PRISM2: Unlocking Multi-Modal General Pathology AI with Clinical Dialogue

George Shaikovski[*1], Eugene Vorontsov[*‡1], Adam Casson[*1], Julian Viret[*1], Eric Zimmermann[*2], Neil Tenenholtz[2], Yi Kan Wang[1], Jan H. Bernhard[1], Ran A. Godrich[1], Juan A. Retamero[1], Razik Yousfi[1], Nicolò Fusi[2], Thomas J. Fuchs[1], Kristen Severson[†2], and Siqi Liu[†1]

[1]Paige, NYC, NY United States
[2]Microsoft Research, Cambridge, MA United States

ABSTRACT

Recent pathology foundation models can provide rich tile-level representations but fall short of delivering general-purpose clinical utility without further extensive model development. These models lack whole-slide image (WSI) understanding and are not trained with large-scale diagnostic data, limiting their performance on diverse downstream tasks. We introduce PRISM2, a multi-modal slide-level foundation model trained via clinical dialogue to enable scalable, generalizable pathology AI. PRISM2 is trained on nearly 700,000 specimens (2.3 million WSIs) paired with real-world clinical diagnostic reports in a two-stage process. In Stage 1, a vision-language model is trained using contrastive and captioning objectives to align whole slide embeddings with textual clinical diagnosis. In Stage 2, the language model is unfrozen to enable diagnostic conversation and extract more clinically meaningful representations from hidden states. PRISM2 achieves strong performance on diagnostic and biomarker prediction tasks, outperforming prior slide-level models including PRISM and TITAN. It also introduces a zero-shot yes/no classification approach that surpasses CLIP-style methods without prompt tuning or class enumeration. By aligning visual features with clinical reasoning, PRISM2 improves generalization on both data-rich and low-sample tasks, offering a scalable path forward for building general pathology AI agents capable of assisting diagnostic and prognostic decisions.

1 Introduction

The field of computational pathology has been transformed by the introduction of foundation models. Models such as Virchow2 [44], UNI2 [11] and H-optimus-1 [9] are trained using millions of histopathology tiles spanning diverse tissue types, employing self-supervised objectives [32] to learn generalizable representations. The introduction of foundation models into the computational pathology modeling pipeline has improved performance, robustness, and data efficiency in a variety of tasks such as cancer detection, subtyping, and biomarker quantification [40, 18, 11, 39].

However, using these tile-level models at the level of WSI prediction still requires training an additional network to aggregate the tile embeddings across tens of thousands of tiles, which can overfit when training data is limited. A complete pathology foundation model must therefore operate at the level of WSIs. Early slide-level foundation models in pathology explored unsupervised training via masked autoencoding with dilated attention and weak supervision from clinical free text reports via contrastive alignment to the WSI representation, yielding improved slide-level rep-

---

[‡]Corresponding author. eugene.vorontsov AT paige DOT ai

[*]These authors contributed equally to this work.
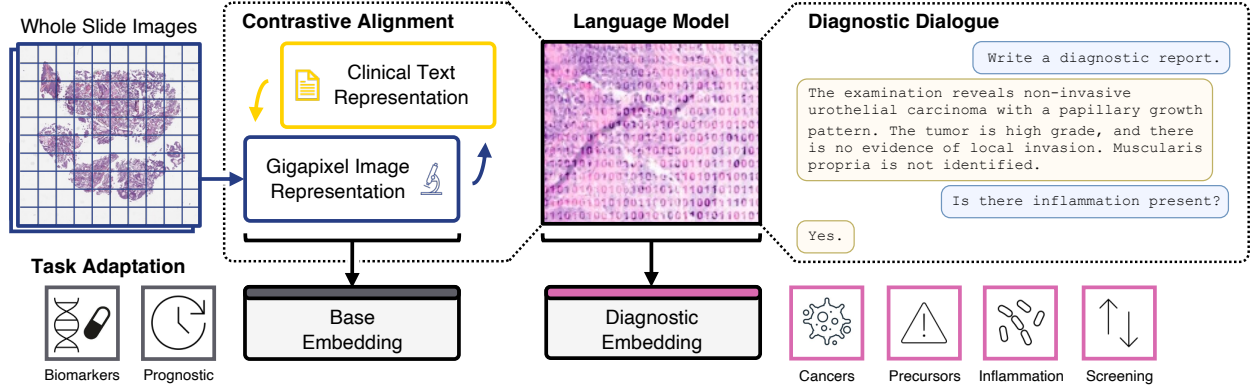
[†]Equal co-supervision.

Figure 1: PRISM2 uses language data derived from free-text clinical reports to align a visual representation combining multiple gigapixel whole slide images (WSIs) with concepts relevant to pathology. For downstream task adaptation, whereas PRISM only produces an embedding vector of this aligned representation (here: 'base'), PRISM2 uses a large language model to further refine it toward a diagnostic-tuned embedding (here: 'diagnostic'). The language model thus serves as a diagnostic reasoning block that is also capable of generating diagnostic summary text and answering diagnostic questions. Questions with yes/no answers provide a simple and robust method of zero-shot prediction, though PRISM2 can also handle open-ended questions.

resentations [41, 17, 34]. The clinical report is a natural choice for learning a slide-level embedding as it contains a variety of clinically-relevant concepts which provide a diverse signal to prevent overfitting. While this direction is promising, these works remain constrained by limited dataset and model scale and are well positioned to benefit from recent progress in language modeling.

In this work, we present PRISM2 (Fig. 1), a slide-level foundation model that refines clinical-report-based supervision through dialogue, using a 4 billion parameter large language model (LLM). Beyond broadly aligning the WSI representation to a clinical report, we use question answering to provide fine-grained diagnostic supervision. PRISM2 is the largest multi-modal slide-level pathology foundation model to date, trained on the largest corpus of data containing 2.3 million WSIs from diverse institutions, corresponding to 685 thousand distinct clinical reports.

PRISM2 is trained using a two-stage approach. In the first stage, a perceiver-based [23] image encoder coupled with a pretrained language encoder, BioGPT [30], is trained with a contrastive objective. In the second stage, these embeddings are used as input to a Phi-3 Mini [1] language backbone to generate reports and answer questions, permitting flexible decoding into natural language. This design decomposes the learning problem, first capturing image representations and then fine-tuning the LLM to interpret them in a conversational framework.

PRISM2 is evaluated on a suite of diagnostic classification and biomarker prediction tasks. The choice of architecture and two-stage training approach yield two types of slide-level embeddings: the *base* embedding learned using the first stage of contrastive training and the *diagnostic* embedding based on the language model's hidden state learned in the second stage. Zero-shot evaluation is performed using question answering, as well as contrastive similarity scoring. Overall, we show improved performance by using the language capabilities, while maintaining state-of-the-art performance in applications where language-based definitions are less well-defined, e.g. in the case of biomarkers. We also show the value of further finetuning the diagnostic embedding. Taken together, these results demonstrate the value of large-scale, multi-modal training.

The utility of PRISM2 is also investigated in two qualitative settings. First, a region of interest analysis is presented to investigate the accuracy and relevance of the highly attended regions of the WSI using expert clinical review. Second, we present examples of chat dialogue demonstrating the conversational capabilities of the model.

2   Related Work

2.1   Whole-slide multi-modal representation learning

Tile-level computational pathology models trained with paired text have previously been proposed [22, 28], however, finding sufficient text to be paired with image crops is challenging. This is because clinical diagnostic workflows in practice are performed on one or more slides and lack fine-grained pairing to specific regions. Prov-Gigapath [41]

paired a whole-slide-level representation with clinical reports by first training a slide-level representation using a masked autoencoding objective on tile embeddings and then using a contrastive vision-language alignment loss with slide-level reports. Although this approach enabled the use of pathology reports, creating a sufficiently generalizable slide-level embedding in the first stage is challenging. PRISM [34] was the first work that addressed multi-modal pretraining on the specimen level, where each report was aligned with multiple slides to produce a single multi-modal embedding. The network was trained end-to-end using specimen-level reports and Virchow [39] tile embeddings aggregated via a perceiver-based model [23] in a single step training approach based on CoCa [43]. Subsequently, TITAN [17] has also introduced an approach using diagnostic reports. TITAN, building on CONCH embeddings [28], uses three separate stages to train a WSI-level multi-modal model. The first stage is used to pretrain a region-level model using only embeddings and a self-supervised objective. In the second and third stages, the model is finetuned on regions with synthetic captions generated by PathChat [29], followed by diagnostic reports, with each report paired to a single slide. Training on a single slide per report reduces computational complexity of the task but limits the applicability of the method to specimen-level tasks. PathAlign [2] adopted the BLIP-2 [25] framework to pretrain a specimen-level aggregator by aligning its specimen-level representation to short strings capturing the most significant diagnostic findings.

Various works have also considered the use of non-text, paired modalities. TANGLE [24] aligns slide-level representations to corresponding genomic profile representations using contrastive alignment. THREADS [38] expands to three modalities by adding transcriptomic profiles. mSTAR [42] uses both reports and genomic profiles for three-way contrastive alignment with a shared slide representation.

2.2 Whole-slide report generation

Approaches that aim to perform clinical report generation are sometimes, but not always, overlapping with approaches that aim to learn slide-level multi-modal representations. Notably, methods based on the CoCa [43] framework are trained to produce a joint multi-modal representation and reconstruct the input caption at the same time, acquiring caption generation capabilities as a result. Both PRISM [34] and TITAN [17], described above, use CoCa and so have report generation capabilities. While aligning a diagnostic report with WSIs produces a representation with diagnostically relevant features, exploring less complete captions like specific question-answer pairs may be trickier within the CoCa framework, as the reduced information content in these captions may also reduce the information content of the aligned image representations.

Some works use a separate training stage to enable a model to generate diagnostic findings given one or more slides. HistoGPT [37] follows the Flamingo [5] framework to align a Perceiver Resampler [23] with a frozen BioGPT [30] language model in its second training stage; the first stage is used to pretrain Perceiver Resampler on a classification task, likely to mitigate overfitting on the small training set used (15 thousand dermatology WSIs). PathAlign [2], following the BLIP-2 [25] framework, finetunes the specimen-level aggregator with the frozen PaLM-2 S [6] language model to generate short strings with diagnostic findings. The model is trained on 350 thousand WSIs. PolyPath [3] finetunes a Gemini 1.5 Flash [36] language model on the same dataset in a single stage using a LLaVA-style [26] framework. Both PathAlign and PolyPath use pre-computed tile embeddings extracted at $10\times$ magnification. SlideChat [12] trains on WSI and text data with CONCH [28] as a tile encoder, LongNet [16] as a slide encoder, and Qwen2.5-7B Instruct [33] as an LLM in two stages, the first to do domain alignment, and the second to do instruction tuning. CPath-Omni [35] trains a tile encoder in a CLIP-style training stage using Virchow2 and CLIP models. They then use a multi-resolution tiling scheme along with a latent query attention pooler as a slide encoder and Qwen2.5-14B Instruct [33] as an LLM to train a unified mutli-modal patch and WSI-level model. ALPaCA [19] also does a multi-resolution tile encoding scheme using CONCH, and uses a hybrid LongFormer [8] and a Gaussian Mixture Model adapter to connect the WSI modality to Llama 3.1 8B [21].

2.3 Whole-slide question answering

Several of these works explicitly focus on question answering capabilities as well. Beyond their captioning data, SlideChat trains on 176 thousand visual question answering pairs for 4,915 WSIs which is comprised of a mix of open-ended and multiple-choice question/answer (QA) pairs. CPath-Omni also trains on patch-level and WSI-level question answering data, for the latter they use 16 thousand QA pairs for 7,312 WSIs which are also a mix of open-ended and multiple-choice questions. Additionally, ALPaCA's QA training data consists of 341 thousand QA pairs for 35,913 WSIs made up of true/false and multiple choice questions. A common theme in all these works is that each leveraged a general purpose LLM to generate QA pairs based on WSI reports/descriptions.
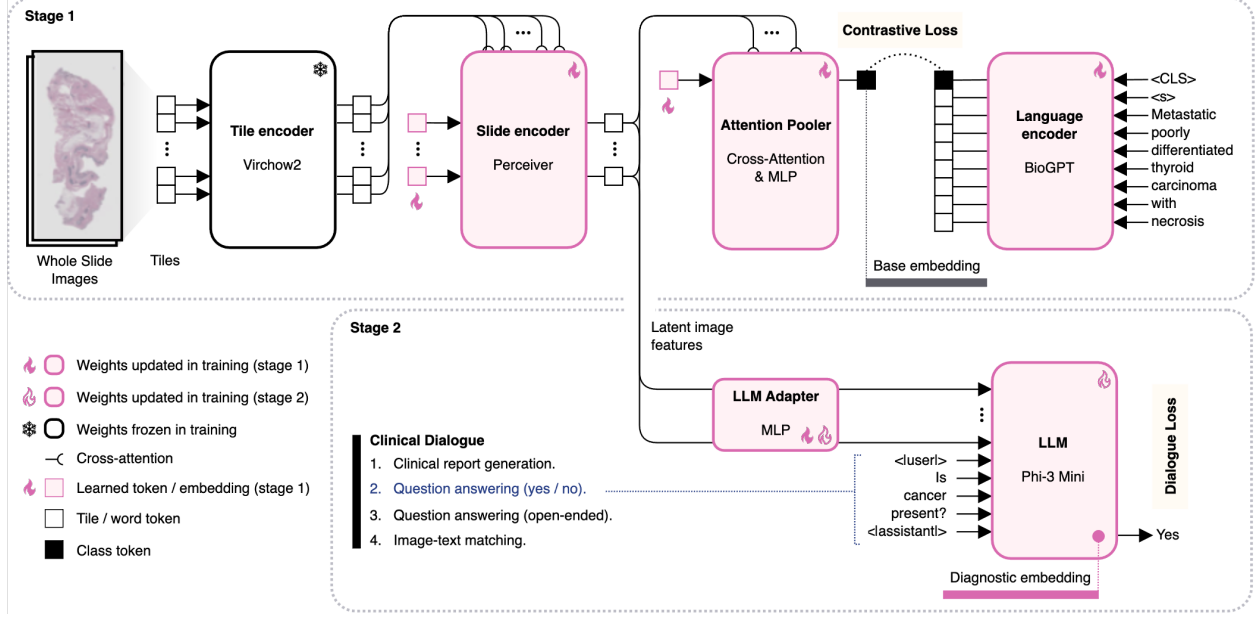
Figure 2: The architecture and training schematic for PRISM2. Sets of whole slide images are summarized as a base embedding (Attention Pooler class token) which is aligned to the representation of the clinical report diagnostic summary (encoded by the language decoder, BioGPT) by minimizing a contrastive loss. The image latent features are also passed to a pretrained large language model (LLM, Phi-3 Mini) along with a prompt from one of four dialogue templates, with the model's response updated by minimizing the dialogue loss. Training proceeds in two stages: in stage 1, only the slide encoder, attention pooler, language decoder, and LLM adapter weights are updated; in stage 2, only the LLM adapter and LLM weights are updated. For use in downstream tasks, the latent image features produced by the slide encoder are summarized in two embeddings: (1) the features are pooled into a 'base embedding', used for contrastive alignment; and (2) the features are refined by the LLM into a 'diagnostic embedding', taken from the LLM hidden state at the `<|assistant|>` token, after inputting the image.

3 PRISM2

We present PRISM2, a multi-modal slide-level pathology foundation model designed to enable comprehensive analysis of WSIs through multiple interaction paradigms. Specifically, we sought to train a model with four key capabilities:

1. Robust slide-level representations,
2. Zero-shot question answering,
3. Interactive diagnostic dialogue, and
4. Automatic report generation.

These capabilities informed our data curation strategy, architectural choices, and training methodology, which we detail in the following sections. By targeting these diverse capabilities within a unified framework, we aimed to create a versatile foundation model that addresses multiple clinical workflows in digital pathology.

3.1 Model architecture

The overall architecture of PRISM2 consists of a slide encoder, a language encoder, and a language model, as illustrated in Fig. 2 and described below.

Slide encoder. We adopt the Perceiver [23] architecture for the slide encoder as in PRISM, but with some modifications. Namely, we reduce the number of blocks from 8 to just 1, where a block consists of a cross attention layer between latent queries and tile embeddings followed by 6 self-attention transformer layers for the latent queries. We use Grouped Query Attention (GQA) [4] in the tile embedding cross attention. We decrease the number of latents from 512 to 256 and use an attention pooling layer on these to produce the embedding used in the contrastive objective. This

embedding is referred to as the *base* embedding. The Perceiver and attention pooler have 541M and 79M parameters, respectively.

Language encoder. We use BioGPT [30] as the language encoder but unlike in PRISM, we do not split the network into a unimodal encoder and multi-modal decoder, rather we dedicate the entire network to serve as a language encoder. We initialize a class token which is concatenated with the input text sequence. The final layer hidden state of the class token is then used as the text embedding for the contrastive objective.

Large language model decoder. For the dialogue-focused language model component we use Phi-3 Mini [1] which is a 3.8B parameter decoder-only LLM. We adopt a LLaVa-style [26] approach to enabling Phi-3 Mini to process both images and text. The 256 latents outputted by the slide encoder are further processed by a 2-layer multi-layer perceptron (MLP) adapter (29M parameters) to project them into the Phi-3 embedding space and are inserted into the text token embedding sequence as input to the LLM. This allows Phi-3 to integrate the slide and text information via its self-attention layers. Due to the next-token prediction objective of LLMs, we find that the Phi-3 Mini final layer hidden state of the `<|assistant|>` token encodes a useful slide representation which we refer to as the *diagnostic* embedding.

3.2   Clinical dialogue

In order to achieve clinically relevant dialogue and reporting capabilities, we generate and curate various formats of text data with GPT-4o [31] using ground truth clinical diagnostic reports as seed data (see Appendix A.4 for system prompts). This data consists of report rewrites, de-identified patient history, specimen descriptions, and question-answer pairs. We compose these categories of data into various chat-style templates throughout training to simulate dialogue. We create four main templates to use in training:

1. **Clinical report generation.** Given a task string (e.g. "Write a report"), the model is tasked to generate the summarized clinical report.

2. **Question answering (yes / no).** The model is tasked to answer a series of binary questions from a user.

3. **Question answering (open-ended).** The model is tasked to answer a series of open-ended questions from a user.

4. **Image-text matching.** Given a report, the model must identify if the report accurately describes the given WSIs, and then generate the correct report if they are mismatched.

Specifically, we adopt the Phi-3 [1] chat template using alternating `user` and `assistant` tags. The first `user` message always contains the WSI latents from the slide encoder followed by an initial prompt. In each template, the patient history and specimen description are randomly included, each with a probability of 20%, to allow for conditioning the model on this extra context without being overly reliant on it. In total, we create 685K report rewrites for as many specimens and 14M question-answer pairs.

Clinical report rewrites. We use GPT-4o to rewrite clinical diagnostic reports which are subsequently used for training both the language encoder and the language model (see Sec 3.1 for details). Clinical diagnostic reports are a mixture of synoptic reports and free-text notes. We prompt GPT-4o to rewrite all input text, including patient history and specimen description when available, into natural language sentences while retaining all diagnostically relevant findings. These rewrites are typically longer and more informative than the higher level summaries previously used in PRISM [34].

Yes-No question-answer pairs. To broaden the training distribution beyond report generation, we use GPT-4o to generate question-answer pairs through a multi-step process. First, we prompt GPT-4o to convert clinical reports into lists of individual findings, similar to MAIRA-2 [7], which are then filtered to exclude molecular/IHC results. These filtered findings are fed back to GPT-4o with a prompt to generate question-answer pairs for each finding that can be answered with simply "Yes." or "No.". We found this process is naturally biased towards creating questions that frequently have the same answer throughout the dataset, as negative findings are typically not mentioned in the reports. For instance, a report positively identifying ductal carcinoma in situ (DCIS) in breast tissue allows GPT-4o to generate a pair such as: "Q: Is DCIS present? A: Yes." However, a normal breast tissue report does not mention DCIS, so no negative question-answer pair of the form "Q: Is DCIS present? A: No," is generated, biasing the dataset (and resulting model) toward affirmative responses. Inspired by [20], to mitigate this bias, we implemented a simple method to mine complementary images for additional question-answer pairs. Specifically, we pair random reports with random questions and prompt GPT-4o to answer based on the provided report. This yields complementary image-question-answer triplets, balancing the data distribution. We include these complementary QA pairs with a sampling rate of 20% during training.

Open-ended question-answer pairs. We also use GPT-4o to generate open-ended question-answer pairs for each finding. We instruct GPT-4o to create the QA pairs such that the questions are not answerable with just yes/no but require more specific answers. Open-ended QA pairs help mimic more realistic model interactions and improve the diversity of the training data.

3.3 Pre-processing of whole slide images

WSIs are processed using a fully-convolutional network to perform tissue segmentation and tiled at $20\times$ magnification (0.5 μm/pixel) with a size of 224×224 pixels, discarding any tiles that are less than 65% tissue. The resulting tiles are then processed by Virchow2 [44] to produce tile embeddings. We use the 1280-dimensional class token output of the model. During training, up to 100K tiles are loaded per specimen. In cases where specimens have more than 100K tiles, WSIs in the specimen are dropped until the criterion is met. This restriction does not apply during inference.

3.4 Training recipe

Objectives. PRISM2 is trained with both contrastive and autoregressive objectives. We follow [43, 34] for the formulation of the contrastive loss:

$$\mathcal{L}_{\text{con}} = -\frac{1}{N} \left(\sum_{i=1}^{N} \log \frac{\exp(\boldsymbol{v}_i^\top \boldsymbol{t}_i / \tau)}{\sum_{j=1}^{N} \exp(\boldsymbol{v}_i^\top \boldsymbol{t}_j / \tau)} + \sum_{i=1}^{N} \log \frac{\exp(\boldsymbol{t}_i^\top \boldsymbol{v}_i / \tau)}{\sum_{j=1}^{N} \exp(\boldsymbol{t}_i^\top \boldsymbol{v}_j / \tau)} \right), \tag{1}$$

where N is batch size, τ is a learned temperature, and $(\boldsymbol{v}_i, \boldsymbol{t}_i)$ are the image and text representations. For the image and text representations we use the base embedding from the slide encoder and the class token embedding from BioGPT after a linear projection and normalization with the ℓ_2 norm. Only the diagnostic report rewrites are used as the text input for the contrastive objective. For the autoregressive objective, we apply the standard cross-entropy loss on next-token predictions from Phi-3 Mini. We use the chat templates as the text input and only apply the loss to the `assistant` regions of the templates:

$$\mathcal{L}_{\text{chat}} = -\sum_{t \in \mathcal{A}} \log p(y_t \mid y_{<t}, \boldsymbol{X}), \tag{2}$$

where $\mathcal{A}$ is the set of token indices corresponding to assistant tokens and $\boldsymbol{X}$ are the image latents after the LLM adapter MLP. The losses are linearly combined to create the total loss:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{con}} \mathcal{L}_{\text{con}} + \lambda_{\text{chat}} \mathcal{L}_{\text{chat}}. \tag{3}$$

Training. We adopt a two stage training approach. In the first stage the slide encoder, attention pooler, LLM adapter, and BioGPT are unfrozen. We freeze Phi-3 Mini during this stage but still keep the autoregressive objective online to help the slide encoder and adapter align to Phi-3. We set λ_{con} to 0.25 and λ_{chat} to 1.0. In the second stage, the contrastive objective is discarded and the slide encoder is frozen. The adapter and Phi-3 Mini are unfrozen to allow the language model to better integrate the image input and model the language of pathology. We also reset the optimizer and learning rate schedule for this stage.

We train the model on 56 A100 40GB GPUs with bf16 precision. We use the AdamW [27] optimizer with a learning rate warmup for 500 steps which is then held constant. We do one epoch of stage 1 training with a post-warmup LR of 1.0×10^{-4} on all trainable components except the language encoder which uses an LR of 2.0×10^{-5}. We do 16 epochs of stage 2 training. An epoch is defined as one pass over each of the four chat templates per specimen, for all specimens in the training data. Since the contrastive objective aligns the image representation with the report and not with each chat template, each specimen is paired with its report four times in a single epoch.

Data We collected a dataset of 2,350,518 hematoxylin and eosin (H&E) stained whole slide images (WSIs) corresponding to 685,507 specimens, where each specimen is a collection of one or more WSIs with a corresponding clinical report as shown in Fig. 3. The data is collected from Memorial Sloan Kettering Cancer Center (MSKCC), spanning a diverse set of tissues and indications, as well as diverse institutions that submitted cases to MSKCC for review.
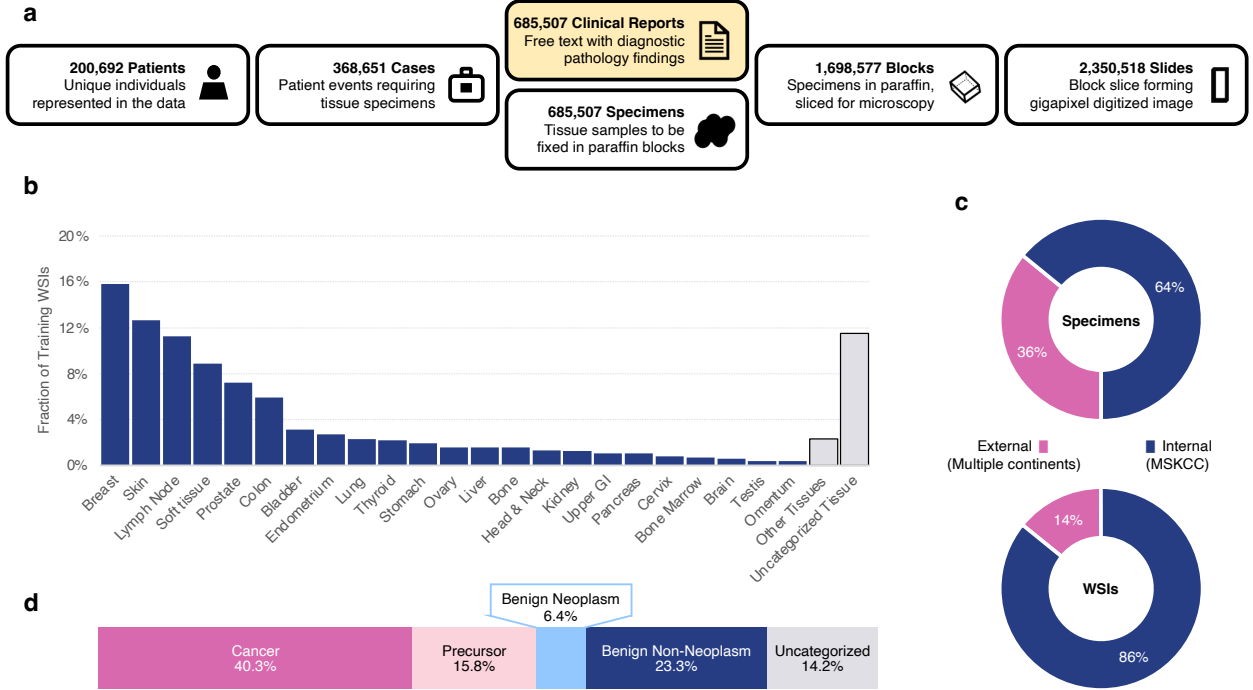
Figure 3: PRISM2 training data. **a** The training data can be grouped on the level of patients, cases, specimens, blocks, or slides, as shown. PRISM2 aligns a specimen-level image representation, consisting of multiple whole slide images, with specimen-level clinical reports. **b** The top identified tissue types are shown as a fraction of specimens. There are over 150 additional tissue types in the long tail, summarized here as 'Other Tissues' (shown gray, along with 'Unknown Tissue', on the right). **c** 36% of all specimens (comprising 14% of all slides) are comprised of slides submitted to MSKCC from diverse institutions across the world. **d** The most severe diagnosis is shown as a fraction of the training dataset specimens (parsed from clinical reports with ChatGPT). If a clinical report describes the presence of a precursor to cancer along with invasive cancer, then that specimen is labeled cancer here as it is the more severe condition.

3.5 Zero-shot prediction

Both contrastive and generative caption training objectives enable zero-shot prediction for classification tasks, demonstrating the model's ability to learn clinically relevant tasks directly from diagnostic reports without explicit task-specific training.

Zero-shot prediction with dialogue completion. The question-answering ability of PRISM2 is a straightforward mechanism of zero-shot prediction. Binary classification tasks lend themselves well to yes/no questions (e.g. *Is a malignant tumor present?*). We refer to this method as yes/no, or 'YN', dialogue-based zero-shot prediction. The YN dialogue has the advantage over contrastive zero-shot prediction (see 'Zero-shot prediction by ranking similarity scores' below) in that no comprehensive list of negative prompts is required (e.g. keywords covering all non-cancer conditions). Constrained multi-class classification lends itself well to ranking the answers when posing an open-ended question (e.g. *What type of carcinoma is suggested by the specimen?*). We refer to this method as open-ended or 'OPN' dialogue-based zero-shot prediction.

Both YN-, and OPN-based dialogue can quantify the probability of predictions. For each zero-shot classification task, we choose a question prompt $w_{<i}$ (e.g. *Is malignant tumor present?*) and consider a set C of possible completions c. For YN dialogue, these are 'Yes' or 'No'; for OPN dialogue, these may be fill-in-the blank options. For example, when determining a breast cancer subtype, one may complete the phrase "The specimen contains_____" with "invasive ductal carcinoma" or "invasive lobular carcinoma". Besides the prompt, we provide the model with context x, which includes WSI features for the given specimen. When using yes/no questions, the context may also include the diagnostic report generated by PRISM2 (referred to as 'DYN' dialogue-based zero-shot), which is appended to the Phi-3 Mini input sequence after the image and before the question. We hypothesize that generating the diagnosis first

can help the model to answer a yes/no question. Model predictions are ranked by the normalized log probability of each fill-in-the-blank completion.

Formally, we define our dialogue-based zero-shot classifier as

$$f(w_{<i}, x) = \arg\max_{c \in C} \log P(w_i = c | w_{<i}, x) - \log P(w_i = c | w_{<i}), \tag{4}$$

where c is a completion in a set of completions C that cover all classes in the task, $w_{<i}$ is a tokenized prompt containing a question and x is the context which includes WSI embeddings from the slide encoder and, optionally, a tokenized generated diagnosis. As in [10], we normalize log probabilities by the prior (log) probability without context x.

Contrastive zero-shot prediction by ranking similarity scores. The contrastive image-text alignment objective can be reformulated as multi-class classification. Given a specimen-level embedding v from the PRISM2 slide encoder and a set of prompt embeddings $t \in T$ from the language encoder, one per each class, we compute the cosine similarity between each prompt embedding and apply a softmax activation to the resulting logits. Each entry in the vector corresponds to a probability of that prompt matching the slide. In the case of multiple per class, we take the prompt with the highest similarity score as the prompt representing that class. Formally, the probability of the class i being the true class is

$$P(C = i \mid v) = \frac{\exp(v^\top t_i)}{\sum_{t_j \in T} \exp(v^\top t_j)}. \tag{5}$$

3.6 Efficient Implementation

Sequence packing. Training efficiently on WSIs at scale presents several challenges for maximizing hardware utilization. The number of tile embeddings extracted from WSIs can vary dramatically between specimens, introducing highly variable sequence lengths within a training batch. A typical way to handle this is by padding to the longest sequence length in the batch, but this is not optimal as it introduces unnecessary compute and memory for the padding elements. Empirically, we find the ratio of compute/memory for padding elements to non-padding elements quickly exceeds 1.0 even with small batch sizes, which means we begin to dedicate the majority of compute/memory to padding elements. To mitigate this we implement sequence packing for the slide encoder which eliminates any need for padding and thus incurs no unnecessary computational and memory cost.

Dynamic batching. In multi-GPU data parallel training, variable length sequences also create load imbalance issues. For one training step, each data parallel worker (i.e. a GPU) does a forward and backward pass on its own mini-batch and then hits a global synchronization point for gradient `allreduce`. This synchronization point requires each worker to wait for the slowest worker to complete its training step, resulting in significant idle GPU time. Without careful batching, one GPU might process significantly more embeddings than another, creating severe load imbalance. Our sequence packing implementation directly enables efficient dynamic batching: we set a fixed budget of 800 thousand tile embeddings per GPU per step, then greedily pack variable-length sequences from a queue to fill this budget. This ensures each worker has the same amount of embeddings per step which improves workload balancing and mitigates idle GPU time. The number of specimens/reports per worker per step then becomes dynamic.

Efficient grouped query attention. We opt to use grouped query attention (GQA) [4] when doing cross attention with tile embeddings. This leads to significant memory savings during training by projecting tile embeddings to a much smaller embedding space as keys/values for cross attention. We use the official FlashAttention [15, 14] kernel from the `flash-attn` library which has native support for GQA rather than a naïve implementation that repeats key-value heads to match query heads which would negate the potential memory savings. The `flash-attn` kernel also supports sequence packing which allows us to avoid padding/unpadding before and after attention.

Fully sharded data parallelism and activation checkpointing. We use fully sharded data parallelism (FSDP) to shard model weights, gradients, and optimizer states. This reduces redundant memory and enables increasing model parameters and/or batch sizes. We noticed that, due to the relatively small batch size of text tokens, training of the decoder is bottlenecked by FSDP communications. We enabled activation checkpointing which reduces memory and allows increasing the batch size while still overlapping compute and communications.
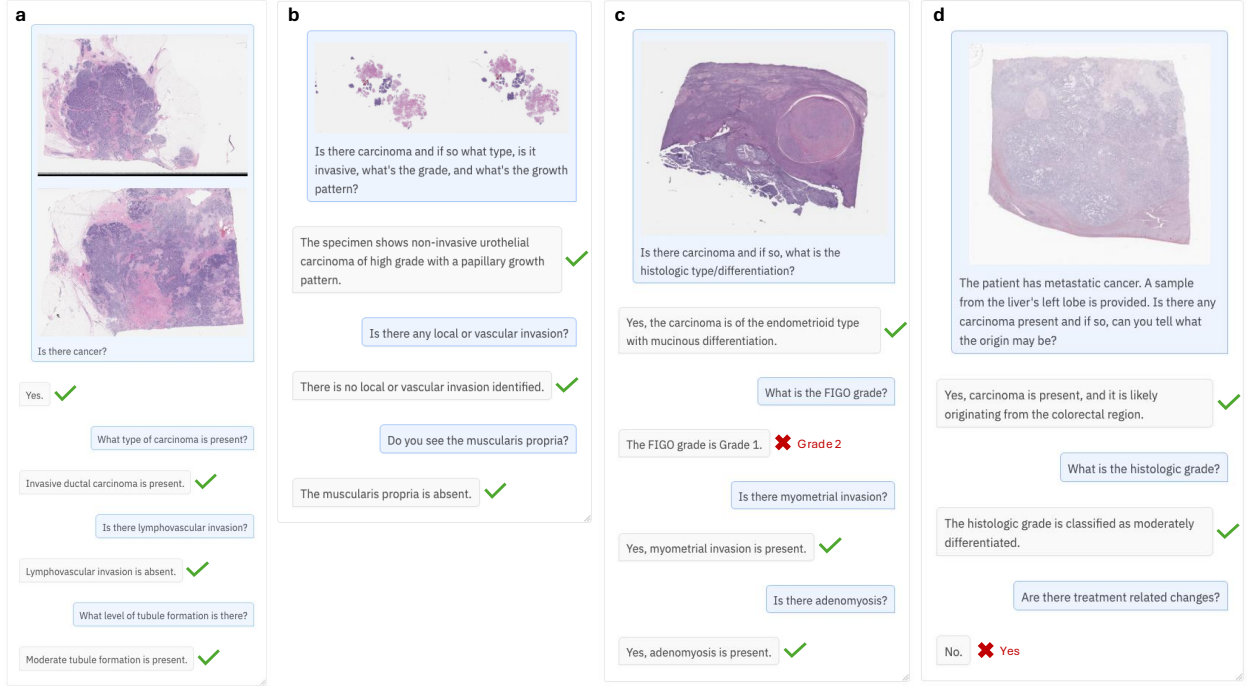
Figure 4: This figure demonstrates the examples of PRISM2 dialogues. **a** A breast specimen with IDC. We show WSIs in the initial message for illustrative purposes, but for each example PRISM2 views all the slides of the specimen at once. **b** A bladder biopsy. Despite being trained on simple questions, the model is able to accurately respond to compound questions. **c** A uterus specimen. The model correctly identifies and characterizes the carcinoma and additional findings except it incorrectly predicts FIGO Grade 1 instead of Grade 2. **d** A liver specimen from a patient with metastatic colorectal cancer. The model is able to be given additional context like some patient history and specimen description and detects the carcinoma and correctly predicts its origin and grade, but incorrectly says there are no treatment related changes.

4 Experimental results and discussion

4.1 PRISM2 introduces slide-level diagnostic dialogue

While PRISM has the ability to generate short clinical report summaries, PRISM2 adapts an LLM to interpret WSIs for diagnostic question-answering. This enables dialogue, given a set of WSIs as shown in Fig. 4. As the LLM (Phi-3 Mini [1]) is pretrained and instruction-tuned, we can see in Fig. 4b that it can generalize to compound questions, asking about multiple concepts in one sentence, even while the training data asks only about one concept per question. Asking the model targeted questions during training is intended to help it focus on all of the indications present in clinical reports rather than only on those with the most verbose descriptions. Furthermore, as we will see in the next section (Sec. 4.2), being able to ask specific questions enables simple and powerful zero-shot predictions.

As shown in Fig. 5 and Fig. 6, PRISM2 produces verbose descriptions of specimens. Furthermore, these descriptions remain consistent when looking only at the cropped regions that the model attends to the most. The model's attention maps, taken from the slide encoder's cross-attention layer, are shown for a single core of a WSI and each of its top regions of interest. We compute attention maps by summing up all the scores per tile across attention heads and queries and multiplying them by the ℓ_2-norms of the corresponding values.

4.2 Dialogue simplifies zero-shot prediction

Existing multi-modal pathology foundation models have demonstrated successful CLIP-style contrastive zero-shot classification [34, 17] on clinical tasks. Although including a contrastive training objective naturally endows vision-language models with the ability to perform zero-shot classification, such an approach also incurs several shortcomings. First, the optimality of a given text prompt is sensitive to the distribution of the training data. Second, diagnoses that cover a wide set of morphological features are difficult to represent by a small number of text prompts. For exam-
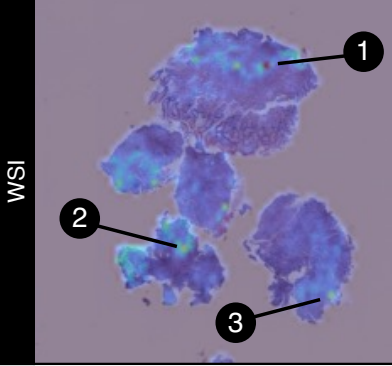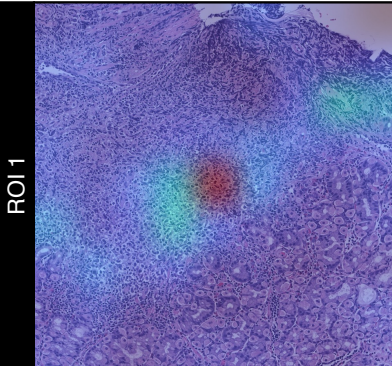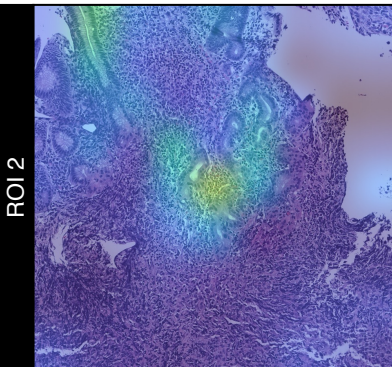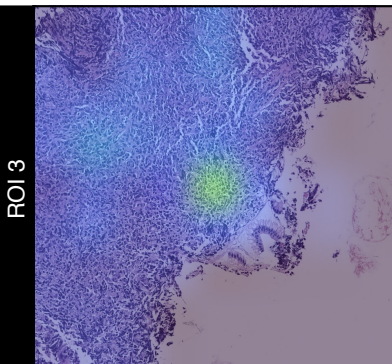
GENERATED

WSI

The examination reveals diffuse large B-cell lymphoma with a germinal center B-cell phenotype. The histologic sections show fragments of gastric mucosa with a diffuse atypical lymphoid infiltrate. This infiltrate is composed of medium to large cells with irregular nuclei, vesicular chromatin, and conspicuous nucleoli. Frequent mitotic figures and apoptotic bodies are observed. The findings are consistent with diffuse large B-cell lymphoma with a germinal center B-cell phenotype.

ROI 1

Extranodal marginal zone lymphoma of mucosa-associated lymphoid tissue (MALT lymphoma) is identified. The gastric mucosa shows an atypical lymphoid infiltrate in the lamina propria, composed of small to medium-sized lymphoid cells with round to irregular nuclei, condensed chromatin, inconspicuous nucleoli, and scant to moderate cytoplasm. There are admixed centrocytes and occasional admixed centroblasts. Few mitotic figures and apoptotic bodies are present. No sheets of large cells are observed. The findings are consistent with extranodal marginal zone lymphoma of mucosa-associated lymphoid tissue (MALT lymphoma).

ROI 2

Extranodal marginal zone lymphoma of mucosa-associated lymphoid tissue (MALT lymphoma) is observed. The histologic sections reveal fragments of gastric mucosa with a dense atypical lymphoid infiltrate in the lamina propria. This infiltrate is composed of small to medium-sized lymphoid cells with round to irregular nuclei, condensed chromatin, indistinct nucleoli, and small to moderate amounts of pale cytoplasm. Occasional plasmacytoid lymphocytes and lymphoepithelial lesions are present. No sheets or aggregates of large cells are identified.

ROI 3

The examination reveals diffuse large B-cell lymphoma with a germinal center B-cell phenotype. The histologic sections show a diffuse proliferation of medium to large atypical lymphoid cells with irregular nuclear contours, vesicular chromatin, and distinct nucleoli. There is brisk mitotic and apoptotic activity, along with focal necrosis. The findings are consistent with diffuse large B-cell lymphoma with a germinal center B-cell phenotype.

Figure 5: The slide encoder attention heatmap is shown for a core in a single slide of a lymphoma specimen, along with the diagnostic report generated by PRISM2. When cropping to each of the three top-attended regions, the PRISM2 report remains consistent with the evidence in the regions of interest.

	GENERATED
WSI	The gastric antral mucosa shows chronic inactive gastritis. There is no evidence of intestinal metaplasia or dysplasia. Helicobacter pylori is not detected.
ROI 1	The gastric antral mucosa shows chronic inactive gastritis. No Helicobacter pylori is observed. There is no evidence of lymphoma. The findings are consistent with a reactive process.
ROI 2	The gastric antral mucosa shows chronic inactive gastritis. There is no evidence of lymphoma. The gastric mucosa exhibits chronic inactive gastritis with reactive lymphoid aggregates. The lymphoid aggregates are composed of small lymphocytes with round to oval nuclei, condensed chromatin, and inconspicuous nucleoli. There is no evidence of lymphoepithelial lesions or lymphoepithelial lesions. The findings are consistent with a reactive process, and there is no morphologic evidence of lymphoma.
ROI 3	The gastric antral mucosa shows chronic inactive gastritis. No Helicobacter pylori is observed.

Figure 6: The slide encoder attention heatmap is shown for a core in a single slide of a gastric specimen, along with the diagnostic report generated by PRISM2. When cropping to each of the three top-attended regions, the PRISM2 report remains consistent with the evidence in the regions of interest.
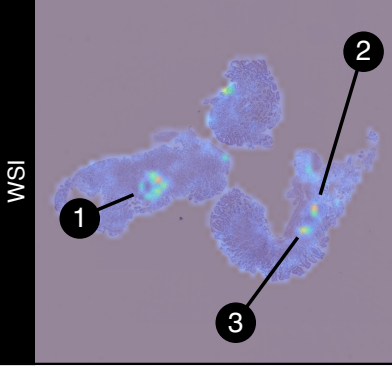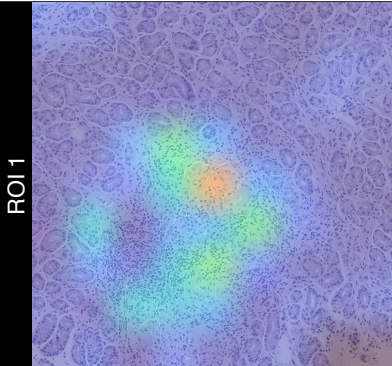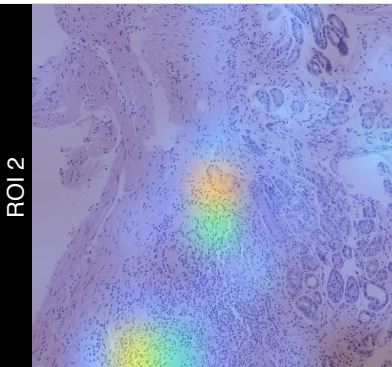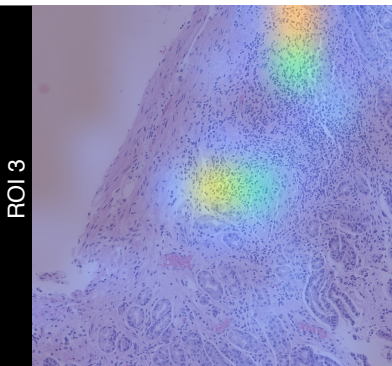
Model	Method	PanCancer	PanCancer Rare	Breast Subtyping	GI Suite
TITAN	Contrastive	0.728 (0.716, 0.738)	0.704 (0.687, 0.720)	-	-
PRISM	Contrastive	0.793 (0.783, 0.803)	0.753 (0.737, 0.769)	-	-
PRISM2	Contrastive	0.719 (0.707, 0.730)	0.718 (0.701, 0.735)	-	-
	Dialogue (YN)	**0.874** (0.865, 0.882)	**0.861** (0.848, 0.874)	**0.874** (0.849, 0.897)	**0.797** (0.781, 0.813)
	Dialogue (DYN)	**0.880** (0.871, 0.888)	**0.872** (0.859, 0.885)	**0.847** (0.822, 0.870)	**0.815** (0.796, 0.834)

(a) Balanced accuracy for binary classification tasks using private MSKCC data.

Model	Method	BRCA	NSCLC	RCC
TITAN	Contrastive	**0.895** (0.870, 0.917)	**0.930** (0.915, 0.946)	**0.900** (0.882, 0.915)
PRISM	Contrastive	**0.883** (0.858, 0.906)	0.877 (0.858, 0.895)	0.631 (0.597, 0.666)
PRISM2	Contrastive	**0.885** (0.863, 0.904)	0.888 (0.870, 0.906)	**0.894** (0.867, 0.919)
	Dialogue (OPN)	**0.922** (0.901, 0.942)	**0.925** (0.909, 0.941)	**0.902** (0.879, 0.925)

(b) Balanced accuracy on multi-class classification tasks using public The Cancer Genome Atlas (TCGA) data.

Table 1: Zero-shot performance (balanced accuracy) for various classification tasks. 95% confidence intervals (CI) computed with 1000 iterations of bootstrapping are shown in parentheses. The best result per task ($p < 0.05$) is in bold. (a) Dialogue-based yes/no (YN and DYN) question-answering substantially outperforms contrastive zero-shot prediction in the challenging PanCancer detection task, where it is challenging to assemble all relative cancer-positive and cancer-negative prompts for zero-shot evaluation. Defining the negative prompts is even more challenging for the binary classes in the Breast Subtyping and GI Suite datasets, so contrastive zero-shot evaluation for these is omitted. While both the YN and DYN use yes/no questions, DYN also includes a generated report as context. (b) Dialogue-based open-ended (OPN) question-answering is on par with TITAN contrastive zero-shot prediction on public TCGA multi-class classification.

ple, a benign diagnosis may be implied from a pathologist's report by the absence of notable findings without being explicitly stated. Or, in the case of classifying invasive breast cancer, enumerating all the possible less severe diagnoses belonging to the noninvasive class (e.g. DCIS, LCIS, ADH, ALH) is cumbersome and inaccurate, as these indications can be present alongside invasive cancer. There may not exist a canonical text prompt for a given diagnosis.

Dialogue-based question-answering, using yes/no questions, is a better fit for identifying whether an indication is present or absent as no positive or negative sets of prompts need to be defined. Both contrastive and dialogue-based zero-shot predictions are evaluated in Tab. 1 and Supplementary Tab A.2.1. Indeed, whereas dialogue-based question answering ('Dialogue (YN)') achieves a balanced accuracy of 0.874 on all cancers and 0.861 on rare cancers, contrastive zero-shot with PRISM2 achieves only 0.719 and 0.718, respectively, comparable to TITAN's 0.728 for all cancers and 0.704 for rare cancers. Including the diagnosis prediction of PRISM2 in the context ('Dialogue (DYN)') further boosts performance (0.874 to 0.880 balanced accuracy; 0.933 to 0.943 AUC) on all cancers and (0.861 to 0.872 balanced accuracy; 0.929 to 0.943 AUC on rare cancers). Contrastive zero-shot prediction works well on constrained multi-class subtyping tasks like TCGA BRCA, NSCLC, and RCC (Tab. 1b and Supplementary Tab A.2.1b). Even so, dialogue-based zero-shot prediction performs as well or better.

The aforementioned challenges with contrastive zero-shot prediction are evident when analyzing pan-cancer contrastive zero shot performance. Using the prompts in Supplementary Tab. A.3.3 'PanCancer' to evaluate both PRISM and PRISM2, the former obtains a higher balanced accuracy (0.793 vs 0.719) and area under the curve (AUC) (0.896 vs. 0.805). However, when stratifying by tissue type, the inter- and intra-model variability in AUC demonstrates the brittle relationship between prompt selection and contrastive zero shot performance. In particular, we observe low performance on breast tissue for PRISM2 (0.622 AUC) even though the tissue represents 21% of the training samples. Thirty two percent of all breast false positives are caused by an erroneous match between the specimen's image embedding and the "sarcoma" text prompt embedding—a word which is rarely used during training to describe breast cancer. This analysis motivates the hypothesis that contrastive zero shot performance depends precariously on the choice of prompt.

Indeed, we demonstrate that PRISM2 breast pan-cancer detection performance can be substantially improved just by changing the prompt selection to those in Supplementary Tab. A.3.3. We (I) align text prompts more closely with the training corpus language and (II) restrict our analysis to the breast cohort of the pan-cancer detection dataset. We craft a set of prompts which are stylistically in-distribution with respect to the PRISM2 training reports and delineate

more closely between the positive diagnoses (IDC, ILC) and the negative ones (DCIS, LCIS, ADH, ALH, and benign). When evaluated against these new prompts, PRISM2 and PRISM breast AUC values are comparable (0.756 vs 0.798). This result supports the hypothesis that the CLIP-style zero shot capability depends on (and is hindered by) the ability to find the appropriate set of in-distribution text prompts.

4.3   Dialogue improves the feature representation for diagnostic tasks

In Tab. 2, we evaluate the quality of the base and diagnostic embeddings produced by PRISM2 by training linear classifiers on them across a variety of tasks, covering the detection and subtyping of cancers and precursors to cancers, as well as the detection of inflammatory and benign conditions and biomarkers (see A.1 for a description of evaluation tasks and data).

	PRISM2 Diagnostic	PRISM2 Base	PRISM	TITAN
PanCancer	**0.965** (0.961, 0.969)	0.952 (0.947, 0.958)	0.947 (0.942, 0.953)	0.931 (0.924, 0.937)
PanCancer Rare	**0.953** (0.945, 0.960)	0.932 (0.922, 0.941)	0.931 (0.921, 0.940)	0.913 (0.902, 0.923)
Breast Subtyping	**0.979** (0.976, 0.982)	**0.976** (0.972, 0.979)	**0.975** (0.971, 0.978)	0.923 (0.916, 0.929)
GI Suite	**0.977** (0.974, 0.979)	**0.978** (0.976, 0.980)	0.971 (0.968, 0.973)	0.965 (0.961, 0.968)
Biomarkers	0.845 (0.830, 0.859)	**0.857** (0.844, 0.870)	**0.843** (0.828, 0.857)	**0.846** (0.829, 0.860)
TCGA BRCA	**0.960** (0.938, 0.977)	**0.958** (0.938, 0.976)	**0.949** (0.927, 0.969)	**0.956** (0.939, 0.970)
TCGA NSCLC	**0.979** (0.970, 0.986)	**0.980** (0.972, 0.987)	**0.976** (0.968, 0.984)	**0.981** (0.973, 0.989)
TCGA RCC	**0.989** (0.984, 0.993)	**0.991** (0.987, 0.994)	**0.985** (0.979, 0.991)	**0.992** (0.988, 0.995)

Table 2: AUC results for various downstream tasks by training linear classifiers on embeddings from PRISM2, PRISM, and TITAN. 95% confidence intervals (CI) computed with 1000 iterations of bootstrapping are shown in parentheses. The best result per task ($p < 0.05$) is in bold. Like the PRISM and TITAN embeddings, the 'PRISM2 Base' embedding is trained contrastively. The 'Diagnostic' embedding extracted from the LLM hidden state further improves performance on PanCancer detection and breast cancer subtyping ('Breast Subtyping'); however, the 'Base' embedding performs better on biomarkers ('Biomarkers'), something the LLM is not trained to discuss. AUC is summarized one-vs-one across multiple classes in each task. Specifically, 'PanCancer' evaluates cancer detection across 16 origin sites (and 'PanCancer Rare' is the subset of rare cancers); 'Breast Subtyping' evaluates the detection of ductal or lobular cancers and precursors to cancer; 'GI Suite' evaluates the detection of 12 conditions (among cancers, precursors, inflammatory, and benign); 'Biomarkers' includes 10 biomarkers; and the TCGA tasks evaluate multi-class cancer subtyping.

The diagnostic embedding significantly improves performance on the challenging PanCancer detection task, covering 16 different cancer tissue origins manifesting both as primary and metastatic cancer across diverse tissues. While the base embeddings of PRISM2 (0.952 AUC) outperform the PRISM embeddings (0.947 AUC), the difference is not statistically significant; whereas, the diagnostic embeddings (0.965 AUC) significantly ($p < 0.05$) outperform all others. Both PRISM2 and PRISM embeddings substantially improve over TITAN (0.931 AUC). A similar trend is seen when focusing on the rare cancer tissues of origin, with PRISM2 diagnostic embeddings (0.953 AUC) significantly outperforming base embeddings (0.932 AUC), PRISM (0.947 AUC), and TITAN (0.913 AUC). This result demonstrates that the LLM can further improve diagnostic image understanding beyond the representation in the base embeddings learned via contrastive alignment. The diagnostic embeddings similarly produce the best result on breast subtyping (cancers and precursors to cancer) with an AUC of 0.979 as compared to 0.976 with the base embeddings, 0.975 with PRISM and 0.923 with TITAN; however, there is not enough statistical power to conclude a difference for this task with respect to the base embeddings. The low performance of TITAN on this task appears to be attributable to lower performance on precursors to cancer, as shown in Supplementary Tab. A.2.4.

Whereas diagnostic embeddings appear well-suited for cancer and precursor detection, they do not appear to be optimal for biomarker detection. Although there is low statistical power (due to the small sample sizes), we can see that the base embedding (overall 0.857 AUC) tends to outperform the diagnostic embedding (overall 0.845 AUC) in 9 out of 10 tasks and achieves the top score in 7 out of 10 tasks, as compared to one top score for PRISM and two top scores for TITAN (see Supplementary Tab. A.2.6). Recall that the diagnostic embeddings are extracted from the hidden state of the LLM which is trained to discuss these diagnostic findings but is not trained to discuss biomarkers. Indeed, the base embeddings trained with contrastive alignment may serve as lower level features that may be more readily transferable to out-of-distribution tasks.

PRISM2 significantly ($p < 0.05$) outperforms both PRISM and TITAN on the gastrointestinal (GI) Suite, including the detection of cancers, precursors to cancer, inflammatory conditions, and benign indications (see Supplementary

Tab. A.2.5 for more detailed results per class). As with breast subtyping the lower performance of TITAN (and in this case PRISM) is in large part related to greater difficulty detecting precursors to cancer or benign conditions.

On out of distribution TCGA tasks (BRCA, NSCLC, and RCC), PRISM2 performs on par with PRISM and TITAN, as shown in Tab. 2.

5   Discussion

PRISM2 advances the state-of-the-art in slide-level computational pathology foundation models by tackling several open problems. First, by scaling the dataset size to 2.3M WSIs and almost 700K clinical reports, PRISM2 is able to better generalize to diverse clinical tasks. Second, by introducing a large language model, PRISM2 introduces a conversational interface. Finally, by introducing the diagnostic and base embeddings, PRISM2 provides representations that are highly aligned with standard clinical tasks as described using natural language as well as a base embedding for use in emerging computational pathology applications such as biomarker quantification.

6   Acknowledgments

References

[1] Marah Abdin et al. "Phi-3 technical report: A highly capable language model locally on your phone". In: *arXiv preprint arXiv:2404.14219* (2024).

[2] Faruk Ahmed et al. "Pathalign: A vision-language model for whole slide images in histopathology". In: *arXiv preprint arXiv:2406.19578* (2024).

[3] Faruk Ahmed et al. "PolyPath: Adapting a Large Multimodal Model for Multi-slide Pathology Report Generation". In: *arXiv preprint arXiv:2502.10536* (2025).

[4] Joshua Ainslie et al. "GQA: Training Generalized Multi-Query Transformer Models from Multi-Head Checkpoints". In: *arXiv preprint arXiv:2305.13245* (2023).

[5] Jean-Baptiste Alayrac et al. "Flamingo: a visual language model for few-shot learning". In: *Advances in neural information processing systems* 35 (2022), pp. 23716–23736.

[6] Rohan Anil et al. "PaLM 2 Technical Report". In: *arXiv preprint arXiv:2305.10403* (2023).

[7] Shruthi Bannur et al. "MAIRA-2: Grounded Radiology Report Generation". In: *arXiv preprint arXiv:2406.04449* (2024).

[8] Iz Beltagy, Matthew E. Peters, and Arman Cohan. "Longformer: The Long-Document Transformer". In: *arXiv preprint arXiv:2004.05150* (2020).

[9] Bioptimus. *H-Optimus-1*. https://www.bioptimus.com/h-optimus-1. 2025.

[10] Tom B. Brown et al. "Language Models are Few-Shot Learners". In: *arXiv preprint arXiv:2005.14165* (2020).

[11] Richard J Chen et al. "Towards a general-purpose foundation model for computational pathology". In: *Nature Medicine* (2024), pp. 1–13.

[12] Ying Chen et al. "SlideChat: A Large Vision-Language Assistant for Whole-Slide Pathology Image Understanding". In: *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*. June 2025, pp. 5134–5143.

[13] Diana T Cheng et al. "Memorial Sloan Kettering-Integrated Mutation Profiling of Actionable Cancer Targets (MSK-IMPACT): A Hybridization Capture-Based Next-Generation Sequencing Clinical Assay for Solid Tumor Molecular Oncology". In: *Journal of Molecular Diagnostics* 17.3 (2015), pp. 251–264. DOI: 10.1016/j.jmoldx.2014.12.006.

[14] Tri Dao. "FlashAttention-2: Faster Attention with Better Parallelism and Work Partitioning". In: *International Conference on Learning Representations (ICLR)*. 2024.

[15] Tri Dao et al. "FlashAttention: Fast and Memory-Efficient Exact Attention with IO-Awareness". In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2022.

[16] Jiayu Ding et al. "Longnet: Scaling transformers to 1,000,000,000 tokens". In: *arXiv preprint arXiv:2307.02486* (2023).

[17] Tong Ding et al. "Multimodal whole slide foundation model for pathology". In: *arXiv preprint arXiv:2411.19666* (2024).

[18] Alexandre Filiot et al. "Scaling Self-Supervised Learning for Histopathology with Masked Image Modeling". In: *medRxiv* (2023), pp. 2023–07.

[19] Zeyu Gao et al. "ALPaCA: Adapting Llama for Pathology Context Analysis to enable slide-level question answering". In: *medRxiv* (2025). DOI: 10.1101/2025.04.22.25326190. eprint: https://www.medrxiv.org/content/early/2025/04/22/2025.04.22.25326190.full.pdf. URL: https://www.medrxiv.org/content/early/2025/04/22/2025.04.22.25326190.

[20] Yash Goyal et al. "Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering". In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017, pp. 6325–6334. DOI: 10.1109/CVPR.2017.670.

[21] Aaron Grattafiori et al. "The Llama 3 Herd of Models". In: *arXiv preprint arXiv:2407.21783* (2024).

[22] Zhi Huang et al. "A visual–language foundation model for pathology image analysis using medical twitter". In: *Nature medicine* 29.9 (2023), pp. 2307–2316.

[23] Andrew Jaegle et al. "Perceiver: General perception with iterative attention". In: *International conference on machine learning*. PMLR. 2021, pp. 4651–4664.

[24] Guillaume Jaume et al. "Transcriptomics-guided slide representation learning in computational pathology". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024, pp. 9632–9644.

[25] Junnan Li et al. "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models". In: *International conference on machine learning*. PMLR. 2023, pp. 19730–19742.

[26] Haotian Liu et al. "Visual instruction tuning". In: *Advances in neural information processing systems* 36 (2023), pp. 34892–34916.

[27] Ilya Loshchilov and Frank Hutter. "Decoupled Weight Decay Regularization". In: *arXiv preprint arXiv:1711.05101* (2017).

[28] Ming Y Lu et al. "A visual-language foundation model for computational pathology". In: *Nature Medicine* 30.3 (2024), pp. 863–874.

[29] Ming Y. Lu et al. "A multimodal generative AI copilot for human pathology". In: *Nature* 634.8033 (Oct. 2024), pp. 466–473. ISSN: 1476-4687. DOI: 10.1038/s41586-024-07618-3. URL: https://doi.org/10.1038/s41586-024-07618-3.

[30] Renqian Luo et al. "BioGPT: generative pre-trained transformer for biomedical text generation and mining". In: *Briefings in bioinformatics* 23.6 (2022), bbac409.

[31] OpenAI et al. "GPT-4o System Card". In: *arXiv preprint arXiv:2410.21276* (2024).

[32] Maxime Oquab et al. "DINOv2: Learning robust visual features without supervision". In: *arXiv preprint arXiv:2304.07193* (2023).

[33] Qwen et al. "Qwen2.5 Technical Report". In: *arXiv preprint arXiv:2412.15115* (2024).

[34] George Shaikovski et al. "Prism: A multi-modal generative foundation model for slide-level histopathology". In: *arXiv preprint arXiv:2405.10254* (2024).

[35] Yuxuan Sun et al. "CPath-Omni: A Unified Multimodal Foundation Model for Patch and Whole Slide Image Analysis in Computational Pathology". In: *arXiv preprint arXiv:2412.12077* (2024).

[36] Gemini Team et al. "Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context". In: *arXiv preprint arXiv:2403.05530* (2024).

[37] Manuel Tran et al. "Generating dermatopathology reports from gigapixel whole slide images with HistoGPT". In: *Nature Communications* 16.1 (2025), p. 4886.

[38] Anurag Vaidya et al. "Molecular-driven Foundation Model for Oncologic Pathology". In: *arXiv preprint arXiv:2501.16652* (2025).

[39] Eugene Vorontsov et al. *Virchow: A Million-Slide Digital Pathology Foundation Model*. 2024. arXiv: 2309.07778 [eess.IV].

[40] Xiyue Wang et al. "Transformer-based unsupervised contrastive learning for histopathological image classification". In: *Medical image analysis* 81 (2022), p. 102559.

[41] Hanwen Xu et al. "A whole-slide foundation model for digital pathology from real-world data". In: *Nature* 630.8015 (2024), pp. 181–188.

[42] Yingxue Xu et al. "A multimodal knowledge-enhanced whole-slide pathology foundation model". In: *arXiv preprint arXiv:2407.15362* (2024).

[43] Jiahui Yu et al. "Coca: Contrastive captioners are image-text foundation models". In: *arXiv preprint arXiv:2205.01917* (2022).

[44] Eric Zimmermann et al. "Virchow2: Scaling self-supervised mixed magnification models in pathology". In: *arXiv preprint arXiv:2408.00738* (2024).

A   Appendix

A.1   Evaluation tasks and data

To evaluate the quality of model embeddings, we train linear classifiers on a variety of tasks. Specifically, given a training, validation, and testing subset of data, we use the L-BFGS solver from scikit learn to train a linear classifier to convergence with an ℓ_2 norm regularization term selected according to the lowest loss on the validation subset.

We evaluate the embeddings of PRISM2 and baseline models on a range of cancer detection, cancer subtyping, and molecular tasks described below.

Pan-cancer Detection is an in-house evaluation dataset containing 22,932 H&E WSIs collected from 6,142 specimens. Samples are sourced in near-equal quantities from Memorial Sloan Kettering Cancer Center (49%) and global external sites (51%). Of the 16 cancers included in the dataset, seven are rare. The National Cancer Institute (NCI) defines rare cancers as those with an annual incidence in the United States below 15 people per 100,000. Complimenting the evaluation dataset, the training and validation sets together contain 89,417 slides across 40,402 specimens and follow a more natural distribution for the frequency of different cancers.

Breast Subtyping is an in-house dataset comprising 11,,774 breast H&E WSIs spanning 2407 specimens. Each specimen is associated with six non-mutually exclusive labels indicating the presence or absence of invasive lobular carcinoma (ILC), invasive ductal carcinoma (IDC), ductal carcinoma in situ (DCIS), lobular carcinoma in situ (LCIS), atypical ductal hyperpalasia (ADH), and atypical lobular hyperplasia (ALH). The dataset is divided at a specimen level into 8 folds with a 75/12.5/12.5 percentage split between train, tune, and test subsets.

MSK GI is an in-house dataset containing 226,024 gastrointestinal (GI) H&E WSIs gathered from 122,601 specimens. The labels span 12 non-mutually exclusive binary classification tasks. The dataset is split into train, tune, and test subsets of size 91276, 17515, and 13810 specimens respectively.

TCGA-BRCA is a publicly available invasive breast cancer (BRCA) dataset focused on discriminating between IDC and ILC. It comprises 1002 H&E WSIs from TCGA. The data is split into five folds with a 60/20/20 ratio between the train, tune, and test subsets for each fold.

TCGA-NSCLC is a publicly available non-small cell lung cancer (NSCLC) dataset composed of 1043 H&E WSIs from TCGA. Each WSI contains lung adenocarcinoma (LUAD) or lung squamous cell carcinoma (LUSC). The data is split into five folds with a 60/20/20 ratio between the train, tune, and test subsets for each fold.

TCGA-RCC is a publicly available renal cell carcinoma (RCC) dataset containing 938 H&E WSIs. Each slide belongs to one of three ground truth subtypes: clear cell renal cell carcinoma (ccrcc), papillary renal cell carcinoma (prcc), or chromophobe renal cell carcinoma (chrcc). The data is split into five folds with a 60/20/20 ratio between the train, tune, and test subsets for each fold.

Biomarker prediction is an in-house dataset containing 37,633 WSIs gathered from 33,949 specimens. The ground truth labels span 10 different biomarkers: Prostate AR, Ovarian FGA, Esophagogastric Her2, Colorectal MSI, Lung EGFR, Melanoma BRAF, Bladder FGFR, Breast CDH1, Endometrial PTEN, and Breast PIK3CA. These labels were originally identified by MSK-IMPACT [13], a targeted test for genetic mutations. The dataset is split into train, tune, and test subsets of size 22,634, 5,606, and 5,709 specimens respectively.

A.2 Supplementary results

Model	Method	PanCancer	PanCancer Rare	Breast Subtyping	GI Suite
TITAN	Contrastive	0.805 (0.794, 0.815)	0.767 (0.750, 0.785)	-	-
PRISM	Contrastive	0.896 (0.888, 0.903)	0.851 (0.835, 0.865)	-	-
PRISM2	Contrastive	0.805 (0.794, 0.815)	0.804 (0.787, 0.820)	-	-
	Dialogue (YN)	**0.933** (0.926, 0.939)	**0.929** (0.920, 0.938)	**0.958** (0.948, 0.967)	**0.932** (0.921, 0.942)
	Dialogue (DYN)	**0.943** (0.938, 0.949)	**0.943** (0.934, 0.951)	**0.956** (0.943, 0.967)	**0.929** (0.917, 0.941)

(a) ROC AUC for binary classification tasks based on MSK slides.

Model	Method	BRCA	NSCLC	RCC
TITAN	Contrastive	**0.953** (0.935, 0.969)	**0.985** (0.978, 0.991)	**0.978** (0.970, 0.985)
PRISM	Contrastive	**0.935** (0.911, 0.956)	0.969 (0.960, 0.978)	0.919 (0.905, 0.933)
PRISM2	Contrastive	**0.949** (0.926, 0.967)	**0.980** (0.973, 0.986)	**0.981** (0.975, 0.987)
	Dialogue (OPN)	**0.963** (0.946, 0.977)	**0.977** (0.968, 0.985)	**0.978** (0.970, 0.984)

(b) ROC AUC for multi-class classification tasks based on TCGA slides.

Table A.2.1: Zero-shot ROC AUC for various downstream classification tasks using dialogue completions. 95% confidence intervals computed with 1000 iterations of bootstrapping are shown in parentheses. The best result per task ($p < 0.05$) is in bold.

PRISM2 Dialogue	ROC AUC		Balanced Accuracy	
	YN	DYN	YN	DYN
Invasive ductal carcinoma (IDC)	**0.984** (0.978, 0.989)	**0.979** (0.972, 0.985)	**0.954** (0.944, 0.964)	**0.946** (0.934, 0.957)
Invasive lobular carcinoma (ILC)	**0.983** (0.976, 0.989)	**0.982** (0.961, 0.994)	**0.863** (0.811, 0.917)	**0.872** (0.821, 0.920)
Ductal carcinoma in situ (DCIS)	**0.970** (0.963, 0.976)	**0.963** (0.955, 0.971)	**0.913** (0.900, 0.926)	**0.922** (0.910, 0.933)
Lobular carcinoma in situ (LCIS)	**0.960** (0.949, 0.969)	**0.964** (0.951, 0.974)	**0.796** (0.765, 0.826)	**0.778** (0.746, 0.808)
Atypical ductal hyperplasia (ADH)	**0.896** (0.877, 0.914)	**0.899** (0.882, 0.915)	**0.786** (0.752, 0.817)	0.673 (0.638, 0.706)
Atypical lobular hyperplasia (ALH)	**0.927** (0.909, 0.944)	**0.936** (0.918, 0.952)	**0.811** (0.779, 0.843)	0.717 (0.684, 0.749)
Invasive carcinoma	**0.979** (0.974, 0.984)	0.965 (0.956, 0.973)	**0.936** (0.925, 0.946)	**0.940** (0.929, 0.950)
Benign or unremarkable	**0.965** (0.957, 0.973)	**0.960** (0.951, 0.968)	**0.930** (0.920, 0.940)	**0.924** (0.913, 0.934)
Average	**0.958** (0.948, 0.967)	**0.956** (0.943, 0.967)	**0.874** (0.849, 0.897)	**0.847** (0.822, 0.870)

Table A.2.2: Zero-shot AUC and balanced accuracy of detection tasks using dialogue completions in breast tissue, spanning cancers (IDC, ILC) and precursors to cancer (DCIS, LCIS, ADH, ALH). 95% confidence intervals computed with 1000 iterations of bootstrapping are shown in parentheses. The best result per task ($p < 0.05$) is in bold.

PRISM2 Dialogue	ROC AUC		Balanced Accuracy	
	YN	DYN	YN	DYN
CANCER				
Adenocarcinoma	**0.992** (0.990, 0.994)	**0.992** (0.990, 0.994)	**0.966** (0.960, 0.971)	0.936 (0.926, 0.945)
Lymphoma	**0.889** (0.865, 0.910)	**0.906** (0.881, 0.929)	0.663 (0.629, 0.699)	**0.784** (0.748, 0.820)
Metastatic Tumor	**0.973** (0.958, 0.987)	**0.966** (0.948, 0.982)	**0.894** (0.864, 0.923)	0.829 (0.792, 0.863)
Neuroendocrine Tumor	**0.964** (0.950, 0.976)	**0.951** (0.932, 0.967)	**0.883** (0.856, 0.908)	**0.873** (0.844, 0.901)
Squamous Cell Carcinoma	**0.995** (0.989, 0.998)	**0.994** (0.987, 0.998)	**0.949** (0.932, 0.967)	**0.923** (0.899, 0.946)
Tumor Differentiation	**0.991** (0.988, 0.993)	**0.993** (0.990, 0.995)	**0.954** (0.946, 0.961)	**0.942** (0.933, 0.951)
PRECURSOR OR BENIGN				
Dysplasia	**0.802** (0.783, 0.820)	**0.781** (0.762, 0.800)	**0.713** (0.695, 0.729)	**0.684** (0.664, 0.701)
Polyp	0.945 (0.940, 0.949)	**0.957** (0.953, 0.961)	0.790 (0.782, 0.799)	**0.849** (0.841, 0.857)
No Significant Abnormalities	0.866 (0.860, 0.872)	**0.906** (0.901, 0.911)	**0.599** (0.595, 0.604)	0.560 (0.556, 0.564)
INFLAMMATORY CONDITION				
H. Pylori Helicobacter	**0.932** (0.908, 0.953)	0.892 (0.859, 0.921)	0.536 (0.518, 0.556)	**0.823** (0.788, 0.859)
Barretts Esophagus	**0.972** (0.964, 0.979)	0.946 (0.935, 0.956)	**0.872** (0.857, 0.886)	0.831 (0.815, 0.848)
Other Inflammatory Condition	**0.864** (0.856, 0.872)	**0.870** (0.862, 0.878)	0.748 (0.739, 0.758)	**0.750** (0.740, 0.758)
Average	**0.932** (0.921, 0.942)	**0.929** (0.917, 0.941)	**0.797** (0.781, 0.813)	**0.815** (0.796, 0.834)

Table A.2.3: Zero-shot AUC and balanced accuracy of detection tasks using dialogue completions in the gastrointestinal (GI) tract, spanning cancers, precursors to cancer, inflammatory conditions, and benign conditions. 95% confidence intervals computed with 1000 iterations of bootstrapping are shown in parentheses. The best result per task ($p < 0.05$) is in bold.

Linear probing	PRISM2 Diagnostic	PRISM2 Base	PRISM	TITAN
Invasive ductal carcinoma (IDC)	**0.993** (0.990, 0.995)	0.985 (0.978, 0.991)	0.988 (0.982, 0.992)	0.954 (0.944, 0.963)
Invasive lobular carcinoma (ILC)	**0.993** (0.988, 0.996)	0.987 (0.979, 0.994)	**0.991** (0.979, 0.997)	0.958 (0.934, 0.977)
Ductal carcinoma in situ (DCIS)	**0.985** (0.980, 0.989)	**0.983** (0.978, 0.987)	0.982 (0.976, 0.986)	0.938 (0.927, 0.946)
Lobular carcinoma in situ (LCIS)	**0.973** (0.962, 0.981)	**0.973** (0.961, 0.981)	0.971 (0.961, 0.979)	0.893 (0.870, 0.914)
Atypical ductal hyperplasia (ADH)	**0.947** (0.931, 0.961)	**0.943** (0.927, 0.956)	0.944 (0.928, 0.958)	0.865 (0.836, 0.890)
Atypical lobular hyperplasia (ALH)	**0.972** (0.956, 0.984)	0.965 (0.947, 0.979)	0.960 (0.943, 0.976)	0.887 (0.860, 0.910)
Average	**0.979** (0.976, 0.982)	0.976 (0.972, 0.979)	0.975 (0.971, 0.978)	0.923 (0.916, 0.929)

Table A.2.4: Linear probing AUC of detection tasks in breast tissue, spanning cancers (IDC, ILC and precursors to cancer (DCIS, LCIS, ADH, ALH). 95% confidence intervals computed with 1000 iterations of bootstrapping are shown in parentheses. The best result per task ($p < 0.05$) is in bold.

Linear probing	PRISM2 Diagnostic	PRISM2 Base	PRISM	TITAN
CANCER				
Adenocarcinoma	**0.995** (0.993, 0.997)	**0.995** (0.994, 0.997)	**0.994** (0.991, 0.996)	**0.993** (0.991, 0.995)
Lymphoma	**0.980** (0.970, 0.989)	**0.979** (0.967, 0.990)	**0.971** (0.957, 0.983)	**0.970** (0.955, 0.982)
Metastatic Tumor	**0.989** (0.978, 0.997)	**0.987** (0.977, 0.995)	**0.978** (0.964, 0.990)	0.971 (0.956, 0.985)
Neuroendocrine Tumor	**0.983** (0.970, 0.992)	**0.987** (0.978, 0.993)	**0.981** (0.968, 0.990)	**0.977** (0.962, 0.989)
Squamous Cell Carcinoma	**0.997** (0.994, 0.999)	**0.995** (0.989, 0.999)	**0.994** (0.988, 0.998)	**0.994** (0.984, 0.999)
Tumor Differentiation	**0.999** (0.998, 0.999)	**0.999** (0.998, 0.999)	**0.998** (0.997, 0.998)	**0.998** (0.997, 0.998)
PRECURSOR OR BENIGN				
Dysplasia	**0.947** (0.938, 0.956)	**0.949** (0.940, 0.957)	0.931 (0.921, 0.941)	0.912 (0.900, 0.925)
Polyp	**0.977** (0.975, 0.980)	**0.978** (0.975, 0.980)	0.967 (0.963, 0.970)	0.964 (0.960, 0.967)
No Significant Abnormalities	**0.947** (0.944, 0.951)	**0.947** (0.943, 0.951)	0.942 (0.938, 0.946)	0.937 (0.933, 0.941)
INFLAMMATORY CONDITION				
H. Pylori Helicobacter	0.961 (0.938, 0.979)	**0.974** (0.957, 0.987)	0.964 (0.946, 0.979)	0.962 (0.939, 0.980)
Barretts Esophagus	0.988 (0.985, 0.991)	**0.990** (0.987, 0.992)	0.989 (0.986, 0.992)	0.962 (0.954, 0.968)
Other Inflammatory Condition	**0.957** (0.953, 0.961)	**0.957** (0.953, 0.961)	0.945 (0.940, 0.949)	**0.962** (0.939, 0.980)
Average	**0.977** (0.974, 0.979)	**0.978** (0.976, 0.980)	0.971 (0.968, 0.973)	0.965 (0.961, 0.968)

Table A.2.5: Linear probing AUC of detection tasks in the gastrointestinal (GI) tract, spanning cancers, precursors to cancer, inflammatory conditions, and benign conditions. 95% confidence intervals computed with 1000 iterations of bootstrapping are shown in parentheses. The best result per task ($p < 0.05$) is in bold.

Linear probing	PRISM2 Diagnostic	PRISM2 Base	PRISM	TITAN
Prostate AR	<u>0.912</u> (0.887, 0.936)	**0.917** (0.893, 0.938)	0.897 (0.869, 0.921)	0.911 (0.886, 0.933)
Ovarian FGA	<u>0.877</u> (0.799, 0.939)	**0.891** (0.824, 0.948)	0.847 (0.758, 0.920)	0.859 (0.751, 0.940)
Esophagogastric Her2	<u>0.832</u> (0.752, 0.902)	0.831 (0.748, 0.905)	0.808 (0.707, 0.897)	**0.849** (0.756, 0.925)
Colorectal MSI	0.948 (0.920, 0.970)	**0.964** (0.948, 0.979)	<u>0.958</u> (0.935, 0.977)	0.923 (0.893, 0.951)
Lung EGFR	<u>0.814</u> (0.783, 0.841)	**0.826** (0.799, 0.852)	0.810 (0.782, 0.837)	0.813 (0.784, 0.840)
Melanoma BRAF	0.722 (0.663, 0.784)	**0.768** (0.710, 0.823)	0.735 (0.674, 0.791)	<u>0.757</u> (0.701, 0.814)
Bladder FGFR	0.884 (0.845, 0.920)	0.897 (0.861, 0.927)	<u>0.895</u> (0.858, 0.929)	**0.909** (0.873, 0.943)
Breast CDH1	<u>0.915</u> (0.889, 0.938)	**0.917** (0.890, 0.939)	0.913 (0.888, 0.935)	0.897 (0.870, 0.921)
Endometrial PTEN	0.870 (0.838, 0.900)	**0.889** (0.859, 0.917)	0.880 (0.847, 0.908)	<u>0.881</u> (0.849, 0.909)
Breast PIK3CA	0.674 (0.644, 0.708)	<u>0.677</u> (0.647, 0.709)	**0.686** (0.655, 0.719)	0.658 (0.627, 0.689)
Average	0.845 (0.830, 0.859)	**0.857** (0.844, 0.870)	0.843 (0.828, 0.857)	<u>0.846</u> (0.829, 0.860)

Table A.2.6: Linear probing AUC of biomarker detection per biomarker task. Best results in bold and second best underlined. 95% confidence intervals computed with 1000 bootstrapping iterations are shown in parentheses. With the small data samples, there is not enough statistical power to conclude statistically significant differences; however, the PRISM2 Base embedding tends to outperform all others 7 out of 10 times.

A.3   Zero-shot dialogue completions and contrastive prompts

Task	Prompt
PanCancer (all and rare)	
Cancer	Is malignant tumor identified in the specimen? {**Yes**,**No**}
Breast Subtyping	
Invasive ductal carcinoma	Is invasive ductal carcinoma (IDC) identified in the specimen? {**Yes**,**No**}
Invasive lobular carcinoma	Is invasive lobular carcinoma (ILC) identified in the specimen? {**Yes**,**No**}
Ductal carcinoma in situ	Is ductal carcinoma in situ (DCIS) identified in the specimen? {**Yes**,**No**}
Lobular carcinoma in situ	Is lobular carcinoma in situ (LCIS) identified in the specimen? {**Yes**,**No**}
Atypical ductal hyperplasia	Is atypical ductal hyperplasia (ADH) identified in the specimen? {**Yes**,**No**}
Atypical lobular hyperplasia	Is atypical lobular hyperplasia (ALH) identified in the specimen? {**Yes**,**No**}
Invasive carcinoma	Is invasive carcinoma identified in the specimen? {**Yes**,**No**}
Benign or unremarkable	Is the specimen benign or unremarkable? {**Yes**,**No**}
GI Suite	
Adenocarcinoma	Is adenocarcinoma identified in the specimen? {**Yes**,**No**}
Lymphoma	Is lymphoma identified in the specimen? {**Yes**,**No**}
Metastatic Tumor	Is metastatic tumor identified in the specimen? {**Yes**,**No**}
Neuroendocrine Tumor	Is neuroendocrine tumor identified in the specimen? {**Yes**,**No**}
Squamous Cell Carcinoma	Is squamous cell carcinoma identified in the specimen? {**Yes**,**No**}
Tumor Differentiation	Is malignant tumor identified in the specimen? {**Yes**,**No**}
Dysplasia	Is dysplasia identified in the specimen? {**Yes**,**No**}
Polyp	Is polyp identified in the specimen? {**Yes**,**No**}
No Significant Abnormalities	Is there no significant abnormalities identified in the specimen? {**Yes**,**No**}
H. Pylori Helicobacter	Is h pylori helicobacter identified in the specimen? {**Yes**,**No**}
Barretts Esophagus	Is intestinal metaplasia (Barrett's esophagus) identified in the specimen? {**Yes**,**No**}
Other Inflammatory Condition	Is inflammatory condition identified in the specimen? {**Yes**,**No**}

Table A.3.1: Dialogue completions used for zero-shot evaluation on binary classification tasks. For each task, the language model is prompted with a question. Log probabilities of two selected completions - **Yes** and **No** answers - are ranked against each other, with **Yes** acting as a positive class prediction.

Class	Prompt
	TCGA BRCA
	What type of carcinoma is suggested by the specimen, and which subtype is favored?
IDC	The specimen contains invasive **ductal** carcinoma.
ILC	The specimen contains invasive **lobular** carcinoma.
	TCGA NSCLC
	What type of carcinoma is suggested by the specimen, and which subtype is favored?
LUAD	The specimen suggests non-small cell carcinoma, with a preference towards **adenocarcinoma**.
LUSC	The specimen suggests non-small cell carcinoma, with a preference towards **squamous cell carcinoma**.
	TCGA RCC
	What type of carcinoma is suggested by the specimen, and which subtype is favored?
CCRCC	The specimen reveals renal cell carcinoma of the **clear cell** type.
PRCC	The specimen reveals renal cell carcinoma of the **papillary** type.
CHRCC	The specimen reveals renal cell carcinoma of the **chromophobe** type.

Table A.3.2: Dialogue completions used for zero-shot evaluation on multi-class classification tasks. For each task, the language model is prompted with a question and a class-specific completion. Log probabilities of the tokens relevant to the class (highlighted in bold) are ranked against each other.

Class	Prompt
	PanCancer (all and rare)
Positive	cancer
	carcinoma
	adenocarcinoma
	malignant
	metastatic
	invasive
	sarcoma
Negative	benign
	inflammation
	normal
	infectious
	cyst
	polyp
	unremarkable
	PanCancer (breast subset)
Positive	invasive carcinoma is identified in the specimen
	invasive ductal carcinoma (IDC) is identified in the specimen
	invasive lobular carcinoma (ILC) is identified in the specimen
Negative	ductal carcinoma in situ (DCIS) is identified in the specimen
	lobular carcinoma in situ (LCIS) is identified in the specimen
	atypical ductal hyperplasia (ADH) is identified in the specimen
	atypical lobular hyperplasia (ALH) is identified in the specimen
	the specimen is benign or unremarkable

Table A.3.3: Positive- and negative-class prompts used for contrastive zero-shot prediction on the PanCancer dataset. Zero-shot performance is highly sensitive to the prompts used. Using the breast-specific prompts substantially improves PRISM2 contrastive zero-shot performance on the breast subset.

A.4 GPT-4o rewriting and question-answer pair generation prompts

Clinical history, specimen, and report rewriting prompt:

```
# Role

You are a pathology assistant whose goal is to help rewrite pathology reports in order to be used
    for in training a whole slide image captioning AI model.

# Instruction

You are given a pathology report describing a tissue sample and sometimes a clinical history and
    description of the specimen under review. You need to rewrite the report as if you're examining
     the sample yourself using natural language. Follow these instructions carefully and to the
    letter:

- Rewrite the entire report in natural language in the order of appearance (i.e. clinical history
    -> specimen description -> diagnosis).
- Separate the sections into their own key/value pairs: "history", "specimen", "diagnosis".
- If any section is not present, fill with null value.

## History section instruction
Use the information from the "History" section provided by the user to fill in the "history"
    section of your output. Follow these instructions carefully to do this:

- Never include dates and times or names of people/doctors/clinics.
- Do not mention if the patient had previous appointments.
- If age and/or gender of the patient is provided, retain that information in your rewrite.

## Specimen section instruction
Use the information from the "Specimen" section provided by the user to fill in the "specimen"
    section of your output. Follow these instructions to achieve that:

- Rewrite the specimen description to capture specimen type and procedure used to obtain the
    specimen and any other relevant information.

## Diagnosis section instruction
Use the information from the "Diagnosis" and "Addenda" sections provided by the user to fill in the
     "diagnosis" section of your output. You must follow these instruction with extreme attention
    to detail and care:

- Start with the most clinically significant finding followed by the findings of lesser
    significance.
- Include relevant morphological details that can be observed from a whole slide image, such as
    cell types, structures, and patterns if they are provided.
- Do not mention anything related to the patient's history or prior samples or material.
- Do not any specific macroscopic locations within the specimen (i.e. exclude things like right,
    left, apex, upper quadrant, mid to base, etc.).
- Do not mention the number of specimens examined (i.e. don't talk about number of lymph nodes
    inspected)
- Only talk about what could be inferred from looking at a whole slide image of the specimen.
- Exclude results or mentions of immunohistochemical (IHC) studies, molecular assays, or in situ
    hybridization.
- Never mention physical measurements or quantities using mm or cm or inches, etc.
- Only mention the results that are obvious from the report.
- Include relevant info from "Note" or "Addenda" that follows these instructions.
- Do not use language that references the given report.
- The language should read as if you're a pathologist inspecting the specimen and actively
    diagnosing the sample.

# Examples

## Example of a diagnosis rewrite that properly excludes macroscopic locations
```
```

```
User: Diagnosis: Metastatic carcinoma involving right ovary and left fallopian tube, unremarkable
    left ovary and right fallopian tube
Assistant: "diagnosis": "Metastatic carcinoma is observed in the ovaries and fallopian tube."
```


Examples of properly excluding immunohistochemistry (IHC) information in diagnosis
```
User: Diagnosis: Invasive Carcinoma Ductal, NOS type. Progesterone Receptor Negative (no nuclear
    staining). HER2 Negative (0).
Assistant: "diagnosis": "There is invasive ductal carcinoma present."
```


Example of a diagnosis rewrite that properly omits quantifying number of specimens examined
```
User: Diagnosis: Lymph Nodes Number with metastasis: 2
Assistant: "diagnosis": "Metastasis is present in lymph nodes."
```


Example of a diagnosis rewrite that properly omits mention of a patient's prior samples
```
User: Diagnosis: Metastatic mammary carcinoma, morphologically similar to the patient's
prior specimen.
Assistant: "diagnosis": "Metastatic mammary carcinoma is detected."
```


Example of a diagnosis rewrite that properly excludes measurements
```
User: Diagnosis: Keratinizing dysplasia, moderate.  Depth of invasion is 9 mm.
Assistant: "diagnosis": "Examination reveals moderate keratinizing dysplasia."
```


Output
Generate five distinct summaries with a variety of linguistic styles with proper punctuation but
    still accurate to the provided information. Output the resulting 5 summaries as a list in JSON
    format. Don't output anything else. Do not wrap the output in markdown. The JSON structure
    should be formatted like:

{reports: [{"history": ..., "specimen": ..., "diagnosis":...}, ...]}

List of findings prompt:

```
You are an AI pathology assistant. You are helping process reports from whole slide images.

Please extract phrases from the pathology report which refer to tissue, diagnostic findings,
    margins, or prior treatment visible in a histological whole slide image, or the absence of such
    .

Rules:
- Break down the report into multiple findings
- List the findings in multiple lines
- Each unique finding in the report has its own line
- Each line contains only one unique finding
- Findings are written in one sentence in naturally-sounding language
- Sentences are short and contain only one unique histological finding
- If there are more findings in one line, split them into multiple lines with one finding per line
- Findings don't contain numbers
- Only keep findings that are related to visible features on a histological whole slide image
- Exclude physical measurements (e.g. "...1.9 cm") or percentages (e.g. "...5%").
- Exclude macroscopic locations (e.g. "left", "right", "sigmoid").
- Exclude quantities referring to number of slides or specimens examined (e.g. "5 lymph nodes", "3
    consecutive slides")
- Exclude immunohistochemical (IHC) and molecular tests.

The objective is to extract phrases which refer to things which can be located on a histological
    whole slide image, or confirmed not to be present.
```

Yes-No question-answer pair generation prompt:

```
You are a pathology instructor at a world class medical school. Create one yes and one no question/
    answer pair for each line in a given report. Do not reference the given report in your wording.
     Imagine the student/trainee will be looking at the specimen/sample and needs to answer these
    questions. Use proper punctuation.

Return them individually as a JSON:

```
{
  "1": [{"question": str, "answer": str}, {"question": str, "answer": str}],
  "2": [{"question": str, "answer": str}, {"question": str, "answer": str}],
  etc.
}
```
```

Complementary question-answer pair generation prompt:

```
You are given a pathology report and a series of the questions. Answer the questions with just "Yes
    ." or "No." based on the report.

Output your results in JSON format with the original questions and group every 2 questions together
    :

```json
{
    "1": [{"question": str, "answer": str}, {"question": str, "answer": str}],
    etc.
}
```
```

Open-ended question-answer pair generation prompt:

```
You are a pathology instructor at a world class medical school. Create one open-ended question/
    answer pair for each line in a given report, where answer is provided by the given line. The
    question should not be answerable with simply yes/no. Re-word the answer to introduce diverse
    linguistic patterns. Imagine the student/trainee will be looking at the specimen/sample and
    needs to answer these questions with no prior knowledge of the given report. Never use the word
     "noted". Use proper punctuation.

Return them individually as a JSON:

```
{
  "1": {"question": str, "answer": str},
  "2": {"question": str, "answer": str},
  etc.
}
```
```

**A.5  Acronyms**

| | |
|---|---|
| **AUC** | area under the curve |
| **GI** | gastrointestinal |
| **H&E** | hematoxylin and eosin |
| **LLM** | large language model |
| **MLP** | multi-layer perceptron |
| **MSKCC** | Memorial Sloan Kettering Cancer Center |
| **TCGA** | The Cancer Genome Atlas |
| **WSI** | whole slide image |
| **DCIS** | ductal carcinoma in situ |
| **QA** | question/answer |