d\times k}$ in S the update is constrained to $W' = W + BA$ (Equation (2)), with $B \in \mathbb{R}^{d\times r}$ and $A \in \mathbb{R}^{r\times k}$ trainable and $r \ll \min(d, k)$. Each matrix has its own factors; indices and the LoRA scaling factor are absorbed into BA. The implementation uses rank 16 and scaling 32 (Section 4.3). Given a supervised D_{repair} of a small number of multi-task samples, the objective is restricted to the low-rank increments of S:

$$W' = W + \Delta W = W + BA \tag{2}$$

$$\min_{A,B} \mathbb{E}_{(x,y)\in D_{repair}}\big[\mathcal{L}(\hat{y}, y)\big], \quad \hat{y} = f(x; \Theta_{frozen}, \{W + BA\}_{W\in S}). \tag{3}$$

Equation (3) replaces every frozen $W \in S$ by $W + BA$ in the forward pass. Restricting updates to a low-rank subspace reduces the trainable parameter count, but does not by itself demonstrate feature realignment or prevent overfitting; the experiment asks whether localized intervention alone can alleviate multi-task degradation while the fused base remains strictly frozen.

3.3 Structural Analysis of Core Network Components

To compare candidate update locations in S, the two tasks of Section 3.1 are contrasted against the main Transformer components: translation requires stable autoregressive control, four-way answer selection requires reliable comparison among options, and their responses to different locations are compared empirically (Figure 2).

Self-attention aggregates context; Qwen2.5 uses grouped-query attention, in which groups of query heads share key and value projections. For an input tensor X, head i is given by Equation (4) and the output projection W^o reconstructs the subspaces (Equation (5)):

$$head_i = \mathrm{Softmax}\left(\frac{XW_i^q(XW_i^k)^T}{\sqrt{d_k}}\right)XW_i^v \tag{4}$$

$$\mathrm{MultiHead}(X) = \mathrm{Concat}(head_1, \ldots, head_h) \cdot W^o \tag{5}$$

The inner products of projected queries and keys determine addressing weights, W^v extracts content features, and W^o maps head outputs back to the model dimension. Gated feed-forward networks (Geva et al., 2021) use W_{gate}, W_{up}, and W_{down} (Equation (6)):

$$\mathrm{FFN}(X) = \big(\mathrm{Activation}(XW_{gate}) \odot XW_{up}\big)W_{down}. \tag{6}$$

W_{up} and W_{down} form the primary linear path, while W_{gate} with a non-linear activation controls

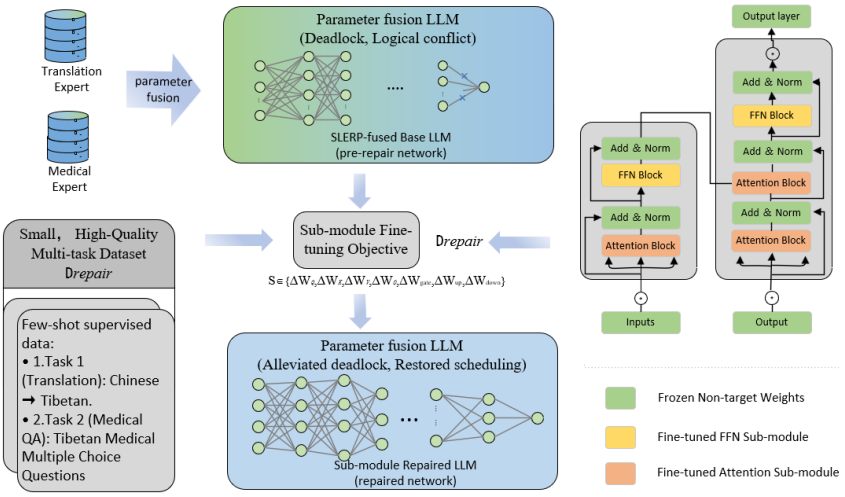
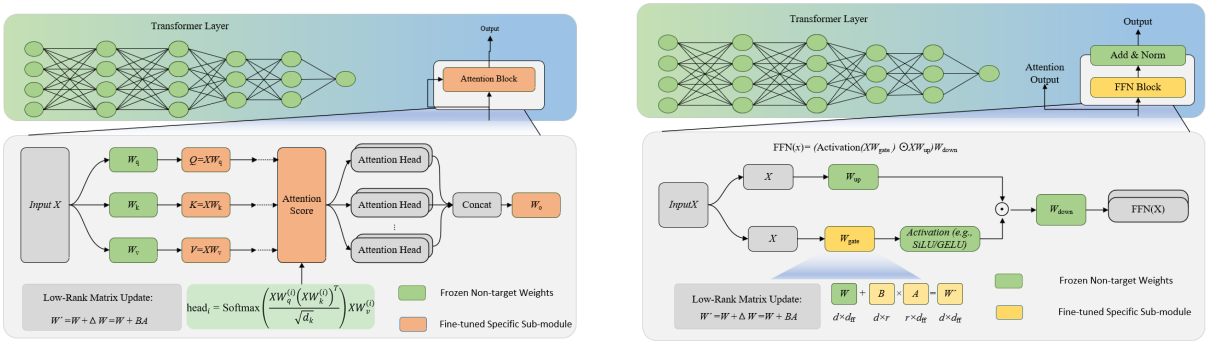


Figure 1: Research framework. Control and scheduling labels are conceptual.



(a) Attention module.

(b) FFN gating matrix.

Figure 2: Localized low-rank fine-tuning of the two target components.

filtering. The ablations below compare the measured responses of both tasks to each component.

These structural roles motivate the two repair targets selected in this study: the FFN gating matrix W_{gate} modulates intermediate features via element-wise gating and is central to translation fluency and the degeneration seen after fusion, whereas the attention value projection W^v extracts the content features used to compare the four answer candidates. This choice is corroborated empirically—among single-matrix updates, GATE yields the strongest translation recovery and V the highest answer-selection accuracy (Tables B-5 and B-6).

4 Experimental Setup

4.1 Datasets and Evaluation Metrics

The publicly released Chinese–Tibetan departmental medical QA dataset is the main resource (Wang et al., 2025): parallel data from hospital departments and Tibetan medical question–answer pairs, whose description reports translation, de-identification, and review by Tibetan-language and medical experts. To train the single-task experts, 30,000 parallel and 30,000 answer-selection examples are sampled into separate training sets, and the translation and answer-selection test sets contain 1,000 and 514 examples respectively; 500 mixed examples form the intervention set. Generation is evaluated with COMET, BLEU, chrF++, ROUGE-L; answer selection with accuracy. A uniform random choice among four options has an expected accuracy of 25%; the task-specific ACC-TB expert scores 26.26%, close to this reference. Throughout, "expert" denotes a model fine-tuned on its designated task.

Three metrics describe answer selection. *ACC-TB* is the raw four-choice accuracy over all 514 answer-selection test items; *Answer Rate* is the proportion of outputs parseable into one of the four options (format compliance); *Answered Accuracy* is accuracy computed only over successfully parsed items, separating genuine task ability from output-format instability. Each candidate answer carries a numeric label, and the parser extracts the chosen digit, so answer parsing is deterministic.

Evaluation protocol. Accuracy is reported as the macro-average of per-department accuracies: the answer-selection test set is organised by hospital department, accuracy is computed within each department, and the reported value is the unweighted mean over departments — which is why the reported values are not multiples of 0.1 (accuracy is macro-averaged across hospital departments). All 514 answer-selection test items enter the statistics; an output that cannot be parsed into one of the four options is counted as incorrect rather than excluded. Translation is scored with COMET using the `Unbabel/wmt22-comet-da` checkpoint, and the returned score is multiplied by 100; BLEU and chrF++ are computed with SACREBLEU under its default settings, with no additional language-specific tokenizer applied to Tibetan or Chinese strings. Unless stated otherwise, translation results are for the Chinese→Tibetan direction; YFB is evaluated in both directions and the tabulated score is the unweighted mean of the two direction scores.

Three benchmarks assess out-of-distribution generalization: FLORES-200 (NLLB Team, 2024), CMedQA-derived Chinese medical answer selection (Zhang et al., 2017), and YFB — 1,000 parallel pairs from entertainment, sports, customer service, and tourism released by a university-affiliated research group — for bidirectional translation (YFB dataset, 2022). Character-level N-gram overlap between the intervention set and the additional test sets is reported to limit contamination risk (Appendix A).

4.2 Model Construction and Baseline Comparison Setup

This study uses Qwen2.5-7B-Instruct as the base model; as an open-source instruction-tuned model it provides a practical starting point for MergeKit-based weight merging (Goddard et al., 2024; Yadav et al., 2025). Low-rank fine-tuning on the two 30,000-example sets yields the TB and ACC-TB experts. Six algorithms are compared — Linear, TIES, DARE-TIES, Model Stock, Franken Merging, and SLERP — all in MergeKit float16, providing several starting points for observing post-merge behavior. Settings: Linear, average addition at a 0.5/0.5 translation-to-answer-selection ratio; TIES, density 0.5 with sign masking and normalization, DARE-TIES adding Bernoulli dropping and rescaling; Model Stock, cosine angles for the central ratio; Franken Merging, interleaved splicing over the 28 layers; SLERP, hyperspherical interpolation with self-attention weights rising 0.0→1.0 with depth, feed-forward falling 1.0→0.0, others at 0.5.

The six strategies produce different task scores and output-format errors. Model Stock has a 100% invalid-option rate, so its 0.00% accuracy cannot be interpreted as evidence of absent medical knowledge. SLERP is retained as one fixed case; since the other merges are not repaired under the same protocol, this choice does not establish SLERP as the most suitable or cheapest starting point.

Controls include dense merged models without local intervention, the two single-task experts (TB, ACC-TB), and Base-FullFT, a full-parameter fine-tuning of the original Qwen2.5-7B base (7.0B) on the same 500 samples; Merged-FullFT denotes the same full-parameter fine-tuning applied to the SLERP-merged checkpoint (formerly FULL). Base-FullFT and Merged-FullFT share the same update scope and differ only in whether the starting point is the unmerged base or the merged checkpoint, which isolates the effect of merging; GATE/V instead apply a low-rank update (≈10M/1.8M parameters) to a single projection of the frozen merged backbone. Dynamic MoE uses two task-specific sparse experts plus a shared expert, activates one at a time, and uses the 30,000-example condition in Table B-4; Router MoE updates only the routing gate on the same 500 samples.

4.3 Training Hyperparameters and Generation Configurations

For the intervention setting, 500 mixed supervised examples are sampled at a 1:1 ratio. Because the dataset is extremely small, no validation set is reserved and early stopping is not used; all recovery configurations use one epoch, so module comparisons are exploratory and do not constitute validation-set selection of an optimal configuration. LoRA is applied to the selected projection type across all 28 layers of the merged base, with trainable modules varied by ablation design (for example the feed-forward gating matrix or the attention value matrix). Rank is 16, scaling 32, dropout 0.05.

AdamW with cosine annealing is used; the peak learning rate is 5×10^{-5} with 10 warm-up steps, on four GPUs with per-device batch size 4, gradient accumulation 2, and maximum sequence length 2048. BF16 and Flash Attention 2 are enabled, and the seed is 42. Inference uses vLLM in bfloat16; to reflect default decoding and avoid masking degradation, near-greedy decoding (tem-

perature = 0.1) with at most 1024 new tokens is used. All comparisons use the fixed seed 42 and are interpreted as single-run observations. The same seed and prompt templates are released with the code at https://github.com/zishu00/suture (model weights excluded) for exact reproducibility. Evaluation is conducted on fixed, held-out benchmarks rather than ad-hoc splits, so reported scores are not subject to evaluation-split variance. Given the modest absolute scale of the answer-selection benchmark, we note that absolute point differences (e.g., the 0.81pp gap between the V and GATE repairs) are inherently subject to seed-level and sampling variance, and are read against this variance rather than in isolation.

The two MoE controls share the compute setting but differ in data budget. Dynamic MoE is built from scratch with two task-specific sparse experts and one shared expert, activates one expert at a time, and is trained on the 30,000-example condition. Router MoE keeps that architecture but receives only the 500 mixed samples, with which only the routing gate is updated; its LoRA hyperparameters match those of the sub-module runs, while architecture and parameter-storage requirements still differ.

5 Results and Analysis

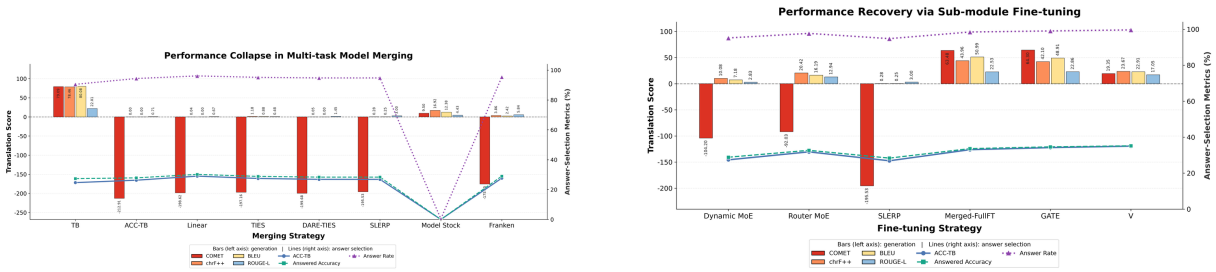
5.1 Multi-Task Conflict and Decoding Degradation in Static Fusion

The tested static fusion configurations show reduced translation scores and output-format problems on this task pair. After fine-tuning the single-task experts, multi-task models are built with Linear, TIES, DARE-TIES, SLERP, Model Stock, and Franken Merging (Figure 3a, Table B-1). All degrade markedly on translation: Model Stock yields the highest fused-baseline COMET (9.50) but obtains 0.00% accuracy with a 100% invalid-option rate on answer selection, while SLERP retains 26.85% accuracy with translation far below the single-task model. These large negative COMET scores reflect decoding degeneration—malformed or truncated outputs rather than erasure of medical knowledge—as corroborated by the decoding diagnostics in Table B-2 and the lexical-overlap analysis in Appendix A. None combines the translation expert's scores with an improvement in valid four-choice answering. The unrepaired SLERP merge retains 26.85% answer-selection accuracy, close to the un-fused ACC-TB expert's own 26.26% (Table B-1), indicating that static fusion did not erase answer-selection knowledge; the ∼25% random baseline is therefore contextualized by the expert's capability, not by a collapse to chance. The decomposition in Appendix B corroborates this reading: Model Stock's 0.00% ACC-TB coincides with a 0.39% Answer Rate, so its zero reflects output-format failure rather than absent medical knowledge, whereas SLERP's 94.75% Answer Rate with 28.34% Answered Accuracy shows that once it emits a valid option its correctness already approaches the ACC-TB expert (27.84%).

Decoding behavior shows the same pattern (Table B-2). Model Stock produces very long outputs (191.59 tokens) and a 100% invalid-option rate, while the SLERP base produces mostly valid option formats on answer selection but has an N-gram repetition rate of 80.07% and an EOS success rate of only 8.75% on translation. These summaries describe output behavior alongside task metrics; they do not identify the cause of each termination or measure representation conflict directly. Single-matrix fine-tuning partly alleviates this: GATE raises the translation EOS success rate to 43.77%, whereas V reduces the invalid-option rate to 0% but leaves its EOS success rate at 21.98%. Both GATE and V have a 0% invalid-option rate, so this diagnostic does not favor V over GATE for answer format.

5.2 Recovery via Sub-module Fine-tuning and System Overhead

Given the degradation after static merging, the next question is whether sub-module fine-tuning recovers multi-task performance (Figure 3b, Table B-3). Fine-tuning the feed-forward gating matrix raises translation COMET to 64.30, slightly above the 63.48 of full-parameter fine-tuning under the same 500-sample budget; for answer selection, the attention value matrix yields the highest accuracy, 34.99%, against 34.18% for GATE and 33.02% for Merged-FullFT. All sub-module repairs reach 98–100% Answer Rate (Appendix B), so the recovered accuracy is genuine task recovery rather than a formatting fix; their Answered Accuracy (33–35%) likewise exceeds the 28–30% of the unrepaired SLERP merge. These are separate trained models; their task-wise best scores cannot be combined. GATE does not dominate Merged-FullFT: BLEU/chrF++ are 48.91/42.10 against 50.99/43.96. Small answer-selection differences are treated as descriptive.



(a) Static fusion baselines.

(b) Main baselines and sub-module fine-tuning.

Figure 3: Static fusion and sub-module fine-tuning. Complete numbers in Tables B-1, B-3. Answer selection is reported by ACC-TB (raw accuracy), Answered Accuracy (accuracy over parsed outputs), and Answer Rate (format compliance); see Appendix B for definitions.

Router MoE is a control for dynamic architectures under low-resource conditions: it reaches 31.81% accuracy on medical answer selection, but its translation scores remain well below Merged-FullFT and GATE, so updating the router alone is not sufficient. The unmerged base cannot be adapted to either task within the 500-sample budget (0.00% ACC, Table B-5), indicating that merging—not merely fine-tuning—supplies the multi-task prior in this run. The Dynamic MoE condition uses a larger budget and takes about 15.0 GPU-hours; this comparison does not establish the minimum data or compute needed for routing convergence.

Table B-4 summarizes training and weight-storage requirements: Dynamic MoE assembles multiple experts, uses 30,000 examples and about 15.0 GPU-hours; full-parameter fine-tuning uses 500 samples but updates roughly 7.0 billion parameters. GATE, V, and Router MoE update approximately 10.08 M, 1.83 M, and 12.5 M parameters; the 500-example budget excludes the earlier training of the two experts. Memory and runtime are specific to the setting used here.

5.3 Fine-Grained Ablation of Sub-module Fine-tuning

To provide a direct-adaptation reference, the original pretrained base without merging is fine-tuned on the same 500 mixed samples; Figure 4a shows it reaches only 3.80 COMET on translation and 0.00% answer-selection accuracy: these values describe the tested run; since update scope differs and answer parsing may contribute to zero accuracy, this comparison does not establish a causal advantage of merged initialization.

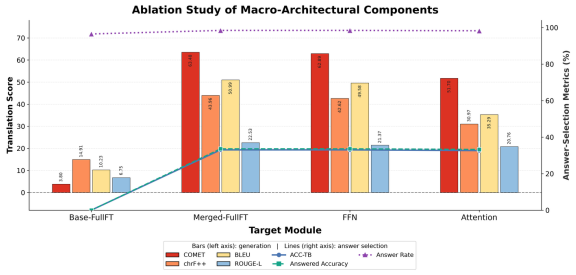
The macroscopic roles of the feed-forward network and attention are then compared (Figure 4a, Table B-5): feed-forward-only yields 62.89 COMET, close to the 63.48 of full-parameter fine-tuning, whereas attention-only yields 51.70; for answer selection, Merged-FullFT and FFN both reach 33.02% while Attention reaches 32.63%. FFN updates obtain higher translation scores than attention-only updates in this run, whereas answer-selection accuracies are close.

Microscopic ablations locate effective sites within each module: for translation, Figure 4b shows GATE yields the highest COMET (64.30), the best attention submatrix is O at 43.27, and V achieves the highest single-matrix answer-selection accuracy (34.99%). The task trade-offs differ across locations: GATE+UP obtains higher BLEU (52.06) and chrF++ (45.13) than GATE despite a lower COMET. V's answer-selection advantage over V+O is only 0.22 points, so neither a universal best matrix nor a reliable answer-selection ranking is established.

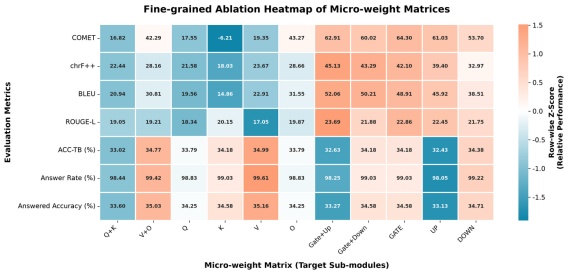
Across micro-matrices the Answer Rate stays at 98–99% (Table B-6), confirming that format compliance is broadly restored once any content projection is updated.

5.4 Interference in Joint Self-Attention/FFN Updates

Given these ablations, the next question is whether joint cross-module fine-tuning yields additive gains. Figure 5 compares V, GATE, V+GATE, O+GATE, and V+O+GATE+UP (Table B-7); gains are not consistent across these configurations. V+GATE reaches 59.53 COMET on translation, below the 64.30 of GATE alone, and 33.57% on answer selection, below the 34.99% of V alone; expanding to O+GATE and V+O+GATE+UP drops COMET to -204.09 and -203.74 and answer-selection accuracy to 28.40% and 27.82%. These combinations do not improve both tasks consistently at the fixed learning rate; the degradation is compatible with optimization instability or interacting updates, but



(a) FFN versus attention.



(b) Fine-grained heatmap.

Figure 4: Ablation comparisons. Same 500 mixed samples throughout; complete numbers in Tables B-5, B-6. ACC-TB is accompanied by Answer Rate and Answered Accuracy in Appendix B.

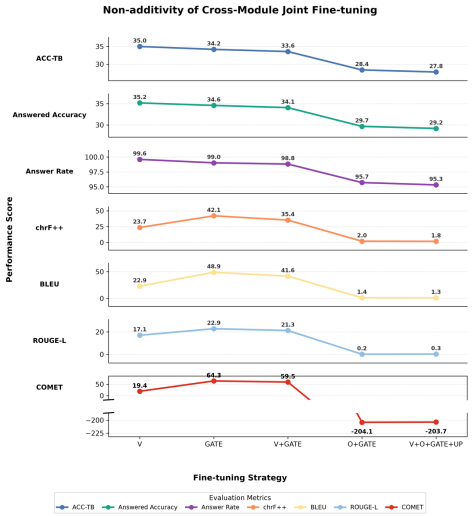


Figure 5: Results of cross-module joint fine-tuning. ACC-TB, Answered Accuracy, and Answer Rate are reported; see Appendix B.

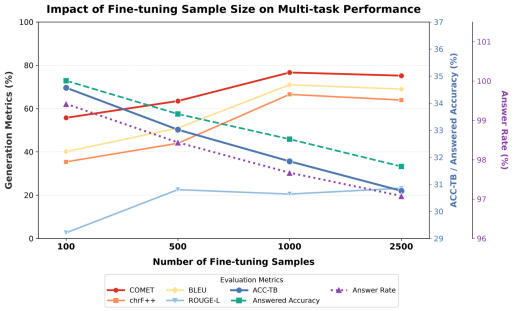
the experiments do not isolate these explanations. The within-FFN GATE+UP gains in Table B-6 also rule out a general conclusion that multi-module updates are ineffective. The format-stability view is further supported by Answer Rate: it falls from 99.6% (V) to 95.7% and 95.3% for O+GATE and V+O+GATE+UP (Table B-7), tracking the ACC-TB decline and suggesting that jointly updating more modules reintroduces output-format instability.

5.5 Data Sensitivity and Cross-Domain Generalization

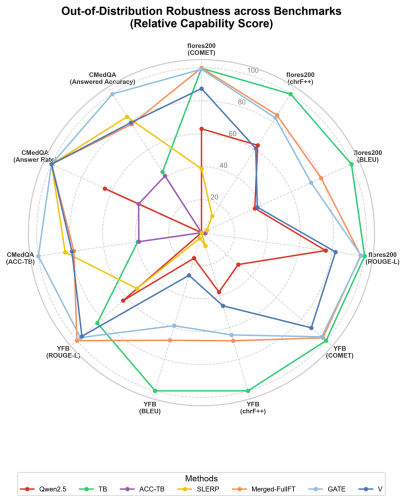
To investigate data sensitivity, full-parameter fine-tuning is examined under varying sample budgets. Figure 6a and Table B-8 show translation COMET rising from 55.73 at 100 samples to 63.48 at 500 and 76.66 at 1,000, then easing to 75.20 at 2,500, while answer-selection accuracy declines from 34.57% at 100 to 30.76% at 2,500. The two tasks respond differently to additional supervision; these trends are descriptive and do not identify the causes of the changes. The COMET dip between 1,000 and 2,500 samples is not explained by the concurrent answer-selection decline. Because all budgets use one epoch and the same 10 warm-up steps, changing the sample count also changes the optimization duration and warm-up fraction, so the curve describes the combined data-and-training schedule rather than isolating data quantity. It concerns Merged-FullFT, not GATE/V. Under the present setting, 500 mixed samples are a practical operating point rather than an identified optimum. Answer Rate remains near 97–99% across all budgets (Table B-8), so the accuracy decline with more data is genuine task degradation rather than worsening format compliance.

Out-of-distribution behavior is evaluated on general translation, domain-specific translation, and Chinese medical answer selection. Character-level N-gram overlap between the mixed training set and all independent test sets is computed to control for contamination; at the 4-gram level overlap remains low (Appendix A), which describes limited lexical overlap with the repair set but does not rule out shared source items or overlap with expert training. Figure 6b summarizes the scores (Table B-9); all tables are in Appendix B.

On the general and domain-specific translation benchmarks, local updates improve scores over the unrepaired SLERP model, but recovery is incomplete. On FLORES-200 the GATE-updated model reaches 55.77 COMET, close to the 57.26 of the translation expert and far above the -102.80 of the unrepaired base; on YFB it reaches 78.24 versus 87.58. The corresponding BLEU scores are 16.69 versus 22.82 and 15.12 versus 25.70; these residual gaps should be retained in the assessment of recovery. On the Chinese medical answer-selection benchmark, V yields 48.60% accuracy while GATE reaches 59.80%, the best among the compared controls, so transfer also depends on where the update



(a) Data scale versus Merged-FullFT performance.



(b) FLORES-200, YFB, CMedQA (relative axes).

Figure 6: Data-scale sensitivity and out-of-distribution results. In (b) each axis is the min-max normalized value of that metric across the seven compared models, rescaled to 0–100, where 0 and 100 are the lowest and the highest value of that metric among them and a metric on which all models tie is set to 50. Axis values are therefore relative and are not comparable across axes; absolute values are in Table B-9; the three CMedQA axes in (b) are ACC-TB, Answer Rate, and Answered Accuracy.

is applied; the different ranking on this Chinese task limits extrapolation from the Tibetan answer selection result. The experiments do not identify whether the gains arise from knowledge recovery, additional learning, or output-format changes. On CMedQA the unadapted Qwen2.5 base scores 5.35% ACC-TB but 96.50% Answer Rate and only 5.60% Answered Accuracy (Table B-9), so its low score warrants further checks of labels, candidates, and scoring rather than being read as genuine inability, whereas SLERP and the repaired models attain both high Answer Rate (≈99.8%) and high Answered Accuracy (48–60%).

6 Conclusion

This paper studies a deployment problem in Chinese–Tibetan medical expert systems: a translation expert and the ACC-TB answer-selection expert merged by static parameter fusion lose most of their translation ability, and the question is whether the merged model can be partly recovered without full-parameter retraining. The results support partial recovery under the conditions examined here. Keeping the merged backbone frozen and applying low-rank updates to a single sub-module partly recovers both tasks (Section 5.2): the gating matrix gives the strongest translation recovery and the value matrix the highest answer-selection accuracy among single-matrix updates, as reported above.

Two observations may be useful for low-resource minority-language MT. First, effective repair is task-specific: measured task trade-offs differ across update locations, and the preferred configuration depends on the metric; the single-seed results do not establish a universally best intervention point. Second, the repair is bounded: some tested cross-module combinations show renewed degradation, although GATE+UP improves several translation metrics over GATE. Because data and GPU memory are both scarce, a local update to an already-merged checkpoint offers a recovery option complementing the data-augmentation and continued-pre-training approaches more common for this language pair. Future work should test whether this pattern holds for other low-resource pairs and scales, and should replace manual selection of the update location with a predictor based on task properties.

Limitations

Although this study reports a low-cost intervention baseline, its conclusions remain bounded by the present setting. First, the observed sub-module effects are based on Qwen2.5; whether the same pattern holds for other families and scales remains untested. Second, the experiments cover one heterogeneous task pair, and recovery may depend in part on representational alignment already present in the base model, so languages without comparable pretraining support may prove harder. Third, only a limited set of cross-module combinations is examined rather than an exhaustive search; com-

patibility boundaries among functional matrices remain open. Fourth, the fixed-seed comparisons and absence of a validation set limit claims about small score differences and configuration selection. Fifth, the direct-adaptation control does not match the GATE/V update scope, and the other merging algorithms are not repaired under the same protocol. We also lack a parameter-matched LoRA baseline trained from the original base on the same modules and mixed data, so the effect of merging per se is not isolated from standard adaptation. Sixth, quality is assessed with COMET and lexical-overlap metrics, so the conclusions are restricted to these automatic measures. Seventh, our answer-selection results concern constrained answer selection and do not establish performance on open-ended medical QA. Eighth, the experiments do not separate recovery of retained knowledge from learning induced by the additional supervision.

Declaration of Interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (Grant No. 62466052) and the Fundamental Research Funds for the Central Universities of Northwest Minzu University (Grant No. 31920260103). We thank the scholars who constructed and released the datasets used in this study, whose openly available data made this research possible. We also acknowledge the contributors to the open-source Qwen2.5, MergeKit, and vLLM projects.

References

Zewen Chi, Li Dong, Shaohan Huang, Damai Dai, Shuming Ma, Barun Patra, Saksham Singhal, Payal Bajaj, Xia Song, Xian-Ling Mao, Heyan Huang, and Furu Wei. 2022. On the representation collapse of sparse mixture of experts. In *Advances in Neural Information Processing Systems*, volume 35, pages 34600–34613.

William Fedus, Barret Zoph, and Noam Shazeer. 2022. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2021. Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5484–5495.

Charles Goddard, Shamane Siriwardhana, Malikeh Ehghaghi, Luke Meyers, Vlad Karpukhin, Brian Benedict, Mark McQuade, and Jacob Solawetz. 2024. Arcee's MergeKit: A toolkit for merging large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 477–485.

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 *The Tenth International Conference on Learning Representations*.

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

Dong-Hwan Jang, Sangdoo Yun, and Dongyoon Han. 2024. Model stock: All we need is just a few fine-tuned models. In *European Conference on Computer Vision*.

Sneha Kudugunta, Yanping Huang, Ankur Bapna, Maxim Krikun, Dmitry Lepikhin, Minh-Thang Luong, and Orhan Firat. 2021. Beyond distillation: Task-level mixture-of-experts for efficient inference. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3577–3599.

Yunshui Li, Yiyuan Ma, Shen Yan, Chaoyi Zhang, Jing Liu, Jianqiao Lu, Ziwen Xu, Mengzhao Chen, Minrui Wang, Shiyi Zhan, Jin Ma, Xunhao Lai, Yao Luo, Xingyan Bin, Hongbin Ren, Mingji Han, Wenhao Hao, Bairen Yi, LingJun Liu, and 5 others. 2025. Model merging in pre-training of large language models. In *Advances in Neural Information Processing Systems*.

NLLB Team. 2024. Scaling neural machine translation to 200 languages. *Nature*, 630(8018):841–846.

Ken Shoemake. 1985. Animating rotation with quaternion curves. In *Proceedings of the 12th Annual Conference on Computer Graphics and Interactive Techniques*, pages 245–254.

George Stoica, Daniel Bolya, Jakob Bjorner, Pratik Ramesh, Taylor Hearn, and Judy Hoffman. 2024. ZipIt! merging models from different tasks without training. In *The Twelfth International Conference on Learning Representations*.

Dejie Wang, Chuanjie Wen, and Ning Ma. 2025. XBMU Chinese–Tibetan Multi-Department Medical QA Dataset. figshare [Data set].

Hao Wang, Yongbin Yu, Tashi Nyima, Rinchen Dongrub, Favour Ekong, Mengwei Ai, Kalzang Gyatso, Yong Cuo, and Qun Nuo. 2022. Life is short, train it less: Neural machine Tibetan–Chinese translation based on mRASP and dataset enhancement. In *Machine Translation: 18th China Conference, CCMT 2022, Lhasa, China, August 6–10, 2022, Revised Selected Papers*, volume 1671 of *Communications in Computer and Information Science*, pages 54–59. Springer.

Mitchell Wortsman, Gabriel Ilharco, Samir Yitzhak 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 *International Conference on Machine Learning*, pages 23965–23998.

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

Prateek Yadav, Tu Vu, Jonathan Lai, Alexandra Chronopoulou, Manaal Faruqui, Mohit Bansal, and Tsendsuren Munkhdalai. 2025. What matters for model merging at scale? *Transactions on Machine Learning Research*.

Enneng Yang, Li Shen, Zhenyi Wang, Guibing Guo, Xiaojun Chen, Xingwei Wang, and Dacheng Tao. 2024. Representation surgery for multi-task model merging. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 56332–56356. PMLR.

YFB dataset. 2022. A Chinese–Tibetan Multi-domain Machine Translation Third-party Test Dataset (V1.0). National Basic Science Data Center [Data set].

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*.

Sheng Zhang, Xin Zhang, Hui Wang, Jiajun Cheng, Pei Li, and Zhaoyun Ding. 2017. Chinese medical question answer matching using end-to-end character-level multi-scale CNNs. *Applied Sciences*, 7(8):767.

Wenhao Zhuang, Dawa Cairen, Pengmao Cairang, and Yuan Sun. 2025. CCMT 2024 Tibetan–Chinese machine translation evaluation technical report. In *Machine Translation: 20th China Conference, CCMT 2024, Xiamen, China, November 8–10, 2024, Proceedings*, volume 2365 of *Communications in Computer and Information Science*, pages 119–127. Springer.

Barret Zoph, Irwan Bello, Sameer Kumar, Nan Du, Yanping Huang, Jeff Dean, Noam Shazeer, and William Fedus. 2022. ST-MoE: Designing stable and transferable sparse expert models. *arXiv preprint arXiv:2202.08906*.

A Lexical Overlap between Repair Data and Additional Test Sets

The contamination check compares the 500-example repair set with CMedQA, YFB, and FLORES-200 at the character level, after removing whitespace. CMedQA: 3-gram overlap 0.0261% (134,187 N-grams in the test set, 35 hits) and 4-gram overlap 0.0063% (190,190; 12). YFB: 3-gram 7.5864% (8,568; 650) and 4-gram 1.4851% (9,023; 134). FLORES-200: 3-gram 6.6491% (34,486; 2,293) and 4-gram 1.5525% (37,489; 582). CMedQA thus shows negligible overlap of continuous segments. The 3-gram rates of YFB and FLORES-200 are higher, which may include common expressions; the present summary does not identify the contribution of specific linguistic categories. However, as the matching granularity increases to 4-gram, the overlap rates of both translation test sets drop to approximately 1.5%, without establishing whether complete source items overlap across task variants.

These statistics concern only lexical overlap between the 500-example repair set and the additional benchmarks. They do not establish separation from the experts' training data, rule out cross-language duplicates, or distinguish knowledge transfer from other causes of improved scores.

B Experimental Data Tables

Evaluation Metric		Single model TB	Single model ACC-TB	Linear	TIES	DARE-TIES	SLERP	MODEL_STOCK	FRANKEN
Translation	COMET	79.09	-212.91	-198.62	-197.16	-199.68	-195.53	9.50	-175.45
Translation	chrF++	78.46	0.00	0.04	1.18	0.05	0.28	16.92	3.86
Translation	BLEU	80.08	0.00	0.00	0.88	0.00	0.25	12.38	2.42
Translation	ROUGE-L	22.01	0.71	0.67	0.48	1.45	3.00	4.43	5.84
Tibetan medical answer selection (ACC-TB)	ACC	24.65%	26.26%	28.99%	27.43%	26.85%	26.85%	0.00%	27.63%
Tibetan medical answer selection (ACC-TB)	Answer Rate	90.47%	94.36%	96.11%	95.14%	94.75%	94.75%	0.39%	95.33%
Tibetan medical answer selection (ACC-TB)	Answered Accuracy	27.31%	27.84%	30.16%	28.83%	28.34%	28.34%	0.00%	28.98%

Table B-1: Baseline performance of static fusion methods on Chinese–Tibetan translation and Tibetan medical answer selection.

Model Strategy	Linear	Model Stock	SLERP	Dynamic MoE	Merged-FullFT	GATE	V
Avg Tokens - TB	2.6	191.59	13.97	12.98	131.41	130.08	161.74
Rep Rate % - TB	79.04%	77.14%	80.07%	28.15%	22.85%	25.48%	68.65%
EOS Success % - TB	0.39%	3.31%	8.75%	70.48%	43.87%	43.77%	21.98%
Answer Rate % - ACC-TB	96.11%	0.39%	94.75%	95.14%	98.44%	99.03%	99.61%

Table B-2: Sequence Decoding Feature Diagnosis. Rows: output length, repetition rate, EOS success rate in translation, and answer-selection parseability (Answer Rate).

Evaluation Metric		Dynamic MoE	Router MoE	SLERP	Merged-FullFT	GATE	V
Translation	COMET	-104.20	-92.03	-195.53	63.48	64.30	19.35
Translation	chrF++	10.08	20.42	0.28	43.96	42.10	23.67
Translation	BLEU	7.18	16.19	0.25	50.99	48.91	22.91
Translation	ROUGE-L	2.83	12.94	3.00	22.53	22.86	17.05
Tibetan medical answer selection (ACC-TB)	ACC	27.38%	31.81%	26.85%	33.02%	34.18%	34.99%
Tibetan medical answer selection (ACC-TB)	Answer Rate	95.14%	97.67%	94.75%	98.44%	99.03%	99.61%
Tibetan medical answer selection (ACC-TB)	Answered Accuracy	28.83%	32.67%	28.34%	33.60%	34.58%	35.16%

Table B-3: Core baselines and sub-module fine-tuning on the two benchmarks.

Evaluation Metric	Dynamic MoE	Router MoE	SLERP	Merged-FullFT	GATE	V
Samples	30,000	500	0	500	500	500
Trainable parameters	$\sim$21.0 B	$\sim$12.5 M	0	7.0 B	$\sim$10.08 M	$\sim$1.83 M
Peak VRAM per GPU	42 GB	42 GB	-	32 GB	18 GB	18 GB
Active Params	7 B	7 B	7 B	7 B	7 B	7 B
Memory Footprint	46 GB	46 GB	14 GB	14 GB	14 GB (merged)	14 GB (merged)
GPU-Hours	$\sim$15.0 h	$<$ 0.2 h	0	$\sim$0.5 h	$<$ 0.3 h	$<$ 0.3 h

Table B-4: Computing resources and VRAM overhead of the compared strategies.

Evaluation Metric		Base-FullFT	Merged-FullFT	FFN	Attention
Translation	COMET	3.80	63.48	62.89	51.70
Translation	chrF++	14.91	43.96	42.62	30.97
Translation	BLEU	10.23	50.99	49.58	35.29
Translation	ROUGE-L	6.75	22.53	21.37	20.76
Tibetan medical answer selection (ACC-TB)	ACC	0.00%	33.02%	33.02%	32.63%
Tibetan medical answer selection (ACC-TB)	Answer Rate	96.50%	98.44%	98.44%	98.25%
Tibetan medical answer selection (ACC-TB)	Answered Accuracy	0.00%	33.60%	33.60%	33.27%

Table B-5: Ablation: feed-forward versus attention fine-tuning.

Evaluation Metric		Q+K	V+O	Q	K	V	O	Gate+up	Gate+down	GATE	UP	DOWN
Translation	COMET	16.82	42.29	17.55	-6.21	19.35	43.27	62.91	60.02	64.30	61.03	53.70
Translation	chrF++	22.44	28.16	21.58	18.03	23.67	28.66	45.13	43.29	42.10	39.40	32.97
Translation	BLEU	20.94	30.81	19.56	14.86	22.91	31.55	52.06	50.21	48.91	45.92	38.51
Translation	ROUGE-L	19.05	19.21	18.34	20.15	17.05	19.87	23.69	21.88	22.86	22.45	21.75
Tibetan medical answer selection (ACC-TB)	ACC	33.02%	34.77%	33.79%	34.18%	34.99%	33.79%	32.63%	34.18%	34.18%	32.43%	34.38%
Tibetan medical answer selection (ACC-TB)	Answer Rate	98.44%	99.42%	98.83%	99.03%	99.61%	98.83%	98.25%	99.03%	99.03%	98.05%	99.22%
Tibetan medical answer selection (ACC-TB)	Answered Accuracy	33.60%	35.03%	34.25%	34.58%	35.16%	34.25%	33.27%	34.58%	34.58%	33.13%	34.71%

Table B-6: Fine-grained comparison of projection types and selected combinations.

Evaluation Metric		V	GATE	V+GATE	O+GATE	V+O+GATE+UP
Translation	COMET	19.35	64.30	59.53	-204.09	-203.74
Translation	chrF++	23.67	42.10	35.40	2.01	1.84
Translation	BLEU	22.91	48.91	41.63	1.42	1.33
Translation	ROUGE-L	17.05	22.86	21.26	0.2	0.32
Tibetan medical answer selection (ACC-TB)	ACC	34.99%	34.18%	33.57%	28.40%	27.82%
Tibetan medical answer selection (ACC-TB)	Answer Rate	99.61%	99.03%	98.83%	95.72%	95.33%
Tibetan medical answer selection (ACC-TB)	Answered Accuracy	35.16%	34.58%	34.06%	29.67%	29.18%

Table B-7: Performance of the selected combinations of joint updates to self-attention and FFN projections.

Evaluation Metric		100 Samples	500 Samples	1000 Samples	2500 Samples
Translation	COMET	55.73	63.48	76.66	75.20
Translation	chrF++	35.38	43.96	66.55	63.92
Translation	BLEU	40.10	50.99	70.98	69.00
Translation	ROUGE-L	2.60	22.53	20.50	23.17
Tibetan medical answer selection (ACC-TB)	ACC	34.57%	33.02%	31.85%	30.76%
Tibetan medical answer selection (ACC-TB)	Answer Rate	99.42%	98.44%	97.67%	97.08%
Tibetan medical answer selection (ACC-TB)	Answered Accuracy	34.83%	33.60%	32.67%	31.66%

Table B-8: Data-scale comparison for full-parameter fine-tuning under the one-epoch schedule.

Evaluation Dataset	Evaluation Metric	Qwen2.5	TB	ACC-TB	SLERP	Merged-FullFT	GATE	V
flores200	COMET	-39.15	57.26	-204.48	-102.80	57.89	55.77	24.94
flores200	chrF++	17.74	28.08	0.09	3.44	23.81	23.27	17.12
flores200	BLEU	8.14	22.82	0.04	0.87	18.22	16.69	8.52
flores200	ROUGE-L	11.54	14.79	1.52	1.20	14.46	14.5	12.37
YFB	COMET	-100.28	87.58	-177.86	-178.88	80.59	78.24	55.94
YFB	chrF++	12.32	32.76	0.03	2.77	22.40	21.19	15.16
YFB	BLEU	4.15	25.70	0.01	0.93	17.49	15.12	6.94
YFB	ROUGE-L	2.08	2.60	0.50	1.80	3.01	2.94	2.91
CMedQA	ACC	5.35%	26.65%	26.26%	50.85%	47.95%	59.80%	48.60%
CMedQA	Answer Rate	96.50%	90.50%	94.40%	99.80%	99.80%	99.80%	99.80%
CMedQA	Answered Accuracy	5.60%	29.39%	27.86%	51.00%	48.10%	59.92%	48.70%

Table B-9: Out-of-domain translation and Chinese medical answer-selection performance. Translation is scored in the Chinese→Tibetan direction; the YFB rows give the unweighted mean of the two translation directions.