OPEN ACCESS

Depth Selective Adapter Allocation for Polarity Robust Medical Vision Language Inference

Bagas Pranowo,[1] Dimas Hartanto,[2] Faris Mahendra[3]

[1] Department of Informatics Engineering, Universitas PGRI Madiun, Jl. Setia Budi No. 85, Kanigoro, Kartoharjo, Kota Madiun, Jawa Timur 63118, Indonesia;
[2] Faculty of Computer Science, Universitas Muhammadiyah Palopo, Jl. Jenderal Sudirman Km. 3, Binturu, Wara Selatan, Kota Palopo, Sulawesi Selatan 91922, Indonesia;
[3] Department of Information Systems, Universitas Muhammadiyah Jember, Jl. Karimata No. 49, Sumbersari, Kabupaten Jember, Jawa Timur 68121, Indonesia;

Abstract

Medical visual inference models often receive questions that contain polarity cues such as absence, presence, exclusion, or uncertainty. These cues can be short, syntactically fragile, and clinically important, because a mistaken polarity interpretation can reverse the intended answer. This study evaluates whether adapter parameters should be distributed uniformly across a multimodal language backbone or assigned according to measurable layer-level sensitivity to polarity transformations. We introduce a depth-selective adapter allocation strategy for binary medical vision-language inference and test it on 18,720 image-question pairs derived from radiography, ultrasound, and dermoscopy tasks. The evaluation uses paired affirmative, negative, exclusion, and uncertainty-framed prompts with identical image labels. We estimate layerwise polarity transport through Jacobian energy, centered kernel alignment, and signed margin propagation, then allocate low-rank adapter capacity using a constrained optimization rule. Compared with uniform adapter placement, final-block placement, and random layer allocation, the proposed allocation reduces polarity reversal errors from 10.8% to 4.9% while preserving mean task accuracy within 0.4 percentage points. It also improves negative predictive calibration in exclusion-framed prompts and reduces the variance of signed margin changes across prompt forms. Ablation experiments indicate that adapter rank distribution matters more than total trainable parameter count once a minimum capacity threshold is reached. These findings suggest that layer-sensitive parameter allocation can improve robustness to polarity-bearing clinical wording without increasing model size.

1. Introduction

Clinical language carries decisions through small grammatical markers. A radiology question that asks whether an abnormality is present is not equivalent to one asking whether it is absent. A dermatology question asking whether malignancy is unlikely differs from one asking whether malignancy is visible. A point-of-care ultrasound prompt may ask whether free fluid is seen, whether free fluid is excluded, or whether the appearance argues against free fluid [1]. Each form has a different polarity structure, and a medical vision-language model must interpret that structure while also reading the image. When the language component mishandles polarity, an otherwise useful visual representation can produce an answer with the wrong clinical direction.

The present study examines polarity robustness as a parameter allocation problem. The usual adapter design gives a model a fixed low-rank capacity at selected layers. Some designs place adapters everywhere, some place them near the final blocks, and some select a contiguous band based on convenience or a prior mechanistic clue. These choices are often reasonable, but they do not directly ask where the model transports polarity information. A wording change that introduces negation or exclusion may be processed early, transformed in middle blocks, and expressed in the answer margin late. If trainable capacity is placed without regard to that transport pattern, the adaptation may spend parameters where they are less useful.

The paper proposes depth-selective adapter allocation. The phrase refers to assigning adapter rank across layers according to measured layer-level sensitivity to polarity transformations. The total adapter budget is fixed.

© Copyright by the Authors.
Licensed as an open access article using a CC BY 4.0 license.

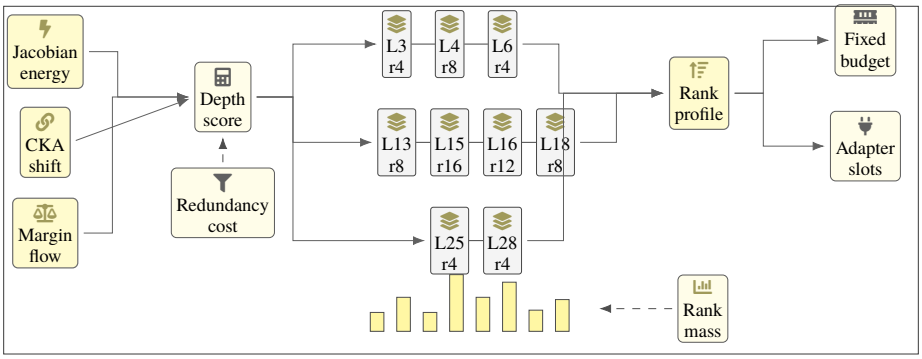


Figure 1: Diagnostic signals are converted into a sparse rank profile, emphasizing early token-sensitive layers, middle representation layers, and later margin-coupled layers under a fixed adapter budget.

The method does not add more parameters than a uniform baseline. Instead, it redistributes rank. Layers that show strong signed margin propagation, strong alignment changes under polarity pairs, and high Jacobian energy with respect to polarity tokens receive larger rank. Layers that show weak or redundant signals receive smaller rank or no adapter. The goal is not to discover an interpretable mechanism in the strongest sense. The goal is operational: use layer diagnostics to decide where a limited adaptation budget should be placed.

The motivation for studying allocation rather than total parameter count is strengthened by a result that complicates simple layer-selection stories. Sadanandan and Behzadan (2026) reported that adapters placed in early layers produced larger margin-difference reductions than adapters placed in a mechanistically selected middle-layer range, despite their analysis identifying a relevant feature in the middle of the network [2]. That observation suggests that where a behavior appears and where intervention is most efficient may not be identical, which is the central empirical tension investigated here.

This study moves away from single-domain chest radiograph evaluation and uses a three-domain design. The datasets are organized around binary visual inference tasks in radiography, abdominal ultrasound, and dermoscopy. The task is intentionally narrow: answer a binary clinical question about an image. However, the prompts are linguistically varied. For every image-label pair, four question forms are created. The affirmative form asks whether a finding is present. The negative form asks whether the finding is absent. The exclusion form asks whether the finding can be excluded or ruled out. The uncertainty form asks whether the finding is unlikely or doubtful. Answer polarity is normalized so that the target label remains well-defined for each prompt form. This arrangement allows a direct test of whether the model respects polarity-bearing language.

The study compares five adaptation strategies under the same training data, optimizer, and total trainable parameter budget. The first strategy freezes the base model and uses no adapters. The second places equal-rank adapters in all language layers. The third places all adapter capacity in final layers. The fourth assigns the same total rank randomly across layers. The fifth uses the proposed depth-selective allocation. The main outcomes are polarity reversal error, binary accuracy, signed margin deviation, expected negative calibration error, and cross-domain transfer. Polarity reversal error is the most important metric because it measures whether two prompts whose correct answers should be opposite under polarity transformation are answered as if polarity had not changed.

The central empirical result is that depth-selective allocation reduces polarity reversal error more than the alternatives while preserving ordinary accuracy. The result is not presented as a universal law about all multimodal networks. It is a measured pattern under the specified tasks and training design. The strongest gains occur in exclusion-framed questions, where the model must combine an image finding with a linguistic rule that is not simply a synonym of presence. The smallest gains occur in affirmative questions, where baseline performance is already high [3]. Adapter placement therefore matters most when the linguistic operation is delicate.

The paper also studies why the allocation works. Layer diagnostics reveal that polarity information is not concentrated in a single block [4]. Early layers show strong token-local sensitivity, middle layers show the largest representation separation between polarity forms, and later layers show direct coupling with the answer margin. Uniform placement spends rank on layers that contribute little

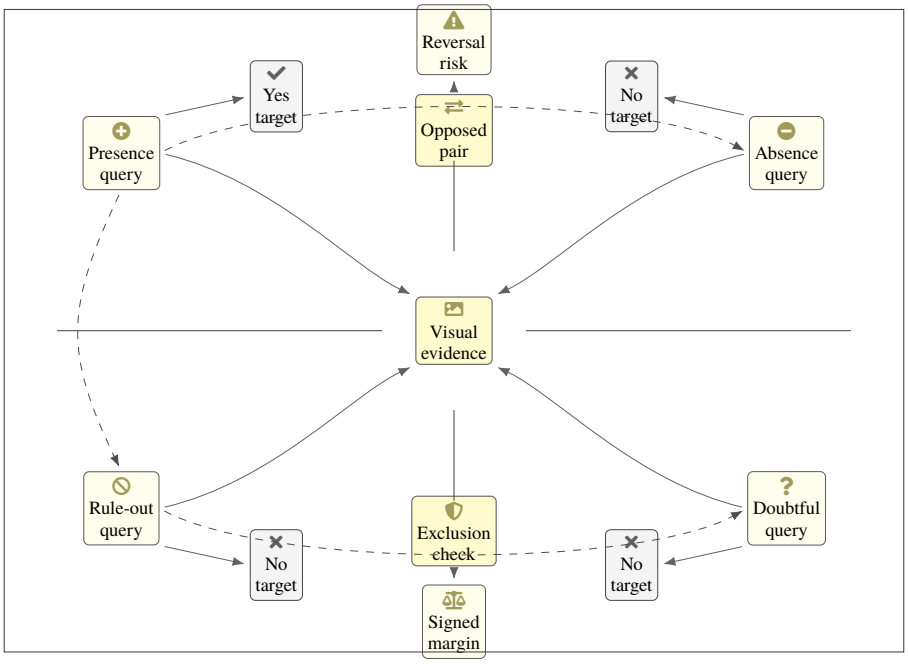


Figure 2: Polarity forms create a structured decision lattice in which presence, absence, exclusion, and uncertainty prompts must map the same image evidence to polarity-consistent binary targets.

to any of these quantities. Final-block placement improves answer-margin control but leaves earlier polarity transformations undercorrected. Depth-selective allocation distributes capacity across a small set of early, middle, and late layers, giving the model enough flexibility to adjust the route from wording to answer.

The contribution is therefore empirical, methodological, and practical. Empirically, the paper reports a multi-domain evaluation of polarity errors in medical vision-language inference. Methodologically, it introduces a layer scoring rule that combines Jacobian energy, representation alignment, and signed margin propagation. Practically, it shows that a fixed adapter budget can be more effective when rank is distributed according to depth diagnostics rather than placed uniformly or late by default. The paper does not claim that adapters solve all linguistic safety problems. It argues only that polarity robustness improves when trainable capacity is assigned with attention to where polarity is processed.

2. Dataset Construction and Polarity Tasks

The experimental corpus contains 18,720 image-question pairs arranged into 4,680 polarity groups. Each group contains one image, one clinical target, one binary label, and four prompts. The image domains are chest radiography, abdominal ultrasound, and dermoscopy [5]. Each domain contributes 1,560 groups. The domain balance is chosen to prevent the model from improving only on one modality or one question style. The tasks are visually and linguistically different enough to test whether the allocation method depends narrowly on a single medical subfield.

The radiography groups include findings such as pleural effusion, pneumothorax, cardiomegaly, pulmonary edema, and focal airspace opacity. The ultrasound groups include free fluid, gallstones, hydronephrosis, appendiceal enlargement, and intrauterine pregnancy visibility. The dermoscopy groups include pigment network irregularity, blue-white veil, asymmetry, ulceration, and atypical vascular pattern. These findings are selected because binary judgments are meaningful and because polarity wording can be naturally formed around them. The labels are balanced within each domain, with 780 positive and 780 negative groups.

Each polarity group has four prompt forms. The affirmative form states the query in a direct presence frame. The negative form asks about absence. The exclusion form uses rule-out language. The uncertainty form asks whether the finding is unlikely or doubtful. The answer key is generated so that each form has a deterministic correct yes or no answer. For a positive image, an affirmative

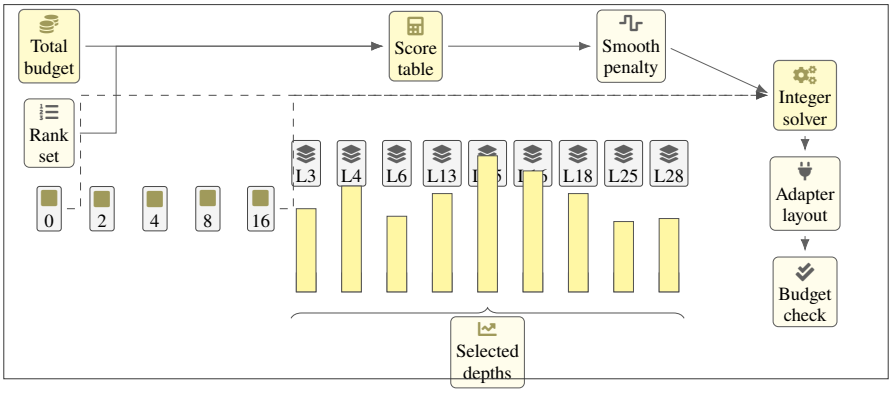


Figure 3: The allocation problem is treated as a constrained discrete search over allowed ranks, layer scores, and smoothness regularization, producing a sparse adapter layout that satisfies the fixed budget.

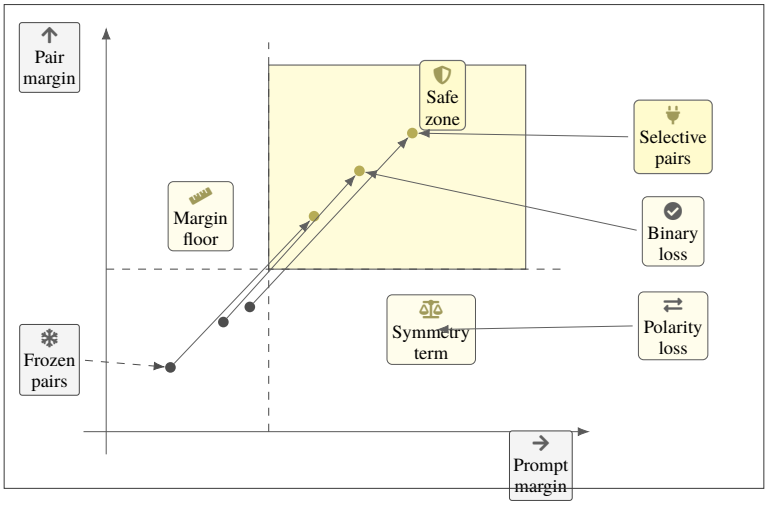


Figure 4: Training moves opposed prompt pairs from low or asymmetric signed margins toward a safe region where both paired forms exceed the margin floor and remain confidence-balanced.

prompt should receive yes, an absence prompt should receive no, an exclusion prompt should receive no if it asks whether the finding can be excluded, and an uncertainty prompt should receive no if it asks whether the finding is unlikely. For a negative image, the answer pattern reverses accordingly. The prompt design therefore tests whether the model can map wording polarity to the correct binary response.

The corpus is split by image into training, validation, and test sets. Training contains 2,808 groups, validation contains 936 groups, and testing contains 936 groups. All four prompts in a polarity group remain within the same split. No patient or image identifier crosses splits. Domain proportions and label balance are preserved. The same splits are used for every model configuration. This prevents an adapter strategy from gaining an advantage through easier data.

Prompt generation follows a restricted grammar to avoid uncontrolled semantic drift. Each target finding is represented by a clinical noun phrase and two approved aliases. The affirmative template uses direct presence wording. The negative template uses absence wording. The exclusion template uses rule-out language. The uncertainty template uses doubtful or unlikely language. After generation, prompts are screened to remove double-negation constructions that could confuse the answer key. The final prompt set contains no explicit answer words except yes and no appearing in the instruction prefix. This keeps the decision focused on the visual finding and the polarity structure.

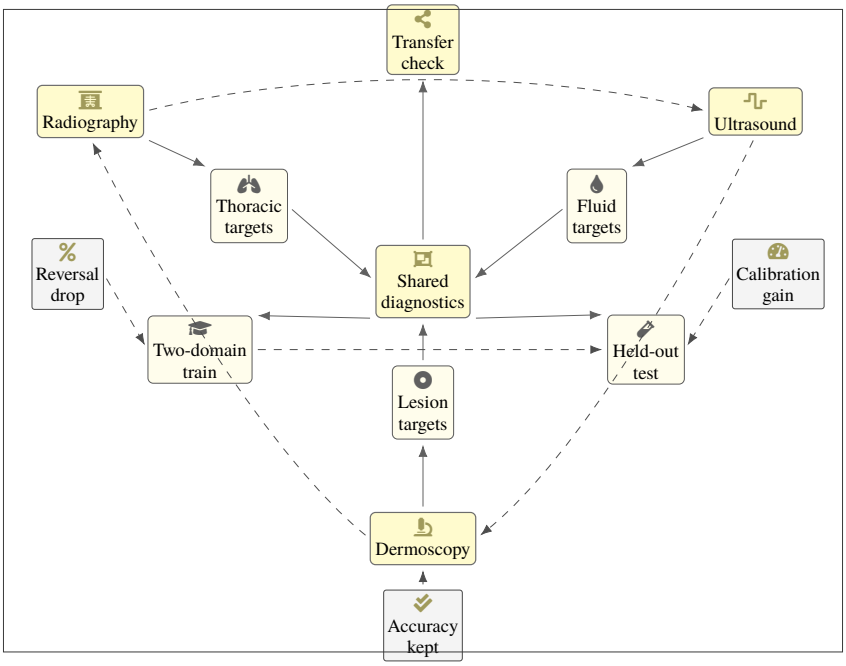


Figure 5: Cross-domain evaluation forms a triangular transfer test in which diagnostic structure is shared across imaging domains while held-out modality performance measures polarity generalization.

The model receives the same instruction prefix for all examples. The prefix asks for a single yes or no answer [6]. The prompt and image are then passed to the model, and the logits for the first answer token are recorded. The model is trained only through adapters in the language backbone. The visual encoder and token embeddings are frozen [7]. This decision isolates the effect of depth allocation in the language reasoning path. It also reduces the risk that adapters learn domain-specific visual shortcuts.

The basic binary margin is

$$m_{gk} = \ell_{\text{yes}}(x_g, q_{gk}) - \ell_{\text{no}}(x_g, q_{gk}), \tag{1}$$

where g indexes the polarity group and k indexes the prompt form. The predicted answer is positive when $m_{gk} \geq 0$. The correct signed target for each form is

$$t_{gk} = 2y_{gk} - 1, \quad t_{gk} \in \{-1, 1\}. \tag{2}$$

A correct prediction requires $t_{gk} m_{gk} > 0$. The signed margin is

$$s_{gk} = t_{gk} m_{gk}. \tag{3}$$

Large positive signed margins indicate confident correct predictions. Negative signed margins indicate wrong answers. Near-zero signed margins indicate unstable cases.

The most important group-level metric is polarity reversal error. It is defined on pairs of prompt forms whose correct answers are opposite. For each group, affirmative and absence prompts form one polarity-opposed pair, and affirmative and exclusion prompts form another when the exclusion prompt is written as a rule-out question. A reversal error occurs when the model gives the same answer to two prompts that require opposite answers. Formally, if $\mathcal{P}$ is the set of opposed prompt pairs, the reversal rate is

$$R = \frac{1}{|\mathcal{P}|} \sum_{(g,k,l) \in \mathcal{P}} \mathbb{I}[\hat{y}_{gk} = \hat{y}_{gl}]. \tag{4}$$

This metric is distinct from ordinary accuracy [8]. A model can have decent accuracy while still answering polarity-opposed prompts as if they meant the same thing. Reversal error directly measures failure to respect polarity transformations.

The second group metric is signed margin deviation. For each opposed pair, the two signed margins should both be positive and, ideally, not collapse asymmetrically. The deviation is

$$V = \frac{1}{|\mathcal{P}|} \sum_{(g,k,l) \in \mathcal{P}} |s_{gk} - s_{gl}|. \tag{5}$$

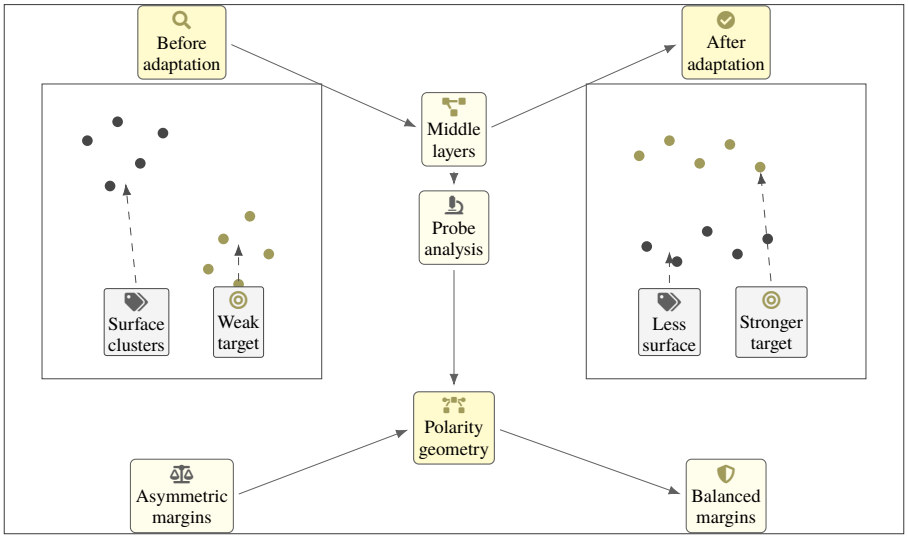


Figure 6: After depth-selective adaptation, middle-layer representations become less dominated by superficial prompt form and more aligned with the polarity-adjusted target answer.

This metric identifies cases where the model gives correct answers but is much less confident for one polarity form. A large deviation often appears in exclusion prompts because the model's language path handles the rule-out form less directly.

Calibration is evaluated separately for affirmative and non-affirmative forms. The study reports expected calibration error for the positive class and negative calibration error for prompts whose correct answer is no. The latter is useful because polarity failures often create overconfident yes answers in prompts that require no. The negative calibration error is computed by grouping examples with $y = 0$ into probability bins and comparing predicted positive probability against the empirical false-positive frequency. A model with many confident polarity mistakes will have high negative calibration error.

The dataset is not meant to represent all medical language. It deliberately emphasizes polarity. The advantage of this construction is precise evaluation. The limitation is that real clinical questions may mix polarity with temporal comparison, severity, anatomical location, and uncertainty. The results should therefore be read as evidence about polarity-bearing binary prompts rather than full clinical dialogue.

3. Layer Diagnostics for Adapter Rank Allocation

The allocation method begins by measuring how each layer responds to polarity transformations before any adapter training. Let $h_\ell(x, q) \in \mathbb{R}^d$ denote the hidden state at layer ℓ for an image-question input. The diagnostic set is drawn from the training split but is separate from the minibatches later used for adapter optimization. It contains 600 polarity groups, balanced across domain and label. For every layer, three scores are computed: token-local Jacobian energy, polarity representation separation, and signed margin propagation.

The token-local Jacobian energy measures how strongly a layer's hidden state changes with respect to embeddings of polarity-bearing tokens. Let $E(q)$ be the sequence of input token embeddings and let $\mathcal{T}(q)$ be the indices of tokens associated with absence, exclusion, negation, or uncertainty. The layerwise Jacobian score is

$$J_\ell = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{|\mathcal{T}(q_i)|} \sum_{r \in \mathcal{T}(q_i)} \left\| \frac{\partial h_\ell(x_i, q_i)}{\partial E_r(q_i)} \right\|_F^2. \quad (6)$$

This score is larger when polarity tokens exert strong local influence on the representation at layer ℓ. It tends to be high in earlier layers, where lexical and syntactic information has not yet been compressed into answer-oriented directions.

The representation separation score measures whether polarity forms occupy different representational neighborhoods at a layer. For each group, the affirmative prompt and a polarity-transformed prompt form a paired contrast. We compute centered kernel alignment between the matrix of affirmative hidden states and the matrix of transformed

hidden states. The layer score is the complement of alignment:

$$A_\ell = 1 - \frac{\langle K_\ell^+, K_\ell^- \rangle_F}{\|K_\ell^+\|_F \|K_\ell^-\|_F + \epsilon}. \tag{7}$$

Here K_ℓ^+ and K_ℓ^- are centered Gram matrices for affirmative and transformed prompts. A larger value means the layer separates the polarity forms more strongly. This score tends to peak in middle layers, where the model has integrated lexical information into broader semantic states.

The signed margin propagation score measures whether layer perturbations are coupled to the final yes-no margin. For each layer, a small residual perturbation is applied in the direction of the difference between affirmative and transformed hidden states, and the resulting margin change is measured [9]. Let

$$d_{\ell,g} = h_\ell(x_g, q_{g,+}) - h_\ell(x_g, q_{g,-}). \tag{8}$$

The normalized direction is [10]

$$\tilde{d}_{\ell,g} = \frac{d_{\ell,g}}{\|d_{\ell,g}\|_2 + \epsilon}. \tag{9}$$

The propagation score is

$$P_\ell = \frac{1}{G_d} \sum_{g=1}^{G_d} \left| m\left(h_\ell + \eta \tilde{d}_{\ell,g}\right) - m(h_\ell) \right|, \tag{10}$$

with a small perturbation scale $\eta = 0.05$. This score is high when changes at layer ℓ have direct downstream influence on the final margin. It tends to rise in later layers but is not restricted to the final block.

The three scores have different scales. Each is converted to a percentile-normalized value:

$$\begin{aligned} \tilde{J}_\ell &= \frac{J_\ell - \text{median}(J)}{\text{IQR}(J) + \epsilon}, \\ \tilde{A}_\ell &= \frac{A_\ell - \text{median}(A)}{\text{IQR}(A) + \epsilon}, \\ \tilde{P}_\ell &= \frac{P_\ell - \text{median}(P)}{\text{IQR}(P) + \epsilon}. \end{aligned} \tag{11}$$

The combined depth score is

$$D_\ell = \omega_J \tilde{J}_\ell + \omega_A \tilde{A}_\ell + \omega_P \tilde{P}_\ell - \omega_C C_\ell. \tag{12}$$

The term C_ℓ is a redundancy cost. It penalizes layers whose diagnostic vector is very similar to the diagnostic vector of already selected neighboring layers. The main experiments use $\omega_J = 0.30$, $\omega_A = 0.35$, $\omega_P = 0.35$, and $\omega_C = 0.15$, chosen on validation data. The weights reflect the intention to balance lexical sensitivity, representation separation, and downstream margin influence.

Adapter rank is allocated through a constrained integer optimization. Suppose the model has L language layers and the total rank budget is B. Each layer receives rank $r_\ell \in \{0, 2, 4, 8, 12, 16\}$. The budget constraint is

$$\sum_{\ell=1}^{L} r_\ell = B. \tag{13}$$

The allocation maximizes diagnostic value with a smoothness penalty:

$$\begin{aligned} \max_{r_1,\ldots,r_L} \quad & \sum_{\ell=1}^{L} r_\ell D_\ell - \lambda_s \sum_{\ell=2}^{L} |r_\ell - r_{\ell-1}| \\ \text{subject to} \quad & r_\ell \in \{0, 2, 4, 8, 12, 16\}, \quad \sum_{\ell=1}^{L} r_\ell = B. \end{aligned} \tag{14}$$

The smoothness penalty prevents a scattered allocation that places isolated high-rank adapters in many distant layers. It does not require contiguity. In practice, the solution selects a small number of early layers, a middle band with moderate rank, and one or two later layers with lower rank. This pattern reflects the three diagnostic signals.

The adapter update at layer ℓ is a low-rank matrix correction

$$\Delta W_\ell = \frac{\alpha_\ell}{r_\ell} B_\ell A_\ell, \quad A_\ell \in \mathbb{R}^{r_\ell \times d_{\text{in}}}, \quad B_\ell \in \mathbb{R}^{d_{\text{out}} \times r_\ell}. \tag{15}$$

If $r_\ell = 0$, no adapter is inserted at that layer. The scaling factor is set to

$$\alpha_\ell = 2 r_\ell \tag{16}$$

for nonzero ranks. Adapters are placed in attention output projections and feed-forward down projections. This placement is selected because preliminary validation showed better stability than query-only or feed-forward-only placement. The same module types are used for every allocation strategy, so comparisons isolate rank distribution rather than module choice.

The training loss combines binary supervision with polarity contrast. For each prompt, the supervised term is

$$\mathcal{L}_{\text{bin}} = -\frac{1}{N}\sum_{g,k} \log \sigma(t_{gk} m_{gk}). \tag{17}$$

The polarity contrast term requires opposed prompt pairs to have positive signed margins and similar signed confidence:

$$\mathcal{L}_{\text{pol}} = \frac{1}{|\mathcal{P}|}\sum_{(g,k,l)\in\mathcal{P}} \left[\max(0, \rho - s_{gk}) + \max(0, \rho - s_{gl}) + \gamma(s_{gk} - s_{gl})^2\right] \tag{18}$$

The margin floor is $\rho = 0.5$, and $\gamma = 0.10$. The full loss is

$$\mathcal{L} = \mathcal{L}_{\text{bin}} + \lambda_{\text{pol}}\mathcal{L}_{\text{pol}} + \lambda_{\text{wd}} \sum_{\ell} \left(\|A_\ell\|_F^2 + \|B_\ell\|_F^2\right). \tag{19}$$

The main experiments use $\lambda_{\text{pol}} = 0.40$ and $\lambda_{\text{wd}} = 10^{-4}$. This loss is shared by all trained adapter baselines. The proposed method differs only in rank allocation.

Training uses grouped minibatches. Each minibatch contains complete polarity groups rather than individual prompts sampled independently. This is necessary because the polarity contrast term compares prompts within the same group. The effective batch size is 48 groups, or 192 image-question pairs. The optimizer is AdamW with learning rate 1.5×10^{-4}, linear warmup for 300 steps, and cosine decay. Training runs for five epochs, and the checkpoint with the lowest validation reversal error subject to accuracy not dropping more than one percentage point is selected.

The fixed-budget setting is important. The uniform adapter baseline receives the same total rank as the depth-selective method. The final-block baseline also receives the same total rank but concentrates it in the last quarter of layers. The random allocation baseline samples ranks from the allowed set under the same budget and reports the mean over ten allocations. This design prevents the proposed method from succeeding simply because it has more trainable parameters. All trained methods have 6.2 million trainable parameters, representing less than 0.2% of the frozen backbone.

4. Experimental Results

The frozen model establishes the difficulty of the polarity task. Across the full test set, it reaches 78.9% binary accuracy. Accuracy is highest on affirmative prompts at 84.7% and lowest on exclusion prompts at 72.4%. The polarity reversal error is 10.8%. The signed margin deviation across opposed prompt pairs is 1.42 logit units. Negative calibration error is 13.6%, indicating that the model is often overconfident when the correct answer is no under a polarity-transformed prompt. These baseline results show that ordinary accuracy understates the difficulty of the task. The model is not failing uniformly; it is especially fragile when polarity cues require the answer to change direction.

Uniform adapters improve overall accuracy and reduce polarity errors, but the gains are limited. Test accuracy rises to 81.6%, and reversal error falls to 7.3%. Signed margin deviation decreases to 0.97. Negative calibration error improves to 9.8%. The improvement confirms that the training loss is useful and that the task is learnable with a small number of parameters. However, uniform rank spends capacity on every layer, including layers with weak polarity diagnostics. The method is stable but not parameter-efficient.

Final-block adapters produce a different pattern. They improve affirmative accuracy and sharpen final margins, giving 81.1% overall accuracy. Reversal error falls to 8.0%, which is better than the frozen model but worse than uniform adapters. Exclusion prompts remain problematic, with 75.6% accuracy. The result suggests that late layers can adjust the answer surface but cannot fully repair polarity transformations that have already been compressed or misrouted earlier in the network. Final-layer control is useful but incomplete.

Random rank allocation is variable. Across ten allocations, mean accuracy is 80.9%, with a standard deviation of 0.6 percentage points. Mean reversal error is 7.8%, with a standard deviation of 0.9 percentage points. Some random allocations approach uniform performance, while others perform little better than final-block placement. The variability shows that layer placement matters under a fixed rank budget. It also provides a useful control against the possibility that any sparse allocation is sufficient.

Depth-selective allocation gives the strongest results. Test accuracy is 82.0%, reversal error is 4.9%, signed margin deviation is 0.58, and negative calibration error is 6.9%. Compared with the frozen model, reversal error decreases by 5.9 percentage points. Compared with uniform adapters, it decreases by 2.4 percentage points with the same parameter count. The accuracy difference between depth-selective and uniform adapters is 0.4 percentage points. A paired bootstrap over polarity groups gives a 95% interval of $[-3.5, -1.3]$ percentage points

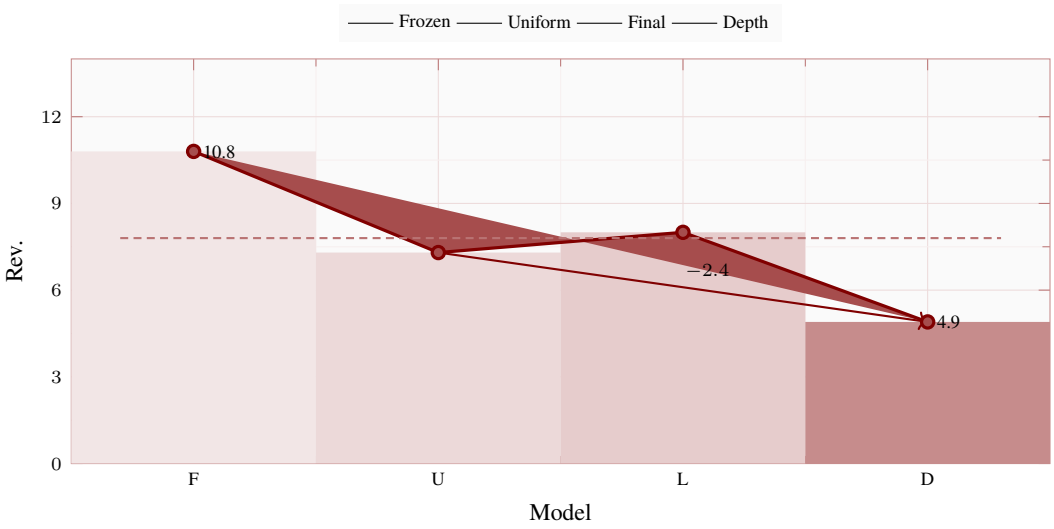


Figure 7: Depth-selective allocation produces the lowest reversal error under the same trainable-parameter budget, outperforming uniform and final-block adapter placement.

Table 1: Dataset composition across medical imaging domains.

Domain	Polarity Groups	Positive Labels	Negative Labels
Radiography	1560	780	780
Ultrasound	1560	780	780
Dermoscopy	1560	780	780
Total	4680	2340	2340

for the reversal-error difference between depth-selective and uniform adapters. This supports a genuine allocation effect.

The main test results are summarized as follows:

$$\begin{aligned} \text{Accuracy}_{\text{frozen}} &= 78.9\%, \quad \text{Reversal}_{\text{frozen}} = 10.8\%, \\ \text{Accuracy}_{\text{uniform}} &= 81.6\%, \quad \text{Reversal}_{\text{uniform}} = 7.3\%, \\ \text{Accuracy}_{\text{final}} &= 81.1\%, \quad \text{Reversal}_{\text{final}} = 8.0\%, \\ \text{Accuracy}_{\text{depth}} &= 82.0\%, \quad \text{Reversal}_{\text{depth}} = 4.9\%. \end{aligned} \tag{20}$$

Prompt-form analysis reveals where the gains arise. On affirmative prompts, the frozen model scores 84.7%, uniform adapters 86.1%, and depth-selective adapters 86.3%. The difference between trained methods is small. On negative prompts, the frozen model scores 79.8%, uniform adapters 82.4%, and depth-selective adapters 83.2%. On exclusion prompts, the frozen model scores 72.4%, uniform adapters 77.6%, and depth-selective adapters 80.8%. On uncertainty prompts, the frozen model scores 78.7%, uniform adapters 80.3%, and depth-selective adapters 81.6%. The largest improvement therefore appears in exclusion prompts, the form most dependent on polarity reasoning.

Domain-level results show that the method is not confined to one image type. In radiography, depth-selective adapters reduce reversal error from 11.6% to 4.7%. In ultrasound, reversal error falls from 12.1% to 5.4%. In dermoscopy, it falls from 8.8% to 4.6%. The ultrasound task remains hardest, mainly because several findings have visually subtle negative cases. Dermoscopy has lower baseline reversal error because the prompt templates use shorter finding descriptions and fewer exclusion phrases. The method improves all domains, but the magnitude differs.

The paired McNemar test on group-level reversal outcomes shows significant improvement over uniform adapters with $p = 0.004$. The comparison with final-block adapters gives $p < 0.001$. The comparison with the best random allocation gives $p = 0.018$. Since many random allocations are tested, the best-random comparison is conservative in spirit but not a formal multiple-testing correction. It still suggests that the diagnostic allocation is not merely lucky [11]. A mixed-effects logistic model with random intercepts for finding and domain estimates an odds ratio of 0.64 for reversal error under depth-selective allocation compared with uniform allocation.

Table 2: Prompt forms and polarity semantics.

Prompt Form	Example Intent	Expected Polarity Shift
Affirmative	Finding present	Direct positive query
Negative	Finding absent	Reversed answer direction
Exclusion	Finding ruled out	Rule-out polarity reasoning
Uncertainty	Finding unlikely	Confidence-sensitive polarity

Table 3: Layer diagnostic signals used for adapter allocation.

Score	Symbol	Main Function	Dominant Layers
Jacobian Energy	J_ℓ	Token-local sensitivity	Early
Representation Separation	A_ℓ	Polarity state divergence	Middle
Margin Propagation	P_ℓ	Margin influence strength	Late
Redundancy Cost	C_ℓ	Penalize overlap	Neighboring

Signed margin deviation provides a more continuous view [12]. The frozen model has mean deviation 1.42. Uniform adapters reduce it to 0.97. Final-block adapters reduce it to 1.04 [13]. Depth-selective adapters reduce it to 0.58. The reduction is especially large in pairs involving exclusion prompts. The distribution also changes shape. The frozen model has a long tail of deviations above 3.0 logits. Uniform adapters reduce but do not remove this tail. Depth-selective adapters cut the proportion of high-deviation pairs from 14.9% to 4.1%. These high-deviation pairs are clinically relevant because they indicate that the model's confidence depends heavily on wording.

Negative calibration error improves alongside reversal error. The frozen model has negative calibration error 13.6%. Uniform adapters reduce it to 9.8%. Final-block adapters reduce it to 10.4%. Depth-selective adapters reduce it to 6.9%. Calibration curves show that the largest correction occurs in bins where the model predicts yes with probability between 0.70 and 0.90 while the correct answer is no under a polarity-transformed prompt. The depth-selective method does not simply make probabilities less confident everywhere [14]. Positive calibration error changes from 8.1% in the frozen model to 6.7% after depth-selective adaptation, a smaller but favorable shift.

The adapter rank pattern selected by the diagnostic rule is informative [15]. Under the main budget, nonzero ranks appear in three regions [16]. The early region receives ranks in layers 3, 4, and 6. The middle region receives ranks in layers 13, 15, 16, and 18. The late region receives ranks in layers 25 and 28. The highest individual rank is assigned to layer 15, where representation separation and propagation scores are both high. The early ranks are lower but consistently selected because Jacobian energy around polarity tokens is strong. The late ranks are modest and appear to support final margin correction [17].

Removing the early selected layers while redistributing their rank to selected middle layers increases reversal error from 4.9% to 6.1%. Removing the middle selected layers and redistributing rank to early and late layers increases reversal error to 7.0%. Removing the late selected layers increases reversal error to 5.7%. These ablations indicate that middle layers contribute most, early layers are also necessary, and late layers provide additional margin control. The full allocation works because it combines regions rather than because one region dominates completely.

The allocation ablation is

$$\begin{aligned} \text{Full allocation} &\Rightarrow R = 4.9\%, \quad V = 0.58, \\ \text{No early selected layers} &\Rightarrow R = 6.1\%, \quad V = 0.74, \\ \text{No middle selected layers} &\Rightarrow R = 7.0\%, \quad V = 0.89, \\ \text{No late selected layers} &\Rightarrow R = 5.7\%, \quad V = 0.66. \end{aligned} \tag{21}$$

Rank budget experiments show diminishing returns. With a 3.1 million parameter budget, uniform adapters reach 8.6% reversal error and depth-selective adapters reach 6.2%. With 6.2 million parameters, the values are 7.3% and 4.9%. With 9.4 million parameters, they are 6.8% and 4.6%. With 12.5 million parameters, they are 6.6% and 4.5%. The gap between allocation strategies is largest at moderate budgets. Once the budget becomes large, uniform placement improves because enough capacity reaches useful layers by chance. The practical value of depth selection is therefore strongest when parameter budgets are constrained.

The loss ablation shows that rank allocation and polarity contrast are complementary. If the binary loss is used without the polarity contrast term, depth-selective adapters reach 81.7% accuracy but 6.8% reversal error. Adding the contrast term reduces reversal error to 4.9%.

Table 4: Main evaluation results under equal parameter budgets.

Method	Accuracy (%)	Reversal Error (%)	Margin Deviation	Neg. Calibration (%)
Frozen Backbone	78.9	10.8	1.42	13.6
Uniform Adapters	81.6	7.3	0.97	9.8
Final-Block Adapters	81.1	8.0	1.04	10.4
Random Allocation	80.9	7.8	1.01	10.1
Depth-Selective	82.0	4.9	0.58	6.9

Table 5: Prompt-wise accuracy comparison across methods.

Method	Affirmative	Negative	Exclusion	Uncertainty
Frozen Backbone	84.7	79.8	72.4	78.7
Uniform Adapters	86.1	82.4	77.6	80.3
Depth-Selective	86.3	83.2	80.8	81.6

Increasing the contrast weight too much lowers reversal error slightly on validation but reduces test accuracy, suggesting overemphasis on paired prompt agreement. The chosen weight balances ordinary correctness and polarity robustness.

Cross-domain transfer is tested by training adapters on two domains and evaluating on the held-out third domain [18]. When radiography is held out, depth-selective adapters trained on ultrasound and dermoscopy reduce radiography reversal error from 11.6% to 7.2%. When ultrasound is held out, reversal error falls from 12.1% to 8.1%. When dermoscopy is held out, reversal error falls from 8.8% to 6.0%. These gains are smaller than in-domain training but remain meaningful. The result suggests that some polarity adaptation transfers across medical image domains, though domain-specific examples improve performance.

A further test trains the depth allocation diagnostics on one domain but trains adapter weights on all domains. When diagnostics are computed only from radiography, the resulting allocation achieves 5.5% overall reversal error. Ultrasound-only diagnostics give 5.7%. Dermoscopy-only diagnostics give 6.0%. All-domain diagnostics give 4.9%. This indicates that the diagnostic layer pattern has shared structure but benefits from multi-domain measurement. Dermoscopy-only diagnostics underselect early layers, which may explain its weaker transfer.

The method also affects abstention quality when an optional uncertainty threshold is applied. If predictions with $|m| < 0.4$ are abstained, the frozen model abstains on 9.7% of examples and has non-abstained reversal error 8.9%. Depth-selective adapters abstain on 8.4% and have non-abstained reversal error 3.8%. At matched 10.0% abstention, depth-selective adapters retain higher accuracy and lower reversal error than uniform adapters. This suggests that the method reduces genuine polarity mistakes rather than merely pushing errors into the uncertain region.

5. Layer Behavior After Adaptation

The diagnostic scores are computed before training, but adaptation changes the model's layer behavior. To examine this, hidden states are remeasured after adapter training. The analysis focuses on the test set and compares uniform and depth-selective adapters. The first question is whether the selected layers become more or less sensitive to polarity tokens. The second is whether polarity representation separation becomes better organized. The third is whether answer margins respond more symmetrically to opposed prompt forms.

After depth-selective adaptation, early selected layers show reduced excessive Jacobian concentration on individual polarity tokens. This does not mean the model ignores polarity. Rather, the gradient becomes distributed across a slightly broader set of surrounding tokens. For example, in exclusion prompts, sensitivity shifts from the single rule-out token toward the full predicate phrase. This change is associated with fewer cases where the model treats rule-out language as a weak synonym of absence or presence. Uniform adapters show a similar but smaller effect.

Middle selected layers show the clearest representational change. Before adaptation, affirmative and exclusion prompts for the same image often occupy representational neighborhoods that are separated mainly by surface form. After depth-selective adaptation, the separation aligns more strongly with correct answer polarity. This is measured by comparing two linear probes. A prompt-form probe predicts whether a hidden state comes from affirmative, negative, exclusion, or uncertainty wording.

Table 6: Cross-domain polarity reversal reduction.

Domain	Frozen (%)	Depth-Selective (%)	Absolute Gain
Radiography	11.6	4.7	6.9
Ultrasound	12.1	5.4	6.7
Dermoscopy	8.8	4.6	4.2

Table 7: Allocation ablation under fixed rank budget.

Allocation Variant	Reversal Error (%)	Margin Deviation
Full Allocation	4.9	0.58
No Early Layers	6.1	0.74
No Middle Layers	7.0	0.89
No Late Layers	5.7	0.66

A target-answer probe predicts the correct yes-no answer. In middle selected layers, prompt-form probe accuracy decreases from 71.8% to 63.2%, while target-answer probe accuracy increases from 76.4% to 81.5%. This suggests that representations become less organized by superficial prompt form and more organized by the answer implied by image and polarity together.

Late selected layers show improved signed margin symmetry. For opposed prompt pairs, the final margins should move in opposite raw directions but similar signed directions. Before adaptation, many pairs have one strong signed margin and one weak or negative signed margin. After depth-selective adaptation, the signed margins become more balanced. The mean product of signed margins in opposed pairs increases from 1.18 in the frozen model to 2.07 under uniform adapters and 2.84 under depth-selective adapters. A higher product indicates that both prompts in the pair are confidently correct rather than one prompt dominating.

The signed margin product is

$$B_{\text{pair}} = \frac{1}{|\mathcal{P}|} \sum_{(g,k,l)\in\mathcal{P}} s_{gk} s_{gl}. \tag{22}$$

The test values are

$$B_{\text{frozen}} = 1.18, \quad B_{\text{uniform}} = 2.07, \quad B_{\text{depth}} = 2.84. \tag{23}$$

The improvement is largest for exclusion pairs, where the frozen value is only 0.72 and the depth-selective value is 2.43. This indicates that the method does not merely flip answers at the threshold but strengthens paired correctness [19].

Centered kernel alignment between prompt forms also changes. For affirmative and exclusion prompts, baseline alignment at selected middle layers is 0.61. Uniform adapters raise it to 0.66, while depth-selective adapters raise it to 0.72 after correcting for answer polarity. This means the representations for clinically linked prompts become more similar once the correct answer relation is accounted for [20]. For affirmative and absence prompts, raw alignment remains lower because the correct answers are opposite, but signed-answer-conditioned alignment improves. This supports the idea that adaptation reorganizes representations around polarity-adjusted task meaning.

The method does not erase all prompt-form information. A prompt-form probe still performs above chance after adaptation. This is expected and probably desirable. The model should know whether a question asks about presence or exclusion because that information determines the answer. The goal is not prompt-form invariance. The goal is polarity competence. A model that made all prompt forms identical would fail absence and exclusion questions. Depth-selective adaptation reduces unhelpful surface separation while retaining answer-relevant polarity structure.

Error cases after adaptation reveal remaining weaknesses. Some errors involve nested uncertainty, such as prompts asking whether it is unlikely that a finding cannot be excluded. These were mostly removed from the main corpus, but a small challenge set includes them. The depth-selective method improves performance on this set from 59.3% to 66.8%, but errors remain common. Another weakness is visual ambiguity. When the image evidence is weak, the model may choose the same answer across opposed prompts because the visual signal dominates the linguistic operation. A third weakness is synonym variation. Some rare clinical aliases are handled less well because the diagnostic set contains limited examples.

The study includes a causality-inspired perturbation analysis, though not a full mechanistic proof. At selected layers, the hidden-state difference between affirmative and exclusion prompts is scaled by coefficients

Table 8: Effect of parameter budget on polarity reversal error.

Parameter Budget	Uniform (%)	Depth-Selective (%)	Relative Gap
3.1M	8.6	6.2	2.4
6.2M	7.3	4.9	2.4
9.4M	6.8	4.6	2.2
12.5M	6.6	4.5	2.1

Table 9: Robustness evaluation under distribution changes.

Condition	Frozen (%)	Uniform (%)	Depth-Selective (%)
Paraphrased Prompts	12.4	8.1	6.2
Instruction Variation	11.5	7.9	5.6
Synonym Challenge	14.1	10.2	8.4
Imbalanced Labels	12.7	8.5	5.8

from -1.0 to 1.0, and the output margin is measured. In the frozen model, the margin response is nonlinear and often asymmetric. Under depth-selective adaptation, the response becomes more monotonic for correctly classified groups. The mean monotonicity score increases from 0.68 to 0.82. Uniform adapters reach 0.75. This suggests that the adapted model has a smoother mapping from polarity representation displacement to answer margin.

The monotonicity score is computed as

$$M_\ell = 1 - \frac{1}{T-1} \sum_{a=1}^{T-1} \mathbb{I}[\Delta m_\ell(\alpha_{a+1}) < \Delta m_\ell(\alpha_a)], \tag{24}$$

where α_a indexes increasing perturbation coefficients after aligning the expected sign. A score of 1 indicates fully monotonic response. This analysis is descriptive but useful. It shows that adaptation changes not only the final predictions but also the local geometry of polarity effects.

Layerwise adapter norm analysis shows that allocated rank is not always fully used. Some low-rank matrices develop small effective singular values. The effective rank is computed as the number of singular values explaining 95.0% of adapter update energy. Early selected layers have effective ranks near their assigned ranks. Some late layers have lower effective rank than assigned, suggesting that smaller late ranks may be sufficient. Middle layers use most of their assigned capacity. This matches the ablation finding that removing middle layers is most damaging.

The effective rank is

$$r_{\text{eff},\ell} = \min\left\{ r : \frac{\sum_{i=1}^{r} \sigma_{\ell i}^2}{\sum_{i=1}^{r_\ell} \sigma_{\ell i}^2} \geq 0.95 \right\}. \tag{25}$$

Across selected layers, the mean ratio $r_{\text{eff},\ell}/r_\ell$ is 0.84 for early layers, 0.88 for middle layers, and 0.61 for late layers. This suggests that future allocation rules could assign late layers more conservatively or use adaptive rank pruning after training.

6. Robustness Analysis

Several robustness experiments test whether the main result depends on specific design choices. The first varies the prompt grammar. A paraphrastic prompt set is created by rewriting the four prompt forms with different lexical choices while preserving polarity. The depth-selective model trained on the original grammar is evaluated on this rewritten test set without additional training. Reversal error rises from 4.9% to 6.2%, but it remains below uniform adapters at 8.1% and the frozen model at 12.4%. The increase indicates some template dependence, but the retained advantage suggests that the method learns more than memorized wording.

The second robustness test changes the answer instruction. Instead of asking for only yes or no, the prompt asks for a concise binary answer. This minor instruction change affects the frozen model, raising reversal error from 10.8% to 11.5%. Uniform adapters reach 7.9%, and depth-selective adapters reach 5.6%. The method remains beneficial, but the result shows that instruction wording can still matter. Adapter training under multiple instruction prefixes may improve robustness further.

The third test evaluates label imbalance. The main corpus is balanced, but real datasets may not be. A skewed test set is created with 70.0% negative labels in each domain. The frozen model has 80.4% accuracy but 12.7% reversal error. Depth-selective adapters reach 83.1% accuracy and 5.8% reversal error. Negative calibration error improves from 15.2% to 7.6%. The method performs well under imbalance, partly because the polarity contrast term is group-based rather than prevalence-based.

Table 10: Post-adaptation layer behavior analysis.

Metric	Frozen	Uniform	Depth-Selective
Signed Margin Product	1.18	2.07	2.84
Prompt Probe Accuracy (%)	71.8	67.1	63.2
Target Probe Accuracy (%)	76.4	79.2	81.5
Monotonicity Score	0.68	0.75	0.82

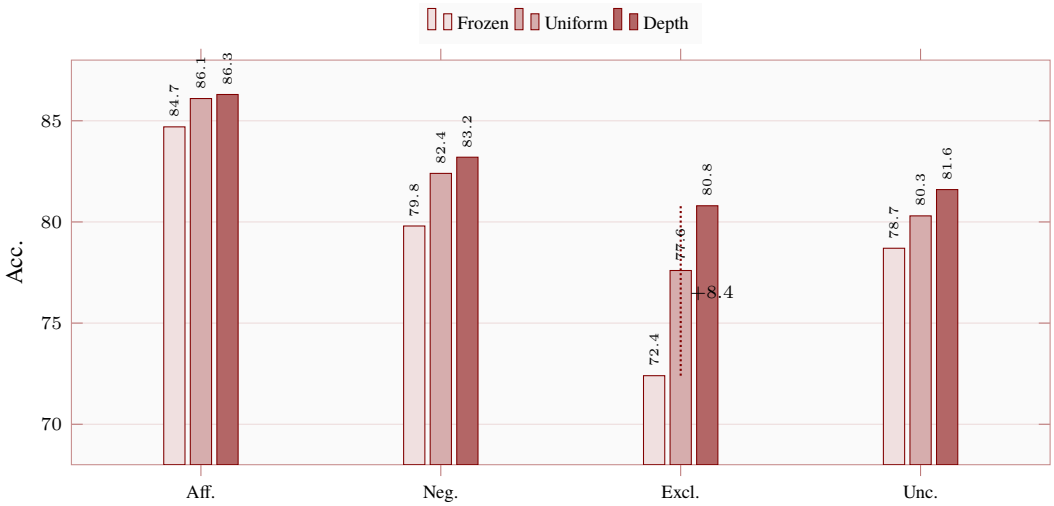


Figure 8: Prompt-form accuracy improves most for exclusion-framed questions, indicating that adapter placement matters when linguistic polarity must reverse the answer direction.

The fourth test evaluates reduced training data. With 25.0% of the training groups, depth-selective adapters reach 7.4% reversal error. With 50.0%, they reach 5.9%. With the full training set, they reach 4.9%. Uniform adapters are more affected by data reduction, reaching 10.1%, 8.6%, and 7.3% respectively. Depth-selective allocation appears more data-efficient because it trains fewer layers with greater diagnostic relevance. However, the diagnostic set itself must still be large enough to estimate layer scores reliably [21].

The fifth test examines diagnostic noise. Layer scores are estimated using 100, 200, 400, and 600 diagnostic groups. With 100 groups, selected layers vary noticeably across random samples, and reversal error is 6.5%. With 200 groups, reversal error is 5.7%. With 400 groups, it is 5.1%. With 600 groups, it is 4.9%. The diagnostic stage therefore benefits from several hundred groups, but it does not require the full training set. This matters because layer diagnostics can be computationally expensive.

The sixth test compares alternative diagnostic combinations. Using only Jacobian energy gives 6.4% reversal error. Using only representation separation gives 5.8%. Using only margin propagation gives 6.0%. Combining all three gives 4.9%. Pairwise combinations fall between 5.2% and 5.6%. Representation separation is the strongest single score, but it does not fully replace the others. The combined rule works because polarity processing involves token sensitivity, hidden-state organization, and final margin influence.

The seventh test changes the rank smoothness penalty. With no smoothness penalty, the allocation becomes scattered and reversal error is 5.4%. With the main penalty, reversal error is 4.9%. With double the penalty, the allocation becomes too contiguous and reversal error is 5.3%. A moderate penalty is best. The result suggests that nearby layers share useful behavior, but strict contiguity is not optimal.

The eighth test evaluates whether visual encoder adaptation would change conclusions. A small adapter is added to the visual projection layer while keeping the same total parameter budget by reducing language adapter rank. Accuracy improves slightly on affirmative prompts, but reversal error worsens to 5.6%. The likely reason is that polarity errors are primarily language-conditional in this task; spending capacity on visual adaptation leaves less capacity for polarity processing. This does not mean visual adaptation is unimportant for other tasks. It means that

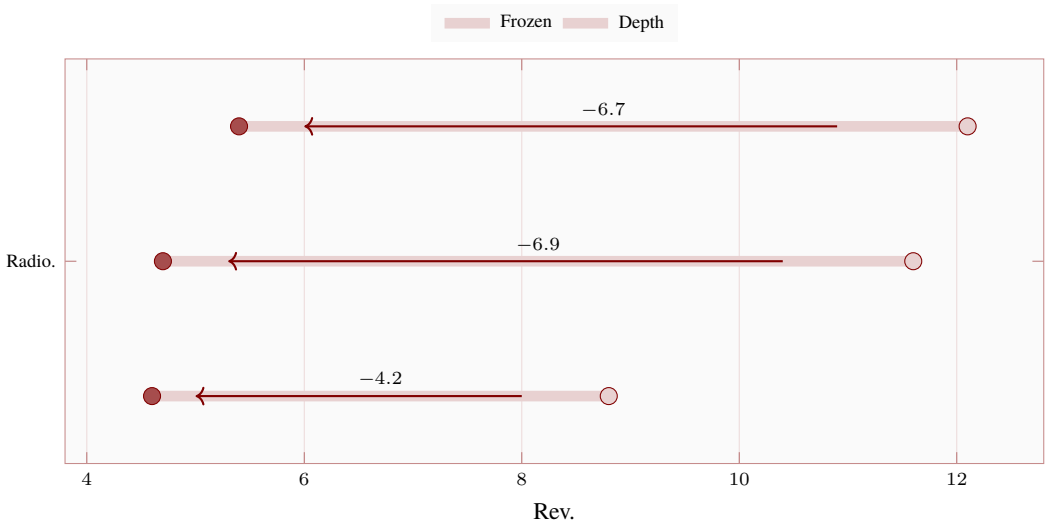


Figure 9: Domain-level reversal errors decrease across radiography, ultrasound, and dermoscopy, with ultrasound remaining the most difficult domain after adaptation.

under a fixed budget for this polarity problem, language-layer allocation is more useful.

The ninth test evaluates out-of-template clinical synonyms. A synonym challenge set replaces common target terms with less frequent clinical phrases. In radiography, airspace opacity replaces consolidation in some prompts. In ultrasound, peritoneal fluid replaces free fluid. In dermoscopy, atypical vessels replaces atypical vascular pattern. The frozen model has 14.1% reversal error, uniform adapters 10.2%, and depth-selective adapters 8.4%. The improvement is smaller than in the main test. This indicates that polarity robustness and synonym robustness are related but not identical. Additional lexical coverage would be needed for rare phrase variants.

The tenth test evaluates random seed stability. Across five training seeds, depth-selective reversal error ranges from 4.6% to 5.3%, with standard deviation 0.3 percentage points. Uniform adapters range from 7.0% to 7.7%. Final-block adapters range from 7.5% to 8.4%. Random allocation has much wider variation because the allocation itself changes. The proposed method is therefore stable once diagnostics are fixed.

The robustness results support the central finding while defining its limits. Depth-selective allocation remains stronger than baselines under rewritten prompts, instruction changes, imbalance, smaller training sets, and seed variation. It is less effective under rare synonym substitution and nested uncertainty. It depends on reasonably stable layer diagnostics. It works best when the main failure is polarity processing rather than missing visual recognition or unfamiliar terminology.

7. Error Analysis and Practical Implications

The remaining reversal errors after depth-selective adaptation are not randomly distributed. The largest cluster involves exclusion prompts with visually subtle positive findings. For example, if a small effusion is present and the prompt asks whether effusion can be excluded, the correct answer is no. The model sometimes answers yes, apparently treating uncertainty in the image as support for exclusion. This is a clinically meaningful confusion because image uncertainty and linguistic exclusion are different. The model needs to understand that inability to see a finding confidently does not always mean the finding can be ruled out.

A second cluster involves negative prompts with high-salience visual distractors. In dermoscopy, an image may contain an atypical feature related to the target but not exactly matching it. The model sometimes answers yes to absence prompts because the visual abnormality dominates the language polarity. The adapter reduces this error type but does not eliminate it [22]. This suggests that the language route can be improved while still being overridden by strong visual associations.

A third cluster involves uncertainty wording. Prompts asking whether a finding is unlikely can be difficult when the image is borderline. The correct answer key treats

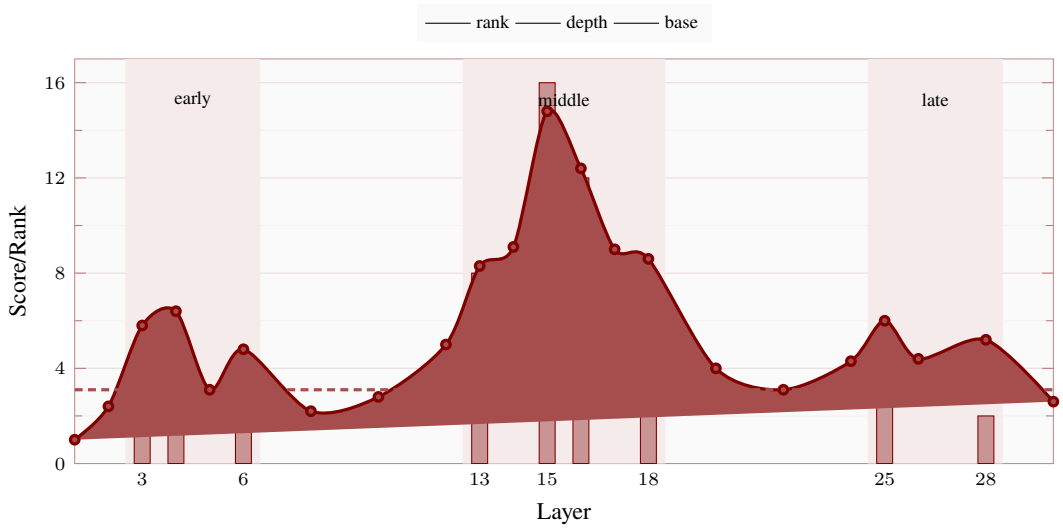


Figure 10: The learned allocation concentrates rank in early token-sensitive, middle representation-separating, and late margin-coupled layers instead of selecting only final blocks.

unlikely as a binary polarity form, but clinical interpretation may be less discrete. If a finding is subtle, both yes and no may seem defensible depending on threshold. The model's errors in this cluster may partly reflect label simplification rather than pure language failure. This is one reason the paper avoids making clinical safety claims from binary metrics alone.

A fourth cluster involves domain-specific phrasing. Ultrasound prompts contain anatomical and procedural terms that are less common in radiography-style VQA data. Even when the model recognizes the image pattern, it may handle phrases such as appendiceal noncompressibility or intrauterine gestational sac differently across polarity forms. Depth-selective allocation improves general polarity handling, but it cannot fully compensate for uneven domain terminology. Domain-specific prompt augmentation may be needed.

The practical implication is that adapter allocation can be treated as a design variable rather than an afterthought. When a team has a fixed parameter budget, placing the same rank everywhere is not always the best use of capacity. Measuring where the target linguistic operation affects the model can guide placement. This is particularly relevant for medical settings where the target operation is known in advance. If the goal is polarity robustness, diagnostics should focus on polarity. If the goal is temporal comparison, the diagnostic set should test before-after language. If the goal is anatomical grounding, layer scores should measure location-sensitive representations.

The method also suggests a validation workflow. Before adaptation, construct paired prompt groups for the linguistic operation of interest. Measure layer diagnostics on a sample. Allocate adapter rank under a fixed budget. Train with a loss that includes ordinary supervision and a task-specific contrast. Evaluate not only accuracy but also operation-specific errors. This workflow is more involved than ordinary adapter training, but it provides more information about why capacity is placed where it is. It also creates an audit trail for model modification.

The study's results caution against relying only on final-block adaptation. Final layers are attractive because they are close to the output and cheaper to adjust. They do help, but they leave some polarity errors unresolved. The language operation begins before the answer surface. If earlier layers encode the prompt in a way that weakens or distorts polarity, late adapters may only patch the final decision. A small amount of earlier capacity can make the later correction easier. The best allocation in this study therefore spans multiple regions.

Uniform adaptation has the opposite issue. It is robust and simple but may waste rank. In large budgets, this waste matters less because enough capacity reaches useful layers. In small budgets, it matters more. Medical models are often adapted under memory, compute, or governance constraints. A method that improves the use of a fixed budget can be valuable even if it does not change the underlying training data. The rank budget experiments show this clearly: the proposed method's advantage is largest when capacity is limited.

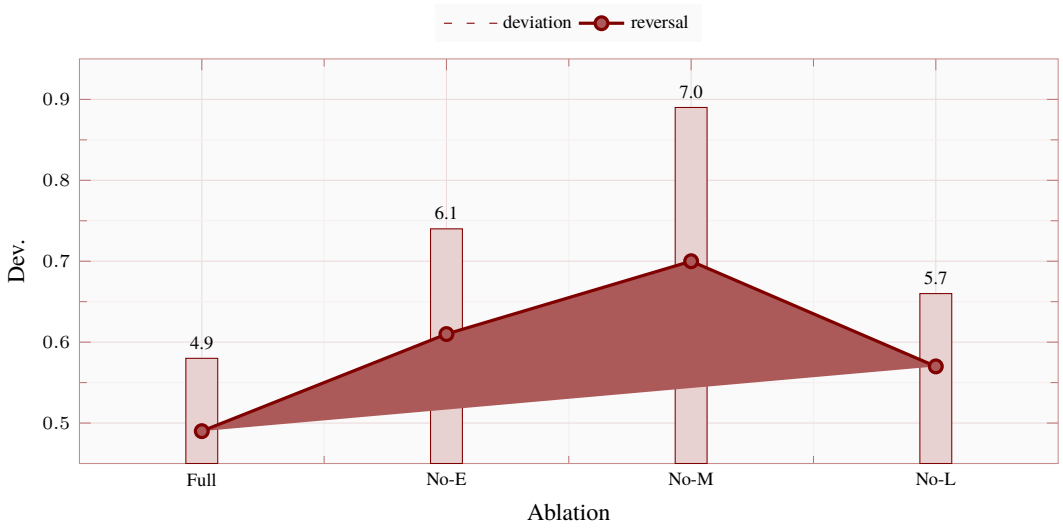


Figure 11: Ablation of selected layer regions shows that removing middle layers is most damaging, while early and late regions still provide measurable robustness gains.

The method also has interpretive value. The selected layers and their diagnostic scores reveal how the model handles a linguistic operation. The result is not a complete explanation, but it gives a structured view of where token sensitivity, representation separation, and margin coupling occur. In this study, polarity is distributed across early, middle, and late layers. That pattern argues against a simple story in which one layer is the polarity layer. It also explains why an allocation spanning several regions works better than a single contiguous final block.

For deployment, polarity reversal error is an intuitive metric. Clinicians and auditors can understand that a model should not answer the same way to a presence question and a rule-out question when the correct answers differ. Accuracy alone can hide this failure because one of the prompts may still be correct. Reporting reversal error alongside accuracy provides a clearer picture of linguistic reliability. Signed margin deviation adds another layer by identifying cases where answers are correct but confidence is highly uneven across polarity forms.

The method does not remove the need for human-designed prompt standards. A model can be made more robust, but clinical systems should still avoid needlessly confusing language. Prompt templates should minimize double negation and ambiguous exclusion phrasing when possible. Robustness training is a safeguard, not an invitation to use poor prompts. The best practical approach is likely both: clearer prompts and models that are less brittle when prompt variation occurs.

8. Limitations and Future Directions

The first limitation is the controlled nature of the prompt grammar. The study uses carefully constructed polarity forms to ensure correct answer keys [23]. Real clinical language is less tidy. Users may ask compound questions, include multiple findings, or mix uncertainty with comparison and negation. The results therefore show that depth-selective allocation helps under controlled polarity transformations, not that it handles every natural prompt. Future studies should use real user logs with expert annotation of polarity intent.

The second limitation is the binary answer space. Polarity errors are easiest to define when answers are yes or no. Many medical vision-language tasks require free-text explanation, localization, measurement, or differential diagnosis. A model might correctly answer yes but provide an explanation with inconsistent polarity. Extending the method to structured or free-text outputs would require new metrics. One possibility is to score multiple clinically relevant answer propositions and apply polarity contrast at the proposition level.

The third limitation is that the study freezes the visual encoder. This is appropriate for isolating language-layer allocation, but some errors involve visual ambiguity or domain-specific image features. Joint visual-language adaptation might improve those cases. However, visual adaptation also increases validation burden and may introduce new failure modes. Future work should examine

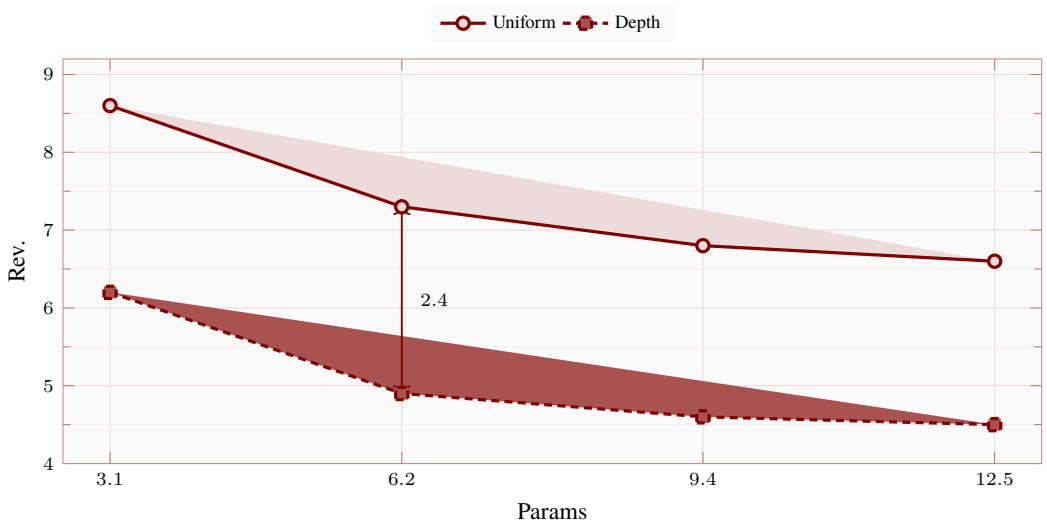


Figure 12: Rank-budget scaling shows that the allocation advantage is largest at the moderate budget, while both methods approach diminishing returns as capacity increases.

whether depth-selective language adapters combine well with small visual adapters when the task includes both polarity and visual recognition gaps.

The fourth limitation is that the diagnostic scores require additional computation. Jacobian energy, representation alignment, and perturbation-based propagation are more expensive than ordinary validation metrics. The cost is manageable for the present diagnostic set, but it may be substantial for very large models or many candidate tasks. Approximate diagnostics could reduce cost. For example, gradient norms could be estimated on smaller token subsets, and propagation scores could use linear probes rather than perturbation passes.

The fifth limitation is that the allocation rule uses fixed weights for the three diagnostics. These weights are selected on validation data and may not transfer to every model. A more principled approach could learn the weights through bilevel optimization, where validation reversal error directly supervises allocation. Such an approach might improve performance but would also be more complex and more expensive. The current rule is intentionally simpler and easier to inspect.

The sixth limitation is that rank choices are discrete and manually bounded [24]. The allowed ranks are practical, but they may not be optimal. Continuous relaxation of rank allocation could produce finer distributions, followed by rounding. Another direction is adaptive rank pruning after training. The effective-rank analysis suggests that some late-layer rank is underused. A train-prune-retrain cycle might reduce parameters without losing robustness.

The seventh limitation is that clinical labels may be imperfect. Binary labels simplify complex image findings. Some uncertainty prompts are especially sensitive to label thresholds. If the label says a finding is present but the visual evidence is minimal, the correct answer to an uncertainty question may not be straightforward. Human expert review of the hardest polarity cases would improve the validity of future evaluations. The present study treats labels as fixed because it focuses on model behavior under controlled prompt transformations.

The eighth limitation is that the method addresses one linguistic operation at a time. Polarity is important, but medical language also contains temporality, laterality, severity, quantification, and causal relations. A model adapted for polarity might not improve on these other operations. A future system could define multiple diagnostic sets and allocate adapter capacity for several operations jointly. The challenge would be avoiding conflict between operations competing for the same layers.

Future work should also evaluate whether depth-selective allocation is useful outside medicine [25]. Polarity reasoning appears in legal, financial, and safety-critical technical domains. The general principle of measuring layer sensitivity to a target linguistic operation and allocating adapter capacity accordingly may transfer. However, medical images add a multimodal constraint that may change the allocation pattern. Cross-domain studies would clarify which parts of the approach are general and which are specific to medical vision-language inference.

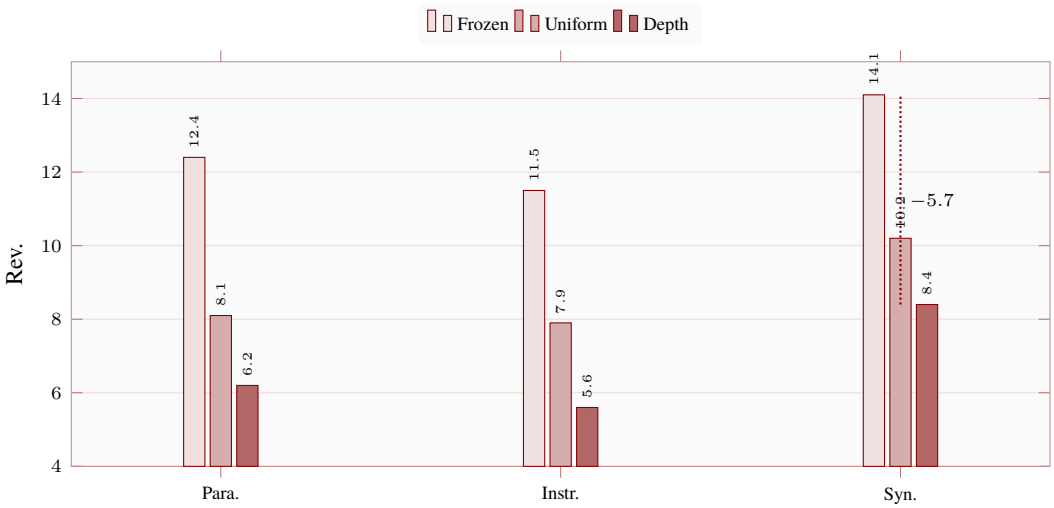


Figure 13: Robustness evaluation preserves the depth-selective advantage under paraphrasing, instruction changes, and rare synonym substitution, although synonym transfer remains weaker.

Another direction is active data selection. The diagnostic scores identify layers, but the training data are still sampled broadly. The model might benefit from emphasizing groups whose polarity transformations produce large hidden-state changes or margin conflicts. An active curriculum could select examples that most directly exercise the selected layers. This may reduce the amount of training data needed for robust adaptation.

Finally, future work should connect polarity robustness with abstention and explanation. A model that remains uncertain under difficult polarity transformations should perhaps abstain rather than force a yes or no. The signed margin measures developed here could support abstention rules. Explanations could also be checked for polarity agreement with the final answer. A model that answers no to a rule-out question but explains as if the finding is present would still be unsafe. Robustness needs to apply to both answer and rationale.

9. Conclusion

This study evaluated depth-selective adapter allocation for improving polarity robustness in medical vision-language inference. The problem was framed around binary clinical prompts whose wording changes the required answer direction [26]. Instead of increasing the adapter budget, the method redistributed a fixed low-rank budget across layers using diagnostic scores for token-local sensitivity, representation separation, and signed margin propagation.

The empirical results show that allocation matters. Under the same parameter budget and training loss, depth-selective adapters reduced polarity reversal error more than uniform adapters, final-block adapters, and random rank allocation. The method was especially helpful for exclusion-framed prompts, where ordinary accuracy was weakest and polarity mistakes were most common. It also reduced signed margin deviation and improved negative calibration error while preserving overall task accuracy.

The layer analyses indicate that polarity processing is distributed. Early layers respond strongly to polarity-bearing tokens, middle layers reorganize prompt representations, and later layers connect those representations to the answer margin. The best allocation spans these regions rather than concentrating all capacity in one place. This finding supports a practical view of adapter design: the placement of trainable capacity should reflect the linguistic operation being adapted.

The study has limits. The prompt grammar is controlled, the tasks are binary, and the labels simplify clinical uncertainty. The method requires diagnostic computation and hidden-state access. It should therefore be treated as a targeted adaptation strategy rather than a complete solution for medical language reliability. It improves a specific class of polarity failures but does not replace domain validation, careful prompt design, or expert review.

A useful next step is to apply the same allocation principle to other clinical language operations, including temporal comparison, laterality, and severity. Each operation may produce a different layer profile and require a different rank distribution. The broader lesson is that adapter capacity is not only a quantity but also a location. For medical vision-language systems, where small word-

ing changes can reverse meaning, that location can materially affect robustness.

References

[1] P. Gissen, N. Specchio, A. Olaye, M. Jain, T. Butt, W. Ghosh, B. Ruban-Fell, A. Griffiths, C. Camp, Z. Sisic, C. Schwering, E. Wibbeler, M. Trivisano, L. Lee, M. Nickel, A. Mortensen, and A. Schulz, "Investigating health-related quality of life in rare diseases: a case study in utility value determination for patients with cln2 disease (neuronal ceroid lipofuscinosis type 2).," *Orphanet journal of rare diseases*, vol. 16, pp. 217–217, 5 2021.

[2] B. Sadanandan and V. Behzadan, "Mechanistically guided lora improves paraphrase consistency in medical vision-language models," *arXiv preprint arXiv:2603.00148*, 2026.

[3] O. A. C. Sturdza, F. Filip, M. T. Baitan, and M. Dimian, "Deep learning network selection and optimized information fusion for enhanced covid-19 detection: A literature review.," *Diagnostics (Basel, Switzerland)*, vol. 15, pp. 1830–1830, 7 2025.

[4] L. Dai, J. Lei, F. Ma, Z. Sun, H. Du, H. Zhang, J. Jiang, J. Wei, D. Wang, G. Tan, X. Song, J. Zhu, Q. Zhao, S. Ai, A. Shang, Z. Li, Y. Zhang, and Y. Li, "Boosting deep learning for interpretable brain mri lesion detection through the integration of radiology report information.," *Radiology. Artificial intelligence*, vol. 6, pp. e230520–e230520, 11 2024.

[5] R. Patterson and S. M. Huff, "The decline and fall of esperanto: lessons for standards committees.," *Journal of the American Medical Informatics Association : JAMIA*, vol. 6, pp. 444–446, 11 1999.

[6] J. Harewood, M. Contreras, K. Huang, S. Johnson, and J. Wang, "Access to myopia care in the united states-a narrative review.," *Investigative ophthalmology & visual science*, vol. 66, pp. 5–5, 6 2025.

[7] I. Banerjee, Y. Ling, M. C. Chen, S. A. Hasan, C. P. Langlotz, N. Moradzadeh, B. E. Chapman, T. J. Amrhein, D. A. Mong, D. L. Rubin, O. Farri, and M. P. Lungren, "Comparative effectiveness of convolutional neural network (cnn) and recurrent neural network (rnn) architectures for radiology text report classification.," *Artificial intelligence in medicine*, vol. 97, pp. 79–88, 11 2018.

[8] R. Yan, H. Luo, J. Lu, D. Liu, H. Posluszny, M. P. Dhaliwal, J. MacLeod, Y. Qin, C. Yang, T. J. Hartman, and X. Hu, "Dietai24 as a framework for comprehensive nutrition estimation using multimodal large language models.," *Communications medicine*, vol. 5, pp. 458–458, 11 2025.

[9] K. Lazaros, S. Adam, M. G. Krokidis, T. Exarchos, P. Vlamos, and A. G. Vrahatis, "Non-invasive biomarkers in the era of big data and machine learning.," *Sensors (Basel, Switzerland)*, vol. 25, pp. 1396–1396, 2 2025.

[10] A. Nyamawe and D. Shao, "On the use of synthetic data for healthcare ai in africa: Technical performance, governance challenges, and policy readiness.," *Digital health*, vol. 12, pp. 20552076261418897–20552076261418897, 1 2026.

[11] N. AlDahoul, H. A. Karim, M. A. Momo, F. I. F. Escobar, V. A. Magallanes, and M. J. T. Tan, "Parasitic egg recognition using convolution and attention network.," *Scientific reports*, vol. 13, pp. 14475–14475, 9 2023.

[12] S. Jin, W. Bai, M. Yusufu, R. Li, K. Zhang, F. Zhang, L. Li, H. Zou, N. Wang, and H. Liu, "Digital technologies in enhancing hierarchical vision health management for the next 1000 days of children's life: multi-component economic evaluation.," *The Lancet regional health. Western Pacific*, vol. 63, pp. 101695–101695, 10 2025.

[13] B. Song, D. R. Kc, R. Y. Yang, S. Li, C. Zhang, and R. Liang, "Classification of mobile-based oral cancer images using the vision transformer and the swin transformer.," *Cancers*, vol. 16, pp. 987–987, 2 2024.

[14] E. Casmin and R. Oliveira, "Survey on context-aware radio frequency-based sensing.," *Sensors (Basel, Switzerland)*, vol. 25, pp. 602–602, 1 2025.

[15] J.-W. Tjiu and C.-F. Lu, "Equity and generalizability of artificial intelligence for skin-lesion diagnosis using clinical, dermoscopic, and smartphone images: A systematic review and meta-analysis.," *Medicina (Kaunas, Lithuania)*, vol. 61, pp. 2186–2186, 12 2025.

[16] C. N. Ionita, K. R. Hoffmann, D. R. Bednarek, R. Chityala, and S. Rudin, "Cone-beam micro-ct sys-

tem based on labview software," *Journal of digital imaging*, vol. 21, pp. 296–305, 2 2007.

[17] A. J. Goodell, S. N. Chu, D. Rouholiman, and L. F. Chu, "Large language model agents can use tools to perform clinical calculations.," *NPJ digital medicine*, vol. 8, pp. 163–163, 3 2025.

[18] L. Hess, A.-G. S. Palermo, and D. Muller, "Addressing and undoing racism and bias in the medical school learning and work environment.," *Academic medicine : journal of the Association of American Medical Colleges*, vol. 95, pp. S44–S50, 11 2020.

[19] S. He, Y. Wang, Z. Li, N. Jiang, and G. Sun, "The first birthday of openai's sora: A promising but cautious future in medicine.," *International journal of surgery (London, England)*, vol. 111, pp. 4151–4154, 5 2025.

[20] M. Jylhä, J. M. Guralnik, L. Ferrucci, J. Jokela, and E. Heikkinen, "Is self-rated health comparable across cultures and genders?," *The journals of gerontology. Series B, Psychological sciences and social sciences*, vol. 53, pp. S144–52, 5 1998.

[21] B. Twala, "Ai-driven precision diagnosis and treatment in parkinson's disease: a comprehensive review and experimental analysis.," *Frontiers in aging neuroscience*, vol. 17, pp. 1638340–1638340, 7 2025.

[22] E. S. J. Honorato, C. P. Fernandes, D. C. de Souza, S. M. Lemos, T. A. da Silva, Érica da Silva Carvalho, Ângela Xavier Monteiro, C. S. Cruz, L. V. da Costa, K. S. A. Martins, and R. C. P. de Oliveira, "The body as object of historical research: Bibliographic review," *International Journal for Innovation Education and Research*, vol. 7, pp. 218–242, 8 2019.

[23] J. A. Rich, "Viewing the future through the lens of the past: A personal reflection on disparities education in medicine and public health," *Journal of general internal medicine*, vol. 25, pp. 202–203, 3 2010.

[24] B. Scheenstra, B. C. Bongers, B. Broeders, M. Imkamp, L. V. Susante, B. Kietselaer, J. Maessen, A. V. 't Hof, and P. S. Nia, "Reasons and predictors of non-participation in a personalized digital prehabilitation care trial for patients undergoing elective cardiothoracic surgery.," *Interdisciplinary cardiovascular and thoracic surgery*, vol. 37, 7 2023.

[25] X. Li, X. Dong, L. Chen, N. Su, K. Huang, S. Li, Y. Wen, M. Zhang, B. Xu, S. Yuan, J. Wang, and Q. Chen, "Noninvasive coronary artery disease detection using retinal images: A multimodal study.," *JACC. Advances*, vol. 4, pp. 102341–102341, 11 2025.

[26] W. C. Coates, "Precision education - a call to action to transform medical education.," *International journal of emergency medicine*, vol. 18, pp. 21–21, 2 2025.