Bottled Brilliance: Gated Mixture of Experts for Biomedical Explainability

Sidharrth Nagappan
sn666@cam.ac.uk
University of Cambridge

Abstract

Explainability is critical for deploying deep-learning models for computational medicine, where black-box predictions can undermine trust. Concept Bottleneck Models (CBMs) offer a transparent, in-model pipeline by forcing networks to route predictions through human-readable concepts. In this work, we introduce a Mixture-of-Experts (MoE) extension to the "Language In a Bottle" (LaBO) pipeline, combining a generalist vision-language model (CLIP) with a specialist counterpart (BioMedCLIP) in a gated framework that adaptively weighs their respective contributions during inference. Paired with submodular concept selection, each expert's bottleneck remains lean yet semantically specialized. On two medical imaging datasets – HAM10000 for dermatoscopy and COVID-QU-Ex for radiography – MoE outperforms single-expert baselines by up to 20% on radiography and 4.6% for dermatoscopy under full supervision, although few-shot settings reveal instability in the gating mechanism due to sparse representations. This work confirms the hypothesis that fusing domain-specific and generalist knowledge via concept bottlenecks can yield both performant and granularly auditable deep learning pipelines for biomedicine. Source code is made available at https://github.com/sidharrth2002/biomedical-moe.[1]

Keywords Biomedical Explainability · CLIP · BioMedCLIP · Mixture of Experts

1 Introduction

The adoption of deep learning in highly sensitive domains like computational medicine has led to increased calls for robust explainability mechanisms, that medical practitioners can use to trace the reasoning behind specific decisions [1]. While post-hoc interpretability methods are rampant in the literature, the detachment from the internal workings of the model can result in incomplete explanations [2], and the "completeness" paradigm is a crucial part of building trust in these automated systems. Concept Bottleneck Models (CBM) organically incentivise models to route decisions through an interpretable concept layer, where each neuron in the bottleneck corresponds to a human-understandable concept [3]. However, annotating concepts can be costly, leading to Language in a Bottle (LaBO) introducing an end-to-end pipeline that leverages Large Language Models (LLMs) to build and enrich concept bottlenecks, before using Vision Language Models (VLMs) such as CLIP to align images and textual concepts [4, 5].

[1] Word-Count: 3813, computed using Typst's wordometer

Although effective against a range of datasets, performance on the one biomedical dataset they used is among the lowest, having been outperformed by a simple linear probe.

This raises the question of whether domain knowledge can be implicitly plugged into these models, and whether it can enhance their ability to form robust representations of nuanced datasets—particularly those that rely on subtle morphological cues beyond standard visual descriptors. Following this line of reasoning, we explore whether the need for domain expertise can be addressed through a mixture of these experts, with each building their own bottlenecks and harmonising representations. Specifically, we question whether combining generalist and specialist models can yield concept bottlenecks that are both performant in end-to-end classification and capable of offering fine-grained, semantically grounded interpretability.

A learned gating module adaptively combines individual experts, namely the *generalist* CLIP and the *specialist* BioMedCLIP. Two representative biomedical datasets for dermatoscopy and radiology are chosen to investigate two fundamental research questions:

1. **RQ1**: Does domain-specific expertise improve both the interpretability and classification performance across both biomedical domains?
2. **RQ2**: Can this mixture-of-experts outperform single-expert baselines, in both fully supervised and few-shot scenarios?

The findings of this work suggest that MoE-based combinations can produce remarkable boosts in performance during full supervision, while building nuanced, independent bottlenecks that select expert-specific features; this is especially profound in the radiology dataset ($\approx 20\% \uparrow$ improvement from the generalist baseline). This work also makes methodological improvements to the LaBO pipeline via more structured prompt engineering and straightforward concept extraction.

While evaluating such architectures in few-shot settings introduces complexity, especially given the limited data available to train a robust gating mechanism, it remains a valuable diagnostic tool to see if a model can leverage prior relationships stored in its bottlenecks to act under constrained supervision. In this setting, the specialist expert consistently outperforms its counterparts, with the MoE occasionally learning unrepresentative gating policies. Though regularisation techniques encourage more balanced expert usage in few-shot settings, further work is needed to enable stable few-shot deployment, such as by transferring gating priors from adjacent biomedical tasks in ways that do not induce leakage.

2 Related Work

Concept Bottleneck Models (CBMs) improve interpretability by incentivising models to predict human-understandable concepts as an intermediate step before the final prediction [3]. In medical imaging tasks like diagnosing arthritis from an X-Ray, a CBM would first predict clinical concepts (e.g. presence of spurs) and then use those concepts to compute severity. Medical practitioners can inspect and intervene on the model's concept predictions [3]. However, traditional CBMs require training labels for each concept and often lag in accuracy compared to their black-box counterparts [6]. "Label-free" CBMs transform any network into a CBM without per-concept annotations using rudimentary LLMs for concept discovery [6–8]. Language In a Bottle (LaBO) extended this paradigm with submodular optimisation [9] to filter relevant and discriminative concepts in the same way a human expert would [4]. Crucially, LaBO leverages the Vision Language Model (VLM) CLIP [5], which is a foundation model that constrastively

learns joint image-text representations with little to no task-specific data. Orthogonally, these VLMs have been used in specialised domains such as biomedicine through expert variants like MedCLIP [5], PubMedCLIP [10] and BioMedCLIP [11] (with the latter achieving state-of-the-art results across a range of biomedical benchmarks).

On the paradigm of having "expert" models, the Mixture-of-Experts architecture is a long-standing proposition in deep learning, that dynamically combines the strengths of multiple specialised models using a divide-and-conquer approach [8, 12]. Recent work has applied MoEs to fuse generalist and specialist knowledge, which is particularly relevant in biomedical imaging where a model, much like a doctor, would require both broad and fine-grained expertise. Med-MoE introduced a mixture-of-experts design for medical VL tasks using multiple domain-specific experts alongside a global meta-expert, replicating how different medical specialties unite to form robust diagnoses; it attained state-of-the-art performance by activating only a few relevant experts instead of the entire model [13]. Furthermore, because gating decisions reveal which experts were consulted and how much importance was given to their analysis, a clinician can trace deeper intuitions. An Interpretable MoE (IME) uses linear models as experts, with each prediction being accompanied by an exact explanation of which linear expert was used and how it arrived at the outcome [14]. Impressively, this IME approach maintains accuracy comparable to black-box networks, showing that MoE architectures can incorporate interpretability without sacrificing predictive capacity. A tangentially relevant direction uses a hybrid neuro-symbolic design, routing samples down a tree of interpretable experts to explain a black box [6].

While most prior efforts apply mixture-of-experts to fully supervised, end-to-end deep networks, we explore its extension to concept bottleneck models (using LaBO as an architectural baseline) — specifically probing whether we can align class-concept association matrices rather than purely combining neural embeddings, and whether this remains effective under few-shot constraints.

3 Method

3.1 Biomedical Data

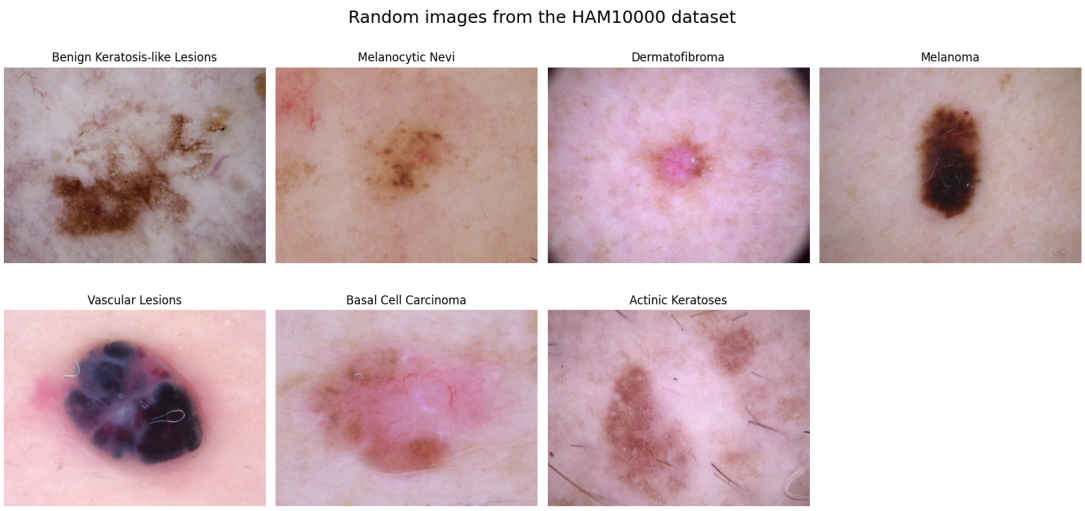


Figure 1: Sample Images from the HAM10000 dataset

We select (i) HAM10000 (dermatoscopy) and (ii) COVID-QU-Ex (X-Rays) as two representative datasets in the biomedical domain.

HAM10000 is a collection of 10,015 dermatoscopic images representing seven variations of skin lesions (Keratosis-Like Lesions, Melanocytic Nevi, Dermatofibroma, Melanoma, Vascular Lesions, Basal Cell Carcinoma and Actinic Keratoses), that are compiled from various populations [15], and is commonly used as a benchmark dataset for medical vision encoders. We use the same training, validation and testing splits as the dataset providers.

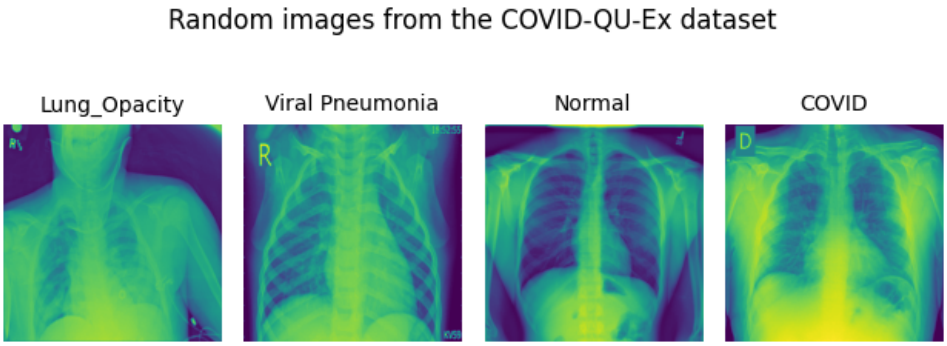


Figure 2: Sample Images from the COVID-QU-Ex dataset

The COVID-19 Radiography Database comprises 33,920 posterior–anterior chest X-ray images, covering COVID-19, viral/bacterial pneumonia, and normal cases [16, 17]. It integrates multiple datasets, including COVID-19 cases from Qatar, Italy, Spain, and Bangladesh, alongside pre-pandemic pneumonia datasets from the USA. Training, validation and test splits are not provided, so a 70%-15%-15% label-stratified split is used. While we initially considered the NIH ChestX-ray14 dataset [18], its multi-label nature required sigmoid-activated dot products on the association matrices — leading to gradient explosion during training, making it incompatible with our current architectural setup.

3.2 A Representative Bottleneck

LaBO employed sentence parsing using a T5 to extract semantic concepts from LLM-generated sentences [19]. We conjecture that this approach is suboptimal, because it leads to information loss and the quality of the final model is dependent on the accuracy of the trained parsing model. Instead, we propose enforcing JSON structure via Pydantic in prompts we send to our LLM suite (LLAMA, DeepSeek, Meditron and OpenAI's 4o), directly extracting phrasal concepts without intermediate parsing [20, 21].

Concepts generated for the generalist are augmented with the phrase "You can be a bit technical."[2] Our enhanced prompt engineering outperforms the manual parsing algorithm.

Technical Prompt Generation

Prompt: Describe the {*feature*} of the {*disease*} disease that can be used for Skin Cancer classification. *You can be slightly technical.*

[2]After several rounds of prompt engineering, this produced the best results.

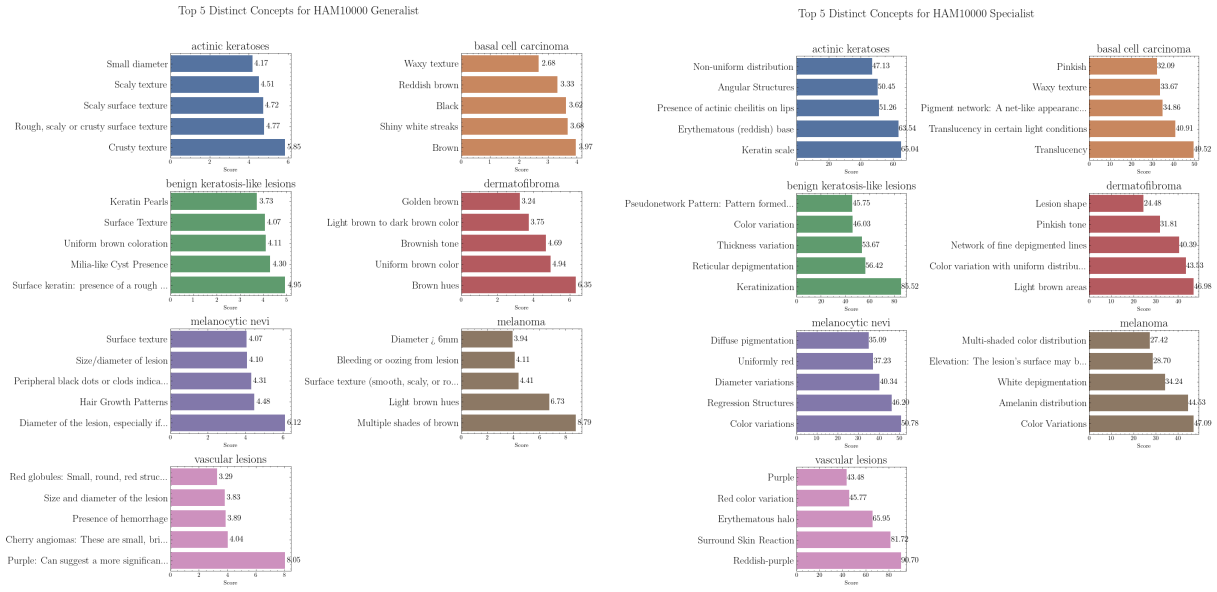
3.3 Multi-Expert Submodular Optimisation

Submodular optimisation is used to select a discriminative subset of concepts that maximise both class-specific coverage while minimising redundancy. Candidate concepts $c \in \mathbb{C}$ are projected onto the same latent space as the image representations, using each expert's encoder. For a given class $y \in Y$, similarity scores between concept and image features are computed using a dot product:

$$\text{Sim}(c, y) = \frac{1}{|X_y|} \sum x_i \in X_y z_c^T x_i \tag{1}$$

This is then used to estimate mutual information (MI between each concept and the underlying class distribution, encoding how well a concept is discriminative across class boundaries. The goal is to find a subset $S \subseteq \mathbb{C}$ that maximises the set function $f(S) = \alpha \cdot \text{coverage}(S) - \beta \cdot \text{redundancy}(S)$ ($\alpha + \beta$ are hyperparameters) using greedy selection. As an improvement to the original paper's algorithm:

1. We incorporate both CLIP + BioMedCLIP embeddings into separate selection processes. This also implicitly makes sure that only textual concepts that are semantically understood by the VLM are part of the final selection. We modify this architecture for the mixture-of-experts scenario, doing expert-wise bottleneck maintenance.
2. Concepts are stemmed, filtered for stopwords and pruned to remove redundant descriptions and any that contain morphological variants of class names — done to reduce semantic leakage and prevent the model from trivially associating concepts with their target classes.



(a) (b)

Figure 3: Concept Bottleneck Scores for the (a) *Generalist* and (b) *Specialist* in HAM10000 - Note the use of increasingly technical concepts in the specialist bottleneck, although several general, visual descriptors are represented in both bottlenecks.

This mechanism proves particularly valuable for advanced biomedical terminology — such as *telangiectasia*, *ovoid* or *keratinization* — which are well-represented in BioMedCLIP's domain corpus but may not be meaningfully encoded by CLIP. By filtering concepts through this embedding-informed scoring process that ranks concept embeddings based on a mutual information score, we obtain a lean and discriminative concept set that adapts to each expert

model. As seen in the MI-ranked concept list in Figure 3, the generalist leans towards visual descriptors like "waxy texture" and "brownish tone", while the specialist uses very specific terminology like "reticular depigmentation" or "erythematous halo", whose visual context is implicitly encoded due to the specialised training corpus, that already maps the relationship between certain visual patterns and a known explanation of the phenomena. In many ways, this method is reminiscent of a dermatologist's intuition, that first looks at visual patterns and then begins connecting to prior knowledge. It is unsurprising that common descriptors such as "presence", "brown" are selected by both experts, meaning neither is overly generic or overly specialised[3].

During MoE, we freeze the concept selection bottlenecks, and use those selected during their corresponding uni-expert training cycles. This allows for fair comparisons and ensures that the data distribution is the only independent factor.

3.4 Mixture-of-Experts

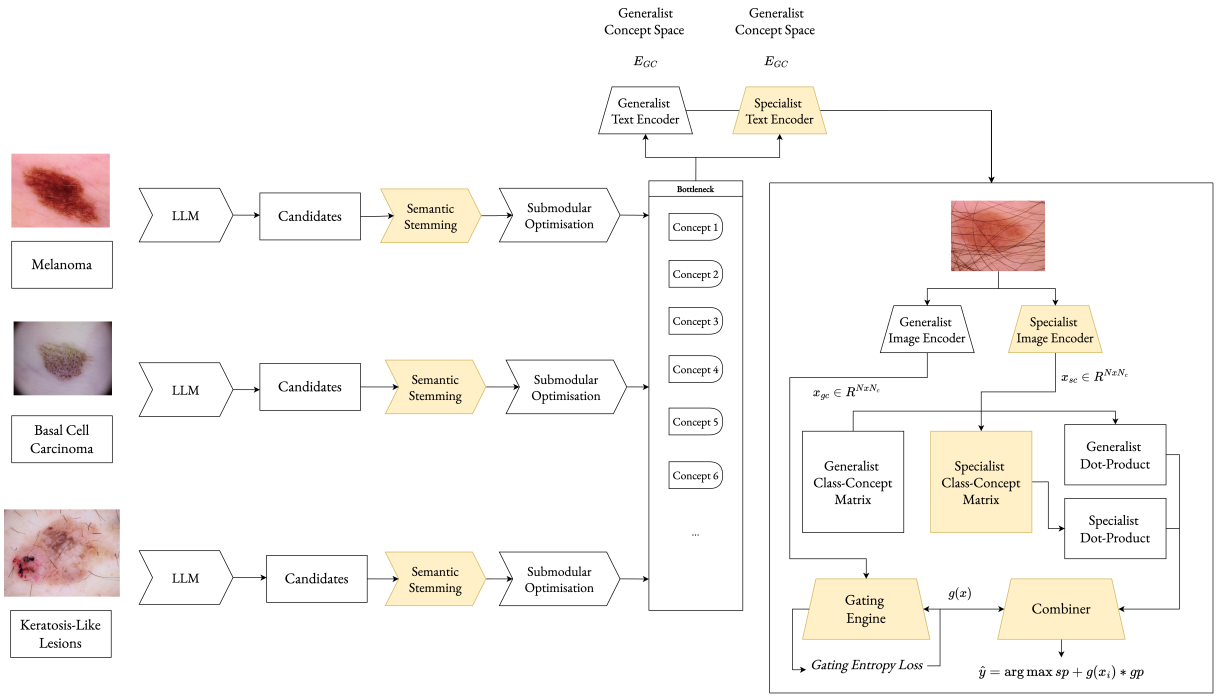


Figure 4: Bottled Brilliance Neural Architecture. Additions to LaBO are highlighted in yellow.

3.4.1 The Case for Expertise

As seen in the selection process, some lesions are distinguishable by general visual attributes like colour patterns and asymmetry, which CLIP catures well [5]. However, we hypothesise that there may be useful biomedical descriptors, that help when attempting to interpret decisions. For instance, a "multilobular pattern" has a distinct morphological shape, that when presented to an ordinary person, would appear obscure, but could highlight a clear analogy to a medical professional who would assess that analogy to build a deeper understanding of the diagnosis. The elucidated scenario comprises of a specialist audience, and a "common man" generalist. When representing this neurally, we find that it is reminiscent of the Mixture-of-Experts architecture [12].

[3]distributions computed using the nltk toolkits and the CLIP similarity scores

3.4.2 The Experts

CLIP is the choice architecture in LaBO; it learns a transferable visual representation by contrastively training image and text encoders on 400 million (image, text) pairs [5]. Its wide training corpus and general understanding of worldly knowledge makes it a suitable candidate for the generalist. BioMedCLIP is a multimodal biomedical foundation model trained on PubMed Central (PMC), a large-scale dataset comprising 15 million biomedical image-text pairs extracted from scientific publications [11]. The corpus taxonomy includes dermatoscopy images, microscopy, histopathology and radiography.

To ensure a fair comparison, we standardise the architecture by using ViT-B/16 for both experts, instead of the ViT-L/14 used in LaBO. While ViT-L/14 outperforms the base variant, large-scale BioMedCLIP models are not publicly available; however, neural scaling laws suggest that performance would improve proportionally by increasing transformer complexity [22].

3.4.3 Formulation

Our Gated Mixture-of-Experts approach combines similarity embeddings from both CLIP (E_C) and BioMedCLIP (E_B). Our approach uses precomputed image-to-concept dot products from each expert, and learns a concept-to-class association matrix for both. Formally, given an input image vector x_i, we pre-compute the generalist and specialist dot products based on their image and concept vectors:

$$\left\{D^g \in \mathbb{R}^{B\times m_g}, D^s \in \mathbb{R}^{B\times m_s}\right\} \tag{2}$$

where m_g and m_s denote the number of generalist and specialist concepts respectively. $A^g \in \mathbb{R}^{K\times m_g}$ and $A^s \in \mathbb{R}^{K\times m_s}$ are learnable association matrices that map concepts to class logits. To encourage semantically meaningful class–concept associations, the individual association matrices are initialised using language model priors by selecting the closest concepts to each class name in the CLIP embedding space.

$$\begin{pmatrix} 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 \end{pmatrix} \tag{3}$$

Weight initialisation for association matrix, where rows = classes, and columns = concepts.

Class-level predictions from each expert are computed as $S^g = D^g \times (A^g)^T$ and $S^s = D^s \times (A^s)^T$.

The gating network, tuned to inhibit over-parametrisation, takes in the generalist's 512-dimension latent image embedding. The generalist's image embedding is chosen because of it's broad coverage, and to inhibit any specialist bias at the outset. This is then passed through a two-layer neural network with layer normalisation, a *LeakyReLU* activation and a sigmoid output, defined as:

$$g(x_i) = \sigma(W_2(\text{LeakyReLU}(W_1(\text{LayerNorm}(x_i))))) \tag{4}$$

$g(x_i) \in [0, 1]$ dynamically determines the cross-expert weighting for each input and produces a weighted combination:

$$S_i = g(x_i) \cdot S_i^s + (1 - g(x_i)) \cdot S_i^g \tag{5}$$

The Gated MoE model's primary classification loss employs cross-entropy. The original paper also proposed two regularisation terms: a diversity loss to prevent uniform similarity scores across classes (encouraging sharper decision boundaries).

$$\mathcal{L}_{\text{classification}} = \text{CrossEntropy}(S, y) + \lambda_{\{\text{div}\}} \cdot \left(-E_{\{i\}}\left[\text{Var}_{\{k\}}\left(S_{\{i,k\}}\right)\right]\right) \tag{6}$$

In addition to these, we propose a gate-entropy loss that discourages the gating network from converging to overly deterministic decisions and completely depending on one of the experts (by collapsing $g(x_i) \to \{0, 1\}$). An ablation to demonstrate the utility of this addition is done.

$$\mathcal{L}_{\text{gate\_entropy}(g_i)} = -(g_i \cdot \log(g_i + \varepsilon) + (1 - g_i) \cdot \log(1 - g_i + \varepsilon)) \tag{7}$$

$$\mathcal{L} = \mathcal{L}_{\text{classification}} + \lambda_{\text{ge}} * \mathcal{L}_{\text{gate\_entropy}} \tag{8}$$

3.5 Experimental Infrastructure

All experiments were run on a single NVIDIA L40S GPU in the Department of Computer Science's GPU server, while Weights and Biases is used for experimental tracking [23]. We trained all few-shot models for a maximum of 5000 epochs, and run tests based on the best validation performance. To tackle the cold start issue for noisy gate parameters under few-shot scenarios, a 500 epoch warm start is allowed, where the only trainable parameters are the gate. We reuse the submodular optimisation hyperparameters α and β from the original paper across the few-shot runs.

4 Results

4.1 Loss Ablation

Primary hyperparameter tuning is done on HAM10000, with the best configurations immediately ported over to COVID-QU-Ex.

Variation	**Validation Accuracy**					
	1-shot	2-shot	4-shot	8-shot	16-shot	All
MoE	31.4	46.4	24.8	**43.9**	46.4	**79.2**
$\text{MoE}_{\text{entropy}}$ ($\lambda_{\text{ge}} = 0.2$)	**48.2**	**49.4**	24.8	34.6	**53.9**	78.6

Table 1: Shot-by-Shot Results

The use of our gating entropy loss provides performance boosts in three of five shots (with an average improvement of 6.19%), representative of its utility in stabilising gate estimates, and discouraging the gating network from collapsing too early to a single expert. The weighting λ_{entropy} for the entropy loss component is set at 0.2, to avoid saturating the loss computation.

4.2 Fully Supervised Baselines

We first evaluate the models in a fully supervised setting. The original paper uses `ViT-L/14` in their architecture. However, for the sake of fair comparisons and reasons mentioned earlier, we use `ViT-B/16` across all expert encoders.

Expertise	**Model Variant**	**Val. Acc. (%)**	**Val. Loss**	**Test Acc. (%)**	**Test Loss**
$E_{\{G\}}$	ViT-B/16	79.0	0.5501	76.8	0.6126
$E_{\{S\}}$	BioMedCLIP	77.3	0.69656	75.03	0.7916
MoE	$g(x_i)c \cdot E_S + (1 - g(x_i))c \cdot E_G$	**79.2**	4.442	**78.61**	0.714

Table 2: Individual Model Performance Results - Note that $E_{\{G\}}$ is the generalist and $E_{\{C\}}$ is the specialist

Under fully supervised conditions, standard CLIP still outperforms BioMedCLIP, likely because ample supervision allows for a sufficiently comprehensive representation—lessening the advantage of specialized domain expertise. The best performance is attained by the Mixture of Experts (MoE), outperforming both uni-expert counterparts by $\approx 4.66\%$. However, the elevated loss suggests that, while the model achieves strong accuracy, its occasional errors are highly confident misclassifications — potentially exacerbated by the gating mechanism's sharp routing decisions. When analysing the average gate distribution, we find that the gate's $g(x_i)$ begins to fluctuate before reaching a batch-wise average of 0.8 by the final training epoch. This shows increased dependence on the specialist in the majority of cases.

4.3 Skin Lesion Classification

Architecture	**Validation Accuracy (%)**					**Test Accuracy (%)**				
	1-shot	2-shot	4-shot	8-shot	16-shot	1-shot	2-shot	4-shot	8-shot	16-shot
$E_{\{G\}}$	23.0	35.9	24.4	34.5	53.0	22.4	38.1	23.4	31.9	52.6
$E_{\{S\}}$	25.0	38.8	**49.4**	**63.0**	52.8	27.2	40.0	**48.8**	**61.7**	52.3
MoE	31.4	46.4	24.8	43.9	46.4	28.1	48.1	27.3	42.9	45.1
$\text{MoE}_{\text{entropy}}$	**48.2**	**49.4**	24.8	34.6	**54.6**	**45.8**	**50.2**	23.5	34.2	**52.8**

Table 3: Shot-by-Shot Results (in %)

The specialist expert outperforms its counterparts by a substantial margin in ultra-low-shot settings (1-2 shots), showing that domain-specific knowledge provides strong performance boosts when there is minimal labelled data. However, under high-supervision and full-supervision, the specialist advantage diminishes, with the generalist outperforming it as soon as there is enough data to learn broad visual features and enrich the association matrix.

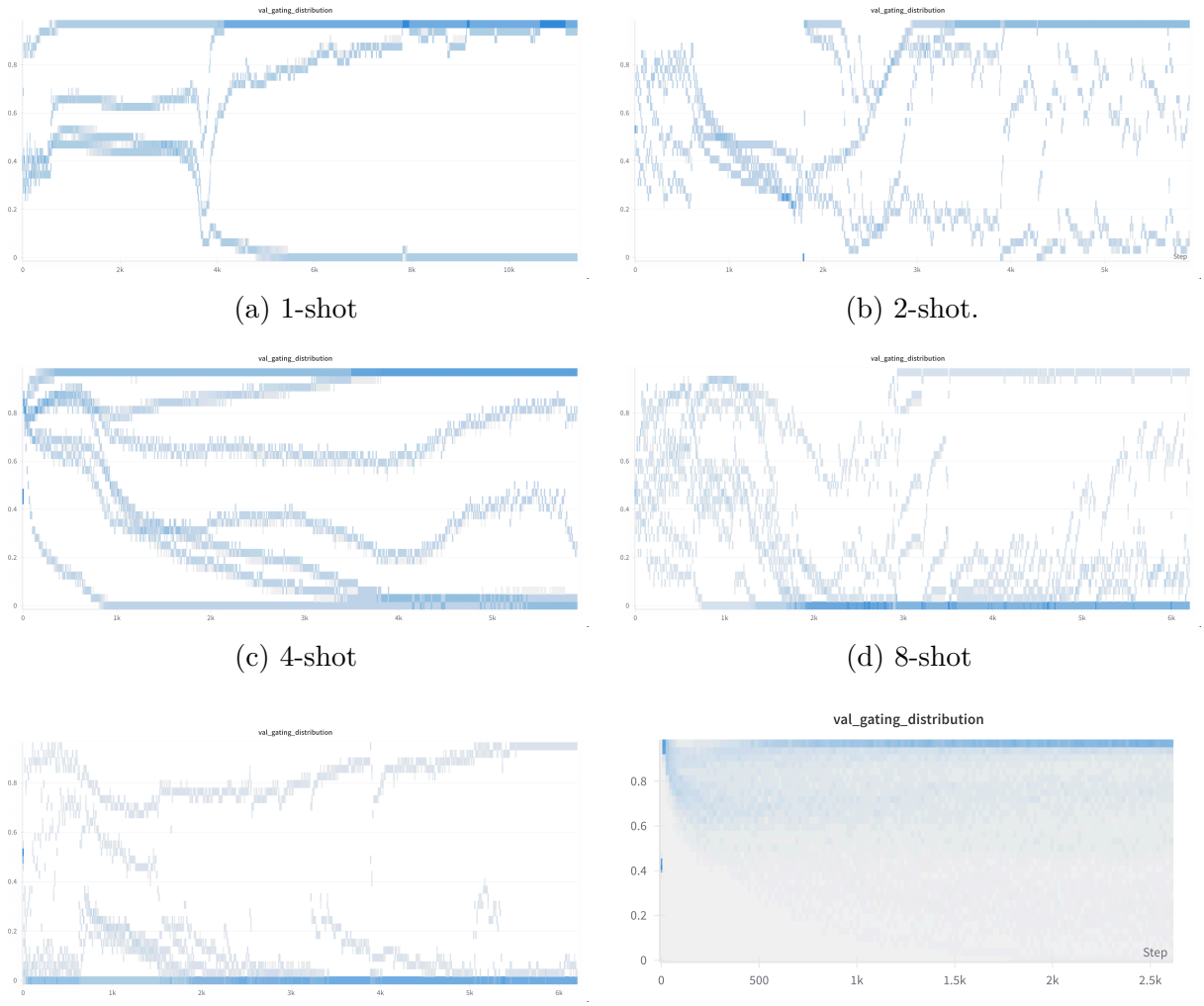
Interestingly, the Mixture-of-Experts excels at ultra-low-shot scenarios, where MoE partial insights are fused from both experts to provide surprisingly impressive performance ($> 70\%$) improvement. It must however be noted that the weights learned by gate may be suboptimal when there isn't sufficient supervision to enrich weighting decisions; in early epochs, the average gate weight $g(x_i)$ wavers between $0.4 \leftrightarrow 0.6$. Under mid-range supervision, the results do not consistently favour MoE, since partial supervision is likely insufficient for the gate to converge on an optimal blend, adversely hurting performance instead.

A general observation is that the mixture-of-experts lacks sufficient supervision in few-shot settings to train a sufficiently rich gate, with the full benefits of combining experts only visible during full supervision. There is also near-random and unexplainable fluctuations in 4- and 8-shots, possibly as a result of this phenomena.

4.3.1 Intuition about Foundation Gates

One solution to this problem would be to train the gate on a tangentially similar task and port the weights, so the starting representation would have some intuition about which types of images would be suitable for which expert. However, this would require clear cross-task alignment, and there is little formal proof to support this conjecture.

4.4 Gating Fluctuations



(a) 1-shot (b) 2-shot.

(c) 4-shot (d) 8-shot

(e) 16-shot (f) All-shot

Figure 5: Gating Distribution Across Different Shots

The gate provides fascinating insights into the inner workings of the model. Across 1- to 2-shot settings, there is a dramatic fluctuation in $g(x)$ as it lacks sufficient training examples to learn a stable preference, bouncing back between near-0 and near-1. In the mid-shot graphs (4-shot and 8-shot), the gate still oscillates in phases but displays a modestly smoother pattern – though there are steep dips towards 0 or 1 at certain epochs; likely indicative of preference collapses. By 16-shot, the trend stabilises, with a surprising number of samples favouring the generalist over the specialist. Under fully supervised training, the network has enough labels to make more confident routing decisions, with the histogram staying relatively consistent across epochs. Unsurprisingly, the specialist is favoured when all training samples are considered. Overall, it seems clear that sufficient supervision is necessary for the gate to build a stable representation.

4.5 Radiography

Arch.	Validation Accuracy						Test Accuracy					
	1-shot	2-shot	4-shot	8-shot	16-shot	All	1-shot	2-shot	4-shot	8-shot	16-shot	All
$E_{\{G\}}$	44.4	48.09	46.94	59.20	64.93	87.36	43.96	47.86	45.81	58.94	63.57	86.6
$E_{\{S\}}$	39.0	40.48	**69.65**	**68.79**	67.7	89.93	38.53	41.06	**69.41**	**68.04**	69.22	89.89
MoE	**45.62**	**65.87**	51.76	61.88	**78.62**	**93.51**	**47.20**	**63.69**	50.06	61.90	**77.61**	**93.67**

Table 4: Shot-by-Shot Results (in %) for COVID-QU-Ex

In extreme low-shot cases (1-2 shots), MoE significantly outperforms both the generalist and the specialist, suggesting that fusing domain-specific cues help overcome the data scarcity. Furthermore, we observe in the gating histogram that there is an even balance between the generalist and specialist weights ($0.2 < g(x) < 0.6$), suggesting that both experts are equally consulted in low-shot scenarios. However, at 4-8 shots, the specialist begins to outperform its counterparts, eclipsing gains from the gated combination of both experts. By 16-shots, there is sufficient supervision for the gating network to converge on a better routing algorithm for each individual sample, with MoE taking the lead and outperforming the specialist by $\approx 11\%$.

Under full supervision, MoE attains the highest accuracy of all, with an outstanding accuracy of 93.67%, consistent with the hypothesis that enough labelled data can produce remarkable synergies. The gating distribution histogram in Figure 6 shows a sparse distribution of gate values, with a significant number prioritising the specialist towards the end of training, reflecting the value of domain-specific representations for X-Ray data.

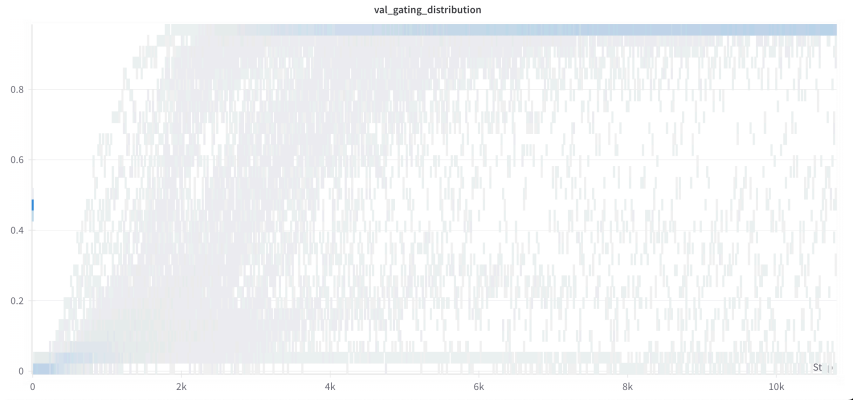


Figure 6: Fully Supervised Gating Distribution for COVID-QU-Ex

The specialist is also consistently strong across all shots, with incremental improvements as more data is appended – unlike in dermatoscopy where the generalist eventually catches up under full supervision. BioMedCLIP was trained with an extensive volume of X-Ray data (10^6 samples), an order of magnitude higher than dermatoscopic samples. Crucially, X-Ray classification often hinges on subtle clinical markers (e.g. faint opacities, lesions) which are difficult to identify on a purely visual or textural level in bitonal images, likely requiring domain-specific context and known prior connections in the scientific literature to help the model make accurate diagnoses. Meanwhile, the maximum gate value (≈ 0.71) in Figure 7 shows that the gating network never collapses to rely exclusively on the specialist—indicating that, even under full supervision, the model still draws on generalist cues when beneficial.

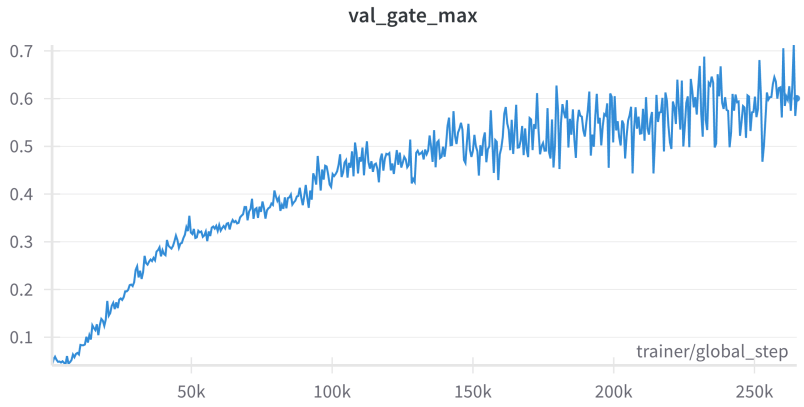


Figure 7: The maximum $g(x)$ value as training progresses for

5 Conclusion and Limitations

In this work, we introduce a Mixture-of-Experts extension to concept bottleneck models, designed to fuse domain-specific knowledge with a more general visual understanding. The architecture learns a gate between experts to make weighting decisions, and is particularly effective under full supervision, where there is sufficient data to allow the gate to converge on a robust representation. The two RQs posed are validated across HAM10000 (dermatoscopy) and COVID-QU-Ex (chest X-rays):

RQ1: Does domain-specific expertise improve both interpretability and classification performance across both biomedical domains? - The specialist excels in data-scarce scenarios by leveraging domain-focused representations. This is especially profound in X-Ray classification where nuanced morphological cues are critical, being preferred over simple visual predictors that are often insufficiently rich in bitonal images [24]. Under sufficient supervision, the Mixture-of-Experts boosts both granular interpretability via auditable, specialised bottlenecks (as seen in the diverse concept selection made by each expert) and high end-to-end performance.

RQ2: Can this mixture-of-experts outperform single-expert baselines, in both fully supervised and few-shot scenarios? - Under full supervision, the MoE consistently outperformed single-expert baselines ($\uparrow 20\%$ in radiology and $\uparrow 4\%$) in dermatology. In ultra-low-shot conditions, the MoE occasionally outperforms either expert alone by blending granular cues, but it is clear that the gate is often unstable in few-shot scenarios due to limited supervision. Despite the use of regularisation techniques and initialisation from vision-language model (VLM) priors, effectively learning to balance expert representations remains challenging when supervision is scarce.

Future work can pre-train the gating network on related biomedical tasks (X-Ray Gate $\leftrightarrow$ Dermatoscopy Gate), initialising it in our MoE framework with a general intuition for expert selection. This could include recognising subtle visual characteristics of an image via self-supervised reconstruction, and using the trained projection to make weighting decisions. Furthermore, due to computational constraints, the individual experts have their encoders frozen in the current architecture. An intriguing next step would be to allow partial fine-tuning or cross-expert transfer learning through residual connections [25], potentially achieving the three-fold objective of (i) further aligning the experts' representations, (ii) preserving their

distinct domain-specific strengths, and (iii) building a rich weighting algorithm. Lastly, this architecture can be tested on related biomedical domains across a wider taxonomy to validate gains, and evaluate its unique advantages beyond granular interpretability.

Bibliography

[1] H. Xu and K. M. J. Shuttleworth, "Medical artificial intelligence and the black box problem: a view based on the ethical principle of "do no harm," *Intelligent Medicine*, vol. 4, no. 1, pp. 52–57, 2024, doi: https://doi.org/10.1016/j.imed.2023.08.001.

[2] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature Machine Intelligence*, vol. 1, no. 5, pp. 206–215, May 2019, doi: 10.1038/s42256-019-0048-x.

[3] P. W. Koh *et al.*, "Concept Bottleneck Models," in *Proceedings of the 37th International Conference on Machine Learning*, H. D. III and A. Singh, Eds., in Proceedings of Machine Learning Research, vol. 119. PMLR, 2020, pp. 5338–5348. [Online]. Available: https://proceedings.mlr.press/v119/koh20a.html

[4] Y. Yang, A. Panagopoulou, S. Zhou, D. Jin, C. Callison-Burch, and M. Yatskar, "Language in a bottle: Language model guided concept bottlenecks for interpretable image classification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 19187–19197.

[5] A. Radford *et al.*, "Learning Transferable Visual Models From Natural Language Supervision," in *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, M. Meila and T. Zhang, Eds., in Proceedings of Machine Learning Research, vol. 139. PMLR, 2021, pp. 8748–8763. [Online]. Available: http://proceedings.mlr.press/v139/radford21a.html

[6] S. Ghosh, K. Yu, F. Arabshahi, and K. Batmanghelich, "Dividing and Conquering a BlackBox to a Mixture of Interpretable Models: Route, Interpret, Repeat," in *Proceedings of the 40th International Conference on Machine Learning*, A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, Eds., in Proceedings of Machine Learning Research, vol. 202. PMLR, 2023, pp. 11360–11397. [Online]. Available: https://proceedings.mlr.press/v202/ghosh23c.html

[7] T. Oikarinen, S. Das, L. M. Nguyen, and T.-W. Weng, "Label-free Concept Bottleneck Models," in *International Conference on Learning Representations*, 2023.

[8] X. Yang *et al.*, "Mixture of Experts Made Intrinsically Interpretable." [Online]. Available: https://arxiv.org/abs/2503.07639

[9] J. Schreiber, J. Bilmes, and W. S. Noble, "apricot: submodular selection for data summarization in Python," *J. Mach. Learn. Res.*, vol. 21, no. 1, Jan. 2020.

[10] S. Eslami, C. Meinel, and G. De Melo, "PubMedCLIP: How Much Does CLIP Benefit Visual Question Answering in the Medical Domain?," in *Findings of the Association for Computational Linguistics: EACL 2023*, 2023, pp. 1151–1163.

[11] S. Zhang *et al.*, "A Multimodal Biomedical Foundation Model Trained from Fifteen Million Image–Text Pairs," *NEJM AI*, vol. 2, no. 1, p. AIoa2400640, 2025, doi: 10.1056/AIoa2400640.

[12] D. Eigen, M. Ranzato, and I. Sutskever, "Learning Factored Representations in a Deep Mixture of Experts." [Online]. Available: https://arxiv.org/abs/1312.4314

[13] S. Jiang, T. Zheng, Y. Zhang, Y. Jin, L. Yuan, and Z. Liu, "Med-MoE: Mixture of Domain-Specific Experts for Lightweight Medical Vision-Language Models," in *Findings of the Association for Computational Linguistics: EMNLP 2024*, Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, Eds., Miami, Florida, USA: Association for Computational Linguistics, Nov. 2024, pp. 3843–3860. doi: 10.18653/v1/2024.findings-emnlp.221.

[14] A. A. Ismail, S. Ö. Arik, J. Yoon, A. Taly, S. Feizi, and T. Pfister, "Interpretable Mixture of Experts for Structured Data," *ArXiv*, 2022, [Online]. Available: https://api.semanticscholar.org/CorpusID:249394928

[15] P. Tschandl, C. Rosendahl, and H. Kittler, "The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions," *Scientific Data*, vol. 5, no. 1, Aug. 2018, doi: 10.1038/sdata.2018.161.

[16] T. Rahman *et al.*, "Exploring the effect of image enhancement techniques on COVID-19 detection using chest X-ray images," *Computers in Biology and Medicine*, vol. 132, p. 104319, 2021, doi: https://doi.org/10.1016/j.compbiomed.2021.104319.

[17] M. E. H. Chowdhury *et al.*, "Can AI Help in Screening Viral and COVID-19 Pneumonia?," *IEEE Access*, vol. 8, no. , pp. 132665–132676, 2020, doi: 10.1109/ACCESS.2020.3010287.

[18] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, "ChestX-ray8: Hospital-scale Chest X-ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases," *CoRR*, 2017, [Online]. Available: http://arxiv.org/abs/1705.02315

[19] C. Raffel *et al.*, "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer," *CoRR*, 2019, [Online]. Available: http://arxiv.org/abs/1910.10683

[20] H. Touvron *et al.*, "LLaMA: Open and Efficient Foundation Language Models." [Online]. Available: https://arxiv.org/abs/2302.13971

[21] DeepSeek-AI *et al.*, "DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning." [Online]. Available: https://arxiv.org/abs/2501.12948

[22] X. Zhai, A. Kolesnikov, N. Houlsby, and L. Beyer, "Scaling Vision Transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2022, pp. 12104–12113.

[23] L. Biewald, "Experiment Tracking with Weights and Biases." [Online]. Available: https://www.wandb.com/

[24] C. Sitaula and S. Aryal, "New bag of deep visual words based features to classify chest x-ray images for COVID-19 diagnosis," *Health Information Science and Systems*, vol. 9, no. 1, p. 24, Jun. 2021, doi: 10.1007/s13755-021-00152-w.

[25] H. Zhao, Z. Qiu, H. Wu, Z. Wang, Z. He, and J. Fu, "HyperMoE: Towards Better Mixture of Experts via Transferring Among Experts." [Online]. Available: https://arxiv.org/abs/2402.12656