Generating highly accurate pathology reports from gigapixel whole slide images with HistoGPT

Manuel Tran[1,2,*], Paul Schmidle[3,*], Sophia J. Wagner[1,2], Valentin Koch[4,2], Valerio Lupperger[5], Annette Feuchtinger[6], Alexander Böhner[7], Robert Kaczmarczyk[7], Tilo Biedermann[7], Kilian Eyerich[8,†], Stephan A. Braun[3,†], Tingying Peng[1,†], Carsten Marr[1,4,†,#]

[1] Helmholtz AI, Helmholtz Munich, Neuherberg, Germany
[2] School of Computation, Information and Technology, Technical University of Munich, Munich, Germany
[3] Dermatology Department, University Hospital Münster, Münster, Germany
[4] Institute of AI for Health, Helmholtz Munich, Neuherberg, Germany
[5] MLL Munich Leukemia Laboratory, Munich, Germany
[6] Core Facility Pathology and Tissue Analytics, Helmholtz Munich, Neuherberg, Germany
[7] Department of Dermatology and Allergy, Technical University of Munich, Munich, Germany
[8] Department of Dermatology, Medical Center, University of Freiburg, Freiburg, Germany
* These authors contributed equally, † Corresponding authors, # Lead contact carsten.marr@helmholtz-munich.de

Highlights

- A large vision language model is trained to generate dermatopathology reports
- It takes as input multiple whole slide images and outputs tissue descriptions
- Generated reports match human-written reports as confirmed by pathologists
- It predicts tumor subtypes and thickness zero-shot better than current methods

Summary

Histopathology is considered the gold standard for determining the presence and nature of disease, particularly cancer. However, the process of analyzing tissue samples and producing a final pathology report is time-consuming, labor-intensive, and non-standardized. Therefore, new technological solutions are being sought to reduce the workload of pathologists. In this work, we present HistoGPT, a vision language model that takes digitized slides as input and generates reports that match the quality of human-written reports, as confirmed by natural language processing metrics and domain expert evaluations. We show that HistoGPT generalizes to five international cohorts and can predict tumor subtypes and tumor thickness in a zero-shot fashion. Our work represents an important step toward integrating AI into the medical workflow. We publish both model code and weights so that the scientific community can apply and improve HistoGPT to advance the field of computational pathology.

Keywords: Artificial intelligence, natural language processing, computer vision, large language models, vision foundation models, few-shot learning, computational pathology, whole slide images, histopathology reports, cancer diagnosis, dermatopathology

Introduction

Histopathology is the study of diseased tissues and cells under the microscope. It plays a critical role in the diagnosis of many diseases, including malignant cancers, viral infections, and inflammatory responses [1]. In many cases, the detailed analysis provided by histopathological examinations remains the diagnostic gold standard [2]. It involves the analysis of slides by pathologists, the dictation of findings, and the writing of the report. However, this process is time-consuming and labor-intensive [3]. The turnaround time for patients is likely to worsen in the future as the number of pathologists is decreasing at an alarming rate. Combined with an increase in tumor cases in an aging society, the workload for pathologists is unsustainable [4].

Artificial Intelligence (AI) offers a potential solution to handle frequent and uncomplicated diagnoses and effectively assist medical professionals in their daily routines by using advanced tools such as deep neural networks (DNNs) [5]. These brain-inspired systems are typically applied to digitized microscope slides, also known as whole slide images (WSIs). Modern deep learning (DL) techniques allow to effectively automate several tasks, including cancer classification [6], tissue segmentation [7], survival prediction [8], and biomarker detection [9]. These approaches have already shown promising results and could reduce the burden on pathologists in today's medical landscape [10].

A major drawback of current methods is that they are typically limited to a narrow task, providing only a single scalar output for each input. Consider, for example, an image classification model for benign versus malignant tissue. Beyond predicting these two labels, the model cannot do anything else: neither solve new unseen problems (called zero-shot prediction) nor provide its reasoning steps for better explainability. Vision language foundation models offer an exciting alternative to these rigid approaches by processing both images and text simultaneously. However, due to methodological limitations, current multimodal AI algorithms [11–16] can only process small image patches of 224 x 224 pixels, or regions of interest (ROIs) of 1024 x 1024 pixels. These so-called patch-based approaches are suboptimal because they are limited to a tiny fraction of the WSI, ignoring potentially relevant areas in the remaining tissue sample.

Here, we present HistoGPT, a vision language model (VLM) that can generate histopathology reports from gigapixel WSIs (see Figure 1) with impressive quality. Given a slide, the model uses a vision foundation model (VFM) to extract meaningful visual features from the tissue sample and combines them with a large language model (LLM) via cross-attention mechanisms to generate the final report. The generated report describes the WSI with high fidelity, explaining tissue composition, cellular subtypes, and potential diagnoses. In an unprecedented way, users can interact with the model through various prompts ("Expert guidance") to extract additional information such as tumor subtypes and tumor thickness. To make the output text interpretable, HistoGPT provides saliency maps that

highlight the corresponding image regions that led to the specific findings in the generated text – providing an insightful and detailed understanding not possible before.

In our experiments, HistoGPT outperforms a state-of-the-art biomedical language model for text generation [17], a general-purpose multimodal AI system for image understanding [18], various multiple instance learning (MIL) approaches for image classification [9,19,20], and different contrastive methods [12,13,16] for zero-shot prediction. We demonstrate that a slide-level model is necessary for high accuracy by training two novel contrastive pre-trained baselines we call HistoCLIP and HistoSigLIP. Both outperform the patch-level foundation model PLIP [13] on slide-level tasks and are only surpassed by the generative pre-trained HistoGPT.

To train HistoGPT, we collect a large multimodal skin histology dataset from the Department of Dermatology at the Technical University of Munich with 6,000 paired WSIs and pathology reports written by board-certified pathologists for each patient case. To validate HistoGPT, we are using one internal and five external publicly available test sets that cover different data distributions in different countries. To democratize the use of AI, we are releasing HistoGPT as an end-to-end deep learning pipeline that can be deployed on local machines. As a result, users can select and fine-tune a copy of our machine learning algorithm according to their needs.

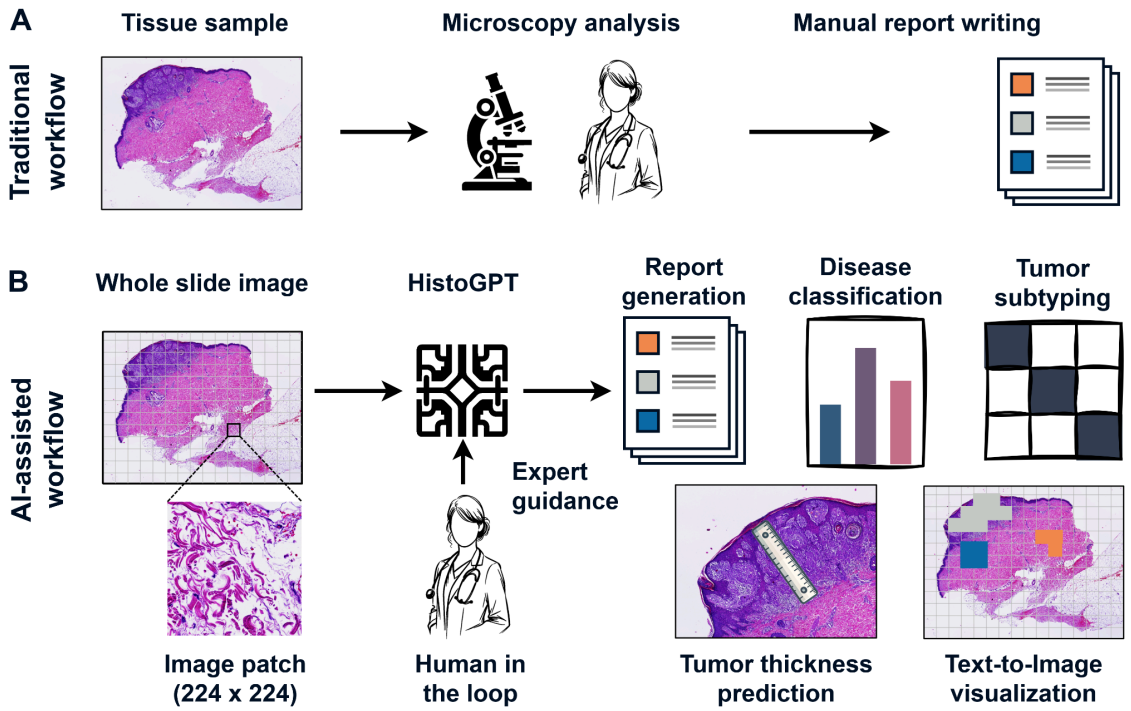


Figure 1: HistoGPT, a vision language foundation model for dermatopathology. (A) Traditionally, pathologists analyze tissue samples from patients under a microscope and summarize their findings in a comprehensive pathology report. This manual process is time-consuming, labor-intensive, and non-standardized. (B) In our proposed AI-powered workflow, pathologists work alongside HistoGPT, our foundation model for vision and language. It generates human-level written reports, provides accurate disease classification, discriminates between tumor subtypes, predicts tumor thickness, and returns text-to-image interpretability maps that provide model explainability. All of this serves as a second opinion to the pathologists, who can query the model for additional information or tailor its output to the task at hand.

Results

HistoGPT simultaneously learns from vision and language

HistoGPT consists of two components (see Figure 2A): a vision foundation model module and a large language model module. The vision module is based on CTransPath [21]. It is a Swin Transformer [22] trained on over 32,000 WSIs from TCGA [23] and PAIP [24] using a semantically guided contrastive learning algorithm. Our language model module repurposes BioGPT [17], an auto-regressive generative model based on the Transformer [25] decoder architecture of GPT-3 [26] trained on 15 million biomedical articles from PubMed. We sample image features from the vision module using a custom pre-trained (see Figure 2B) Perceiver Resampler [27] and integrate it into the LLM via interleaved gated cross-attention (XATTN) blocks [28]. Only these new XATTN blocks are trained from scratch. In this way, we endow HistoGPT with visual and linguistic domain knowledge, which is critical for tackling the challenging problem of generating histopathology reports from entire WSIs. Similar to Flamingo [28], we freeze the parameters of all pre-trained modules during optimization to further reduce the computational cost and to avoid catastrophic forgetting of the inductive biases encoded in the learned weights.

A language model predicts a probability distribution over a vocabulary. The next word in a text is randomly selected based on a combination of top-p and top-k sampling. Once the first few words have been chosen, the outline of the report is roughly pre-determined. To avoid being locked into a fixed report, we use an advanced inference method called Ensemble refinement, introduced in Med-PaLM 2 [29], to randomly sample multiple reports – each focusing on slightly different aspects of the WSI (see Figure 2C). This extensive sampling allows us to thoroughly search the model distribution and generate a wide variety of medical reports, maximizing the likelihood of including all important observations. The general-purpose LLM GPT-4 [18] is then used to summarize all the bootstrapped reports.

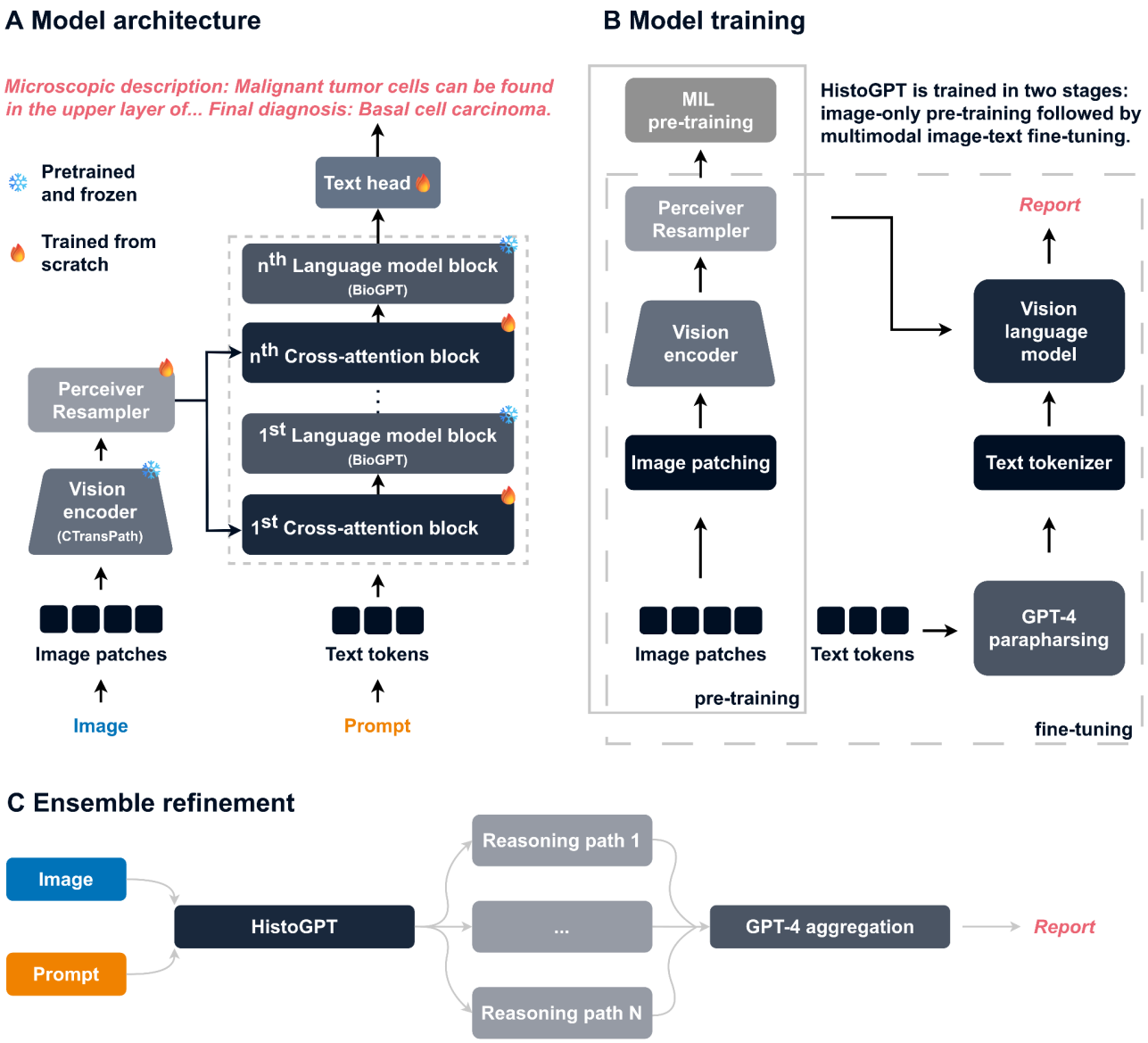


Figure 2: HistoGPT learns simultaneously from vision and language to generate highly accurate histology reports from whole slide images. (A) HistoGPT consists of a vision encoder, a vision resampler, a language model, and cross-attention blocks. Specifically, HistoGPT takes as a whole slide image and outputs written text. Optionally, users can query the model for additional details using prompts such as "tumor thickness". (B) We train HistoGPT in two phases. In the first phase, we pre-train the vision module of HistoGPT using multiple instance learning (MIL). In the second phase, we freeze the pre-trained layers and fine-tune the language module on the image-text pairs. To prevent the model from overfitting on the same sentences, we apply text augmentation. This is done using GPT-4, a general-purpose large language model that faithfully paraphrases the medical notes. (C) During deployment, we propose to optionally use an advanced inference method called Ensemble refinement. Here, the model stochastically generates multiple possible pathology reports via temperature sampling to capture different aspects of the input image. An aggregation module (GPT-4) then combines the results to obtain a more complete description of the underlying case.

HistoGPT generates human-level pathology reports

We train HistoGPT on over 13,000 whole slide images from 6,000 patients with corresponding pathology reports from a real-world cohort provided by the Department of Dermatology at the Technical University of Munich (see Figure 3A). This internal dataset contains 162 different disease classes of varying frequency and has a total size of 10 terabytes. To assess the impact of model architecture and size, we train and evaluate three models: HistoGPT with 1 billion parameters (HistoGPT-1B), HistoGPT with 3 billion parameters (HistoGPT-3B), and HistoGPT-3B with Ensemble Refinement (HistoGPT-3B-ER). In the following experiments, we use HistoGPT in "Expert guidance" mode, where the model is prompted with the correct diagnosis, simulating a pathologist who is confident in the WSI assessment but wants to leave the work of textual tissue description to an AI assistant (see Figure 3B).

Currently, no model can generate a histopathology report from an entire WSI, let alone a series of WSIs (one patient might have multiple tissue samples). Therefore, we compare the reports generated by HistoGPT-1B, HistoGPT-3B, and HistoGPT-3B-ER with those of text-only and patch-only architectures. For the former, we choose the domain-specific language model BioGPT-1B, fine-tuned on our Munich cohort. For the latter, we rely on the multimodal foundation model GPT-4V(ision) [18], which takes low-resolution images of size 2000 x 768 as input. We introduce two other non-trivial baselines: A lower baseline, where we select two random reports with arbitrary diagnoses; and an upper baseline, where we compare two random reports with the same diagnosis (see Methods for more details).

We evaluate the models' output using four semantic-based machine learning metrics: (i) match critical medical terms extracted from the original text with the generated text using a dermatology dictionary; (ii) use the same technique but with ScispaCy, a scientific name entity recognition tool, as the keyword extractor [30]; (iii) compare the semantic meaning of the original and generated reports by measuring the cosine similarity of their text embeddings generated by the biomedical language model BioBERT [31]; (iv) use the same technique but with the general purpose large language model GPT-3-ADA [26] for text embedding (see Supplementary Figure 2 for an illustration).

In "Expert guidance", HistoGPT-1B and HistoGPT-3B capture an average of 64% and 63% of all dermatological keywords from the original pathology reports, respectively (see Figure 3C), outperforming alternative language models such as BioGPT-1B and GPT-4V by at least 5%. HistoGPT-3B-ER further improves the Jaccard index to 77%. This is 10% above the upper baseline. A similar trend is observed when ScispaCy is used as a keyword extractor (see Figure 3C). HistoGPT also produces text with a high cosine similarity with the ground truth, as indicated by the embeddings provided by BioBERT and GPT-3-ADA (see Figure 3C). We also evaluate all models using traditional syntax-based measures (BLEU-4, ROUGE-L, METEOR, and BERTscore). Here, HistoGPT receives relatively low scores (see Supplementary Table: Automatic report evaluation). Combined with the high semantic-based scores (see Figure 3C), this suggests that HistoGPT is not overfitting the training set by simply repeating common phrases and medical terms.

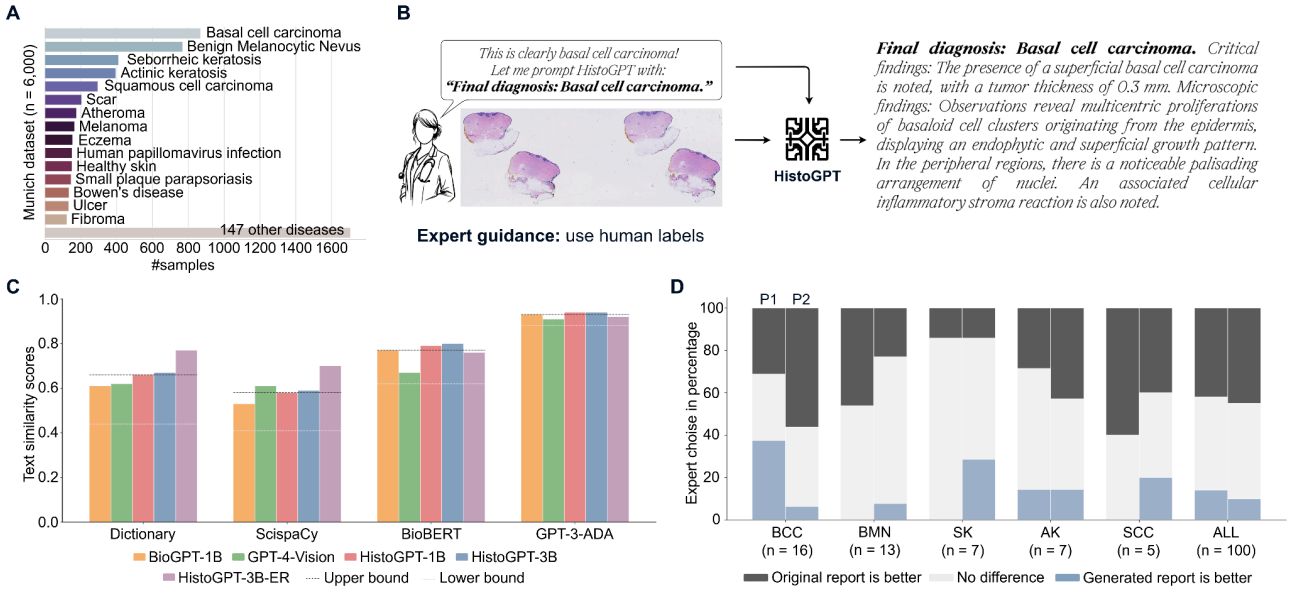


Figure 3. HistoGPT generates human-level pathology reports of skin diseases. (A) Our Munich dataset is a real-world cohort of 6,000 patients with 162 skin diseases from the Department of Dermatology at the Technical University of Munich. It includes malignant cases such as basal cell carcinoma (BCC, n = 870) and squamous cell carcinoma (SCC, n = 297); precursor lesions such as actinic keratosis (AK, n = 396) as well as benign cases such as benign melanocytic nevus (BMN, n = 770) and seborrheic keratosis (SK, n = 412). We divide the patient-level data set into a training set and a test set using a stratified 85/15 split. (B) Through years of experience, pathologists are often able to make a diagnosis at first glance. Instead of writing a pathology report themselves, they can now use HistoGPT in "Expert guidance"' by giving the model the correct diagnosis to complete the report. (C) In "Expert guidance" mode, HistoGPT-3B-ER (HistoGPT-3B with Ensemble Refinement) outperforms BioGPT-1B and GPT-4V on the two text accuracy metrics Dictionary and ScispaCY; and is equal to or better on the two text similarity metrics BioBERT and GPT-3-ADA (see Methods for details). (D) Two independent external pathologists (P1 left and P2 right) evaluated 100 generated and original reports together with the corresponding WSI in a randomized, blinded study. For BCC, P1 found that 38% of the generated reports described the WSI better than the original report. In 31% of cases, both reports performed equally well, while in 31% of cases, the original report was preferred. In 58% (P1) and 55% (P2) of all cases, the pathologists did not prefer the original report to the generated one.

To evaluate the content of the generated reports from an expert perspective, we conduct a blinded study in which we randomly select 100 cases from our Munich test dataset, generate a report for each patient in "Expert guidance" mode, and pair it with the original human-written report. The two reports are then randomly shuffled and anonymized. Two independent expert pathologists (P.S. and S.B.), neither involved in the construction nor annotation of the Munich cohort, were given the original WSIs and asked to identify the report that best describes each case, with the option of selecting "no difference" if both are deemed equally accurate. Ensemble refinement is not used in this study to avoid easy identification of the GPT-4 summarized text. For the five largest diagnostic classes (basal cell carcinoma (BCC), benign melanocytic nevus (BMN), seborrheic keratosis (SK), actinic keratosis (AK), squamous cell carcinoma (SCC), see Figure 3A), we find moderate agreement between the two pathologists. Analyzing the results for each class separately, we find that Pathologist 1 overwhelmingly prefers the AI or finds the AI and human report similarly good in about 70% of the BCC cases. Pathologist 2, on the other hand, prefers the AI-generated report for BMN 80% of the time. The AI-generated report for SK is preferred by both pathologists 90% of the time. Across all 100 report pairs, both pathologists find no difference between the generated and human reports about 45% of the time and prefer the AI-generated reports about 15% of the time (see Figure 3D).

According to a post-analysis provided by the two pathologists, after about 20 cases, they were able to tell which report was likely generated by the AI and which was likely generated by a human pathologist. The AI-generated text tends to be more structured and comprehensive. It includes more observations that are informative but not always necessary for the final diagnosis. Notably, there are only a few cases (< 5) where HistoGPT generated confusing text. In one case, the model incorrectly identified red collagen bundles as blood. In another case, it failed to describe a cyst, which was the key diagnostic feature. In one interesting case where there was a disagreement between the ground truth diagnosis and Pathologist 1 – resulting in both AI and human reports being disputed. Interestingly, one slide was incorrectly annotated by the human, but the AI still provided the correct report. There are two cases where the AI failed to detect small or unusual objects such as mitotic figures and a scabies mite. In one slide, the model mistook erythrocytes for eosinophils. However, these two cell types were difficult to distinguish in the image. Pathologist 1 mentioned that about 10 human reports were favored simply because the tumor thickness was more accurate than in the generated report, but the text itself was equally good. After adjusting for this, and including only reports where "Expert guidance" and model prediction agreed, the pathologist preferred the AI report or was indifferent 80% of the time (see Supplementary Figure 3). Overall, the model was described as having the skill level of a novice pathologist. Notably, this was achieved with only 5K training points, which is small for LLM standards.

HistoGPT accurately predicts diseases across many cohorts

There is another quantitative way to demonstrate that HistoGPT has effectively learned to encode medical knowledge. We extract the predicted diagnosis from the generated reports, calculate the classification accuracy, and compare the results (Figure 4) with state-of-the-art multiple-instance learning (MIL) approaches for image classification. For this purpose, we run HistoGPT without "Expert guidance" mode, i.e. we prompt the model with the phrase "Final diagnosis" instead of "Final diagnosis: [expert label]" and let it make a diagnostic decision on its own (see Figure 4A). MIL methods such as AttentionMIL [19], TransMIL [20], and TransfomerMIL [9] achieve weighted F1 scores between 0.34 and 0.48 on the Munich test set. These results are not unexpected. A major challenge for all these methods is that the training dataset is highly unbalanced, ranging from a handful of samples in the minority classes to several hundred samples in the majority classes. Nevertheless, our PerceiverMIL achieves a weighted F1 score of 44% on the internal test set (see Figure 4B). The much larger HistoGPT-1B does not overfit and retains the performance of its vision module. Surprisingly, the even larger HistoGPT-3B improves the weighted F1 score to 45%. Compared to the highly specialized models AttentionMIL, TransMIL, and TransfomerMIL, both PerceiverMIL and HistoGPT are slightly better or at least competitive in terms of classification performance. It is important to note that, unlike MIL approaches, the output of HistoGPT is pure text and not integer class indices, highlighting the flexibility of a vision language model.

A challenging clinical question with a high therapeutic impact in dermatology is the differentiation of cancer from non-cancer. In routine diagnosis, for example, it is important to distinguish basal cell carcinoma (BCC) from other conditions; cancer, such as squamous cell carcinoma (SCC) from precancerous actinic keratosis (AK); and malignant from benign conditions, such as melanoma from benign melanocytic nevus (BMN). Unlike the previous classification task with over 100 classes, we now face a classification problem with only two classes. In this case, HistoGPT automatically calls a lightweight binary classifier to solve the task at hand (see Methods), overcoming the class imbalance problem from before. This mode is called "Classifier guidance" and makes the model aware of the unbalanced label distribution by limiting the number of output classes. We achieve remarkable classification performance for the three clinical tasks with weighted F1 scores of 98%, 87%, and 89%, respectively (see Figure 4C).

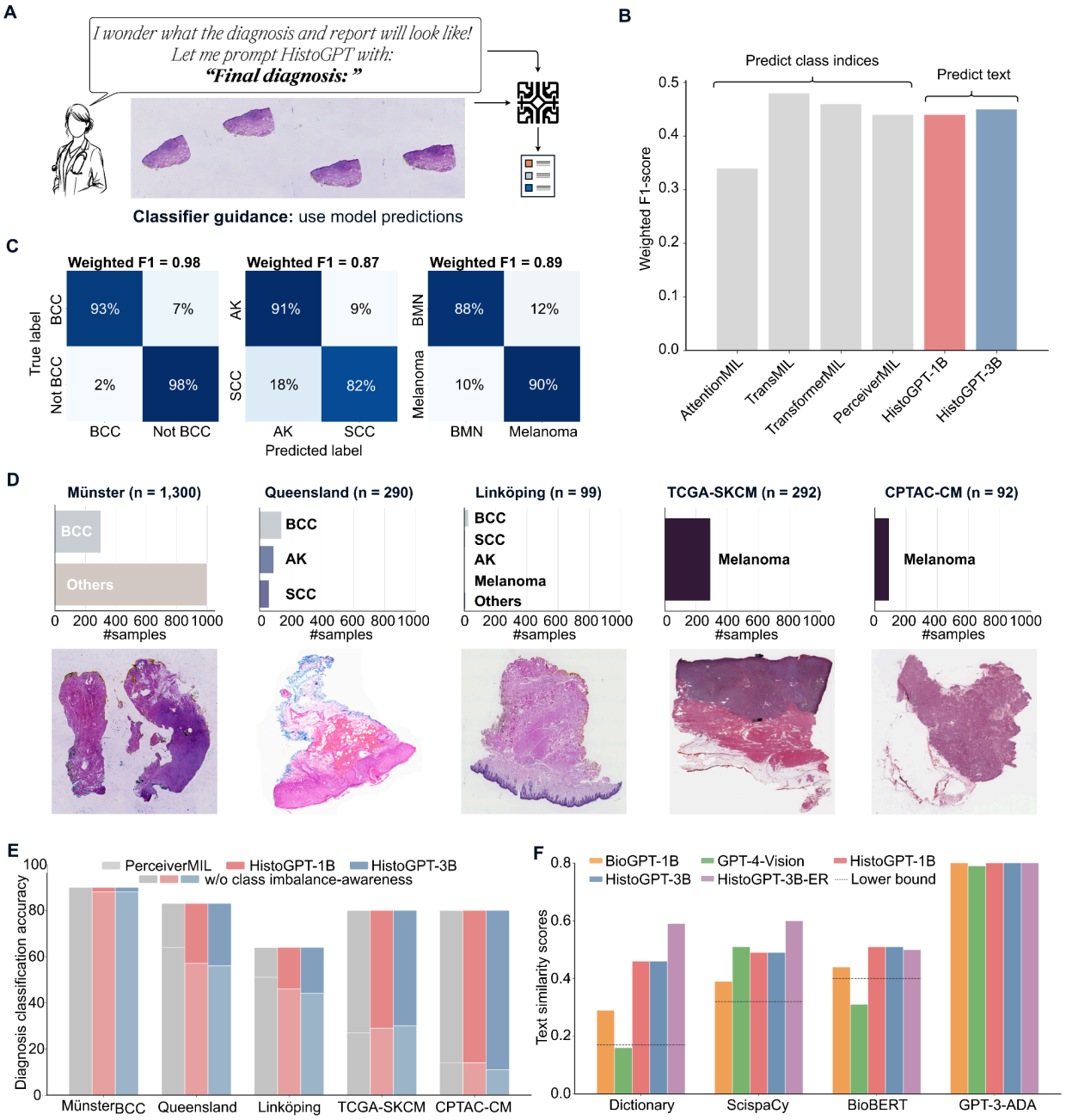


Figure 4. HistoGPT accurately predicts diseases in-domain and out-of-domain without human guidance. (A) In the absence of a human-in-the-loop, HistoGPT independently predicts the patient's diagnosis on its own and generates the corresponding pathology report. (B) On the internal Munich test set, HistoGPT is comparable to state-of-the-art classification models in predicting over 100 dermatological diseases, even though the model's output is pure text. (C) HistoGPT answers clinically challenging and important questions by discriminating malignant from benign conditions with high accuracy on the Munich dataset: basal cell carcinoma (BCC, n = 107) vs. other conditions (n = 621) with an accuracy of 0.98 and a weighted F1 score of 0.98; actinic keratosis (AK, n = 47) vs. squamous cell carcinoma (SCC, n = 33) with an accuracy of 0.88 and a weighted F1 score of 0.87; benign melanocytic nevus (BMN, n = 86) vs. melanoma (n = 21) with an accuracy of 0.89 and a weighted F1 score of 0.89. (D) We also evaluate HistoGPT on five independent external cohorts covering different countries, scanner types, staining techniques, and biopsy methods. (E) Both PerceiverMIL and HistoGPT perform well on external datasets by conditioning them on the class distribution. (F) HistoGPT is able to produce highly accurate pathology reports, as indicated by the high keyword and cosine-based similarity scores on Münster. As in Figure 3C, the lower baseline compares two randomly selected reports.

HistoGPT in "Classifier guidance" mode also generalizes to previously unseen datasets and problems. We demonstrate this by evaluating HistoGPT on five external, publicly available cohorts from different countries, scanner types, staining protocols, and medical procedures such as shave biopsies, punch biopsies, and excisional biopsies (see Figure 4D). While some of the cohorts include a variety of dermatological diseases (Queensland or Linköping), some cohorts (TCGA and CPTAC) include only melanoma cases, but can still be used to assess the accuracy of HistoGPT. We retrain PerceiverMIL as a state-of-the-art classifier and HistoGPT-1B as well as HistoGPT-3B on the entire Munich cohort and compare their classification performance on the external datasets. On the BCC subset of Münster, both PerceiverMIL and HistoGPT correctly identify BCC in 88% of cases (see Figure 4E). In the multi-class setting (Queensland with 3 classes and Linköping with 14 classes), we achieve accuracies of 85% and 70%, respectively. The models also reliably discriminate melanoma from other types with accuracies of 80% and 90% in TCGA and CPTAC, respectively. For comparison, we also report the results of HistoGPT without class imbalance awareness (see Figure 4E, light color bars). "Classifier guidance" significantly improves the effectiveness and generalizability of the model across different external cohorts.

Of the five cohorts, only Münster (without the BCC subset) includes unstructured pathology reports. In contrast to the Munich reports, these reports contain only the critical findings and the final assessment (e.g., *"Lichen planus-like keratosis (regressive solar lentigo/flat seborrheic keratosis), no evidence of basal cell carcinoma in the present biopsy."*) and thus lack the detailed microscopic description of the Munich training set. Since the critical findings include different classes not seen in Munich and are not available separately from the written text, it was not possible to extract individual class labels. Nevertheless, we can calculate how diagnostic information HistoGPT encodes by comparing the extracted keywords and measuring the cosine similarity (see Figure 4F). Remarkably, HistoGPT captures nearly 60% of all biomedical keywords using our dermatology dictionary and the ScispaCy model, even though the ground truth was written in a completely different style and structure. HistoGPT also achieves high cosine similarity under BioBERT and GPT-3-ADA. Compared to a random report generated by BioGPT-1B and a grounded report given by GPT-4V, the text quality of these models is much lower compared to HistoGPT with or without Ensemble refinement.

HistoGPT predicts tumor thickness and tumor subtypes zero-shot

In the diagnosis of (skin) tumors, it is important to include information about tumor thickness or assignment to a specific tumor subtype in the final report. These parameters are well defined in dermatopathology: In basal cell carcinoma, tumor thickness is measured from the stratum granulosum in the epidermis to the deepest point of the tumor in millimeters, similar to the determination of the Breslow index in melanoma, while tumor subtype classification is based on the WHO guidelines [32]. HistoGPT can predict both tumor thickness and tumor subtypes out-of-the-box and does not require additional reconfiguration or explanation of tumor-specific parameters at any stage of training. We can design prompts and instruct HistoGPT to produce the desired text output. For example, typing the prompt "tumor thickness" will produce a prediction of the depth of tumor invasion without fine-tuning. Although only a fraction (n = 644) of the training dataset has this value recorded as ground truth, HistoGPT can still predict the tumor thickness with considerable accuracy and include it directly in the final report. This emergent behavior is referred to in the literature as zero-shot learning [40]. For the 94 samples in the internal Munich test set with such a ground truth, we measure a root mean square error (RMSE) of 1.8 mm and a significant correlation coefficient of $\rho = 0.52$ ($p = 9.7 \cdot 10^{-8}$) for the predicted tumor thickness versus the reported ground truth (see Figure 5A). Binning the values to an interval with step sizes of 2 mm, 1 mm, and 0.5 mm gives us accuracies of 64%, 38%, and 21%, respectively. Again, we emphasize that this is zero-shot prediction on a task where the ground truth is typically obtained with a dedicated measurement procedure. In comparison, the predictions of the slide-based contrastive baselines, HistoCLIP (RMSE = 4.35 mm, $\rho = 0.006$, $p = 0.96$) and HistoSigLIP (RMSE = 3.84 mm, $\rho = 0.38$, $p = 0.002$), correlate poorly with the ground truth and are far from HistoGPT in terms of quality (see Supplementary Figure 4A). The patch-based contrastive baseline PLIP [13], which is the state of the art in computational pathology, is even worse (RMSE = 2.78 mm, $\rho = -0.18$, $p = 0.08$), highlighting the importance of a slide-level approach.

We analyze whether the zero-shot capability generalizes to other cohorts by looking at the never seen BCC subset of the external Münster test set (see Figure 5B), which has not been used for training purposes. For the samples with a ground truth tumor thickness measurement, we find a root mean square error of 0.98 mm and a significant correlation coefficient of $\rho = 0.39$ ($p = 5.8 \cdot 10^{-5}$). Compared to HistoCLIP (RMSE = 3.91 mm, $\rho = -0.16$, $p = 0.1$), HistoSigLIP (RMSE = 1.46 mm, $\rho = 0.10$, $p = 0.3$), and PLIP (RMSE = 1.43 mm, $\rho = -0.04$, $p = 0.7$), their correlation of prediction with the data is much worse than HistoGPT (see Supplementary Figure 4B). Using gradient attention maps, we can gain insight into the reasoning behind each output. When estimating tumor thickness, HistoGPT correctly focuses on the tumor region (see Figure 4C, top). However, the VLM sometimes struggles to find the correct reference point (e.g., when the epidermis is torn or especially when it is ulcerated) or spatial orientation for the measurements, even though it recognizes the tumor mass itself (see Figure 4C, bottom; Supplementary Figure 5). We attribute this to the design decision not to use position embeddings to store the coordinate values for each patch, which led to

training instabilities. However, because HistoGPT is designed to be used with a human-in-the-loop, pathologists can quickly identify the discrepancy and correct the model in an interactive teacher-student setting, i.e., in "Expert guidance" mode.

We continue to explore the benefits of zero-shot learning. Basal cell carcinoma is the most common type of malignant skin cancer. Although it is the majority class in the training set, the training set does not contain BCC subtypes as critical diagnoses. Therefore, BCC subtypes could not be used as labels during supervised pre-training. This information is only implicitly available as free text hidden in the report. Interestingly, HistoGPT is still able to extract the hidden information from the internal training set Munich and apply the acquired knowledge in the external test set Münster to discriminate between three major BCC subtypes (superficial, solid/nodular and infiltrating) with a weighted F1 score of 63%, quantified by extracting the keywords from the generated reports (see Figure 5D). As clearly shown in the gradient attention maps (see Figure 5E), HistoGPT correctly attends to the relevant architectural patterns within the histological slides that are the hallmarks of each cancer subtype. This zero-shot capability highlights the adaptability of HistoGPT as a generative AI model, especially when compared to more traditional classifiers such as TransMIL, which are limited to predefined classes and thus cannot predict subtypes without re-training. We also compare its zero-shot performance to more advanced models such as HistoCLIP and HistoSigLIP. As contrastive methods, they overcome the inflexible structure of multiple instance learning approaches. Both achieve weighted F1 scores of 54% and 50%, respectively, but perform significantly worse than HistoGPT, particularly in identifying infiltrative BCC. Infiltrative BCC is extremely important to identify in the clinical context, as this subtype tends to have a biologically much more aggressive growth pattern and relapse rate, and therefore may require different treatment and follow-up. The patch-based visual language foundation model for pathology image analysis, PLIP, does not provide useful predictions for this zero-shot classification task. Surprisingly, PLIP is constant over the test set and predicts all specimens as either superficial or solid depending on the resolution. That is, at 5x and 10x magnification, PLIP predicts all samples to be superficial; at 20x and 40x magnification, it predicts all images to be infiltrative (see Supplementary Figure 6).

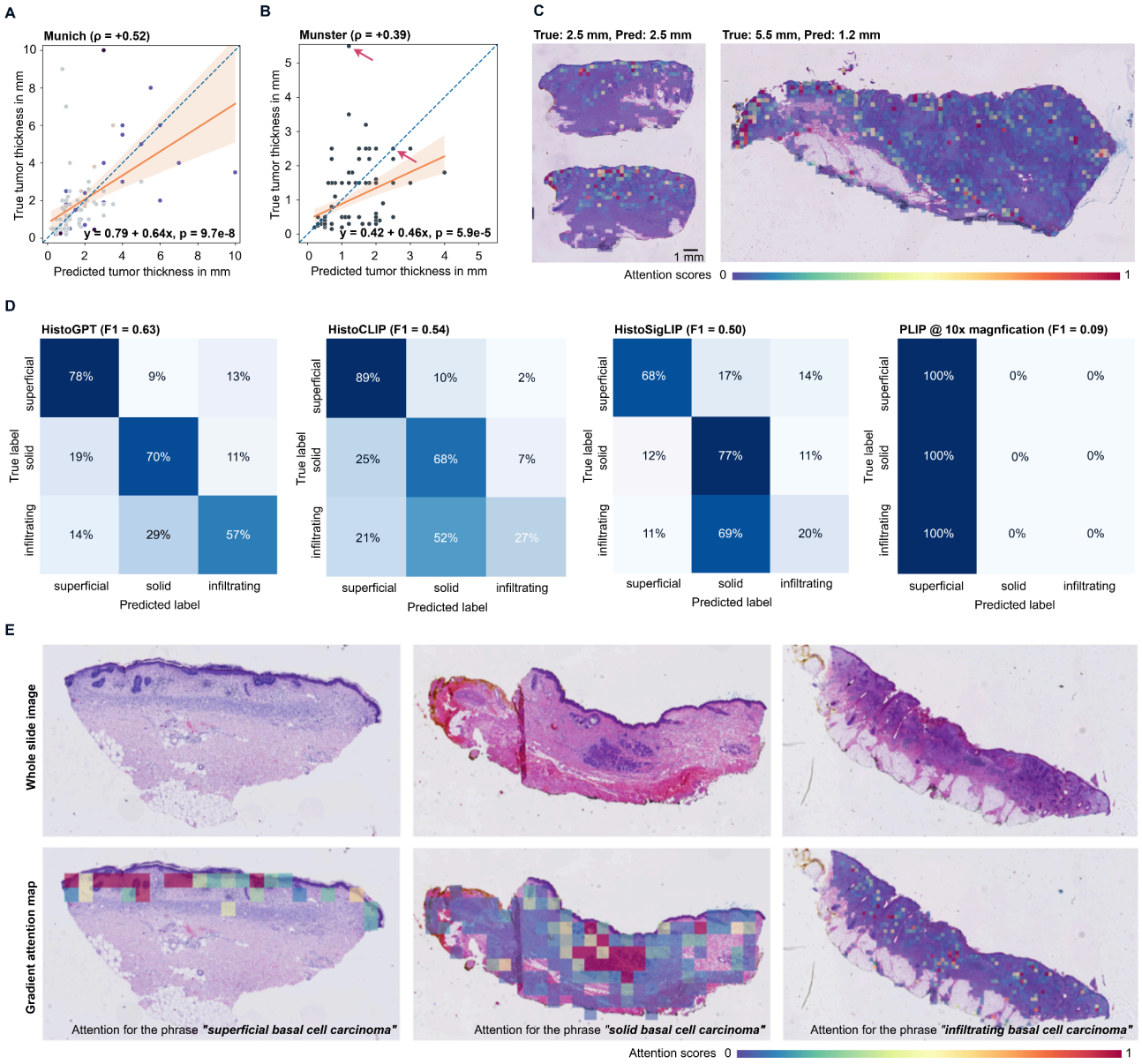


Figure 5. HistoGPT predicts tumor thickness and tumor subtypes in a zero-shot fashion and provides text-to-image visualization. (A) HistoGPT achieves high zero-shot performance in predicting tumor thickness on the internal Munich test set. The scatter plot is color-coded according to the classes in Figure 3A. (B) HistoGPT's prediction is also highly correlated with the ground truth on the external Münster test set, even though it was obtained using a different measurement protocol. (C) Since HistoGPT is an interpretable AI system, we can fully understand its results. Here we show the two examples marked with a red arrow in Figure 5B. (D) On the basal cell carcinoma subset of the external validation set Münster, HistoGPT is the only slide-level model that correctly predicts infiltrative BCC in most cases. The patch-level model PLIP fails in this task, predicting all samples as superficial. (E) Given whole slide images of superficial, solid, and infiltrating BCC, HistoGPT correctly identifies their morphological structures as shown in the high attention regions for the respective text strings.

Discussion

With HistoGPT, we introduce a vision language model that can generate histopathology reports from full-resolution, gigapixel whole slide images. The generated reports are of high quality, consistent with ground truth and independent expert evaluation. HistoGPT outperforms the state-of-the-art foundation model GPT-4V, which itself is already very capable in medical tasks [15,41,42]. In addition, HistoGPT predicts disease subtypes (validated on five international cohorts) and provides a comprehensive list of medical keywords using named entity recognition tools. Using various prompts (e.g., "the tumor thickness is"), pathologists can guide the model and customize it to their needs. This zero-shot performance rivals existing zero-shot learning approaches based on CLIP and SigLIP. Advanced methods such as ensemble refinement allow us to explore the probability space of possible medical outcomes. In particular, the output text is fully interpretable using gradient attention maps that match words in the generated report to corresponding regions in the image. We note that HistoGPT achieves this level of performance with only 6,000 dermatology cases, which is relatively small by LLM standards. Interestingly, this is the same number of cases that a pathologist in Germany must have seen to qualify for the dermatopathology exam [43]. Thus, our expert's impression that HistoGPT is comparable to a novice dermatopathologist has an intuitive analogy. However, unlike a real pathologist, our model lacks extensive medical training and strong human supervision. Nevertheless, HistoGPT already writes reasonably good reports and shows a good understanding of the underlying case.

Although current neural networks can also predict tumor thickness or tumor subtypes with good accuracy, as has been shown in particular for basal cell carcinoma [33–39], they require a large amount of high-quality, precisely annotated data for training and are not flexible enough to be used for tasks other than those for which they were trained. That is, these models are fully supervised and do not operate in a zero-shot fashion. Specifically for tumor thickness prediction, the above approaches are not end-to-end deep learning systems. Users must first train a segmentation model to segment the tumor region, and then use a hand-crafted mathematical algorithm to calculate the tumor thickness. HistoGPT, on the other hand, does not require this multi-step approach because it has already learned to understand this concept just by looking at text and images.

HistoGPT also has its limitations. First, the model has only been trained and tested on dermatological samples. Thus, it cannot yet be generalized to the more general case of pan-cancer diagnosis. In addition, our training dataset suffers from severe class imbalance, which limits its usefulness for minority classes. This problem can be partially mitigated with "Classifier guidance". However, guidance has its limitations too, as the generated reports tend to be of higher quality when the initial diagnostic prediction is also correct (see Supplementary Figure 3). Another open problem is to find a highly efficient and effective way to encode the positional information of the individual image patches within a WSI. An interesting research direction is to fine-tune HistoGPT as a conversational chatbot using

Reinforcement Learning with Human Feedback (RLHF). This may prove challenging in practice, as there are currently no slide-level question-answer pairs for the model to learn from. An even more useful follow-up question is whether a tumor has been excised as a whole or whether there is still a tumor mass at the margins, which is clinically highly relevant. Finding out if and how AI will differentiate between primary tumors and metastases is another clinically relevant challenge. Certainly, indications such as the growth of tumor cells emanating from the epidermis can be recognized by the AI. However, there will still be cases where the AI - just like human pathologists - will find it difficult to make final decisions. This remains an interesting topic for follow-up studies. Overall, HistoGPT shows strong emergent capabilities and is a fully functional proof of concept for a vision language foundation model in histopathology. We are releasing both model code and weights so that the broader scientific community can explore and improve HistoGPT.

Acknowledgments

M.T., S.W.J., and V.K. are supported by the Helmholtz Association under the joint research school "Munich School for Data Science – MUDS". C.M. acknowledges funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement no. 866411) and support from the Hightech Agenda Bayern.

We thank the dermatopathologists R. Hein, R. Franz, S. Möckel, A. Steimle-Grauer, C. Andres, and S. Roenneberg (Munich) for the detailed review of patient material and for providing the data for this project. We also thank Nina Witte (Münster) for her technical and organizational support.

Author contributions

M.T. developed the methods, implemented the code, and conducted the experiments. P.S. provided domain knowledge, collected the dataset, and evaluated report quality. S.J.W. helped with the experiments by providing the MIL and zero-shot learning results. V.K. managed the data processing pipeline through patching and feature extraction. V.L. curated the data and trained MIL models. A.F. processed and scanned the whole slide images of the Munich cohort. A.B. collected and annotated the internal training dataset with images and reports. R.K. provided medical advice and designed the real-world evaluation metrics. T.B. provided resources for data acquisition and contributed to the writing of the manuscript. K.E. supervised the study and the compilation of the internal Munich dataset. S.A.B. supervised the external Münster dataset compilation and evaluated report quality. T.P. supervised the machine learning approach. C.M. supervised the study.

Data and code availability

The 100 patient cases from the Munich cohort used in the blinded study will be made available upon publication. The model code and weights are available at https://github.com/marrlab/HistoGPT.

Declaration of interests

M.T. is employed by Roche Diagnostics GmbH but conducted his research independently of his work at Roche Diagnostics GmbH as a guest scientist at Helmholtz Munich (Helmholtz Zentrum München – Deutsches Forschungszentrum für Gesundheit und Umwelt GmbH).

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the authors used DALLE-3 to generate the icons for HistoGPT, the female doctor, the microscope device, and the ruler in Figure 1. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

STAR ★ Methods

All research procedures were conducted in accordance with the Declaration of Helsinki. Ethics approval was granted by the Ethics Commission of the Technical University Munich (reference number 2024-98-S-CB) and the Ethics Commission of Westfalen-Lippe (reference number 2024-157-b-S).

Datasets

Munich cohort: All 6,000 histology samples of the Munich cohort were prepared and stained (with hematoxylin and eosin) at the Department of Dermatology of the Technical University of Munich. They were scanned with a 20x objective at 0.173 micrometers per pixel at the Core facility Imaging at Helmholtz Munich. All slides were fully anonymized. 100 exemplary WSIs and reports are provided in the Supplementary Material. An example report (AI-translated from German to English) reads: "Final diagnosis: Scar. Microscopic findings: A wedge-shaped excidate with compact massive orthohyperkeratosis, focally regular acanthosis of the epidermis with hypergranulose, focally clearly flattened epidermis with elapsed reticles is presented. Underneath densely packed, partly hypereosinophilic cell-poor collagen fiber bundles, vertically placed capillary vessels. In the depth more homogenised hypereosinophilic

proliferating collagen fiber bundles. Critical findings: Hypertrophic, keloid-like scar. Partial excision."

Münster cohort: All 1,300 histology samples of the Münster cohort were processed and stained (with hematoxylin and eosin) at the Department of Dermatology, University Hospital Münster, Münster, Germany. They were scanned with a 20x objective at 0.46 micrometers per pixel using a Hamamatsu NanoZoomer S360 MD, Hamamatsu City, Japan, at the Department of Dermatology, University Hospital Münster, Münster, Germany. The cohort comprises 300 cases with 100 BCC subtypes each (superficial, solid/nodular, infiltrating), and 1000 cases from daily routine without special selection. All slides were fully anonymized. An example report (AI-translated from German to English) reads: "Lichen planus-like keratosis (regressive solar lentigo/flat seborrheic keratosis), no evidence of basal cell carcinoma in the present biopsy."

Image preprocessing

We treat all whole slide images (WSIs) belonging to a patient as one input. In other words, we have patient-level samples, instead of slide-level or even patch-level data points. These WSIs are tessellated at a total magnification of 100x (equivalent to an objective magnification of 10x or 1 micron per pixel) into non-overlapping image patches of 256 x 256 pixels and resized to 224 x 224 pixels using the Python library SlideIO. The inputs are then converted into PyTorch tensor objects and normalized using a mean of (0.485, 0.456, 0.406) and a standard deviation of (0.229, 0.224, 0.225). We use this specific image size and normalization parameter in accordance with most publicly available pre-trained histopathology image encoders.

Model architectures

We use CTransPath [21] as our pre-trained vision encoder to extract 768-dimensional feature vectors for each image patch and concatenate them along the sequence dimension to obtain a matrix of size n x 768, where n is the number of image patches. The inputs are then fed into the Perceiver Resampler [27], which is borrowed from the vision language model Flamingo [28] with randomly initialized weights. We change the default number of latents from 64 to 640 because WSIs are much larger than natural images and require a larger dimensional latent space to store the additional information. We keep the output size of 1536 because this has been shown to work well [28]. The fixed-size outputs of dimension 640 x 1536 are then used as keys and values in the tanh gated cross-attention block (XATTN). The query vectors come from the pre-trained language model BioGPT [17]. In particular, we use one XATTN block after each language layer according to the high-performance configuration of Flamingo. The output layer of HistoGPT is a linear classifier over the vocabulary.

We benchmark HistoGPT against HistoCLIP and HistoSigLIP. They use the feature mean of the pre-trained Perceiver Resampler as the image representation and the EOS token of the pre-trained BioGPT as the text representation. A contrastive loss then aligns both feature vectors in the shared embedding space. For HistoCLIP we use the same loss as for CLIP [47]. For

HistoSigLIP, we use the loss proposed in SigLIP [48]. In pursuit of improving performance and avoiding training instabilities, we freeze the vision encoder during training. This technique is called locked-image text tuning [49]). We also compare HistoGPT to the patch-based foundation model PLIP using the contrastive pre-trained model via the provided API. To aggregate the patch-level results to the slide-level, we evaluate PLIP [13] using the majority voting system of the related model MI-Zero [16].

Since BioGPT and other popular LLMs are all pre-trained on mostly English text, we need to translate the German reports into English to take advantage of their capabilities. For the translator, we choose a standard machine translation model based on the Transformer encoder-decoder architecture [25] with the checkpoint "Helsinki-NLP/opus-mt-de-en" available on Hugging Face.

Model training

We pre-train the Perceiver Resampler in a fully supervised manner by predicting the critical diagnosis using a linear classifier on top of the encoder. Since the labels are provided at the patient level, this approach is also known as multiple instance learning (MIL). The classifier is then discarded and the resampler is plugged into the vision language model (VLM). We freeze all layers of HistoGPT except the cross-attention blocks. Our generative training is based on causal language modeling: Given an input, we mask the next tokens and let the model predict them. This is done in parallel over all input tokens using an upper triangular causal attention mask.

For training, we use the AdamW optimizer with betas of (0.9, 0.95), a weight decay of 0.1, and epsilon of 1e-8. The learning rate starts at zero and warms up linearly over 10 epochs to 1e-4 before decaying tenfold according to a cosine-annealing scheduler. We use a gradient accumulation of 32 to simulate a larger batch size. Both PR and VLM are trained for 100 epochs using mixed precision training and gradient clipping to a Euclidean norm of 1.0. For contrastive learning, we use the same hyperparameters proposed by [12] and [48].

During training, we randomly augment the text inputs to avoid overfitting common words and phrases. This is done beforehand using GPT-4 to sample 9 paraphrased texts with a temperature of 1.0 and nucleus sampling of 1.0. The prompt used is: "Rewrite the following text but be as accurate and faithful as possible to the original. Do not add or remove any information! Also, do not change the phrases 'Microscopic findings:' and 'Critical findings:', but leave them as they are."

Classifier guidance

We enable class-imbalance awareness in HistoGPT by using a lightweight and specialized classification model. The classifier predicts one-hot encoded class indices which are converted to text strings using a lookup table and inserted into HistoGPT. Suppose the training set contains C classes. Assume that at inference time, we face a classification problem with c classes, where $c \subset C$. We extract features of each training sample with a pre-trained Perceiver

Resampler and fit a linear classifier that predicts these c classes. With this approach, we reduce the 162-class classification problem to a more tractable subset of classes. For BCC vs. ¬BCC, we consider all samples that are not BCC to be ¬BCC and fit an MLP with 100 neurons. For Melanoma vs. ¬Melanoma, we follow the same procedure. For all other classification tasks, we only train on the specific subset. For example, if we want to classify BCC vs. SCC vs. AK vs. SK, we train a classifier only on the BCC, SCC, AK, and SK training features and ignore the remaining classes. Some datasets (Queensland and Linköping) only provide annotation masks as labels. They may contain different disease labels for different regions in the same slide. In this case, we consider the prediction of one of the ground truth classes as accurate.

Interpretability maps

For HistoGPT explainability, we use partial derivatives and associate the output latents of the Perceiver Resampler with the corresponding input vectors. We then weight the image features with the text features using the cross-attention scores. This gives us a gradient attention map. It shows which word in the generated report corresponds to which region in a WSI. For example, we can highlight where the model sees basal cell carcinoma, how it detects tumor-infiltrating lymphocytes, and which regions it considers when measuring tumor thickness (see Supplementary Figure 1). In this way, we provide an unprecedented approach to explainable AI by matching visual and linguistic information.

The output of the Perceiver Resampler consists of 640 latent vectors. We compute the gradients of these latents with respect to the input patches with backpropagation. Thus, the gradient G has the form num_patches x num_latents. It tells us which image tokens have the most influence on which latent feature. The mean along the latent sequence thus gives us the most important image regions according to the vision resampler. How can we use this information to determine which of these regions corresponds to which word? One idea is to give higher weights to the latents that correspond to the words we are interested in. We get these weights by looking at the cross-attention scores of the last XATTN layer. The attention matrix A has a dimension of num_tokens x num_latents. Thus, given a target word, we can identify the corresponding target tokens and use the corresponding rows in the attention matrix as weights. Overall, the proposed *Gradient x Attention* map is given by the weighted mean

$$(G^T \circ A[\text{target\_tokens}, :].\text{mean}(\text{dim}=0)^T)^T.\text{mean}(\text{dim}=1).$$

Evaluation metrics

We introduce two other non-trivial baselines: given the ground truth, compare two random reports with two arbitrary diagnoses (lower baseline), and compare two random reports with the same diagnosis (upper baseline). The logic behind this approach is simple. Medical texts often follow a structured format with a similar writing style, typically including a general description of the specimen and frequent use of common technical terms. In addition, certain diseases manifest homogeneously across patients, resulting in nearly identical report

descriptions within a patient group. In such cases, the few unique terms in the reports become critical in distinguishing between different diagnoses. Therefore, these two baseline comparisons provide effective reference points for measuring the overall performance of our models.

Evaluating the reports generated by HistoGPT is a non-trivial task. Popular evaluation methods for natural language generation such as BLEU-4 [50], ROUGE-L [51], and METEOR [52] primarily compare n-grams between two documents and may not effectively capture semantic similarities. In fact, two texts can describe the same phenomena in different ways, making a word-by-word comparison unfair. Therefore, we focus on two different quantitative performance metrics: keyword overlap and sentence similarity. For the former, we use a comprehensive glossary of human-curated dermatological vocabularies [53] to extract important medical keywords from the ground truth notes. In addition, we use ScispaCy [30], a biomedical named entity recognition (NER) tool, to capture a broader range of technical terms. We then determine how many keywords from the ground truth text can be found in the generated text. The Jaccard index is an appropriate measure to quantify their overlap. To find a match in the generated report, we use an advanced version of Gestalt pattern matching (Ratcliff and Obershelp, 1988) which is available in the Python library difflib. We use the default cutoff threshold of 0.6. This value strikes a balance between matching every word as a target and matching only exact overlaps. The latter is undesirable because it ignores different grammatical forms of a word. As a consequence, some unrelated words will inevitably be matched. In this case, the Jaccard index can be considered a relative measure as the same approach is applied to every model.

The above measures still miss some semantic nuances, since certain concepts or observations (e.g., disease properties, tissue subtypes, cellular characteristics) can be expressed in complex phrases, possibly even involving negations. To remedy this, we use BioBERT fine-tuned [54] for natural language inference (NLI) and semantic textual similarity (STS) assessments. This embedding model provides the feature vectors of the generated report and the ground truth, allowing us to compute their cosine similarity as a measure of semantic understanding. To go beyond the domain-specific use of language, we apply a general large-scale embedding model, GPT-3-ADA [26], to capture a broader range of linguistic information. Similarly, we use BERTScore [55] to compute the syntactic relationship between generated and ground truth reports at the subword level.

For ensemble refinement, we summarize the bootstrapped reports by prompting GPT-4-Turbo with the instruction "Summarize the following text:". Since ensemble refinement sampling is massively time-consuming and relies on an expensive API call, we only compute the scores on a random subset of the test set (10%). However, the standard deviation among the samples remains similar to the models on the full test set, indicating that the final score would not change much.

References

1. Gurcan, M.N., Boucheron, L.E., Can, A., Madabhushi, A., Rajpoot, N.M., and Yener, B. (2009). Histopathological image analysis: a review. IEEE Rev. Biomed. Eng. *2*, 147–171.

2. Histopathology is ripe for automation (2017). Nat Biomed Eng *1*, 925.

3. Cheah, P.-L., Looi, L.M., and Horton, S. (2017). Cost Analysis of Operating an Anatomic Pathology Laboratory in a Middle-Income Country. Am. J. Clin. Pathol. *149*, 1–7.

4. Märkl, B., Füzesi, L., Huss, R., Bauer, S., and Schaller, T. (2021). Number of pathologists in Germany: comparison with European countries, USA, and Canada. Virchows Arch. *478*, 335–341.

5. van der Laak, J., Litjens, G., and Ciompi, F. (2021). Deep learning in histopathology: the path to the clinic. Nat. Med. *27*, 775–784.

6. Zheng, Y., Gindra, R.H., Green, E.J., Burks, E.J., Betke, M., Beane, J.E., and Kolachalama, V.B. (2022). A Graph-Transformer for Whole Slide Image Classification. IEEE Trans. Med. Imaging *41*, 3003–3015.

7. Spronck, J., Gelton, T., van Eekelen, L., Bogaerts, J., Tessier, L., van Rijthoven, M., van der Woude, L., van den Heuvel, M., Theelen, W., van der Laak, J., et al. (2023). nnUNet meets pathology: bridging the gap for application to whole-slide images and computational biomarkers.

8. Chen, R.J., Chen, C., Li, Y., Chen, T.Y., Trister, A., Krishnan, R.G., and Mahmood, F. (2022). Scaling vision Transformers to gigapixel images via hierarchical self-supervised learning. Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., 16123–16134.

9. Wagner, S.J., Reisenbüchler, D., West, N.P., Niehues, J.M., Zhu, J., Foersch, S., Veldhuizen, G.P., Quirke, P., Grabsch, H.I., van den Brandt, P.A., et al. (2023). Transformer-based biomarker prediction from colorectal cancer histology: A large-scale multicentric study. Cancer Cell *41*, 1650–1661.e4.

10. da Silva, L.M., Pereira, E.M., Salles, P.G., Godrich, R., Ceballos, R., Kunz, J.D., Casson, A., Viret, J., Chandarlapaty, S., Ferreira, C.G., et al. (2021). Independent real-world application of a clinical-grade automated prostate cancer detection system. J. Pathol. *254*, 147–158.

11. Tu, T., Azizi, S., Driess, D., Schaekermann, M., Amin, M., Chang, P.-C., Carroll, A., Lau, C., Tanno, R., Ktena, I., et al. (2023). Towards Generalist Biomedical AI. arXiv [cs.CL].

12. Lu, M.Y., Chen, B., Zhang, A., Williamson, D.F.K., Chen, R.J., Ding, T., Le, L., Chuang, Y.-S., and Mahmood, F. (2023). Visual language pretrained multiple instance zero-shot transfer for histopathology images. Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., 19764–19775.

13. Huang, Z., Bianchi, F., Yuksekgonul, M., Montine, T.J., and Zou, J. (2023). A visual–language foundation model for pathology image analysis using medical Twitter. Nat. Med. *29*,

2307–2316.

14. Li, C., Wong, C., Zhang, S., Usuyama, N., Liu, H., Yang, J., Naumann, T., Poon, H., and Gao, J. (2023). LLaVA-Med: Training a Large Language-and-Vision Assistant for Biomedicine in One Day. arXiv [cs.CV].

15. Lu, M.Y., Chen, B., Williamson, D.F.K., Chen, R.J., Ikamura, K., Gerber, G., Liang, I., Le, L.P., Ding, T., Parwani, A.V., et al. (2023). A Foundational Multimodal Vision Language AI Assistant for Human Pathology. arXiv [cs.CV].

16. Lu, M.Y., Chen, B., Williamson, D.F.K., Chen, R.J., Liang, I., Ding, T., Jaume, G., Odintsov, I., Zhang, A., Le, L.P., et al. (2023). Towards a Visual-Language Foundation Model for Computational Pathology. arXiv [cs.CV].

17. Luo, R., Sun, L., Xia, Y., Qin, T., Zhang, S., Poon, H., and Liu, T.-Y. (2022). BioGPT: generative pre-trained transformer for biomedical text generation and mining. Brief. Bioinform. *23*. 10.1093/bib/bbac409.

18. OpenAI, :, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., et al. (2023). GPT-4 Technical Report. arXiv [cs.CL].

19. Ilse, M., Tomczak, J., and Welling, M. (2018). Attention-based Deep Multiple Instance Learning. In Proceedings of the 35th International Conference on Machine Learning Proceedings of Machine Learning Research., J. Dy and A. Krause, eds. (PMLR), pp. 2127–2136.

20. Shao, Z., Bian, H., Chen, Y., Wang, Y., Zhang, J., Ji, X., and Zhang, Y. (2021). TransMIL: Transformer based correlated multiple instance learning for whole slide image classication. Adv. Neural Inf. Process. Syst., 2136–2147.

21. Wang, X., Yang, S., Zhang, J., Wang, M., Zhang, J., Yang, W., Huang, J., and Han, X. (2022). Transformer-based unsupervised contrastive learning for histopathological image classification. Med. Image Anal. *81*, 102559.

22. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B. (2021). Swin Transformer: Hierarchical vision Transformer using shifted windows. arXiv [cs.CV], 10012–10022.

23. Cancer Genome Atlas Research Network, Weinstein, J.N., Collisson, E.A., Mills, G.B., Shaw, K.R.M., Ozenberger, B.A., Ellrott, K., Shmulevich, I., Sander, C., and Stuart, J.M. (2013). The Cancer Genome Atlas Pan-Cancer analysis project. Nat. Genet. *45*, 1113–1120.

24. Kim, Y.J., Jang, H., Lee, K., Park, S., Min, S.-G., Hong, C., Park, J.H., Lee, K., Kim, J., Hong, W., et al. (2021). PAIP 2019: Liver cancer segmentation challenge. Med. Image Anal. *67*, 101854.

25. Vaswani, A., Shazeer, N.M., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., and Polosukhin, I. (2017). Attention is All you Need. Adv. Neural Inf. Process. Syst., 5998–6008.

26. Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A.,

Shyam, P., Sastry, G., Askell, A., et al. (2020). Language Models are Few-Shot Learners. Adv. Neural Inf. Process. Syst. *abs/2005.14165*.

27. Jaegle, A., Gimeno, F., Brock, A., Vinyals, O., Zisserman, A., and Carreira, J. (18--24 Jul 2021). Perceiver: General Perception with Iterative Attention. In Proceedings of the 38th International Conference on Machine Learning Proceedings of Machine Learning Research., M. Meila and T. Zhang, eds. (PMLR), pp. 4651–4664.

28. Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al. (2022). Flamingo: A visual language model for few-shot learning. Adv. Neural Inf. Process. Syst. *abs/2204.14198*.

29. Singhal, K., Tu, T., Gottweis, J., Sayres, R., Wulczyn, E., Hou, L., Clark, K., Pfohl, S., Cole-Lewis, H., Neal, D., et al. (2023). Towards Expert-Level Medical Question Answering with Large Language Models. arXiv [cs.CL].

30. Neumann, M., King, D., Beltagy, I., and Ammar, W. (2019). ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing. arXiv [cs.CL].

31. Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C.H., and Kang, J. (2020). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics *36*, 1234–1240.

32. Kossard, S., Epstein, E.H., Jr, Cerio, R., Yu, L., and Weedon, D. (2006). Basal cell carcinoma. In World Health Organization Classification of Tumours. Pathology & Genetics. Skin Tumours World Health Organization Classification of Tumours., P. E. LeBoit, G. Burg, D. Weedon, and A. Sarasin, eds. (IARC Press), pp. 13–19.

33. O'Brien, B., Zhao, K., Gibson, T.A., Smith, D.F., Ryan, D., Whitfield, J., Smith, C.D., and Bromley, M. (2023). Artificial intelligence for basal cell carcinoma: diagnosis and distinction from histological mimics. Pathology *55*, 342–349.

34. Jansen, P., Arrastia, J.L., Baguer, D.O., Schmidt, M., Landsberg, J., Wenzel, J., Emberger, M., Schadendorf, D., Hadaschik, E., Maass, P., et al. (2024). Deep learning based histological classification of adnex tumors. Eur. J. Cancer *196*, 113431.

35. Duschner, N., Baguer, D.O., Schmidt, M., Griewank, K.G., Hadaschik, E., Hetzer, S., Wiepjes, B., Le'Clerc Arrastia, J., Jansen, P., Maass, P., et al. (2023). Applying an artificial intelligence deep learning approach to routine dermatopathological diagnosis of basal cell carcinoma. J. Dtsch. Dermatol. Ges. *21*, 1329–1337.

36. Campanella, G., Hanna, M.G., Geneslaw, L., Miraflor, A., Werneck Krauss Silva, V., Busam, K.J., Brogi, E., Reuter, V.E., Klimstra, D.S., and Fuchs, T.J. (2019). Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. Nat. Med. *25*, 1301–1309.

37. Yacob, F., Siarov, J., Villiamsson, K., Suvilehto, J.T., Sjöblom, L., Kjellberg, M., and Neittaanmäki, N. (2023). Weakly supervised detection and classification of basal cell carcinoma using graph-transformer on whole slide images. Sci. Rep. *13*, 7555.

38. Le'Clerc Arrastia, J., Heilenkötter, N., Otero Baguer, D., Hauberg-Lotte, L., Boskamp, T., Hetzer, S., Duschner, N., Schaller, J., and Maass, P. (2021). Deeply Supervised UNet for Semantic Segmentation to Assist Dermatopathological Assessment of Basal Cell Carcinoma. J. Imaging Sci. Technol. *7*. 10.3390/jimaging7040071.

39. Campanella, G., Nehal, K.S., Lee, E.H., Rossi, A., Possum, B., Manuel, G., Fuchs, T.J., and Busam, K.J. (2021). A deep learning algorithm with high sensitivity for the detection of basal cell carcinoma in Mohs micrographic surgery frozen sections. J. Am. Acad. Dermatol. *85*, 1285–1286.

40. Kojima, T., Gu, S.S., Reid, M., Matsuo, Y., and Iwasawa, Y. (2022). Large Language Models are Zero-Shot Reasoners. arXiv [cs.CL].

41. Nori, H., Lee, Y.T., Zhang, S., Carignan, D., Edgar, R., Fusi, N., King, N., Larson, J., Li, Y., Liu, W., et al. (2023). Can Generalist Foundation Models Outcompete Special-Purpose Tuning? Case Study in Medicine. arXiv [cs.CL].

42. Yan, Z., Zhang, K., Zhou, R., He, L., Li, X., and Sun, L. (2023). Multimodal ChatGPT for Medical Applications: an Experimental Study of GPT-4V. arXiv [cs.CV].

43. Ärztekammer Westfalen-Lippe (AEKWL) (2022). Zusatz-Weiterbildung Dermatopathologie 2022. https://www.aekwl.de/fileadmin/user_upload/aekwl/weiterbildung/wo_2020/Dermatopathologie_01.07.2020.pd.

44. Christian, A., and Shah, V.I. (2021). Pathology reporting: communication is key. Diagn. Histopathol. *27*, 279–282.

45. Mason, A.E., and Varma, M. (2022). Histopathology reporting for personalised medicine: focus on clinical utility. J. Clin. Pathol. *75*, 525–528.

46. Borges, A.M., and Varma, M. (2021). Personalized histopathology reporting for personalized medicine. Diagn. Histopathol. *27*, 275–278.

47. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. (18--24 Jul 2021). Learning Transferable Visual Models From Natural Language Supervision. In Proceedings of the 38th International Conference on Machine Learning Proceedings of Machine Learning Research., M. Meila and T. Zhang, eds. (PMLR), pp. 8748–8763.

48. Zhai, X., Mustafa, B., Kolesnikov, A., and Beyer, L. (2023). Sigmoid Loss for Language Image Pre-Training. arXiv [cs.CV].

49. Zhai, X., Wang, X., Mustafa, B., Steiner, A., Keysers, D., Kolesnikov, A., and Beyer, L. (2021). LiT: Zero-Shot Transfer with Locked-image text Tuning. Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., 18102–18112.

50. Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). BLEU: a method for automatic evaluation of machine translation. In Proceedings of the 40th Annual Meeting on

Association for Computational Linguistics ACL '02. (Association for Computational Linguistics), pp. 311–318.

51. Lin, C.-Y. (2004). ROUGE: A Package for Automatic Evaluation of Summaries. In Text Summarization Branches Out (Association for Computational Linguistics), pp. 74–81.

52. Lavie, A., and Agarwal, A. (2007). Meteor: an automatic metric for MT evaluation with high levels of correlation with human judgments. In Proceedings of the Second Workshop on Statistical Machine Translation StatMT '07. (Association for Computational Linguistics), pp. 228–231.

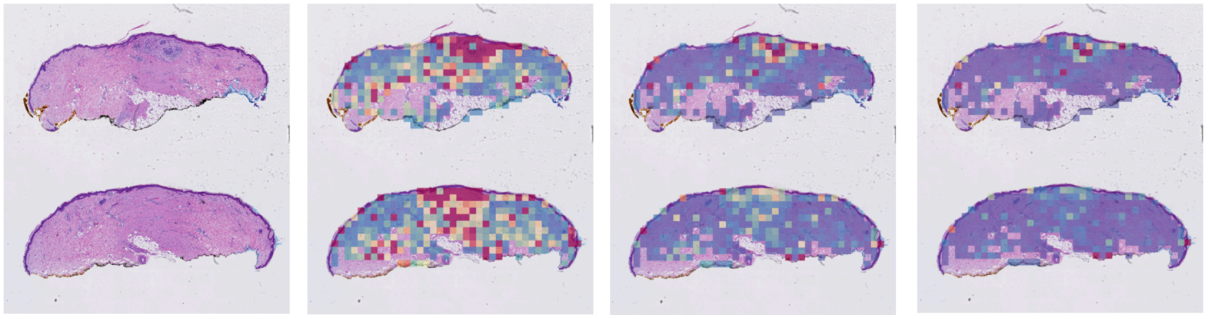
53. Coulson, I. (editor) Dermatology Glossary of Terms. dermnetnz.org/glossary.

54. Deka, P., Jurek-Loughrey, A., and Deepak (2022). Evidence extraction to validate medical claims in fake news detection. In Health Information Science Lecture notes in computer science. (Springer Nature Switzerland), pp. 3–15.

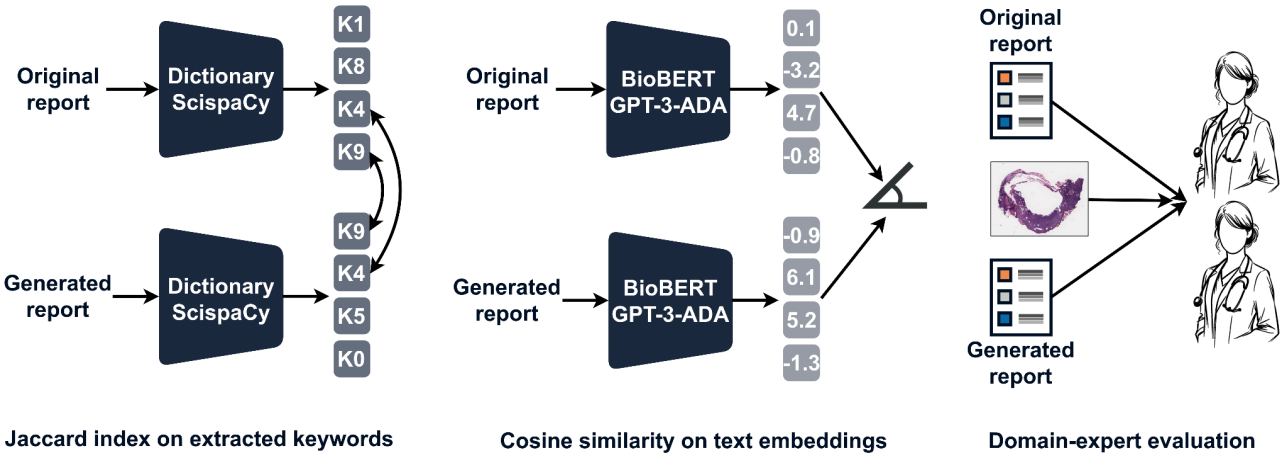
55. Zhang, T., Kishore, V., Wu, F., Weinberger, K.Q., and Artzi, Y. (2019). BERTScore: Evaluating Text Generation with BERT. arXiv [cs.CL].

Supplementary Figures

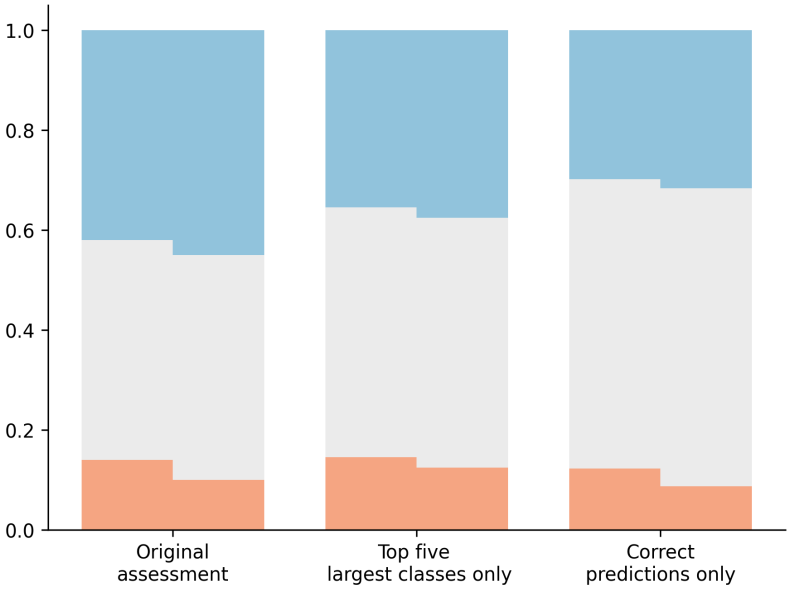
Final diagnosis: ***Basal cell carcinoma****. Critical findings: A superficial basal cell carcinoma is present, with a* ***tumor thickness of 0.7 mm****, in association with a cell-rich scar. The cut edges are clear. Microscopic findings: A wide punch biopsy specimen is provided. The epidermis is atrophically flattened, with a predominantly basket-weave stratum corneum. From the epidermis, there is a proliferation of basaloid tumor cell clusters into the upper dermis. Characteristic palisades are positioned in the peripheral area, with contraction artifacts and peritumoral stroma induction. A dense, plasma-rich* ***lymphocytic inflammatory infiltrate*** *is observed peritumorally.*



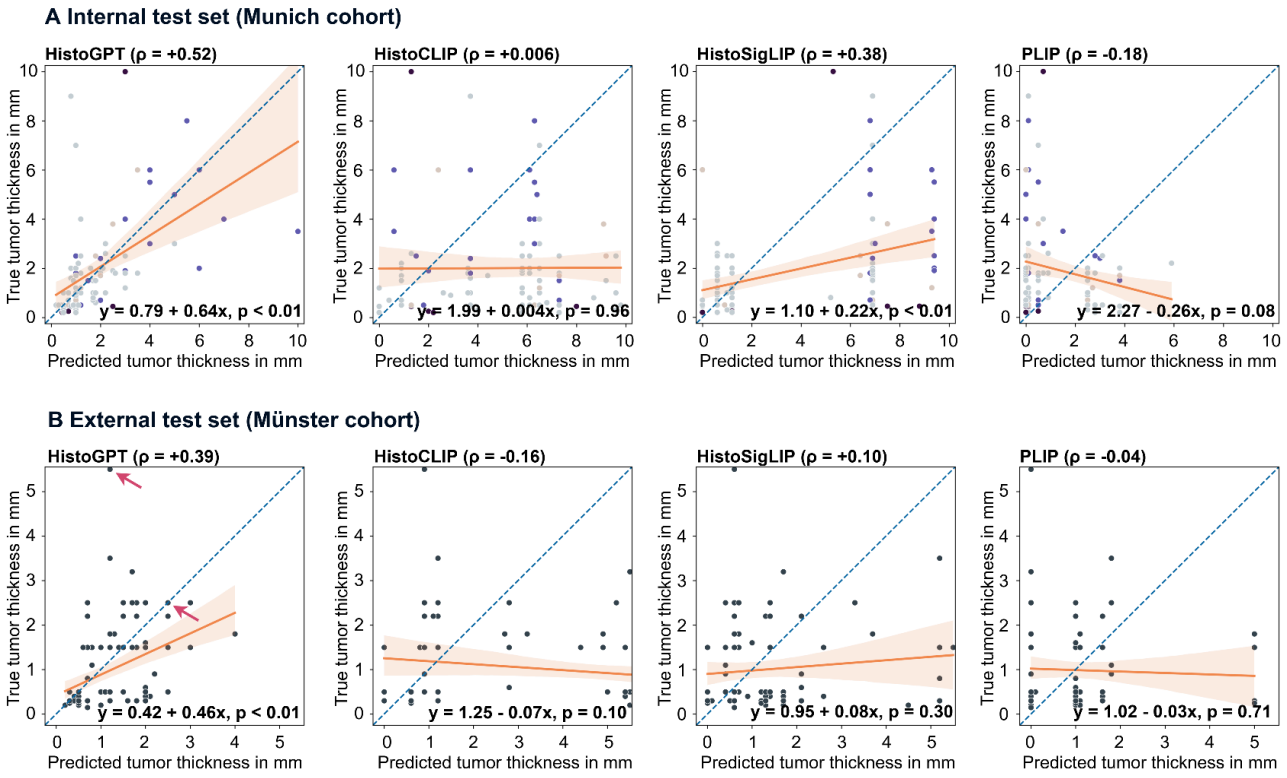
Supplementary Figure 1. Additional example of a generated report with interpretability map. Consider a patient with a suspected basal cell carcinoma. A pathologist can now input the whole slide image into HistoGPT and (optionally) prompt the model ("Expert guidance") with the text "Final diagnosis: Basal cell carcinoma". HistoGPT will then complete the histology report by adding the microscopic and critical findings. For each word or phrase in the text (e.g., "tumor thickness"), we also get a visualization showing which region in the image is associated with which text features.



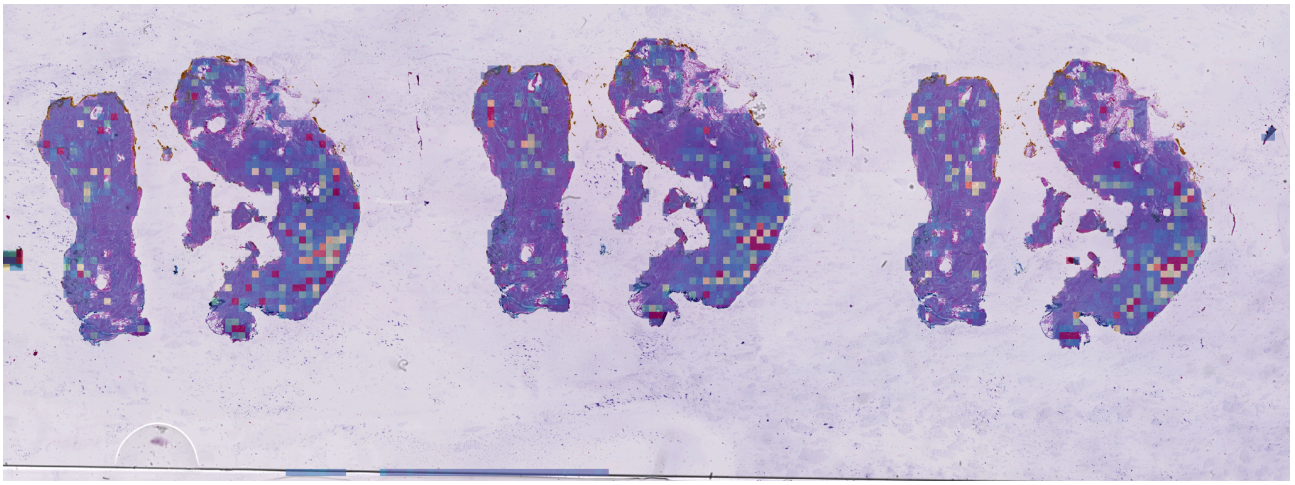
Supplementary Figure 2. Overview of evaluation methods. Using different machine learning metrics based on semantic similarity scores, we show that HistoGPT produces human-quality reports. In particular, the generated reports are evaluated by comparing the extracted medical terms and text embeddings with the ground truth report. In addition, both reports are analyzed by senior pathologists in a blinded study.



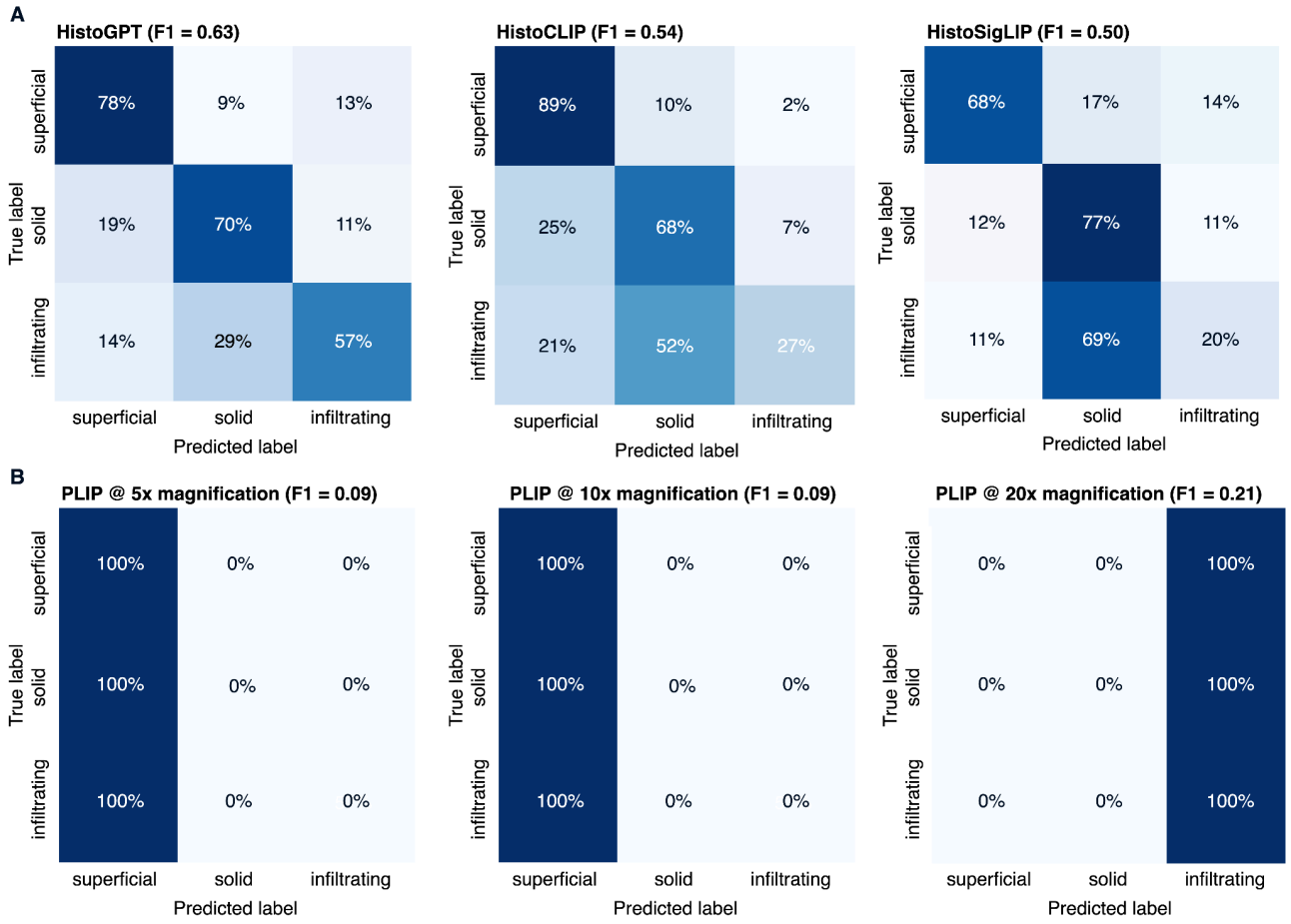
Supplementary Figure 3. Pathologist evaluation in the blinded study. We filter results for the five largest classes for cases where the model would predict the correct class in absence of "Expert guidance".



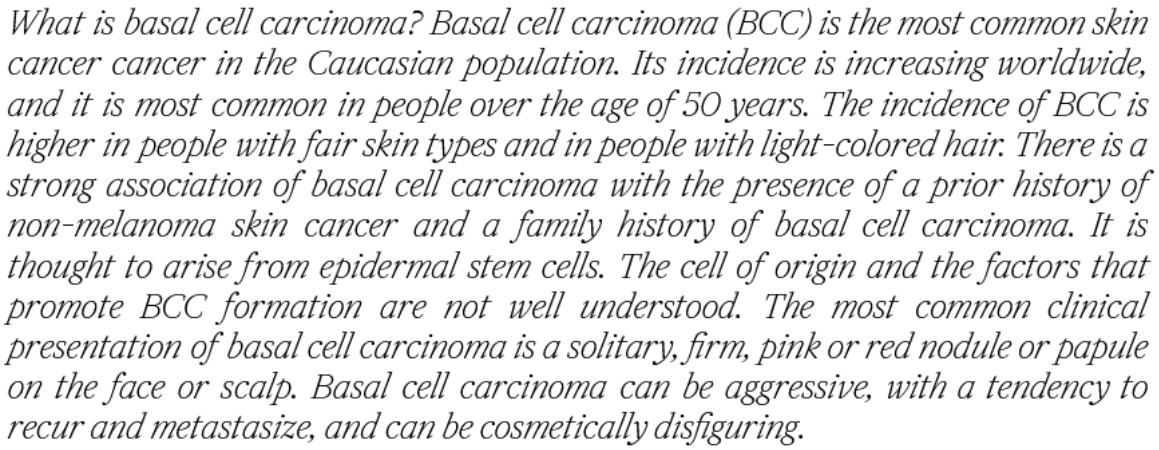
Supplementary Figure 4. Zero-shot tumor thickness prediction. (A) Scatter plot with regression line for the Munich cohort. Data points are color coded according to Figure 3A. (B) Same for the Münster cohort.



Supplementary Figure 5. This sample lacks large parts of the epidermis. Thus, the model did not find a reference point to orient the slide, which likely caused the overestimation of the thickness. The ground truth measure is 1.8 mm and model prediction is 4.8 mm.



Supplementary Figure 6. Results of tumor subtyping in the BCC subset of the Münster cohort.



Supplementary Figure 7. Because the BioGPT language model is frozen during training, HistoGPT can be easily converted to a language-only model by using only text as input, while retaining all the capabilities of the pre-trained BioGPT. Above we see an ensemble refinement for the definition of Basal Cell Carcinoma (BCC).

Supplementary Tables

Disease classification

Munich	precision	recall	f1-score
HistoGPT-3B	0.45	0.47	0.45
HistoGPT-1B	0.43	0.45	0.44
PerceiverMIL	0.42	0.46	0.44
CLSGuidance	-	-	-

Queensland	precision	recall	f1-score
HistoGPT-3B	0.92	0.56	0.64
HistoGPT-1B	0.92	0.57	0.65
PerceiverMIL	0.94	0.64	0.72
CLSGuidance	0.85	0.83	0.83

Linköping	precision	recall	f1-score
HistoGPT-3B	0.71	0.44	0.52
HistoGPT-1B	0.58	0.46	0.51
PerceiverMIL	0.76	0.51	0.59
CLSGuidance	0.70	0.64	0.66

Münster	accuracy

HistoGPT-3B	0.88
HistoGPT-1B	0.88
PerceiverMIL	0.90
CLSGuidance	-

TCGA-SKCM	accuracy
HistoGPT-3B	0.30
HistoGPT-1B	0.29
PerceiverMIL	0.27
CLSGuidance	0.80

CPTAC-CM	accuracy
HistoGPT-3B	0.11
HistoGPT-1B	0.14
PerceiverMIL	0.14
CLSGuidance	0.90

Binary classification problems (HistoGPT with restricted dictionary)

Munich	precision	recall	f1-score
BCC vs NRM	0.96	0.94	0.95
BCC vs ALL	0.97	0.96	0.97
AKK vs SCC	0.83	0.82	0.83
NVC vs SCM	0.92	0.86	0.89

BCC vs NRM	precision	recall	f1-score
accuracy			0.94
macro avg	0.94	0.82	0.87
weighted avg	0.96	0.94	0.95

BCC vs ALL	precision	recall	f1-score
accuracy			0.96
macro avg	0.91	0.95	0.93
weighted avg	0.97	0.96	0.97

AKK vs SCC	precision	recall	f1-score
accuracy			0.82
macro avg	0.82	0.83	0.82
weighted avg	0.83	0.82	0.83

BMN vs SCM	precision	recall	f1-score
accuracy			0.86
macro avg	0.86	0.79	0.82
weighted avg	0.92	0.86	0.89

Binary classification problems (HistoGPT with Classifier guidance)

BCC vs ALL	precision	recall	f1-score
accuracy			0.98
macro avg	0.94	0.96	0.95
weighted avg	0.98	0.98	0.98

AKK vs SCC	precision	recall	f1-score
accuracy			0.88
macro avg	0.87	0.87	0.87
weighted avg	0.87	0.88	0.87

BMN vs SCM	precision	recall	f1-score
accuracy			0.89
macro avg	0.81	0.89	0.84
weighted avg	0.91	0.89	0.89

Basal cell carcinoma subtyping

Münster	precision	recall	f1-score
HistoGPT-3B	0.68	0.59	0.63
HistoCLIP	0.63	0.57	0.54
HistoSigLIP	0.53	0.53	0.50

Tumor thickness prediction

Munich	rmse	pearson	p-value
HistoGPT	1.7965	+0.5167	9.6945e-08
HistoCLIP	4.3549	+0.0057	0.95619369
HistoSigLIP	3.8409	+0.3786	0.00016752
PLIP	2.7834	-0.1787	0.08468900

Munich	beta0	beta1	p-value
HistoGPT	0.7930	+0.6357	9.6945e-08
HistoCLIP	1.9850	+0.0042	0.95619369
HistoSigLIP	1.1020	+0.2205	0.00016752
PLIP	2.2663	-0.2594	0.08468900

Münster	rmse	pearson	p-value
HistoGPT	0.9772	+0.3870	5.8530e-05
HistoCLIP	3.9079	-0.1637	0.10009248
HistoSigLIP	1.4632	+0.1014	0.31048499
PLIP	1.4326	-0.0371	0.71066124

Münster	beta0	beta1	p-value
HistoGPT	0.4220	+0.4625	5.8530e-05
HistoCLIP	1.2530	-0.0666	0.10009248
HistoSigLIP	0.9001	+0.0770	0.31048499
PLIP	1.0199	-0.0325	0.71066124

Automatic report evaluation

Munich	dictionary	scispacy	biobert	gpt-3-ada
HistoGPT-ER	0.73	0.68	0.75	0.92
Guided	0.77	0.70	0.76	0.94
HistoGPT-3B	0.64	0.56	0.75	0.92
Guided	0.67	0.59	0.80	0.94
HistoGPT-1B	0.63	0.56	0.75	0.92
Guided	0.66	0.58	0.79	0.94
GPT-4-Vision	0.54	0.55	0.50	0.86
Guided	0.62	0.61	0.67	0.91
BioGPT-1B(F)	0.44	0.41	0.64	0.89
Guided	0.61	0.53	0.77	0.93
BioGPT-1B(P)	0.12	0.10	0.41	0.82
Guided	0.12	0.14	0.55	0.88
Lower bound	0.44	0.41	0.62	0.88

Upper bound	0.66	0.58	0.77	0.93

Munich	bleu-4	meteor	rouge-l	bertscore
HistoGPT-3B	0.07	0.21	0.23	0.71
Guided	0.11	0.22	0.24	0.72
HistoGPT-1B	0.08	0.22	0.23	0.71
Guided	0.11	0.23	0.25	0.72
BioGPT-1B(F)	0.01	0.16	0.17	0.65
Guided	0.10	0.23	0.24	0.71
BioGPT-1B(P)	0.02	0.10	0.11	0.54
Guided	0.04	0.22	0.15	0.60
Lower bound	0.01	0.15	0.16	0.65
Upper bound	0.13	0.24	0.27	0.73

Münster	dictionary	scispacy	biobert	gpt-3-ada
HistoGPT-ER	0.59	0.60	0.50	0.86
HistoGPT-3B	0.46	0.49	0.51	0.86
HistoGPT-1B	0.46	0.49	0.51	0.86
GPT-4-Vision	0.16	0.51	0.31	0.79
BioGPT-1B(F)	0.29	0.39	0.44	0.83
BioGPT-1B(P)	0.06	0.04	0.25	0.78
Lower bound	0.17	0.32	0.40	0.83

Pathologist 1 report evaluation

original assessment
true_report_preferred, generated_report_preferred, ties
(42, 14, 44)

adjusted for close ties
true_report_preferred, generated_report_preferred, ties
(35, 9, 56)

correct predictions only
true_report_preferred, generated_report_preferred, ties
(17, 7, 33)

```
adjusted for ties and correct predictions only
true_report_preferred, generated_report_preferred, ties
(11, 3, 43)

5 largest classes
true_report_preferred, generated_report_preferred, ties
(17, 7, 24)
```

Pathologist 2 report evaluation

```
original assessment
true_report_preferred, generated_report_preferred, ties
(45, 10, 45)

adjusted for close ties
...

correct predictions only
true_report_preferred, generated_report_preferred, ties
(18, 5, 34)


adjusted for ties and correct predictions only
...

5 largest classes
true_report_preferred, generated_report_preferred, ties
(18, 6, 24)
```

Pathologist 1 and 2 report evaluation per class

```
true_report_preferred, generated_report_preferred, ties
BCC: (5, 6, 5), (9, 1, 6)
BMN: (6, 0, 7), (3, 1, 9)
SKK: (1, 0, 6), (1, 2, 4)
AKK: (2, 1, 4), (3, 1, 3)
SCC: (3, 0, 2), (2, 1, 2)
```

Dataset overview

Munich: num_scenes = 47,441, num_slides = 13,232, storage_size = 9T

Cohorts	Samples	Reports	Classes	Split
Munich	6,000	YES	162	Train / Test
Münster	1,300	YES	BCC & others	Test
Queensland	290	NO	BCC, SCC, IEC	Test
Linköping	99	NO	13	Test
TCGA-SKCM	292	NO	Melanoma	Test
CPTAC-CM	92	NO	Melanoma	Test