

# Conditional Rank Allocation for Taxonomy-Aware Medical Language Model Adaptation

Guangyuan Dong<sup>1,†,‡</sup>, Ziwei Hong<sup>4,†</sup>, Xuehao Zhou<sup>2,3,†</sup>, Zidong Yu<sup>6</sup>, Bingchen Liu<sup>5,†</sup>  
Kehan Liu<sup>1</sup>, Chuang Liu, Rong Fu, Yuchao Hou<sup>\*</sup>

<sup>1</sup>National University of Singapore, Singapore

<sup>2</sup>Department of Urology, Shanghai General Hospital, Shanghai, China

<sup>3</sup>Shanghai Jiao Tong University School of Medicine, Shanghai, China

<sup>4</sup>University of Pennsylvania, USA <sup>5</sup>Shandong University, China <sup>6</sup>Syracuse University, USA

Email: guangyuan@u.nus.edu <sup>†</sup>These authors contributed equally. <sup>‡</sup>Project Lead. <sup>\*</sup>Corresponding author.

**Abstract**—Medical question answering spans specialties and clinical operations that may benefit from different adaptation directions. We propose ARBOR, a parameter-efficient method that selects rank-one components from a shared low-rank basis for each question. An additive gate combines question representations, specialty tags, operation tags, and their interaction; a learned coefficient scales the adapter residual. An illustrative separation under orthogonal, equiprobable subtasks shows how conditional selection can avoid an approximation floor faced by a fixed update with the same active rank. This result motivates the design without asserting a corresponding bound for medical corpora. On Qwen3-8B across CMB, CMExam, MedQA, and MedMCQA, five-seed experiments yield 69.69% mean accuracy across benchmarks, exceeding LoRA r16 and MoELoRA by 1.26 and 1.30 percentage points, respectively. The reported advantage over LoRA r16 increases from 0.08 to 1.94 points as training expands from one to seven specialties. Tag perturbations and atom masking support the usefulness of clinical routing, while atom clusters align with the supplied specialty labels (adjusted Rand index 0.62). Calibration, transfer, and measured costs further characterize the method. These findings support structured conditional adaptation for medical QA, while leaving clinical safety and broader deployment untested.

**Index Terms**—parameter-efficient fine-tuning, conditional rank allocation, medical question answering, clinical taxonomy

## I. INTRODUCTION

Medical question answering spans specialties and clinical operations that may benefit from different adaptation directions. Standard LoRA applies one low-rank update to every input [1]; a drug-interaction question and an anatomy fact therefore share the same adaptation subspace. Conditional selection offers a way to specialize updates while reusing a stored basis.

Specialty and clinical operation describe complementary aspects of a question. A specialty identifies its knowledge context, while an operation describes the requested task, such as recall, diagnosis, or treatment. Questions within one specialty can require different operations, and the same operation can recur across specialties. This motivates separate routing contributions together with a pair-specific interaction, rather than a single undifferentiated clinical label.

ARBOR (Adaptive Rank allocation with Budgeted On-demand Routing) selects rank-one atoms per question using an additive gate over question representations, specialty,

operation, and their interaction. A learned coefficient scales the residual, and a discrete mask records active components. Our contribution is this clinical structure and its evaluation, building on existing rank-level routing rather than introducing that mechanism.

An orthogonal-subtask construction illustrates a representational advantage at matched active rank. Its assumptions do not imply monotonic gains for real medical corpora or explain comparisons against rank-16 LoRA when only seven specialties are tested. We examine these gains empirically through heterogeneity, tag perturbations, and atom masking.

Across four Chinese and English benchmarks and five seeds, ARBOR reaches 69.69% mean accuracy, exceeding LoRA r16 and MoELoRA by 1.26 and 1.30 percentage points. Clustering evaluates alignment with supplied clinical labels; calibration, transfer, and measured costs characterize the method beyond accuracy. Active rank, parameter count, and runtime are distinguished throughout.

## II. RELATED WORK AND POSITIONING

LoRA uses a shared low-rank update [1], while AdaLoRA reallocates rank across modules [2]. MoELoRA routes among adapters [3]; HydraLoRA shares a down-projection while diversifying expert up-projections [4].

Biomedical post-training also includes BFT, which reweights token- and sample-level losses for knowledge alignment [5]. Knowledge-guided pathway prototypes in scProtoTransformer provide interpretable biological representations [6].

Input-conditioned reuse of learned primitives also appears in PriMD [7], which retrieves semantic primitives to compensate for missing modalities. ARBOR applies conditional selection to rank-one parameter updates for medical question answering.

Arrow [8], MeteorA [9], and LoRA-Mixer [10] compose pretrained adapter pools; MedAdapter [11] verifies candidate answers at test time. These pipelines differ from adapters trained under our common CMB-source budget. Matching active rank, stored rank, or parameter count guarantees neither equal runtime nor identical capacity.

In biomedical named entity recognition, MITE combines code-formatted instruction tuning with entity-level voting [12],

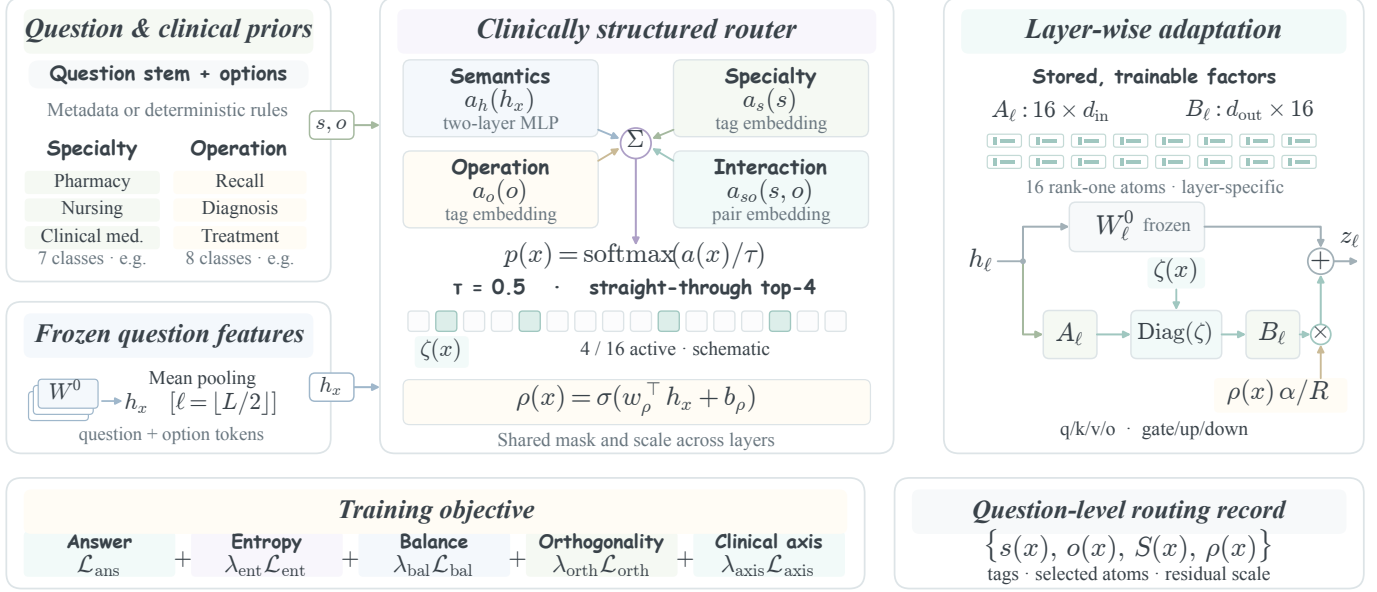


Fig. 1. ARBOR overview. Clinical priors and frozen question features feed four additive routing branches. Straight-through top- $\kappa$  selection activates four of sixteen stored rank-one atoms; a learned coefficient scales their contribution alongside the frozen backbone. Mask and scale are shared across target layers. The training objective and question-level routing record are shown below. Highlighted mask positions are schematic.

while GAMA combines verified annotation guidelines with schema-constrained generation [13]. ARBOR uses clinical tags to condition parameter updates for question answering.

SMoRA treats individual ranks as experts with straight-through selection [14]. ARBOR adds specialty, operation, and interaction logits plus residual scaling, and evaluates their contribution to clinical routing and supplied-label alignment.

### III. WHY CONDITIONAL RANK ALLOCATION HELPS

We give an illustrative linear separation, not a medical-model risk bound. Let  $\Delta_j = B_{:,j} A_{j,:}$ ,  $j \in [R]$ . Fixed selection applies  $\Delta_S = \sum_{j \in S} \Delta_j$  to every  $h(x)$ ; conditional selection applies  $\Delta_{S(x)} h(x)$  with  $|S(x)| = \kappa$ . Constant routing includes fixed supports. Strict inclusion requires nondegenerate atoms and inputs enabling outputs unavailable to constant routing; it need not hold for every basis.

a) *Orthogonal-subtask separation*: Consider  $m$  equiprobable regions with representation  $h(x) = v_i$  and target  $u_i$  in region  $i$ , where both  $\{u_i\}_{i=1}^m$  and  $\{v_i\}_{i=1}^m$  are orthonormal. For a fixed linear update  $\Delta$ , define

$$\mathcal{R}(\Delta) = \frac{1}{m} \sum_{i=1}^m \|\Delta v_i - u_i\|_2^2. \quad (1)$$

The target map  $T = \sum_i u_i v_i^\top$  has  $m$  unit singular values. Consequently, the Eckart-Young theorem [15] gives

$$\inf_{\text{rank}(\Delta) \leq \kappa} \mathcal{R}(\Delta) = \max(0, 1 - \kappa/m). \quad (2)$$

With  $R \geq m$ , store  $\Delta_i = u_i v_i^\top$  and route region  $i$  to atom  $i$ , yielding zero risk at active rank one. Larger budgets remain

exact if sufficient additional orthogonal atoms contribute zero on  $v_i$ . This construction trades stored capacity for active rank.

b) *Scope of the argument*: Orthogonality, equiprobable regions, and correct routing are assumed. Overlap, imbalance, routing errors, and nonlinear representations alter the comparison. More medical specialties need not mean more independent target directions. Rank-16 LoRA has zero floor in Eq. (2) for  $m \leq 7$ , so its heterogeneity comparison is empirical, not bound validation. Matched rank-4 sweeps and multiple specialty combinations would strengthen the connection. A generalization analysis must also account for input-dependent router complexity, not only support counts.

### IV. METHOD

#### A. Conditional update

For frozen projection  $W_\ell^0$  and hidden state  $h_\ell$ , ARBOR computes

$$z_\ell = W_\ell^0 h_\ell + \rho(x) \frac{\alpha}{R} B_\ell \text{Diag}(\zeta(x)) A_\ell h_\ell, \quad (3)$$

where  $A_\ell \in \mathbb{R}^{R \times d_{\text{in}}}$  and  $B_\ell \in \mathbb{R}^{d_{\text{out}} \times R}$  are trainable,  $\zeta(x) \in \{0, 1\}^R$  has  $\kappa$  active entries, and  $\rho(x) \in [0, 1]$ . Atom  $j$  is  $B_{\ell,:,j} A_{\ell,j,:}$ . The same mask and scale are used across target layers for a given question, while each projection retains its own  $A_\ell, B_\ell$ .

#### B. Clinically structured gate

Question representation  $h_x$  is the mean-pooled frozen-backbone hidden state over question and option tokens at layer  $\lfloor L/2 \rfloor$ . The gate combines semantic and clinical logits:

$$a(x) = a_h(h_x) + a_s(s(x)) + a_o(o(x)) + a_{so}(s(x), o(x)). \quad (4)$$

TABLE I  
CLINICAL-PRIOR VALIDATION ON 500 ANNOTATED QUESTIONS.  
AGREEMENT IS INTER-ANNOTATOR  $\kappa$ .

| Axis          | Accuracy | Coverage | Majority | Agreement |
|---------------|----------|----------|----------|-----------|
| Specialty (7) | 91.4%    | 100%     | 28.6%    | 0.86      |
| Operation (8) | 84.6%    | 100%     | 24.2%    | 0.79      |

Here  $a_h$  is a two-layer MLP;  $a_s, a_o, a_{so}$  are embedding tables for specialty, operation, and their pair. This additive form allows each clinical axis to contribute separately while retaining pair-specific effects. We compute

$$p(x) = \text{softmax}(a(x)/\tau), \quad \zeta(x) = \text{ST-top-}\kappa(p(x)), \quad (5)$$

$$\rho(x) = \sigma(w_\rho^\top h_x + b_\rho). \quad (6)$$

The forward pass selects the  $\kappa$  largest entries with a binary mask; a straight-through estimator passes gradients through soft probabilities [16], [17]. The coefficient  $\rho$  learns how strongly to inject the residual. It is neither calibrated uncertainty nor a validated measure of clinical safety. Gate features must be obtained from the frozen backbone before the conditioned adapter pass; their extraction or caching is part of the implementation cost.

### C. Clinical tags

Seven specialty labels and eight operation labels come from available metadata or deterministic rules on the question and options. For example, specialty rules include drug/dose/contraindication for Pharmacy, nursing/care plan for Nursing, and incidence/screening/vaccine for Public Health; the fallback is Clinical Medicine. Operation rules cover diagnosis, treatment, medication, examination, mechanism, nursing action, and prevention, with Recall as fallback. Gold answers are excluded, and target-benchmark labels are not used to tune rules or routing.

Two annotators labeled a 500-item sample (Table I); specialty and operation tags achieve 91.4% and 84.6% accuracy, respectively. Inter-annotator agreement is  $\kappa = 0.86$  and 0.79; rule coverage is 100% through the fallback labels. These measurements describe the sampled items and do not establish equal tag quality across languages or datasets.

### D. Training and computation

Answer-only supervised fine-tuning uses

$$\mathcal{L} = \mathcal{L}_{\text{ans}} + \lambda_{\text{ent}}\mathcal{L}_{\text{ent}} + \lambda_{\text{bal}}\mathcal{L}_{\text{bal}} + \lambda_{\text{orth}}\mathcal{L}_{\text{orth}} + \lambda_{\text{axis}}\mathcal{L}_{\text{axis}}. \quad (7)$$

Entropy and balancing terms regularize routing;  $\mathcal{L}_{\text{orth}} = \|AA^\top \odot (1 - I)\|_F^2$  penalizes correlation between atom directions. Decorrelation also regularizes molecular interaction features in DSIL-DDI [18]. A supervised contrastive  $\mathcal{L}_{\text{axis}}$  encourages questions from the same specialty to have similar routing distributions. Since specialty labels enter both the gate and this objective, clustering evaluates alignment with supplied labels rather than independent discovery of a taxonomy. We

use  $R = 16$ ,  $\kappa = 4$ ,  $\alpha = 32$ ,  $\tau = 0.5$ , and regularizer weights (0.01, 0.001, 0.1, 0.5) in the order above; remaining training settings are in Section V. Rank selection costs  $O(R \log \kappa)$  once per question. Activating four atoms does not imply the parameter count or runtime of rank-4 LoRA: the full basis is stored, and routing, feature extraction, and the adapter implementation contribute to measured cost.

## V. EXPERIMENTAL SETUP

*a) Data and evaluation:* Using Qwen3-8B [19], source training uses a route-balanced CMB [20] subset of 4,200 questions (600 per specialty), disjoint from evaluation. Main evaluation uses Chinese CMB (4,149 questions) and CMExam [21], and English MedQA [22] (700 test questions) and MedMCQA [23] (700 validation questions). PubMedQA [24] provides out-of-family evaluation. Accuracy is primary; macro is the unweighted four-benchmark average, not pooled question accuracy. Exact evaluation manifests and the CMExam split and sample count remain reproducibility limitations.

*b) Training and statistical analysis:* Trainable methods use 400 updates, effective batch size 8, learning rate  $10^{-4}$ , maximum sequence length 768, answer-only supervised fine-tuning, greedy decoding, and a common answer parser and prompt template within each language. ARBOR uses  $R = 16$ ,  $\kappa = 4$ ,  $\alpha = 32$ ,  $\tau = 0.5$ , and  $(\lambda_{\text{ent}}, \lambda_{\text{bal}}, \lambda_{\text{orth}}, \lambda_{\text{axis}}) = (0.01, 0.001, 0.1, 0.5)$ , targeting the q/k/v/o/gate/up/down projections. Hyperparameters were selected using a 400-question held-out source slice, without target-benchmark labels. Runs use an A100-80GB, PyTorch 2.3, and bf16. Results are means and standard deviations over seeds {13, 29, 47, 71, 97}. The paired differences and their 95% confidence intervals in Fig. 2 summarize variation across training seeds; they do not quantify uncertainty from resampling the benchmark population.

*c) Comparators:* LoRA r4 matches the active rank, LoRA r16 the atom-bank size, and LoRA r24 approximately the trainable parameter count. We also compare DoRA [25] r24, HydraLoRA (four heads), and MoELoRA (four rank-8 experts, top-2). The rank-wise control uses an r16/top-4 adapter inspired by SMoRA [14], but removes clinical branches, their interaction, axis contrast, orthogonality, and injection scaling. It therefore compares the complete design with a simplified router and does not isolate the effect of clinical tags alone. Adapter-library results are reported separately because they compose pretrained adapters and are not matched source-training comparisons.

## VI. RESULTS AND ANALYSIS

*a) Main results:* ARBOR achieves  $69.69 \pm 0.12$  macro accuracy and the highest accuracy on each benchmark among the trainable methods in Table II. Its macro advantages are 1.26 percentage points over LoRA r16, 1.30 over MoELoRA, and 1.20 over the rank-wise control. The paired across-seed intervals are [1.15, 1.38], [1.02, 1.58], and [1.03, 1.36], respectively (Fig. 2). Five seeds provide limited information about training variability; these intervals do not include uncertainty from evaluation-set sampling. Adapter libraries sometimes

TABLE II

QWEN3-8B MEDICAL QA ACCURACY (%). TRAINABLE METHODS SHARE THE CMB-SOURCE PROTOCOL; VALUES ARE MEAN  $\pm$  STANDARD DEVIATION OVER FIVE SEEDS. MACRO IS THE UNWEIGHTED BENCHMARK AVERAGE. BOLD MARKS THE BEST TRAINABLE METHOD. ADAPTER LIBRARIES COMPOSE PRETRAINED ADAPTERS AND ARE REPORTED SEPARATELY WITHOUT ACROSS-SEED UNCERTAINTY.

| Method            | Trainable (M) | CMB                               | CMExam                            | MedQA                             | MedMCQA                           | Macro                             |
|-------------------|---------------|-----------------------------------|-----------------------------------|-----------------------------------|-----------------------------------|-----------------------------------|
| Frozen Qwen3-8B   | –             | 71.68                             | 68.95                             | 62.86                             | 54.86                             | 64.59                             |
| LoRA r4           | 10.91         | 77.62 $\pm$ .21                   | 75.67 $\pm$ .24                   | 62.91 $\pm$ .42                   | 57.19 $\pm$ .38                   | 68.35 $\pm$ .18                   |
| LoRA r16          | 43.65         | 77.79 $\pm$ .19                   | 75.98 $\pm$ .22                   | 62.60 $\pm$ .45                   | 57.34 $\pm$ .41                   | 68.43 $\pm$ .21                   |
| LoRA r24          | 65.47         | 77.95 $\pm$ .23                   | 75.88 $\pm$ .25                   | 62.71 $\pm$ .44                   | 57.62 $\pm$ .37                   | 68.54 $\pm$ .22                   |
| DoRA              | 66.87         | 78.48 $\pm$ .20                   | 75.86 $\pm$ .28                   | 62.19 $\pm$ .47                   | 56.90 $\pm$ .44                   | 68.36 $\pm$ .17                   |
| HydraLoRA         | 44.90         | 78.02 $\pm$ .22                   | 75.92 $\pm$ .24                   | 63.15 $\pm$ .46                   | 57.34 $\pm$ .40                   | 68.61 $\pm$ .20                   |
| MoELoRA           | 92.60         | 77.88 $\pm$ .24                   | 75.41 $\pm$ .26                   | 63.02 $\pm$ .51                   | 57.25 $\pm$ .42                   | 68.39 $\pm$ .20                   |
| Rank-wise control | 64.88         | 78.21 $\pm$ .22                   | 75.74 $\pm$ .27                   | 63.28 $\pm$ .48                   | 56.75 $\pm$ .45                   | 68.49 $\pm$ .11                   |
| <b>ARBOR</b>      | 66.59         | <b>78.55 <math>\pm</math> .18</b> | <b>76.71 <math>\pm</math> .21</b> | <b>64.87 <math>\pm</math> .35</b> | <b>58.63 <math>\pm</math> .33</b> | <b>69.69 <math>\pm</math> .12</b> |
| MedAdapter [11]   | –             | 77.61                             | 77.07                             | 64.00                             | 56.86                             | 68.89                             |
| Arrow [8]         | –             | 76.69                             | 77.30                             | 61.57                             | 56.57                             | 68.03                             |
| MeteoRA [9]       | –             | 76.67                             | 75.33                             | 61.57                             | 56.57                             | 67.54                             |
| LoRA-Mixer [10]   | –             | 77.03                             | 75.52                             | 62.00                             | 57.14                             | 67.92                             |

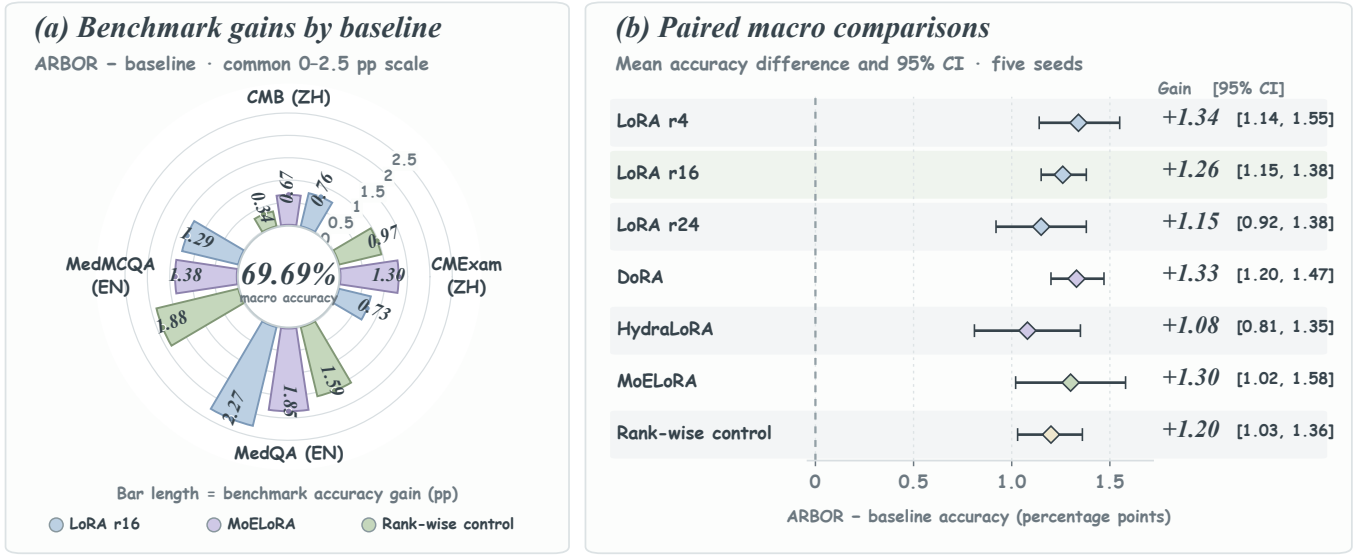


Fig. 2. Accuracy improvements under matched source training. (a) Benchmark gains over three baselines share a 0–2.5 percentage-point radial scale, with zero at the inner ring; the center reports ARBOR’s macro accuracy. (b) Paired macro differences against seven trainable baselines, with reported 95% intervals across five training seeds. All intervals exclude zero.

outperform ARBOR on individual benchmarks: Arrow reaches 77.30 on CMExam, compared with 76.71 for ARBOR.

*b) Transfer:* Relative to LoRA r16, the gains are 0.76 and 0.73 points on the Chinese CMB and CMExam sets, and 2.27 and 1.29 points on the English MedQA and MedMCQA sets. Thus, the improvement extends beyond the training language. On held-out PubMedQA, ARBOR reaches  $74.15 \pm 0.30$ , compared with the reported strongest baseline at 72.61. These benchmark results suggest transfer, without establishing performance on open-ended clinical tasks.

#### A. Routing and Clinical Structure

*a) Training heterogeneity:* We vary the number of training specialties over  $H \in \{1, 2, 3, 4, 5, 7\}$ , with 4,200 training questions, a fixed mixed evaluation set, and five seeds per setting. The reported ARBOR–LoRA r16 advantage increases

from 0.08 points at  $H = 1$  to 1.94 at  $H = 7$ . Figure 3(b) shows these endpoints. The trend is an empirical observation rather than a direct test of the rank- $\kappa$  separation in Section III: LoRA r16 has rank greater than the maximum seven specialties. The experiment does not establish robustness to different specialty combinations. A rank-matched LoRA r4 sweep and repeated combinations would strengthen the interpretation; low- $H$  sampling and evaluation aggregation also require fuller documentation.

*b) Tag interventions:* Rule-derived tags give macro accuracy 69.69; a learned classifier gives 69.61, shuffled-but-consistent tags 68.94, and random tags 68.72 (Fig. 3(a)). Randomization therefore removes 0.97 of the 1.20-point advantage over the rank-wise control. These interventions support the relevance of tag content within the tested model; they do not establish a general causal claim about clinical reasoning.

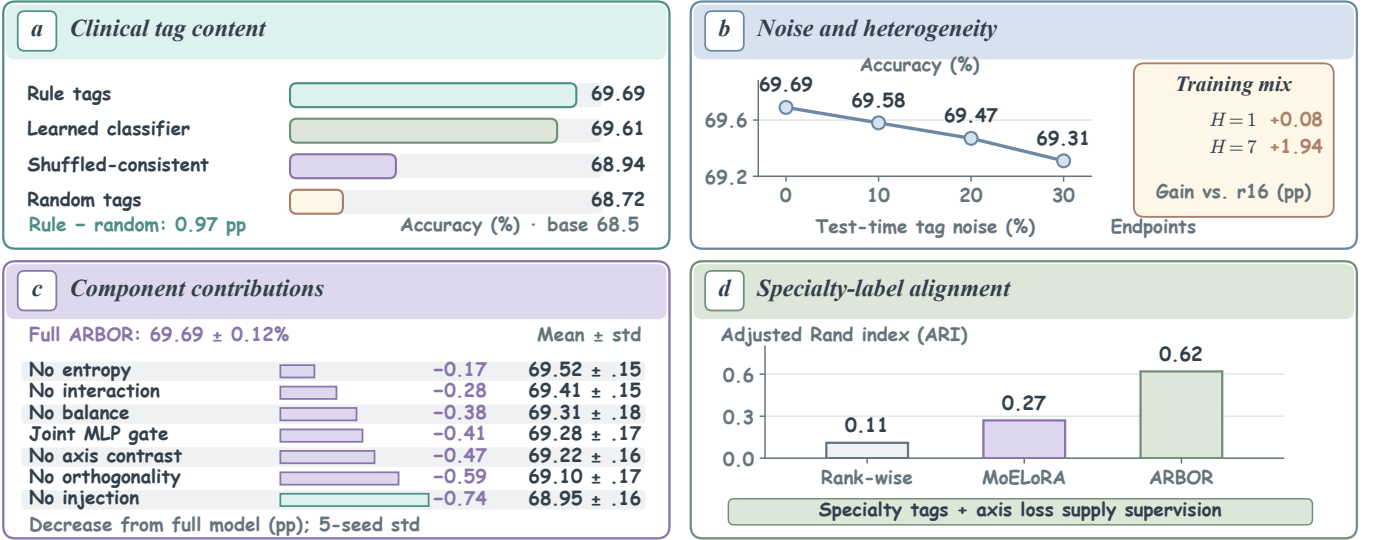


Fig. 3. Clinical routing and component analysis. (a) Tag-source controls with other components held fixed. (b) Aggregate accuracy under test-time tag noise (25 runs per level), with the reported training-heterogeneity endpoints shown separately. (c) Component ablations relative to full ARBOR; annotation uncertainty denotes standard deviation across five seeds. (d) Alignment with supplied specialty labels. Only available numerical summaries are shown; heterogeneity endpoints are not interpolated.

Under 0%, 10%, 20%, and 30% test-time tag noise, aggregate accuracy is 69.69, 69.58, 69.47, and 69.31, respectively. The aggregate trend supports sensitivity to tag quality. Subgroup effects require a defined clean-subset reference and are omitted from the figure.

*c) Atom specialization and taxonomy alignment:* Atom analysis indicates shared and specialty-selective activation and masking effects. Clustering by active-atom signatures gives adjusted Rand index 0.62 against specialty labels (Fig. 3(d)), compared with 0.27 for MoELoRA and 0.11 for the rank-wise control. This is alignment with the supplied taxonomy: specialty tags enter the router, and the axis loss explicitly promotes within-specialty agreement. The result cannot establish independent discovery of a medical taxonomy. Taxonomy metrics without specialty inputs and axis supervision remain an open control. Activation patching and pathway-level interventions have been used to investigate hallucination mechanisms in vision-language models [26]. Our atom-masking analysis measures sensitivity to adapter components, without establishing causal explanations of clinical knowledge.

*d) Component ablations:* Figure 3(c) reports the component ablations. Removing the interaction costs 0.28 points, while removing axis contrast costs 0.47; the additive gate exceeds a joint MLP by 0.41. Removing injection scaling has the largest reported loss, 0.74 points. These results support the full design, but specialty-only and operation-only comparisons are unavailable. The reported budget sweep peaks at  $\kappa = 4$ , without implying a universal optimal rank budget.

*e) Injection strength:* Mean  $\rho(x)$  is 0.55 for recall and 0.83 for diagnosis. It is 0.84 on questions where ARBOR corrects a base-model error and 0.61 where the base is already correct. These associations describe input-dependent update strength;  $\rho$  is not a validated uncertainty or safety measure.

TABLE III  
REPORTED QWEN3-8B COSTS ON ONE A100-80GB. PARAMETERS ARE TOTAL TRAINABLE PARAMETERS; MEMORY IS PEAK INFERENCE MEMORY. LATENCY IS PER QUESTION UNDER GREEDY DECODING.

| Method            | Params (M) | Active rank | Train (min) | Latency (ms) | Memory (GB) |
|-------------------|------------|-------------|-------------|--------------|-------------|
| LoRA r4           | 10.91      | 4           | 38          | 416          | 19.6        |
| LoRA r16          | 43.65      | 16          | 41          | 418          | 19.8        |
| LoRA r24          | 65.47      | 24          | 44          | 421          | 20.0        |
| DoRA              | 66.87      | 24          | 58          | 435          | 21.0        |
| HydraLoRA         | 44.90      | 16          | 49          | 452          | 20.6        |
| MoELoRA           | 92.60      | 16          | 63          | 486          | 22.4        |
| Rank-wise control | 64.88      | 4           | 46          | 428          | 20.1        |
| ARBOR             | 66.59      | 4           | 47          | 431          | 20.1        |

Reported ECE is 4.2 for ARBOR and 6.4 for LoRA r16 [27]; interpreting this difference requires a fully documented confidence and binning protocol.

## VII. TRANSFER AND COMPUTATIONAL COST

*a) Source and backbone checks:* When the Chinese training source changes to CMExam, ARBOR obtains 70.11 macro accuracy versus 69.02 for MoELoRA. Reported five-seed advantages over MoELoRA are 0.97 points on Qwen3-4B and 0.84 on Llama-3.1-8B [28]. These checks extend the evaluation across two model families and scales up to 8B; they do not establish scaling behavior beyond that range.

*b) Active rank and measured cost:* ARBOR activates four atoms per question but stores a 16-atom bank and a router. Table III reports 66.59M trainable parameters, 47 minutes of training, 431 ms per question, and 20.1 GB peak inference memory. Compared with MoELoRA, the total parameter count is 28.1% lower, and the reported training and inference costs are smaller. Relative to LoRA r4, however, ARBOR uses

more parameters and memory and has 3.6% higher latency. Dense masking does not translate a four-atom budget into the execution cost of a rank-4 adapter.

The router uses features from a frozen middle-layer pass before adaptation. The timing protocol does not specify feature caching or whether this pass is included, and parameter counts do not separate the atom bank and router. These implementation details, along with generation lengths and batching, limit conclusions about end-to-end serving cost.

## VIII. DISCUSSION AND LIMITATIONS

The results support clinically structured conditional rank allocation in the evaluated medical QA settings. The illustrative separation assumes orthogonal, equiprobable targets, whereas clinical specialties overlap and may be imbalanced. Taxonomy alignment is conditioned on supplied specialty priors and axis supervision.

Open controls include a LoRA r4 heterogeneity sweep, repeated specialty combinations, single-axis gates, and clinical-input ablations retaining orthogonality and scaling. Full reproducibility requires evaluation identifiers, bilingual tagging rules, router dimensions, annotation sampling, and timing scope. Seed intervals do not replace prediction-level uncertainty. Public QA benchmarks and backbones up to 8B leave open-ended reasoning, prospective clinical validity, and larger-model behavior untested. Selected atoms and injection strength are not validated clinical explanations or safety signals.

## IX. CONCLUSION

ARBOR combines question semantics, specialty and operation priors, top- $\kappa$  rank selection, and learned injection scaling in a shared atom bank. Across five seeds, it improves four-benchmark macro accuracy over fixed-rank and expert-based baselines. Tag interventions and ablations support the structured design, while the heterogeneity trend motivates rank-matched experiments. The orthogonal-task analysis explains a potential benefit of conditional allocation within its assumptions.

## REFERENCES

- [1] E. J. Hu, Y. Shen, P. Wallis *et al.*, “LoRA: Low-rank adaptation of large language models,” in *International Conference on Learning Representations (ICLR)*, 2022.
- [2] Q. Zhang, M. Chen, A. Bukharin *et al.*, “AdaLoRA: Adaptive budget allocation for parameter-efficient fine-tuning,” in *International Conference on Learning Representations (ICLR)*, 2023.
- [3] T. Luo, J. Lei, F. Lei *et al.*, “MoELoRA: Contrastive learning guided mixture of experts on parameter-efficient fine-tuning for large language models,” *arXiv preprint arXiv:2402.12851*, 2024, preprint.
- [4] C. Tian, Z. Shi, Z. Guo *et al.*, “HydraLoRA: An asymmetric LoRA architecture for efficient fine-tuning,” in *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.
- [5] Z. Tang, F. Wang, H. He, J. Zhou, T. Lv, J. Zhu, S. Chen, M. Yang, Y. Wang, J. Wu, Y. Song, Y. Li, J. Huang, B. He, and J. Yao, “Aligning LLMs with biomedical knowledge using balanced fine-tuning,” *arXiv preprint arXiv:2511.21075*, 2025, preprint, version 3, 2026.
- [6] Z. Tang, H. He, S. Chen, J. Zhu, T. Lv, J. Zhou, J. Huang, Y. Li, G. Chen, L. You, and C. Y.-C. Chen, “scProtoTransformer: Scalable reference mapping across molecules, cells, and donors,” *Science Advances*, vol. 12, no. 35, p. eaef0286, 2026.
- [7] J. Zhang, Z. Pang, M. Li, Y. Wang, G. Dong, C. Xue, Y. Wu, Z. Li, H. Phan, S. Zhao, B. W. Schuller, and J. Luo, “Modality disentangled learning for incomplete multimodal emotion recognition: A primitive memory distillation perspective,” *arXiv preprint arXiv:2608.30563*, 2026, preprint.
- [8] O. Ostapenko, Z. Su, E. Ponti *et al.*, “Towards modular LLMs by building and reusing a library of LoRAs,” in *International Conference on Machine Learning (ICML)*, 2024.
- [9] J. Xu, J. Lai, and Y. Huang, “MeteoRA: Multiple-tasks embedded LoRA for large language models,” *arXiv preprint arXiv:2405.13053*, 2024, preprint.
- [10] W. Li, Z. Song, H. Zhou *et al.*, “LoRA-Mixer: Coordinate modular LoRA experts through serial attention routing,” *arXiv preprint arXiv:2507.00029*, 2025, preprint.
- [11] W. Shi, R. Xu, Y. Zhuang *et al.*, “MedAdapter: Efficient test-time adaptation of large language models towards medical reasoning,” *arXiv preprint arXiv:2405.03000*, 2024, preprint.
- [12] S. Li, Y. Zhang, J. Yuan, S. Zhang, F. Zhang, and H. Lin, “Enhancing biomedical named entity recognition via multiple programming languages instruction tuning and ensemble method,” *arXiv preprint arXiv:2610.02949*, 2026, preprint.
- [13] S. Li, Y. Zhang, S. Zhang, J. Yuan, F. Zhang, and H. Lin, “A guideline-augmented multi-agent framework for schema-as-code biomedical named entity recognition,” *arXiv preprint arXiv:2610.02970*, 2026, preprint.
- [14] Z. Zhao, Y. Zhou, Z. Zhang *et al.*, “Each rank could be an expert: Single-ranked mixture of experts LoRA for multi-task learning,” *arXiv preprint arXiv:2501.15103*, 2025, preprint.
- [15] C. Eckart and G. Young, “The approximation of one matrix by another of lower rank,” *Psychometrika*, vol. 1, no. 3, pp. 211–218, 1936.
- [16] Y. Bengio, N. Léonard, and A. Courville, “Estimating or propagating gradients through stochastic neurons for conditional computation,” *arXiv preprint arXiv:1308.3432*, 2013, preprint.
- [17] E. Jang, S. Gu, and B. Poole, “Categorical reparameterization with Gumbel-Softmax,” in *International Conference on Learning Representations (ICLR)*, 2017.
- [18] Z. Tang, G. Chen, H. Yang, W. Zhong, and C. Y.-C. Chen, “DSIL-DDI: A domain-invariant substructure interaction learning for generalizable drug–drug interaction prediction,” *IEEE Transactions on Neural Networks and Learning Systems*, vol. 35, no. 8, pp. 10552–10560, 2024.
- [19] A. Yang, A. Li, B. Yang *et al.*, “Qwen3 technical report,” *arXiv preprint arXiv:2505.09388*, 2025, preprint.
- [20] X. Wang, G. H. Chen, D. Song *et al.*, “CMB: A comprehensive medical benchmark in Chinese,” *arXiv preprint arXiv:2308.08833*, 2024, preprint.
- [21] J. Liu, P. Zhou, Y. Hua *et al.*, “Benchmarking large language models on CMExam: A comprehensive Chinese medical exam dataset,” *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks*, 2023.
- [22] D. Jin, E. Pan, N. Oufattole *et al.*, “What disease does this patient have? a large-scale open domain question answering dataset from medical exams,” *Applied Sciences*, vol. 11, no. 14, p. 6421, 2021.
- [23] A. Pal, L. K. Umapathi, and M. Sankarasubbu, “MedMCQA: A large-scale multi-subject multi-choice dataset for medical domain question answering,” in *Proceedings of the Conference on Health, Inference, and Learning (CHIL)*, ser. Proceedings of Machine Learning Research, vol. 174, 2022, pp. 248–260.
- [24] Q. Jin, B. Dhingra, Z. Liu *et al.*, “PubMedQA: A dataset for biomedical research question answering,” in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2019, pp. 2567–2577.
- [25] S.-Y. Liu, C.-Y. Wang, H. Yin *et al.*, “DoRA: Weight-decomposed low-rank adaptation,” in *International Conference on Machine Learning (ICML)*, 2024.
- [26] J. Liu, D. Zhong, Y. Wang, Z. Yang, Z. Kang, G. Dong, Q. Zhan, P. Fang, and A. Liu, “Dual-Pathway circuits of object hallucination in vision-language models,” *arXiv preprint arXiv:2605.13156*, 2026, preprint.
- [27] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, “On calibration of modern neural networks,” in *International Conference on Machine Learning (ICML)*, ser. Proceedings of Machine Learning Research, vol. 70, 2017, pp. 1321–1330.
- [28] A. Grattafiori, A. Dubey, A. Jauhri *et al.*, “The Llama 3 herd of models,” *arXiv preprint arXiv:2407.21783*, 2024, preprint.