A TRAINING-FREE TRUSTWORTHY FRAMEWORK FOR MEDICAL IMAGE DIAGNOSIS WITH LARGE LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models (LLMs) enable training-free and concept-level interpretable medical image diagnosis. However, two issues prevent their clinical use: hallucination, under which erroneous predictions carry nearly the same high confidence as correct ones and caused the model untrustworthy, and the long-tailed class distribution of medical data, which makes common diseases overwhelm rare ones. To address these issues, in this paper, we propose a novel trustworthy diagnosis framework built entirely on frozen LLMs without any training. To achieve the trustworthy diagnosis, we design two inference pathways: a concept matching pathway that diagnoses based on medical concepts, and a visual perception pathway that diagnoses from visual features. Both pathways are further corrected by a long-tailed prior debiasing module to tackle the long-tailed distribution issue. Then, a Dirichlet evidence fusion module finally converts the two debiased posteriors into evidential concentrations, producing a trustworthy fused diagnosis together with a sample-wise uncertainty. Experiments on benchmark datasets show its effectiveness. Besides, the uncertainty of our method is aligned with the accuracy of the prediction, enabling hallucination detection that traditional confidence alone cannot provide.

1 INTRODUCTION

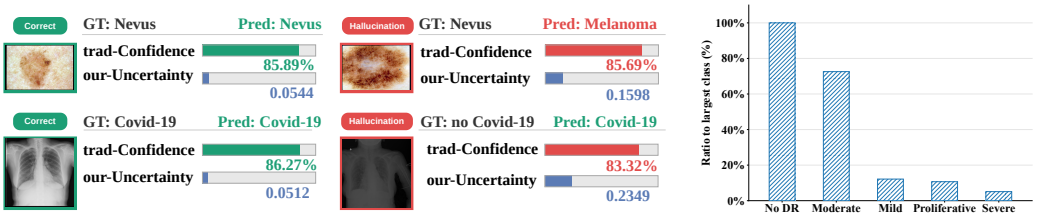
Medical image diagnosis is a crucial component of clinical decision-making. One kind of mainstream medical image diagnosis method is the deep neural network based method (Esteva et al., 2017). They apply the deep neural networks, such as the U-Net (Ronneberger et al., 2015; Ou et al., 2024) and ViT (Park et al., 2022), to achieve medical image classification. Although these methods have achieved promising classification performance, due to the black-box property of deep neural networks, these methods often lack interpretability. Unfortunately, interpretability is a necessary property in medical image diagnosis. Only the interpretable model can earn humans' trust.

To this end, the Concept Bottleneck Model (CBM) (Koh et al., 2020) is proposed. It inserts a human-understandable concept layer between the input image and final diagnosis. Given the medical concepts, it first constructs the relations between concepts and images, and then makes the diagnosis based on concepts, thereby achieving traceable reasoning. Due to its good interpretability, CBM is widely used in medical image diagnosis (Gao et al., 2024; Chowdhury et al., 2024). In recent years, as large language models (LLMs) have been widely studied, some works apply medical-level LLMs to handle medical images (Nam et al., 2025; Fang et al., 2025). In addition, the entire pipeline can run without training, providing flexibility and scalability for clinically-assisted diagnosis (Patrício et al., 2025a; 2024). More detailed related work can be found in Appendix A.

However, although LLM has achieved success in many visual tasks, there are two issues that hinder its use in medical tasks. The first issue is the *hallucination* of LLM. In high-risk areas, such as medical diagnosis, models must be trustworthy. However, hallucination (Zhu et al., 2025) is a well-known problem in LLM, which prevents it from being trustworthy in medical diagnosis. When handling difficult samples or insufficient evidence, the model may output incorrect diagnoses with high confidence. Therefore, hallucinated and non-hallucinated diagnosis may not show significant differences in confidence (Kalai et al., 2025), and thus clinicians cannot rely solely on confidence to

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

distinguish reliable predictions from risky guesses. Figure 1a shows an example. "trad-Confidence" denotes the confidence used in traditional methods, and "our-Uncertainty" denotes the uncertainty computed in our method. Conventional model produces near-identical confidence for correct predictions (left) and hallucinated wrong predictions (right), leading to the hallucination in diagnosis.



(a) Hallucination of LLM.

(b) Long-tailed distribution of medical data.

Figure 1: Two key issues for medical LLMs. (a) The traditional model produces near-identical high confidence score for both correct predictions and hallucinated wrong predictions. Our proposed uncertainty score is substantially higher for hallucinated cases than for correct ones. (b) Class distribution of the DDR dataset, shown as the ratio of each class to the largest class. The severe long-tailed distribution reflects the extreme class imbalance present in medical datasets.

The second issue is the *long-tailed distribution* of medical data. In medical data, a few classes may dominate the dataset, while many other classes may have only a few samples. For example, many rare diseases often have very few samples in the dataset, leading to an extreme imbalance of the dataset (Mayer et al., 2026). Figure 1b shows the class distribution in the DDR dataset. We can see that two classes dominate the whole dataset and the other three classes only have very few samples. Conventional LLMs may rarely use these rare disease samples during training, so they may not handle them well. Moreover, conventional training loss based on accuracy may also systematically ignore samples of rare diseases, also leading to failure of rare disease diagnosis.

To address the two issues of LLM in medical image diagnosis, in this paper, we propose an innovative training-free trustworthy medical image diagnosis framework based on LLMs. To tackle the hallucination issue, we propose two heterogeneous pathways for classification, so that they provide mutually verified evidence rather than a single piece of unchecked evidence. One pathway is the concept matching pathway, which applies the concepts to diagnose. The other pathway is the visual perception pathway, which uses the visual features of image to diagnose. Then, we design a Dirichlet evidence fusion module to integrate the diagnosis of the two pathways, and simultaneously obtain the uncertainty of each pathway for trustworthy diagnosis. Clinicians can easily decide whether to trust the diagnosis of the model by our uncertainty.

To tackle the long-tailed distribution issue, we integrate a long-tailed prior debiasing mechanism into both pathways so that diagnosis is driven by the test image itself rather than class statistical frequency. More specifically, we directly apply the balanced class probability in our module by replacing the empirical class-frequency prior with a uniform prior, so that this explicit factor no longer biases the prediction. By debiasing the class frequency prior to each pathway, minority classes are no longer overwhelmed by common classes in the diagnosis, and thus the long-tailed distribution issue can be effectively alleviated.

To sum up, our main contributions are as follows:

- We propose a novel dual pathway reasoning architecture together with a trustworthy fusion mechanism. The trustworthy fusion mechanism can assign an uncertainty score, which is often aligned with the accuracy of the prediction. Based on the uncertainty score, our method can detect hallucination and achieve a trustworthy diagnosis.
- We design long-tailed prior debiasing within each pathway. This strips the class frequency prior from inference, so diagnosis is driven by the test data itself rather than statistical frequency of the dataset. As a result, rare diseases are not overwhelmed under the long-tailed distribution.
- Extensive experiments on benchmark medical image datasets verify the effectiveness of the framework. Experiments also demonstrate that our uncertainty score can effectively guide the model to achieve the trustworthy diagnosis.

2 METHOD

We follow the conventional setting of concept bottleneck model Koh et al. (2020). In this setting, we have a training set $\mathcal{D}_{\text{train}} = \{(x_i, y_i)\}_{i=1}^{N}$ containing N medical images and K classes, where $x_i \in \mathbb{R}^{H\times W\times C}$ denotes the image with height H, width W, and C channels, and $y_i \in \{1, 2, \cdots, K\}$ denotes the label of x_i. Besides the images, we also have M concepts $c_1(\cdot), \cdots, c_M(\cdot)$ of all training images. Each concept $c_i(\cdot) : \mathbb{R}^{H\times W\times C} \to \mathcal{O}^{(i)}$ can be regarded as a map of an image, where the range $\mathcal{O}^{(i)} = \{o_{i,1}, \cdots, o_{i,r_i}\}$ is the set of values of concept $c_i(\cdot)$ and r_i is the number of possible values of concept $c_i(\cdot)$. For example, in the Covid-19 diagnosis task, given a chest CT image x, we have several concepts of each image, e.g., pleural effusion and pneumothorax. For instance, $c_1(\cdot)$ denotes the concept of pleural effusion. $\mathcal{O}^{(1)} = \{o_{1,1}, o_{1,2}\}$ denotes the two possible values of pleural effusion, i.e., $o_{1,1}$ = "$present$" means pleural effusion is observed in the image; and $o_{1,2}$ = "$absent$" means pleural effusion is not observed in the image. $c_1(x_i)$ = "$present$" means that pleural effusion is observed in the image x_i. Notice that, for all images x_is in the training set, all their concept values $c_j(x_i)$ are already known as inputs. However, in the test phase, all concepts of this test image x_{test} are unknown. We should predict its label y_{test} only by the image x_{test} itself.

Figure 2 shows our method's framework. Notice that, since it is a training-free method, all parameters in the networks or LLMs are frozen, and we directly predict the test image x_{test}. In the following subsections, we will introduce the details of each stage of our method framework.

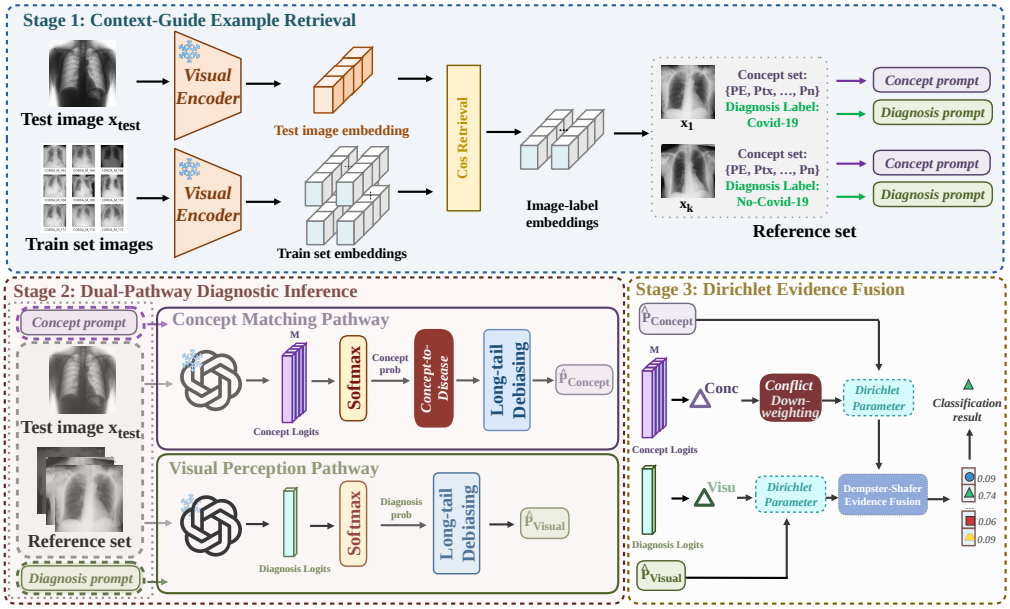


Figure 2: Overview of our framework. Stage 1: Retrieve the top-T most similar training images with a frozen visual encoder to build a reference set and the prompts. Stage 2: Diagnose with the concept matching pathway and the visual perception pathway, both debiased for the long-tailed prior. Stage 3: Convert the two debiased posteriors into Dirichlet evidence and fuse them via Dempster–Shafer to output a diagnosis with the uncertainty.

2.1 STAGE 1: CONTEXT-GUIDE EXAMPLE RETRIEVAL

Since it is a training-free framework, we directly predict test images. Given a test image x_{test}, we first retrieve the top T most relevant images in the training set to guide its prediction. To this end, we apply a frozen visual encoder to extract the features of the test image x_{test} and training images. Then, in the embedding space, we compute the cosine similarity of the test image and all training images. We select the top T most similar images in training set to form a reference set $\mathcal{D}_{\text{reference}}$.

One of the motivations of $\mathcal{D}_{\text{reference}}$ is to provide reference patterns for the current test image x_{test} without updating the LLM parameters, thereby guiding the model to generate predictions. To this end, we design a diagnosis prompt $\mathcal{P}_{\text{Diagnosis}}$ according to the data and labels in $\mathcal{D}_{\text{reference}}$. This will be used in the visual perception pathway to predict the class of the test image x_{test} in Stage 2. Notice that the concepts of test image x_{test} are unknown. Therefore, another motivation of $\mathcal{D}_{\text{reference}}$ is to guide the prediction of the concepts of x_{test}. To this end, we design a concept prompt $\mathcal{P}_{\text{Concept}}$ according to the concepts of the images in $\mathcal{D}_{\text{reference}}$. $\mathcal{P}_{\text{Concept}}$ will be used in the concept matching pathway to predict x_{test}'s concepts in Stage 2. Appendix B provides the detailed examples of the prompts $\mathcal{P}_{\text{Diagnosis}}$ and $\mathcal{P}_{\text{Concept}}$.

2.2 Stage 2: Dual-Pathway Diagnostic Inference

Given a test image x_{test}, after obtaining its reference set, we will diagnose this image. Notice that if we only use the visual features of the image for prediction, the model will lack interpretability. If we only use its concepts to diagnose, we will lose a lot of visual information, limiting the accuracy of diagnosis. To this end, we design two pathways for diagnosis. The concept matching pathway applies the concept information to enhance the interpretability of the model. The visual perception pathway applies the rich visual information for diagnosis. Since medical image datasets often have a long-tailed distribution, which may deteriorate the performance of prediction for the rare diseases, we design a long-tailed debiasing module for both pathways. In the following, we will introduce each part in more detail.

2.2.1 Concept Matching Pathway

Given a test image x_{test}, the values of its concepts are unknown. Therefore, we design a concept matching pathway to predict the values of all concepts of x_{test}. To this end, we feed the test image x_{test} and the concept prompt $\mathcal{P}_{\text{Concept}}$ obtained from its reference set into the LLM. Since the data have M concepts, the LLM will output M logit vectors, where for the concept $c_i(\cdot)$, it outputs a logit vector $z_i^{\text{Concept}} = [z_{i,1}^{\text{Concept}}, \cdots, z_{i,r_i}^{\text{Concept}}] \in \mathbb{R}^{r_i}$ and r_i is the number of possible values of $c_i(\cdot)$. Then we feed it into a Softmax normalization to obtain the probability of each value:

$$P\left(c_i(x_{\text{test}}) = o_{i,j} \mid x_{\text{test}}\right) = \frac{\exp\left(z_{i,j}^{\text{Concept}}\right)}{\sum_{j'=1}^{r_i} \exp\left(z_{i,j'}^{\text{Concept}}\right)}. \tag{1}$$

In the text below, unless otherwise specified, we will abbreviate $P\left(c_i(x_{\text{test}}) = o_{i,j} \mid x_{\text{test}}\right)$ as $P\left(c_i = o_{i,j} \mid x_{\text{test}}\right)$ for simplicity. Applying this process to all M concepts yields M concept probability vectors $\{P(c_i = o_{i,j} \mid x_{\text{test}})\}_{i=1}^{M}$ for the test image x_{test}.

Then, we apply Bayes' Theorem to predict the label of x_{test} using all concept probability vectors. We denote $P(c_i)$ and $P(c_i|y)$ as the prior distribution of the i-th concept and the prior distribution of the i-th concept in class y, respectively. Under the assumptions of independence between concepts, according to Bayes' Theorem, we predict the label of x_{test} with the following formula, where $P(y_{\text{test}})$ is the prior distribution of class y_{test}. The detailed derivation can be found in Appendix C.

$$P_{\text{Concept}}(y_{\text{test}} \mid x_{\text{test}}) = P(y_{\text{test}}) \prod_{i=1}^{M} \Big(\sum_{j=1}^{r_i} \frac{P(c_i = o_{i,j} \mid x_{\text{test}})\, P(c_i = o_{i,j} \mid y_{\text{test}})}{P(c_i = o_{i,j})} \Big). \tag{2}$$

2.2.2 Visual Perception Pathway

Unlike the concept matching pathway, the visual perception pathway does not go through concepts but directly predicts the diagnostic categories from the visual features of images. Given a test image x_{test} and its reference set, under the diagnosis prompt $\mathcal{P}_{\text{Diagnosis}}$, LLM outputs a raw logits $z^{\text{Visual}} = [z_1^{\text{Visual}}, z_2^{\text{Visual}}, \ldots, z_K^{\text{Visual}}]$ for the K categories of x_{test}. Through a Softmax normalization, the probability distribution of the diagnostic category y_{test} is obtained as follows:

$$P_{\text{Visual}}(y_{\text{test}} \mid x_{\text{test}}) = \frac{\exp\left(z_k^{\text{Visual}}\right)}{\sum_{k'=1}^{K} \exp\left(z_{k'}^{\text{Visual}}\right)}. \tag{3}$$

2.2.3 LONG-TAILED PRIOR DEBIASING

In medical image diagnosis tasks, the class distribution of a dataset is often highly skewed. Long-tailed bias causes the model to tend to misclassify rare diseases as common ones. Both pathways of our framework suffer from this bias. Therefore, we need to debias for both pathways. To address the long-tailed distribution issue, inspired by the logit adjustment framework (Menon et al., 2021), we design a long-tailed prior debiasing module for each pathway. In more detail, given a test image x_{test}, conventional Bayesian method predicts its class y_{test} by $P(y_{\text{test}} \mid x_{\text{test}}) \propto P(y_{\text{test}}) \cdot P(x_{\text{test}} \mid y_{\text{test}})$, where $P(y_{\text{test}})$ is the prior. Since the data follows a long-tailed distribution, $P(y_{\text{test}})$ of many classes are very small, leading to inaccurate $P(y_{\text{test}} \mid x_{\text{test}})$. To tackle this problem, the logit adjustment framework replaces prior $P(y_{\text{test}})$ with a balanced prior $P(y_{\text{test}}) = \frac{1}{K}$, leading to unbiased prediction $\hat{P}(y_{\text{test}}|x_{\text{test}})$:

$$\hat{P}(y_{\text{test}} \mid x_{\text{test}}) \propto \frac{1}{K} \cdot P(x_{\text{test}} \mid y_{\text{test}}) \propto P(x_{\text{test}} \mid y_{\text{test}}) \propto \frac{P(y_{\text{test}} \mid x_{\text{test}})}{P(y_{\text{test}})}. \tag{4}$$

Notice that the conventional logit adjustment framework (Menon et al., 2021) has an assumption that the model should be trained to convergence on the current data distribution. Under this assumption, the prior adjustment can precisely offset the influence brought about by the long tail of the data. However, this assumption does not hold in our method. Both pathways of our method are built on a parameter-frozen LLM whose weights have never been updated on the current dataset, and thus the class prior implicitly encoded in its outputs does not match the actual class proportions of the dataset. Consequently, Equation 4 cannot be directly applied to our method.

Since the frozen LLM does not contain any information of our dataset, we should not completely eliminate the prior of dataset as conventional logit adjustment did. To this end, we introduce a debiasing strength parameter $\tau \geq 0$ into the framework, generalizing the prior term $P(y_{\text{test}})$ in the denominator to $P(y_{\text{test}})^{\tau}$, where $\tau = 0$ corresponds to no debiasing, i.e., the prior is fully retained, and $\tau = 1$ corresponds to complete removal of the data prior. When $0 < \tau < 1$, it is a trade-off between retaining the data prior and debiasing the long-tailed distribution.

More specifically, in the concept matching pathway, according to Equation 4, we divide $P(y_{\text{test}}|x_{\text{test}})$ in Equation 2 by $P(y_{\text{test}})^{\tau_1}$, leading to the debiasing probability:

$$\hat{P}_{\text{Concept}}(y_{\text{test}} \mid x_{\text{test}}) = \frac{P_{\text{Concept}}(y_{\text{test}} \mid x_{\text{test}}) \,/\, P(y_{\text{test}})^{\tau_1}}{\sum_{y'=1}^{K} P_{\text{Concept}}(y' \mid x_{\text{test}}) \,/\, P(y')^{\tau_1}} \tag{5}$$

In the visual perception pathway, we apply a similar way to adjust Equation 3. Since $P_{\text{Visual}}(y \mid x)$ is directly given by a frozen LLM followed by a Softmax, the prior is absorbed into the uncalibrated logits and cannot be explicitly separated as in the concept matching pathway. We directly divide the posterior $P_{\text{Visual}}(y_{\text{test}} \mid x_{\text{test}})$ by the debiasing $P(y_{\text{test}})^{\tau_2}$, and obtain the debiasing prediction:

$$\hat{P}_{\text{Visual}}(y_{\text{test}} \mid x_{\text{test}}) = \frac{P_{\text{Visual}}(y_{\text{test}} \mid x_{\text{test}}) \,/\, P(y_{\text{test}})^{\tau_2}}{\sum_{y'=1}^{K} P_{\text{Visual}}(y' \mid x_{\text{test}}) \,/\, P(y')^{\tau_2}}. \tag{6}$$

After the above debiasing, $\hat{P}_{\text{Concept}}$ and $\hat{P}_{\text{Visual}}$ serve as debiased predictions fed into Section 2.3.

2.3 STAGE 3: DIRICHLET EVIDENCE FUSION

Since the LLMs used in both pathways may suffer the hallucination problem, after the two pathways, we design a Dirichlet evidence fusion module to achieve trustworthy fusion by eliminating the hallucination. Classification logits can be regarded as a kind of multinomial distribution. Conventional methods often obtain the logits by Softmax function, which may lead to overconfidence (van Amersfoort et al., 2020). To this end, following (Han et al., 2023), we apply the Dirichlet distribution to predict the logits. Due to the Dirichlet distribution, we can evaluate the uncertainty of the prediction. We apply the Dempster-Shafer evidence theory (Dubois & Prade, 1983) to achieve trustworthy fusion with uncertainty. In the following, we will introduce this process in more detail.

2.3.1 DUAL-PATHWAY EVIDENCE PARAMETER CONSTRUCTION

This step applies the Dirichlet distribution to model the outputs of the two pathways, assigning an evidence mass to the prediction of each class. In (Han et al., 2023), they use a Dirichlet distribution

whose concentration parameters are $\alpha_1, \cdots, \alpha_K$ for prediction. In more detail, $\alpha_k = e_k + 1$ for $1 \leq k \leq K$, where e_k is the evidence mass of the k-th class and is predicted by the network as $e_k = \sigma(z_k)$ and z_k is the output of the network and $\sigma(\cdot)$ is a non-negative activation function.

However, Han et al. (2023) trains the network to directly output the evidence mass e_k, but our frozen LLM is trained to output the classification probability instead. If we directly use the method in (Han et al., 2023), under the same confidence level, different logit combinations may be mapped to different evidence masses. The evidence mass then no longer reflects accurately the true confidence. To this end, we design a new way to construct the evidence mass e_k. Intuitively, e_k should be proportional to the probability that x_{test} belong to the k-th class, i.e., $e_k \propto \hat{P}(y_{\text{test}} = k \mid x_{\text{test}})$ where $\hat{P}()$ is the debiased probability (i.e., $\hat{P}_{\text{Concept}}()$ or $\hat{P}_{\text{Visual}}()$). Hence, we construct e_k by $e_k = \Delta \cdot \hat{P}(y_{\text{test}} = k \mid x_{\text{test}})$, where Δ characterizes the discriminativity of the classifier. In more detail, we define Δ as the difference between the largest logit and the mean of the remaining logits. More formally, let $z_{\max} = \max_k z_k$ be the largest logit. Then, we define Δ. Appendix D provides a theoretical derivation of this definition.

$$\Delta = z_{\max} - \frac{1}{K-1} \sum_{z_k \neq z_{\max}} z_k. \tag{7}$$

Then, we apply Equation 7 to compute the concentration parameter α_k for each pathway. For the visual perception pathway, we define $\alpha_k^{\text{Visual}} = 1 + e_k^{\text{Visual}} = 1 + \Delta^{\text{Visual}} \cdot \hat{P}_{\text{Visual}}(y_{\text{test}} = k \mid x_{\text{test}})$, where Δ^{Visual} is obtained by taking z^{Visual} into Equation 7. For the concept matching pathway, given a test image x_{test}, we obtain M logit vectors $z_1^{\text{Concept}}, \cdots, z_M^{\text{Concept}}$, where each logit vector z_i^{Concept} corresponds to a concept. Hence, according to Equation 7, we can obtain $\Delta^{(1)}, \cdots, \Delta^{(M)}$, where $\Delta^{(i)}$ is obtained from z_i^{Concept}. To this end, we should fuse these $\Delta^{(i)}$s to obtain a consensus Δ^{Concept}, which will be introduced in the next subsection.

Evidence Fusion with Conflict Down-weighting for the Concept Pathway A natural way to fuse $\Delta^{(i)}$s is to compute their average $\Delta^{\text{Concept}} = \frac{1}{M} \sum_{i=1}^{M} \Delta^{(i)}$. However, directly using this Δ^{Concept} ignores the dependencies between concepts. Besides, in Section 2.2.1, we also assume that the concepts are independent. This is often not true in practice. Therefore, in this subsection, when fusing the evidence mass, we take the dependency of two concepts into consideration. Notice that, the dependency of concepts is also often ignored in previous concept based methods (Heidemann et al., 2023; Havasi et al., 2022). The basic idea is to check whether the predicted concepts conflict with the medical knowledge. For example, given a test image x_{test}, if its i-th predicted concept is $o_{i,a}$ (i.e., $c_i(x_{\text{test}}) = o_{i,a}$) and its j-th concept is $o_{j,b}$ (i.e., $c_j(x_{\text{test}}) = o_{j,b}$), but in medical knowledge the co-occurrence of $c_i = o_{i,a}$ and $c_j = o_{j,b}$ is rare, the prediction is not trustworthy and will down-weight the evidence mass.

Specifically, we use mutual information $MI_{ij}^{ab} = \ln \frac{P(c_i = o_{i,a}, c_j = o_{j,b})}{P(c_i = o_{i,a}) \cdot P(c_j = o_{j,b})}$ of c_i and c_j as the medical knowledge to check the prediction. $MI_{ij}^{ab} > 0$ means $c_i = o_{i,a}$ and $c_j = o_{j,b}$ are positively correlated, $MI_{ij}^{ab} < 0$ means they are negatively correlated, and $MI_{ij}^{ab} = 0$ means they are uncorrelated. We compute the conflict score $Conflict_{ij}^{ab}(x_{\text{test}})$ with $P(c_i(x_{\text{test}}) = o_{i,a})$ and $P(c_j(x_{\text{test}}) = o_{j,b})$.

$$Conflict_{ij}^{ab}(x_{\text{test}}) = \begin{cases} 1 - |P(c_i = o_{i,a} \mid x_{\text{test}}) - P(c_j = o_{j,b} \mid x_{\text{test}})|, & if \quad MI_{ij}^{ab} < 0, \\ 0, & if \quad MI_{ij}^{ab} = 0, \\ \left|P(c_i = o_{i,a} \mid x_{\text{test}}) - P(c_j = o_{j,b} \mid x_{\text{test}})\right|, & if \quad MI_{ij}^{ab} > 0. \end{cases} \tag{8}$$

Larger $Conflict_{ij}^{ab}(x_{\text{test}})$ means the prediction is more inconsistent with the medical knowledge and thus the prediction is more unreliable. We define $w = \max_{c_i, c_j, o_{i,a}, o_{j,b}} Conflict_{ij}^{ab}(x_{\text{test}})$ as the maximum over all predicted concepts, and impose an exponential decay on Δ^{Concept} as $\tilde{\Delta}^{\text{Concept}} = \Delta^{\text{Concept}} \cdot \exp(-w)$. Then we obtain the concentration parameters $\alpha_k^{\text{Concept}}$ as:

$$\alpha_k^{\text{Concept}} = 1 + \tilde{\Delta}^{\text{Concept}} \cdot \hat{P}_{\text{Concept}}(y_{\text{test}} = k \mid x_{\text{test}}). \tag{9}$$

2.3.2 DEMPSTER–SHAFER EVIDENCE FUSION

According to Dirichlet evidence theory (Han et al., 2023; Jøsang, 2016), we can construct the belief mass b_k^{Concept} and b_k^{Visual} and the uncertainty mass u^{Concept} and u^{Visual} as follows:

$$b_k^{\text{Concept}}=\frac{\alpha_k^{\text{Concept}}-1}{\tilde{\Delta}^{\text{Concept}}+K}, b_k^{\text{Visual}}=\frac{\alpha_k^{\text{Visual}}-1}{\Delta^{\text{Visual}}+K}, u^{\text{Concept}}=\frac{K}{\tilde{\Delta}^{\text{Concept}}+K}, u^{\text{Visual}}=\frac{K}{\Delta^{\text{Visual}}+K}. \tag{10}$$

According to Dempster-Shafer evidence theory (Dubois & Prade, 1983), we fuse the belief mass and uncertainty mass. In more detail, we define the conflict measure between the two pathways $C = \sum_{i\neq j} b_i^{\text{Concept}} b_j^{\text{Visual}}$. A larger C indicates more severe disagreement between the two pathways. Then we compute the fused belief mass and uncertainty mass as follows:

$$b_k^{\text{Fused}}=\frac{b_k^{\text{Concept}} b_k^{\text{Visual}}+b_k^{\text{Concept}} u^{\text{Visual}} + b_k^{\text{Visual}} u^{\text{Concept}}}{1-C}, \quad u^{\text{Fused}}=\frac{u^{\text{Concept}} u^{\text{Visual}}}{1-C}. \tag{11}$$

According to Dempster-Shafer evidence theory, the fused uncertainty u^{Fused} is lower than u^{Concept} and u^{Visual}, which means the uncertainty is reduced by the fusion and we can obtain a more trustworthy result after this fusion. Similar to Equation 10, we can compute α_k^{Fused} by b_k^{Fused} and u^{Fused} as follows, and the final diagnostic distribution is taken as the Dirichlet mean:

$$\alpha_k^{\text{Fused}}=\frac{b_k^{\text{Fused}}\cdot K}{u^{\text{Fused}}}+1, \quad P_{\text{Fused}}(y_{\text{test}}=k|x_{\text{test}})=\frac{\alpha_k^{\text{Fused}}}{\sum_{k=1}^{K}\alpha_k^{\text{Fused}}}=b_k^{\text{Fused}}+\frac{u^{\text{Fused}}}{K}. \tag{12}$$

The final prediction is obtained by taking the maximum a posteriori: $y_{\text{pred}} = \arg\max_k P_{\text{Fused}}(y_{\text{test}} = k|x_{\text{test}})$. The whole algorithm is introduced in Appendix E.

3 EXPERIMENT

3.1 EXPERIMENTAL SETUP AND IMPLEMENTATION DETAILS

We conducted experiments on three public medical image classification benchmarks, including the dermoscopy image dataset Derm7pt (Kawahara et al., 2019), the chest X-ray image dataset CORDA (Alesina et al., 2023), and the fundus color photograph dataset DDR (Li et al., 2019). Each dataset provided expertly annotated clinical medical concepts. The key information of each dataset is summarized in Appendix F. Then, we compared with 12 methods, including: (1) Concept-based models without LLM: CBM (Koh et al., 2020), ProbCBM (Kim et al., 2023), evi-CEM (Gao et al., 2024); (2) Non-concept-based models without LLM: SOTA(Hu et al., 2026), SS-Text (Silva-Rodríguez et al., 2025), MadCLIP (Shiri et al., 2025). (3) Concept-based models with LLM: MVP-CBM (Wang et al., 2025), CBVLM (Patrício et al., 2025a), Two-Step (Patrício et al., 2025b); (4) Non-concept-based models with LLM: KPL (Liu et al., 2025), BiomedCoOp (Koleilat et al., 2025), MedUnA (Rahman et al., 2025). All results are reproduced with the official code of each method on the same data splits. For a fair comparison, we use the same visual encoder and LLM models for all methods. Since we consider the long-tail distribution of medical data, we report balanced accuracy (BACC) and Macro-F1 in imbalance classification setting. Both metrics are the larger the better.

We use three pre-trained LLMs as the inference backbone of the framework: Qwen3-VL (Bai et al., 2025) (7B), CheXagent (Chen et al., 2024) (8B), and Idefics3 (Laurençon et al., 2024) (8B). All three models maintained parameter freezing during the experiment and extracted probability values through text logits. These three models cover different pre-training data distributions and model architectures, which can verify the adaptability of the framework in this paper to different LLM backbones. Furthermore, in Stage 1, we use the frozen visual encoder of MedImageInsight (Codella et al., 2024) (a vision-language pre-trained model), and select the top $T = 4$ training images as the reference set. In the long-tailed prior debiasing, we search τ_1 and τ_2 in $[0, 1]$ on the validation set.

3.2 EXPERIMENTAL RESULTS

Table 1 shows the comparison results of our method and other compared methods on all datasets. The proposed method achieves best results on Derm7pt and CORDA datasets and competitive results on DDR. On Derm7pt, the BACC of the three LLMs all reach 86.53 - 88.09%, which is an improvement of 1.26 - 2.82 percentage points compared to the strongest baseline CBVLM (85.27%). Among them, Qwen3-VL achieves the highest BACC of 88.09%, and Idefics3 achieves the highest

Macro-F1 of 85.77%. On CORDA, CheXagent achieves the best BACC of 78.70% and Macro-F1 of 78.43%, which are both improvements of 2.8 and 3.0 percentage points compared to the strongest baseline CBVLM. On DDR, CheXagent ranks second with BACC of 64.36% and Macro-F1 of 61.63%. evi-CEM's BACC (67.80%) is slightly higher, but its Macro-F1 is only 52.89%, much lower than ours, reflecting a significant imbalance between precision and recall.

Table 1: Comparison of main results (%). The best, second-best, third-best result is red, blue, green.

Category	**Method**	**LLM**	Derm7pt		CORDA		DDR	
			BACC	Macro-F1	BACC	Macro-F1	BACC	Macro-F1
Concept-based without LLM	CBM	–	74.01	75.66	70.75	70.98	23.62	21.25
	ProbCBM	–	79.14	82.19	63.87	63.32	39.68	38.12
	evi-CEM	–	79.09	63.17	75.57	50.14	**67.80**	52.89
Non-concept without LLM	SOTA	–	83.55	**85.65**	69.21	66.05	60.44	47.26
	SS-Text	–	71.85	71.08	64.76	64.59	43.68	36.91
	MadCLIP	–	74.35	75.72	61.65	61.29	54.68	48.73
Concept-based with LLM	MVP-CBM	Qwen3-VL	75.18	72.38	69.27	66.73	35.55	33.99
		CheXagent	72.36	70.70	70.65	69.64	32.85	21.31
		Idefics3	68.55	67.98	63.28	62.50	41.88	43.18
	CBVLM	Qwen3-VL	79.53	72.99	69.44	68.36	61.27	59.84
		CheXagent	85.27	81.94	75.05	75.40	60.45	**60.33**
		Idefics3	75.64	71.43	**75.90**	**75.43**	**64.23**	**65.32**
	Two-Step	Qwen3-VL	79.78	79.55	69.02	67.22	40.01	27.28
		CheXagent	77.46	76.18	73.72	73.75	34.44	36.53
		Idefics3	78.38	77.92	70.41	70.36	32.58	32.33
Non-concept with LLM	KPL	Qwen3-VL	78.84	73.86	71.98	70.64	44.96	33.48
		CheXagent	71.54	62.49	71.33	70.41	46.39	31.69
		Idefics3	72.68	64.02	70.83	69.63	54.62	43.42
	BiomedCoOp	Qwen3-VL	71.56	72.54	64.78	64.70	36.10	25.95
		CheXagent	77.50	71.56	69.13	68.58	42.36	30.86
		Idefics3	75.94	70.95	67.99	67.96	39.74	28.96
	MedUnA	Qwen3-VL	82.26	80.89	67.92	64.58	38.10	24.61
		CheXagent	81.88	81.16	69.23	67.82	19.91	14.06
		Idefics3	77.43	70.59	69.58	68.05	25.62	22.52
Ours		Qwen3-VL	**88.09**	**85.13**	74.90	75.18	55.39	45.09
		CheXagent	**86.53**	84.51	**78.70**	**78.43**	**64.36**	**61.63**
		Idefics3	**87.44**	**85.77**	**77.13**	**77.04**	64.14	57.67

3.3 Uncertainty Quantification Analysis

In our framework, we introduce the uncertainty to achieve the trustworthy diagnosis. In this subsection, we conduct experiments to analyze the uncertainty. We report the uncertainty of each pathway (u_{Concept} and u_{Visual}) and the fused results (u_{Fused}) in Table 2. We can see that our trustworthy fusion module can indeed reduce the uncertainty. To alleviate hallucination issue, i.e., the model misclassifies samples with high confidence, we report the uncertainty of the samples that we predict correctly (u_{Correct}) and incorrectly (u_{Wrong}) in Table 2. We can see that the uncertainty of the wrong samples is significantly higher than the uncertainty of the correct samples. Therefore, the clinician can decide whether to trust the model according to the uncertainty u_{Fused}. For example, on Derm7pt with CheXagent, if u_{Fused} exceeds 0.3, we should not trust the model and need clinician's manual review. Appendix G provides more experimental results and discussion of the uncertainty.

Table 2: Uncertainty quantification analysis on three datasets and three LLMs.

Dataset	Derm7pt			CORDA			DDR		
LLM	Qwen3-VL	CheXagent	Idefics3	Qwen3-VL	CheXagent	Idefics3	Qwen3-VL	CheXagent	Idefics3
u_{Concept}	0.3281	0.5949	0.6290	0.3196	0.5840	0.4489	0.3990	0.5683	0.5905
u_{Visual}	0.3345	0.3989	0.3875	0.4053	0.5032	0.4947	0.5752	0.6110	0.5384
u_{Fused}	**0.1284**	**0.2698**	**0.2710**	**0.1543**	**0.3184**	**0.2463**	**0.2522**	**0.3961**	**0.3406**
u_{Correct}	0.1167	0.2570	0.2569	0.1370	0.3051	0.2308	0.2313	0.3775	0.3127
u_{Wrong}	0.2019	0.3478	0.3668	0.2091	0.3674	0.2988	0.3201	0.4790	0.4553

Table 3: Ablation studies (%). Three groups of variants: the two pathways used individually, the removal of the long-tail debiasing mechanisms, and alternative fusion strategies. "Full" denotes the complete framework. All variants use the same backbone that performs best for each dataset.

Group	Setting	Derm7pt		CORDA		DDR	
		BACC	Macro-F1	BACC	Macro-F1	BACC	Macro-F1
(a) Pathway comparison	Only concept pathway	84.62	80.79	64.30	62.40	55.26	48.69
	Only visual pathway	82.07	77.75	77.25	77.58	62.82	59.54
(b) Long-tail debiasing	w/o both debias	84.39	83.93	73.91	74.21	54.07	56.81
	Concept debias only	87.17	84.89	77.78	77.44	61.05	61.13
	Visual debias only	85.84	84.58	75.97	75.96	56.83	57.31
(c) Fusion strategy	Fixed $\Delta = 1$	85.49	84.76	74.50	74.75	59.81	56.50
	Linear fusion	86.60	84.02	77.72	77.42	61.71	56.23
	w/o conflict downweighting	87.29	84.95	78.41	78.16	62.10	57.42
Full framework		**88.09**	**85.13**	**78.70**	**78.43**	**64.36**	**61.63**

3.4 Ablation Studies

We conduct three ablation studies on all three datasets to verify the effects of both pathways, the long-tail debiasing, and the fusion strategy, respectively. The results are summarized in Table 3.

(a) Pathway comparison. Group (a) shows the concept matching pathway and the visual perception pathway. On Derm7pt, concept matching pathway outperforms the visual perception pathway, but on CORDA and DDR, the visual perception pathway is more important. Combining both pathways outperforms each individual pathway. It demonstrates the necessity of combining the two pathways.

(b) Long-tail debiasing. Group (b) ablates the long-tail prior debiasing of the two pathways. When the debiasing mechanisms of the two pathways are removed, the performance on the highly imbalanced DDR drops markedly, with BACC dropping from 64.36% to 54.07%, while the relatively balanced Derm7pt only slightly declines from 88.09% to 84.39%. The bias degree and reduction extent of CORDA are between the two, indicating that debiasing is crucial for alleviating long-tail bias. Further observation shows that retaining the debiasing of the concept matching pathway alone can restore most of the performance, while retaining the debiasing of the visual perception pathway alone yields limited gains. This indicates that the long-tail debiasing is mainly undertaken by the concept matching pathway. The combination of the two pathways achieves the best results on all datasets, verifying the complementarity of the two pathways' debiasing mechanisms. The long-tailed debiasing in two pathways is controlled by τ_1 and τ_2, respectively. We also conduct the experiments to show the effects of the two hyperparameters in Appendix H.

(c) Fusion strategy. Group (c) compares the proposed Dirichlet evidence fusion with three variants. In the first variant the sample-wise evidence magnitude is fixed to $\Delta = 1$, which is used in (Han et al., 2023). The second variant replaces the Dirichlet evidence fusion with a fixed linear weighting. Neither variant produces a sample-adaptive uncertainty, so the fusion weights can no longer adapt to the confidence of each pathway. The third variant disables the conflict down-weighting mechanism, i.e., the exponential decay of the concept pathway evidence $\tilde{\Delta}^{\text{Concept}}$ driven by the conflict score w in Equation 8. The results show that no variant exceeds the full framework on any dataset. For example, on DDR dataset, for $\Delta = 1$ and linear fusion, the Macro-F1 drops to 56.50% and 56.23% from 61.63%, indicating that adaptive and trustworthy fusion is indispensable. Besides, the proposed conflict down-weighting mechanism can further improve the performance.

4 Conclusion

In this paper, we proposed a novel trustworthy framework for medical image diagnosis with LLMs. In our framework, we design two pathways for diagnosis based on the concepts and visual features, and proposed a trustworthy fusion module based on Dempster-Shafer evidence theory. Due to the uncertainty in the trustworthy fusion module, our model can effectively alleviate the hallucination issue in LLMs. Besides, to address the long-tailed distribution issue in medical data, we design an innovative long-tail prior debiasing module for each pathway, which can pay more attention to minority classes. Extensive experiments on benchmark medical image datasets demonstrate the effectiveness and superiority of our method.

AI USE STATEMENT

In this work, we used generative AI tools for no tasks requiring mandatory disclosure. We have not used generative AI tools for generating synthetic datasets, developing theoretical models or conceptual frameworks, formulating mathematical claims, providing critical ingredients for proving mathematical claims, assisting in the writing of proofs, proposing or refining hypotheses, designing or providing feedback on research methodology or experiments, implementing methods; supporting qualitative and thematic data analysis, interpreting results, cleaning and reformatting datasets, and assisting with translation are not applicable to this work. Additionally, we used generative AI tools to edit the research paper to improve its readability and to draft and polish parts of the text. We have reviewed all AI-assisted work. All AI-suggested edits and rephrasings were checked and revised by the authors, who verified that the polished text remains faithful to the original meaning, our methods, results, and claims. No AI-generated research ideas, code, or data were used, so other verification procedures were not applicable. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work uses only three public medical image benchmarks: Derm7pt, CORDA, and DDR. All three were collected and de-identified by their original providers, and we accessed them under their published licenses. We collected no new data and recruited no human subjects, so no new ethics approval was required for this study. We did not attempt to re-identify any individual, and we used each dataset with its default split and its expert concept annotations.

Our method performs inference only. Every backbone model we use is publicly released and runs locally with frozen parameters. No patient image left our machines, and no image was used to train or fine-tune any model. We therefore see no privacy risk introduced by this work.

Two risks deserve attention. First, the three benchmarks cover only a narrow range of populations, skin tones, and imaging devices. Dermoscopy datasets in particular are known to under-represent patients with darker skin, and such a gap can be hidden by an aggregate accuracy score. Our reported numbers should be read as evidence on these benchmarks alone, not as evidence of general clinical validity. Second, a system that outputs a diagnosis together with a confidence can be mistaken for an autonomous diagnostic tool. Our method is a research prototype and not a medical device, and it is not intended for clinical decision making. Any clinical use requires review by a qualified clinician, who remains responsible for the final diagnosis; the uncertainty we produce is meant to support that review rather than to replace it.

REPRODUCIBILITY STATEMENT

We have summarized the experimental details necessary to ensure that the experimental results can be replicated and verified, and have indicated where the relevant details are mentioned. Our method is training-free: no fine-tuning is performed on any backbone, no model weight is modified, and inference reads class probabilities directly from the text logits, so there is neither a training step nor a sampling step to reproduce. All external dependencies are publicly available: the three benchmarks and the openly released weights of all four backbone models (Qwen3-VL 7B, CheXagent 8B, Idefics3 8B and the visual encoder of MedImageInsight).

For the method, the full prompt templates for both pathways and their construction are given in Appendix B, the derivations of the posterior weighting and of the debiasing term in Appendix C and Appendix D, and the complete procedure as pseudocode in Algorithm 1 of Appendix E. For the experimental protocol, every baseline is reproduced with the official code of that method, on the same splits and with the same visual encoder and LLMs. The default splits, sizes, concept sets, and class sets of Derm7pt, CORDA, and DDR are listed in Table 4 of Appendix F. The supporting analyses of the uncertainty output and of the two hyperparameters are reported in Appendix G and Appendix H.

REFERENCES

Marta Alesina, Carlo Alberto Barbano, Claudio Berzovini, Marco Busso, Marco Calandri, Agostino De Pascale, Attilio Fiandrotti, Paolo Fonio, Marco Grangetto, Marco Grosso, Hafiza Ayesha Hoor Chaudhry, Maria Elisabetta Mancini, Teresa Maria Gallo, Chiara Martini, Gianluca Pontone, Riccardo Renzulli, Chiara Saviolo, Federico Signoretta, Stefano Tibaldi, Andrea Veltri, and Mirko Zaffaroni. CORDA dataset. Zenodo, 2023. URL https://doi.org/10.5281/zenodo.7821611.

Diego Ardila, Atilla P. Kiraly, Sujeeth Bharadwaj, Bokyung Choi, Joshua J. Reicher, Lily Peng, Daniel Tse, Mozziyar Etemadi, Wenxue Ye, Greg Corrado, David P. Naidich, and Shravya Shetty. End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. *Nature Medicine*, 25:954–961, 2019.

Shuai Bai, Yuxuan Cai, Ke Zhu, et al. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Zhihong Chen, Maya Varma, Jean-Benoit Delbrouck, Magdalini Paschali, Louis Blankemeier, Dave Van Veen, Jeya Maria Jose Valanarasu, Alaa Youssef, Joseph Paul Cohen, Eduardo Pontes Reis, Emily B. Tsai, Andrew Johnston, Cameron Olsen, Tanishq Mathew Abraham, Sergios Gatidis, Akshay S. Chaudhari, and Curtis P. Langlotz. CheXagent: Towards a foundation model for chest x-ray interpretation. In *AAAI 2024 Spring Symposium on Clinical Foundation Models*, 2024.

Townim F. Chowdhury, Vu Minh Hieu Phan, Kewen Liao, Minh-Son To, Yutong Xie, Anton van den Hengel, Johan W. Verjans, and Zhibin Liao. AdaCBM: An adaptive concept bottleneck model for explainable and accurate diagnosis. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, volume 15010 of *LNCS*, pp. 35–45, 2024.

Noel C. F. Codella, Ying Jin, Shrey Jain, Yu Gu, Ho Hin Lee, Asma Ben Abacha, Alberto Santamaria-Pang, Will Guyman, Naiteek Sangani, Sheng Zhang, Hoifung Poon, Stephanie Hyland, Shruthi Bannur, Javier Alvarez-Valle, Xue Li, John Garrett, Alan McMillan, Gaurav Rajguru, Madhu Maddi, Nilesh Vijayrania, Rehaan Bhimai, Nick Mecklenburg, Rupal Jain, Daniel Holstein, Naveen Gaur, Vijay Aski, Jenq-Neng Hwang, Thomas Lin, Ivan Tarapov, Matthew Lungren, and Mu Wei. MedImageInsight: An open-source embedding model for general domain medical imaging. *arXiv preprint arXiv:2410.06542*, 2024.

D. Dubois and H. Prade. Upper and lower possibilities induced by a multivalued mapping. *IFAC Proceedings Volumes*, 16(13):147–152, 1983. ISSN 1474-6670. IFAC Symposium on Fuzzy Information, Knowledge Representation and Decision Analysis, Marseille, France, 19-21 July, 1983.

Mateo Espinosa Zarlenga, Pietro Barbiero, Gabriele Ciravegna, Giuseppe Marra, Francesco Giannini, Michelangelo Diligenti, Zohreh Shams, Frederic Precioso, Stefano Melacci, Adrian Weller, Pietro Liò, and Mateja Jamnik. Concept embedding models: Beyond the accuracy-explainability trade-off. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, 2022.

Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115–118, 2017.

Mengjie Fang, Zipei Wang, Sitian Pan, Xin Feng, Yunpeng Zhao, Dongzhi Hou, Ling Wu, Xuebin Xie, Xu-Yao Zhang, Jie Tian, and Di Dong. Large models in medical imaging: Advances and prospects. *Chinese Medical Journal*, 138(14):1647–1664, 2025.

Yibo Gao, Zheyao Gao, Xin Gao, Yuanye Liu, Bomin Wang, and Xiahai Zhuang. Evidential concept embedding models: Towards reliable concept explanations for skin disease diagnosis. In *Proceedings of the 27th International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pp. 308–317, 2024.

Varun Gulshan, Lily Peng, Marc Coram, Martin C. Stumpe, Derek Wu, Arunachalam Narayanaswamy, Subhashini Venugopalan, Kasumi Widner, Tom Madams, Jorge Cuadros, Ramasamy Kim, Rajiv Raman, Philip C. Nelson, Jessica L. Mega, and Dale R. Webster. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal

fundus photographs. *Journal of the American Medical Association (JAMA)*, 316(22):2402–2410, 2016.

Zongbo Han, Changqing Zhang, Huazhu Fu, and Joey Tianyi Zhou. Trusted multi-view classification with dynamic evidential fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(2):2551–2566, 2023.

Marton Havasi, Sonali Parbhoo, and Finale Doshi-Velez. Addressing leakage in concept bottleneck models. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pp. 23386–23397. Curran Associates, Inc., 2022.

Lena Heidemann, Maureen Monnet, and Karsten Roscher. Concept correlation and its effects on concept-based models. In *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 4769–4777, 2023.

Junlin Hou, Jilan Xu, and Hao Chen. Concept-attention whitening for interpretable skin lesion diagnosis. In *Proceedings of the 27th International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)*, 2024.

Zhanxuan Hu, Qiyu Xu, Yu Duan, Yonghang Tai, and Huafeng Li. Sota: Self-adaptive optimal transport for zero-shot classification with multiple foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26624–26634, June 2026.

Audun Jøsang. *Subjective Logic: A Formalism for Reasoning Under Uncertainty*. Springer, 2016. ISBN 978-3-319-42335-7. doi: 10.1007/978-3-319-42337-1.

Adam Tauman Kalai, Ofir Nachum, Santosh S. Vempala, and Edwin Zhang. Why language models hallucinate. *arXiv preprint arXiv:2509.04664*, 2025.

Jeremy Kawahara, Sara Daneshvar, Giuseppe Argenziano, and Ghassan Hamarneh. Seven-point checklist and skin lesion classification using multitask multimodal neural networks. *IEEE Journal of Biomedical and Health Informatics*, 23(2):538–546, 2019.

Eunji Kim, Dahuin Jung, Sangha Park, Siwon Kim, and Sungroh Yoon. Probabilistic concept bottleneck models. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, 2023.

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, 2020.

Taha Koleilat, Hojat Asgariandehkordi, Hassan Rivaz, and Yiming Xiao. BiomedCoOp: Learning to prompt for biomedical vision-language models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14766–14776, 2025.

Hugo Laurençon, Andrés Marafioti, Victor Sanh, and Léo Tronchon. Building and better understanding vision-language models: insights and future directions. *arXiv preprint arXiv:2408.12637*, 2024.

Tao Li, Yingqi Gao, Kai Wang, Song Guo, Hanruo Liu, and Hong Kang. Diagnostic assessment of deep learning algorithms for diabetic retinopathy screening. *Information Sciences*, 501:511–522, 2019. ISSN 0020-0255.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pp. 34892–34916, 2023.

Jiaxiang Liu, Tianxiang Hu, Jiawei Du, Ruiyuan Zhang, Joey Tianyi Zhou, and Zuozhu Liu. KPL: Training-free medical knowledge mining of vision-language models. In *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence (AAAI)*, volume 39, pp. 18852–18860, 2025.

Andrei Margeloiu, Matthew Ashman, Umang Bhatt, Yanzhi Chen, Mateja Jamnik, and Adrian Weller. Do concept bottleneck models learn as intended? *arXiv preprint arXiv:2105.04289*, 2021. Accepted at ICLR 2021 Workshop on Responsible AI.

Leon Mayer, Piotr Kalinowski, Caroline Ebersbach, Marcel Knopp, Tim Rädsch, Evangelia Christodoulou, Annika Reinke, Fiona R. Kolbinger, and Lena Maier-Hein. 6 fingers, 1 kidney: Natural adversarial medical images reveal critical weaknesses of vision-language models. *arXiv preprint arXiv:2512.04238*, 2026.

Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar. Long-tail learning via logit adjustment. In *International Conference on Learning Representations (ICLR)*, 2021.

Y Nam, D. Y. Kim, S Kyung, J Seo, J. M. Song, J Kwon, J Kim, W Jo, H Park, J Sung, S Park, H Kwon, T Kwon, K Kim, and N Kim. Multimodal large language models in medical imaging: Current state and future directions. *Korean Journal of Radiology*, 26(10):900–923, 2025.

Chih-Ying Ou, I-Yen Chen, Hsuan-Ting Chang, Chuan-Yi Wei, Dian-Yu Li, Yen-Kai Chen, and Chuan-Yu Chang. Deep learning-based classification and semantic segmentation of lung tuberculosis lesions in chest x-ray images. *Diagnostics*, 14(9):952, 2024. ISSN 2075-4418.

Sangjoon Park, Gwanghyun Kim, Yujin Oh, Joon Beom Seo, Sang Min Lee, Jin Hwan Kim, Sungjun Moon, Jae-Kwang Lim, and Jong Chul Ye. Multi-task vision transformer using low-level chest x-ray feature corpus for COVID-19 diagnosis and severity quantification. *Medical Image Analysis*, 75:102299, 2022.

Cristiano Patrício, Luis F. Teixeira, and João C. Neves. Towards concept-based interpretability of skin lesion diagnosis using vision-language models. In *IEEE International Symposium on Biomedical Imaging (ISBI)*, 2024.

Cristiano Patrício, Isabel Rio-Torto, Jaime S. Cardoso, Luís F. Teixeira, and Jo ao C. Neves. CB-VLM: Training-free explainable concept-based large vision language models for medical image classification. *Computers in Biology and Medicine*, 198:111145, 2025a.

Cristiano Patrício, Luís F. Teixeira, and João C. Neves. A two-step concept-based approach for enhanced interpretability and trust in skin lesion diagnosis. *Computational and Structural Biotechnology Journal*, 28:71–79, 2025b.

Umaima Rahman, Raza Imam, Mohammad Yaqub, Boulbaba Ben Amor, and Dwarikanath Mahapatra. Can language-guided unsupervised adaptation improve medical image classification using unpaired images and texts? In *IEEE International Symposium on Biomedical Imaging (ISBI)*, 2025.

Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis Langlotz, Katie Shpanskaya, Matthew P. Lungren, and Andrew Y. Ng. CheXNet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv preprint arXiv:1711.05225*, 2017.

Naveen Raman, Mateo Espinosa Zarlenga, Juyeon Heo, and Mateja Jamnik. Do concept bottleneck models respect localities? *Transactions on Machine Learning Research*, 2024.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, volume 9351 of *LNCS*, pp. 234–241. Springer, 2015.

Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019.

Mahshid Shiri, Cigdem Beyan, and Vittorio Murino. MadCLIP: Few-shot medical anomaly detection with CLIP. In *proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2025*, volume LNCS 15965, pp. 416–426. Springer Nature Switzerland, September 2025.

Julio Silva-Rodríguez, Fereshteh Shakeri, Houda Bahig, Jose Dolz, and Ismail Ben Ayed. Few-Shot, now for real: Medical VLMs adaptation without balanced sets or validation. In *proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2025*, volume LNCS 15966, pp. 237–247. Springer Nature Switzerland, September 2025.

Minh-Hao Van, Prateek Verma, and Xintao Wu. On large visual language models for medical imaging analysis: An empirical study. In *2024 IEEE/ACM Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE)*, pp. 172–176, 2024.

Joost van Amersfoort, Lewis Smith, Yee Whye Teh, and Yarin Gal. Uncertainty estimation using a single deep deterministic neural network. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pp. 9690–9700. PMLR, 2020.

Bas H. M. van der Velden, Hugo J. Kuijf, Kenneth G. A. Gilhuijs, and Max A. Viergever. Explainable artificial intelligence (XAI) in deep learning-based medical image analysis. *Medical Image Analysis*, 79:102470, 2022.

Chunjiang Wang, Kun Zhang, Yandong Liu, Zhiyang He, Xiaodong Tao, and S. Kevin Zhou. MVP-CBM: Multi-layer visual preference-enhanced concept bottleneck model for explainable medical image classification. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI-25*, pp. 529–537, 2025.

Zhihong Zhu, Yunyan Zhang, Xianwei Zhuang, Fan Zhang, Zhongwei Wan, Yuyan Chen, Qingqing Long, Yefeng Zheng, and Xian Wu. Can we trust AI doctors? A survey of medical hallucination in large language and large vision-language models. In *Findings of the Association for Computational Linguistics, ACL 2025*, pp. 6748–6769. Association for Computational Linguistics, 2025.

A Related Work

Medical image diagnosis is an important task in medical artificial intelligence and attracts much attention. In the early years, the mainstream approach is to use deep neural networks (DNN) for diagnosis. For example, Gulshan et al. (2016) developed a deep learning algorithm that can match ophthalmologists in detecting diabetic retinopathy from retinal fundus photographs. Esteva et al. (2017) showed that a single DNN can classify skin cancer with accuracy comparable to board-certified dermatologists. Rajpurkar et al. (2017) proposed CheXNet, a 121-layer DenseNet that achieves radiologist-level performance in detecting pneumonia from chest X-rays. Ardila et al. (2019) developed an end-to-end three-dimensional deep learning model for lung cancer screening on low-dose chest CT, outperforming radiologists in both sensitivity and specificity. However, the black box nature of such DNN models raises interpretability concerns in medical image diagnosis. van der Velden et al. (2022) systematically reviewed the interpretability problems and progress of deep learning in medical image analysis. Rudin (2019), from the perspective of high-risk decision making, argued for the necessity of using intrinsically interpretable models rather than post hoc explanations in such scenarios.

To address the core limitation of interpretability, Koh et al. (2020) pioneered the Concept Bottleneck Model (CBM), which achieves traceable reasoning by inserting a human-understandable concept layer between the image and the diagnosis. Kim et al. (2023) extended CBM with Probabilistic CBM (ProbCBM), and Espinosa Zarlenga et al. (2022) further introduced Concept Embedding Models (CEM). Meanwhile, Margeloiu et al. (2021) scrutinized the faithfulness of concept learning in CBM. Raman et al. (2024) examined the validity of the local independence assumption. Beyond these analyses, Hou et al. (2024) proposed Concept Attention Whitening (CAW), which reduces redundancy by orthogonalizing concept features. To further capture reliability, Gao et al. (2024) proposed the Evidential Concept Embedding Model (ECM), which brings evidence quantification into the concept layer to characterize uncertainty. To bridge the domain gap between general pre-training and medical diagnosis, Chowdhury et al. (2024) proposed the Adaptive Concept Bottleneck Model (AdaCBM), which inserts an adaptive module between the vision language model and the concept layer.

The above CBMs achieve interpretable diagnosis, yet they all rely on concept classifiers trained with concept level annotations. The emergence of LLM has changed this landscape. Liu et al. (2023) proposed the Large Language and Vision Assistant (LLaVA), which extended the visual encoder of a contrastively visual-language pre-trained model (VLP) with an LLM decoder through visual instruction tuning. Van et al. (2024) demonstrated the potential of LLMs for medical image diagnosis. Patrício et al. (2025a) proposed the Concept-Based Large Vision Language Model (CBVLM), which combined concept bottlenecks with LLM to achieve training-free interpretable diagnosis. Liu et al. (2025) mined medical knowledge image-relevant descriptions and from LLMs through retrieval augmentation. Patrício et al. (2025b) designed Two-Step CBM, which adopted a two-step architecture decoupling the VLP models and LLMs to achieve training-free concept diagnosis. Wang et al. (2025) builds a concept bottleneck model on a frozen biomedical VLP backbone, where an LLM generates the concepts off-line and each concept is associated with its preferred visual layers. Koleilat et al. (2025) performs prompt learning on a frozen biomedical VLP model, where an LLM is used to synthesize prompt ensembles. Rahman et al. (2025) performed LLM-guided unsupervised adaptation on a medical VLP backbone. However, the above LLM-based medical diagnosis methods lack a targeted design for the two core problems of large model hallucination and long tail distribution bias. In this paper, we propose a novel training-free and trustworthy LLM-based medical image diagnosis framework to tackle these two problems.

B Prompt templates and examples

This section presents all prompts used in this paper and their construction methods. In the prompts, ⟨image: exemplar t⟩ indicates inserting the image of the t-th retrieved sample x_{i_t} at that position, and ⟨image: test⟩ indicates inserting the image of the test sample x_{test}.

B.1 CONCEPT MATCHING PATHWAY PROMPT

$\mathcal{P}_{\text{Concept}}$ is designed for visual concept judgment and is formed by concatenating the following three parts in order:

1. **Concept Instruction** List verbatim the definitions of the 8 chest X-ray signs (pleural effusion, pneumothorax, pneumonia, consolidation, pulmonary edema, lung opacity, cardiomegaly, enlarged cardiomediastinum) as prior knowledge injected into the model. This part is fixed text, identical for all test samples, and does not depend on retrieval results.
2. **Concept Demonstration** For the test image x_{test}, first retrieve T most similar samples $\{x_{i_t}\}_{t=1}^{T}$ from $\mathcal{D}_{\text{reference}}$; then, for each retrieved sample, apply the template "In the ⟨image: exemplar t⟩, *concept* is: A) absent B) present. Choose one option. Do not provide additional information." sequentially to the M concepts, forming M classification question-answer pairs, where the Answer is determined by the concept label c_{i_t} of that sample. This part dynamically changes with retrieval results.
3. **Concept Query** Insert the test image x_{test} at the ⟨image: test⟩ position, and pose the same classification question for each of the above M concepts once, with the Answer left blank to be completed by the model. This part has the same structure for all test samples, differing only in the test image.

FULL PROMPT TEXT

This prompt is concatenated from three parts: [Concept Instruction], [Concept Demonstration], and [Concept Query]. In the demonstration section, each retrieved sample x_{i_t} answers the M concept questions in turn; below, taking the first retrieved sample (exemplar 1) as an example, the M question-answer pairs are fully presented. Before each question-answer pair, the image of that retrieved sample is inserted (indicated by ⟨image: exemplar 1⟩); exemplar 2 to exemplar T have exactly the same structure, only the image and the corresponding Answer are replaced.

```
[Concept Instruction]
1.A pleural effusion corresponds to the accumulation of fluid in the
pleural space between the lung and chest wall, appearing as a blunting
of the costophrenic angles or a meniscus sign on radiographs.
2.A pneumothorax corresponds to the presence of air in the pleural
space, leading to lung collapse, often visible as a dark area on the
radiograph where the lung is not present.
3.Pneumonia is an infection of the lung parenchyma causing
inflammation and consolidation, which appears as areas of opacity on
radiographs.
4.Consolidation corresponds to an area of the lung where normal
air-filled spaces have been replaced with fluid, cells, or other
material, often seen in pneumonia or other infections.
5.Pulmonary edema is the accumulation of fluid in the lung
interstitium or alveoli, causing increased opacity on imaging, often
indicative of conditions such as heart failure or acute respiratory
distress syndrome.
6.Lung opacity corresponds to an area on the radiograph where the
lung appears more opaque than normal, potentially due to fluid,
consolidation, or masses within the lung tissue.
7.Cardiomegaly is an abnormal enlargement of the heart, often
indicating underlying conditions such as heart failure, hypertension,
or cardiomyopathy.
8.An enlarged cardiomediastinum is an increase in the width of the
mediastinum, which may suggest conditions like aortic aneurysm,
lymphadenopathy, or heart failure.
[Concept Demonstration]
Consider the following examples:
1.In the ⟨image: exemplar 1⟩, pleural effusion is: A) absent B)
present. Choose one option. Do not provide additional information.
Answer: A) absent.
```

```
2.In the ⟨image: exemplar 1⟩, pneumothorax is: A) absent B) present.
Choose one option. Do not provide additional information.
Answer: B) present.
3.In the ⟨image: exemplar 1⟩, pneumonia is: A) absent B) present.
Choose one option. Do not provide additional information.
Answer: B) present.
4.In the ⟨image: exemplar 1⟩, consolidation is: A) absent B) present.
Choose one option. Do not provide additional information.
Answer: A) absent.
5.In the ⟨image: exemplar 1⟩, edema is: A) absent B) present. Choose
one option. Do not provide additional information.
Answer: B) present.
6.In the ⟨image: exemplar 1⟩, lung opacity is: A) absent B) present.
Choose one option. Do not provide additional information.
Answer: A) absent.
7.In the ⟨image: exemplar 1⟩, cardiomegaly is: A) absent B) present.
Choose one option. Do not provide additional information.
Answer: A) absent.
8.In the ⟨image: exemplar 1⟩, enlarged cardiomediastinum is: A)
absent B) present. Choose one option. Do not provide additional
information.
Answer: B) present.
[... For exemplar 2 to T, repeat the above 8 concept question-answer
pairs; each time only replace the image and the corresponding answer
c_{i_t} ...]
[Concept Query]
1.Refer to the concept information of examples, in the ⟨image: test⟩,
pleural effusion is: A) absent B) present. Choose one option. Do
not provide additional information. Answer:
2.Refer to the concept information of examples, in the ⟨image: test⟩,
pneumothorax is: A) absent B) present. Choose one option. Do not
provide additional information. Answer:
3.Refer to the concept information of examples, in the ⟨image: test⟩,
pneumonia is: A) absent B) present. Choose one option. Do not
provide additional information. Answer:
4.Refer to the concept information of examples, in the ⟨image: test⟩,
consolidation is: A) absent B) present. Choose one option. Do not
provide additional information. Answer:
5.Refer to the concept information of examples, in the ⟨image: test⟩,
edema is: A) absent B) present. Choose one option. Do not provide
additional information. Answer:
6.Refer to the concept information of examples, in the ⟨image: test⟩,
lung opacity is: A) absent B) present. Choose one option. Do not
provide additional information. Answer:
7.Refer to the concept information of examples, in the ⟨image: test⟩,
cardiomegaly is: A) absent B) present. Choose one option. Do not
provide additional information. Answer:
8.Refer to the concept information of examples, in the ⟨image: test⟩,
enlarged cardiomediastinum is: A) absent B) present. Choose one
option. Do not provide additional information. Answer:
```

The query section ([Concept Query]) takes the test image x_{test} as input and asks about the 8 concepts one by one, with Answer left blank to be completed by the model.

OUTPUT FORMAT

For each concept query, the model returns a dictionary, with keys as option indices and values as the logits scores for that option. The option with the highest score is taken as the prediction for that concept. Example:

```
"0": 45.5625, "1": 36.1875
```

Here, index "0" corresponds to absent, and index "1" corresponds to present; in this example, the score for absent (45.56) is higher than that for present (36.19), so the concept is predicted as absent.

B.2 Visual perception pathway prompt

Structural Description

$\mathcal{P}_{\text{Diagnosis}}$ is designed for final disease category judgment. It shares the same set of retrieved samples $\{x_{i_t}\}_{t=1}^{T}$ with $\mathcal{P}_{\text{Concept}}$, but differs in that it uses the class labels of the retrieved samples instead of concept annotations, and the query target changes from concept classification to diagnosis category classification. It is also concatenated from three parts in order:

1. **[Class Instruction]** Describe in words the radiographic appearance features of the two classes (Covid-19 and no Covid-19) on X-ray, and state that this is a classification task. This part is fixed text and does not depend on retrieval results.
2. **[Class Demonstration]** For each retrieved sample x_{i_t}, construct a question-answer pair "What is the diagnosis shown in this image?" using its class label y_{i_t}, with the Answer determined by y_{i_t}. This part dynamically changes with retrieval results.
3. **[Class Query]** Take the test image x_{test} as input, pose a diagnosis question for the test image once, with Answer left blank to be completed by the model. This part has the same structure for all test samples, differing only in the test image.

Full Prompt Text

This prompt is concatenated from three parts: [Class Instruction], [Class Demonstration], and [Class Query]. In the demonstration section, the first and second retrieved samples show complete question-answer pairs for the two class labels no Covid-19 and Covid-19 respectively (with ⟨image: exemplar 1⟩ and ⟨image: exemplar 2⟩ indicating the corresponding images); exemplar 3 to exemplar T have the same structure, only the image and the corresponding Answer are replaced.

```
[Class Instruction]
Chest X-ray images of patients with Covid-19 usually show bilateral,
peripheral ground-glass opacities and patchy consolidation, which
often predominate in the lower lung zones; chest X-ray images of
patients without Covid-19 appear normal or show only non-Covid-19
related abnormalities and lack the characteristic bilateral
viral-pneumonia pattern described above.  Determine whether the image
shows Covid-19 or not.  Options:  A) no Covid-19, B) Covid-19.
[Class Demonstration]
Consider the following examples:
What is the diagnosis shown in this ⟨image:  exemplar 1⟩?  Options:
A) no Covid-19 B) Covid-19.  Choose one option.  Do not provide
additional information.
Answer:  A) no Covid-19.
What is the diagnosis shown in this ⟨image:  exemplar 2⟩?  Options:
A) no Covid-19 B) Covid-19.  Choose one option.  Do not provide
additional information.
Answer:  B) Covid-19.
[...  For exemplar 3 to T, repeat the above diagnosis question-answer
pairs; each time only replace the image and the corresponding label y_{i_t}
...]
[Class Query]
What is the diagnosis shown in this ⟨image:  test⟩?  Options:  A) no
Covid-19 B) Covid-19.  Choose one option.  Do not provide additional
information.
Answer:
```

The query section ([Class Query]) takes the test image x_{test} as input (indicated by ⟨image: test⟩), with Answer left blank to be completed by the model.

OUTPUT FORMAT

The model returns a dictionary, with keys as option indices (strings) and values as the logits scores for that option. The option with the highest score is taken as the final diagnosis. Example:

```
"0":  33.5819, "1":  62.9715
```

Here, key "0" corresponds to no Covid-19, and key "1" corresponds to Covid-19; in this example, the score for no Covid-19 (62.97) is higher than that for Covid-19 (33.58), so the diagnosis is no Covid-19.

C DERIVATION OF EQUATION 2

Before the derivation, we introduce the following three assumptions:

Assumption 1. *Since in this path we only predict the labels based on the concepts, we assume that these concepts constitute sufficient statistics for the diagnostic category of the image, i.e.,* $P(y \mid c_1, \ldots, c_M, x) = P(y \mid c_1, \ldots, c_M)$.

Assumption 2. *All concepts are marginally independent within the overall framework, and they are also conditionally independent given the disease category y, i.e.,* $P(c_1, \ldots, c_M) = \prod_{i=1}^{M} P(c_i)$ *and* $P(c_1, \ldots, c_M \mid y) = \prod_{i=1}^{M} P(c_i \mid y)$.

Assumption 3. *Under the condition of the given test image* x_{test}*, the model's output for all concepts is conditionally independent, i.e.,* $P(c_1, \ldots, c_M \mid x_{\text{test}}) = \prod_{i=1}^{M} P(c_i \mid x_{\text{test}})$.

Then, applying Assumption 1 at the third equation in the following formula, we have:

$$
\begin{aligned}
& P(y_{\text{test}} \mid x_{\text{test}}) \qquad (13)\\
= & \sum_{c_1(x_{\text{test}}) \in \mathcal{O}^{(1)}} \sum_{c_2(x_{\text{test}}) \in \mathcal{O}^{(2)}} \cdots \sum_{c_M(x_{\text{test}}) \in \mathcal{O}^{(M)}} P(y_{\text{test}}, c_1, \ldots, c_M \,|\, x_{\text{test}}) \\
= & \sum_{c_1(x_{\text{test}}) \in \mathcal{O}^{(1)}} \sum_{c_2(x_{\text{test}}) \in \mathcal{O}^{(2)}} \cdots \sum_{c_M(x_{\text{test}}) \in \mathcal{O}^{(M)}} P(y_{\text{test}} \,|\, c_1, \ldots, c_M, x_{\text{test}}) \cdot P(c_1, \ldots, c_M \,|\, x_{\text{test}}) \\
= & \sum_{c_1(x_{\text{test}}) \in \mathcal{O}^{(1)}} \sum_{c_2(x_{\text{test}}) \in \mathcal{O}^{(2)}} \cdots \sum_{c_M(x_{\text{test}}) \in \mathcal{O}^{(M)}} P(y_{\text{test}} \,|\, c_1, \ldots, c_M) \cdot P(c_1, \ldots, c_M \,|\, x_{\text{test}}) \\
= & \sum_{c_1(x_{\text{test}}) \in \mathcal{O}^{(1)}} \sum_{c_2(x_{\text{test}}) \in \mathcal{O}^{(2)}} \cdots \sum_{c_M(x_{\text{test}}) \in \mathcal{O}^{(M)}} \frac{P(y_{\text{test}}) P(c_1, \ldots, c_M \,|\, y_{\text{test}})}{P(c_1, \ldots, c_M)} P(c_1, \ldots, c_M \,|\, x_{\text{test}})
\end{aligned}
$$

According to Assumptions 2 and 3, we have $P(c_1, \ldots, c_M) = \prod_{i=1}^{M} P(c_i)$, $P(c_1, \ldots, c_M \mid y_{\text{test}}) = \prod_{i=1}^{M} P(c_i \mid y_{\text{test}})$, and $P(c_1, \ldots, c_M \mid x_{\text{test}}) = \prod_{i=1}^{M} P(c_i \mid x_{\text{test}})$. Taking them back into

Equation 13, we have:

$$
\begin{aligned}
& P(y_{\text{test}} \mid x_{\text{test}}) \qquad (14)\\
=& \sum_{c_1(x_{\text{test}})\in\mathcal{O}^{(1)}} \sum_{c_2(x_{\text{test}})\in\mathcal{O}^{(2)}} \cdots \sum_{c_M(x_{\text{test}})\in\mathcal{O}^{(M)}} \frac{P(y_{\text{test}})P(c_1,\ldots,c_M\,|\,y_{\text{test}})}{P(c_1,\ldots,c_M)} P(c_1,\ldots,c_M\,|\,x_{\text{test}})\\
=& \sum_{c_1(x_{\text{test}})\in\mathcal{O}^{(1)}} \sum_{c_2(x_{\text{test}})\in\mathcal{O}^{(2)}} \cdots \sum_{c_M(x_{\text{test}})\in\mathcal{O}^{(M)}} \frac{P(y_{\text{test}})\prod_{i=1}^{M} P(c_i\,|\,y_{\text{test}})}{\prod_{i=1}^{M} P(c_i)} \prod_{i=1}^{M} P(c_i\,|\,x_{\text{test}})\\
=& P(y_{\text{test}}) \sum_{c_1(x_{\text{test}})\in\mathcal{O}^{(1)}} \sum_{c_2(x_{\text{test}})\in\mathcal{O}^{(2)}} \cdots \sum_{c_M(x_{\text{test}})\in\mathcal{O}^{(M)}} \prod_{i=1}^{M} \frac{P(c_i\,|\,x_{\text{test}})\,P(c_i \mid y_{\text{test}})}{P(c_i)}\\
=& P(y_{\text{test}}) \prod_{i=1}^{M} \sum_{c_i(x_{\text{test}})\in\mathcal{O}^{(i)}} \frac{P(c_i\,|\,x_{\text{test}})\,P(c_i\,|\,y_{\text{test}})}{P(c_i)}\\
=& P(y_{\text{test}}) \prod_{i=1}^{M} \sum_{j=1}^{r_i} \frac{P(c_i = o_{i,j}\,|\,x_{\text{test}})\,P(c_i = o_{i,j}\,|\,y_{\text{test}})}{P(c_i = o_{i,j})}.
\end{aligned}
$$

This concludes the derivation.

D Derivation of Equation 7

The core of the Dirichlet framework lies in the construction of the concentration parameters $\boldsymbol{\alpha}$. The relative proportions among the α_k determine the allocation bias of evidence across classes, and the total concentration $\sum_k \alpha_k$ determines the overall amount of evidence. Therefore, the evidence magnitude must be calibrated in both absolute scale and relative proportion to correctly characterize the model's uncertainty.

In Han et al. (2023), the network directly predicts the evidence mass $e_k = \alpha_k - 1$. However, in our framework, the LLMs are not designed to predict the evidence mass, and we do not retrain the LLMs. To this end, we have to design a new way to calculate the evidence mass. More specifically, the LLMs are optimized with cross-entropy objective. The Cross-entropy quantifies the discrepancy between the true label q and the model output p. Here q is the one-hot distribution of the true label, and p is the predicted probability distribution obtained by applying the Softmax transformation to the model output logits z_k, i.e., $p_k = \frac{\exp(z_k)}{\sum_{j=1}^{K}\exp(z_j)}$. Given a data x, suppose its ground-truth label is class c, which means $q_c = 1$ and $q_i = 0$ for any $i \neq c$. The cross entropy loss can be written as:

$$
L_{\text{CE}} = -\sum_{k=1}^{K} q_k \log p_k = -\log \frac{\exp(z_c)}{\sum_{k=1}^{K}\exp(z_k)}. \qquad (15)
$$

The gradient with respect to the logit z_c is:

$$
\frac{\partial L_{\text{CE}}}{\partial z_c} = -1 + \frac{e^{z_c}}{\sum_j e^{z_j}} = \frac{1}{1 + \sum_{k\neq c} e^{z_k - z_c}} - 1 \qquad (16)
$$

For the ground-truth class c, the loss is determined solely by the relative differences among logits $z_k - z_c$. The optimization objective of cross-entropy training is always to widen the logit gap between the ground-truth class and all other classes, imposing no constraint on the absolute scale of the logits. Therefore, the difference between the maximum logit and the mean of the remaining logits is a natural scalar that characterizes the discriminability:

$$
\Delta = z_{\max} - \frac{1}{K-1} \sum_{z_k \neq z_{\max}} z_k. \qquad (17)
$$

where $z_{\max}$ is the maximum logit. Δ depends only on the relative differences among the logits. A larger Δ indicates a stronger ability of the model to distinguish the optimal class from the remaining classes. That is the reason why we define Δ as Equation 7.

Algorithm 1 The algorithm of three-stage framework.

Input: x_{test}, $\mathcal{D}_{\text{train}}$, $\{c_i(x_n)\}_{n=1}^{N}$, T, τ_1, τ_2
Output: y_{pred}, u_{Fused}
#Stage 1:
Given the test image x_{test}, select the top-T similar samples in training set $\mathcal{D}_{\text{train}}$ to form the reference set $\mathcal{D}_{\text{reference}}$.
Generate the concept prompt $\mathcal{P}_{\text{Concept}}$ as introduced in Appendix B.1.
Generate the diagnosis prompt $\mathcal{P}_{\text{Diagnosis}}$ as introduced in Appendix B.2.
#Stage 2:
Obtain the output $z_i^{(\text{Concept})}$ of LLM in the concept matching pathway.
Obtain $P(c_i(x_{\text{test}}) = o_{i,j} \mid x_{\text{test}})$ by Equation 1.
Obtain $P_{\text{Concept}}(y_{\text{test}} \mid x_{\text{test}})$ by Equation 2.
Obtain the debiased probability $\hat{P}_{\text{Concept}}(y_{\text{test}} \mid x_{\text{test}})$ by Equation 5.
Obtain the output $z^{(\text{Visual})}$ of LLM in the visual perception pathway.
Obtain the debiased probability $\hat{P}_{\text{Visual}}(y_{\text{test}} \mid x_{\text{test}})$ by Equation 6.
#Stage 3:
Obtain Δ^{Visual} by Equation 7.
Obtain the Dirichlet parameter $\alpha_k^{\text{Visual}} = 1 + \Delta^{\text{Visual}} \cdot \hat{P}_{\text{Visual}}(y_{\text{test}} = k \mid x_{\text{test}})$.
Obtain $\Delta^{(i)}$ for each concept by Equation 7.
Obtain the Dirichlet parameter $\alpha_k^{\text{Concept}}$ by Equation 9.
Apply the Dempster-Shafer evidence theory to obtain the fused Dirichlet parameter α_k^{Fused} by Equation 12 and the uncertainty u_{Fused} by Equation 11.
Obtain the final diagnostic distribution $P_{\text{Fused}}(y_{\text{test}} = k \mid x_{\text{test}})$ by Equation 12.
Obtain the final prediction $y_{\text{pred}} = \arg\max_k P_{\text{Fused}}(y_{\text{test}} = k | x_{\text{test}})$.

E ALGORITHM

Algorithm 1 summarizes the complete procedure of the proposed framework. All prior distributions are approximated by the empirical distribution obtained from the training set.

F DETAILS OF THE DATASET

We conduct experiments on three widely-used medical image classification benchmarks, covering three modalities: chest X-rays, fundus color photographs, and dermoscopy images. Each dataset provides clinical medical concepts labeled by experts. As shown in Table 4, we use the default dataset split for all datasets.

Derm7pt (Kawahara et al., 2019) is a dermoscopy dataset, labeled with 5 fine-grained clinical dermatoscopic concepts: (1) pigmentation networks with values "absent", "typical", and "atypical"; (2) streaks with values "absent", "regular", and "irregular"; (3) point-like and spherical structures with values "absent", "regular", and "irregular"; (4) blue-white curtain with values "absent" and "exists"; (5) degenerative structures with values "absent" and "exists". The dataset has two classes: Melanoma and benign Nevus.

CORDA (Alesina et al., 2023) is a chest X-ray dataset, labeling 8 chest disease concepts, called Mediastinal widening, enlarged heart, lung opacity, pulmonary edema, pulmonary consolidation, pneumonia, pneumothorax, and pleural effusion, each with a value of "presence" and "absence". The dataset has two classes: Covid-19 and no Covid-19.

DDR (Li et al., 2019) is a diabetic retinopathy grading dataset, providing 4 fundus lesion concepts, called exudates, hemorrhage, microaneurysms, and soft exudates, each with values "presence" and "absence". The dataset has five classes: no DR, mild NPDR, moderate NPDR, severe NPDR, and proliferative DR.

Table 4: Statistics of the datasets used in our experiments.

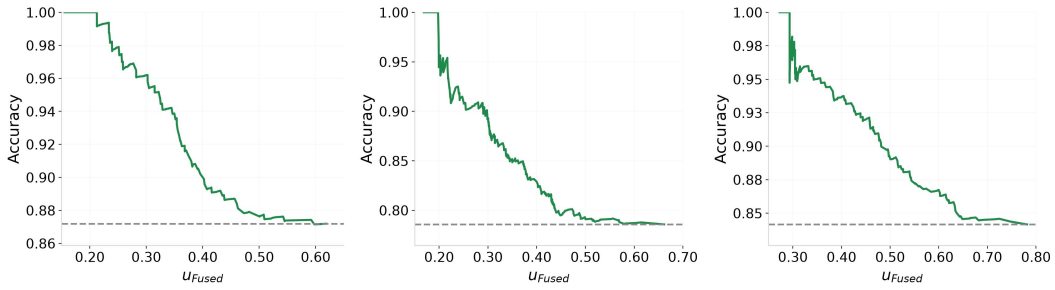
Dataset	Train	Valid	Test	Concepts	Classes	Modalities
Derm7pt	346	161	320	5	2	Dermoscopy
CORDA	967	242	392	8	2	Chest X-ray
DDR	817	279	404	4	5	Fundus color photography

G Uncertainty Analysis

To further analyze the uncertainty u_{Fused}, we sort all test samples in ascending order by u_{Fused}. Then, we predict each test sample in this order and report the prediction accuracy, i.e., more and more uncertainty samples are involved in test. The results are shown in Figure 3. The X-axis indicates the value of the uncertainty threshold u_{Fused}, and the Y-axis indicates the accuracy. The dashed line indicates the accuracy of the whole test set. We can see that, with more and more uncertain samples being tested, the accuracy is decreasing.

This phenomenon indicates that u_{Fused} can effectively characterize the reliability of predictions. Samples with low uncertainty are more likely to be predicted correctly, and samples with high uncertainty are more likely to be predicted incorrectly. Our u_{Fused} is aligned with the accuracy of the prediction. Therefore, in resource-constrained or safety-sensitive clinical scenarios, u_{Fused} can be used as a rejection threshold, only making automatic decisions for low-uncertainty samples, and handing over high-uncertainty samples to clinicians for review, thereby improving the reliability of the diagnosis.

Besides, we further show some examples of individual predictions in Figure 4. The left column shows the examples that are predicted correctly and the right column shows the examples that are predicted incorrectly. "trad-Confidence" indicates the confidence used in conventional methods, and "our-Uncertainty" indicates our uncertainty u_{Fused}. We can see that, the conventional model produces near-identical confidence for correct predictions and incorrect predictions, leading to the hallucination in diagnosis. However, our method often assigns a high uncertainty to those incorrect predictions and a low uncertainty to the correct predictions. Therefore, we can apply u_{Fused} to detect the hallucination and achieve the trustworthy diagnosis.



(a) Derm7pt (b) CORDA (c) DDR

Figure 3: Accuracy curve based on u_{Fused}. The X-axis indicates the value of the uncertainty threshold u_{Fused}, and the Y-axis indicates the accuracy. The dashed line indicates the accuracy of the whole test set.

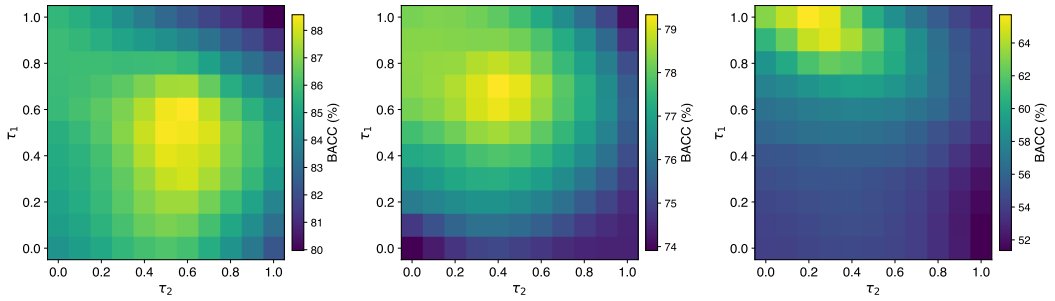
H Hyperparameter Sensitivity Analysis

We conducted the hyperparameter sensitivity analysis on the bias reduction strength τ_1 of the concept matching pathway and the bias reduction strength τ_2 of the visual perception pathway. The results are shown in Figure 5. We can see that the performance is relatively stable in a wide range. Moreover, τ_1 often needs to be larger than τ_2. It shows that the long-tailed distribution issue is more



Figure 4: Qualitative examples of correct (left) and erroneous (right) predictions from Derm7pt, CORDA and DDR (three rows each). For each sample we report its ground-truth label (GT), its predicted label (Pred), the confidence of the prediction (trad-Confidence), and our fused uncertainty (our-Uncertainty). This demonstrates that u_{Fused} can identify unreliable predictions that the confidence alone fails to reveal.

severe in the concept matching pathway than in the visual perception pathway. Therefore, we need a larger τ_1 to debias in the concept matching pathway.



(a) Derm7pt (b) CORDA (c) DDR

Figure 5: Sensitivity of the two debiasing strengths (τ_1, τ_2) on the three datasets. The color of each cell represents the test set BACC (%) value for each combination of hyperparameters.