Quilt-LLaVA: Visual Instruction Tuning by Extracting Localized Narratives from Open-Source Histopathology Videos

Mehmet Saygin Seyfioglu♠* Wisdom O. Ikezogwo ♠ Fatemeh Ghezloo ♠
Ranjay Krishna Linda Shapiro
University of Washington

Abstract

Diagnosis in histopathology requires a global whole slide images (WSIs) analysis, requiring pathologists to compound evidence from different WSI patches. The gigapixel scale of WSIs poses a challenge for histopathology multi-modal models. Training multi-model models for histopathology requires instruction tuning datasets, which currently contain information for individual image patches, without a spatial grounding of the concepts within each patch and without a wider view of the WSI. To bridge this gap, we introduce QUILT-INSTRUCT*, a large-scale dataset of* 107,131 *histopathology-specific instruction question/answer pairs, grounded within diagnostically relevant image patches that make up the WSI. Our dataset is collected by leveraging educational histopathology videos from YouTube, which provides spatial localization of narrations by automatically extracting the narrators' cursor positions.* QUILT-INSTRUCT *supports contextual reasoning by extracting diagnosis and supporting facts from the entire WSI. Using* QUILT-INSTRUCT*, we train* QUILT-LLAVA*, which can reason beyond the given single image patch, enabling diagnostic reasoning across patches. To evaluate* QUILT-LLAVA*, we propose a comprehensive evaluation dataset created from* 985 *images and* 1283 *human-generated question-answers. We also thoroughly evaluate* QUILT-LLAVA *using public histopathology datasets, where* QUILT-LLAVA *significantly outperforms SOTA by over 10% on relative GPT-4 score and 4% and 9% on open and closed set VQA*[1].

1. Introduction

In histopathology, analyzing giga-pixel-level whole slide images (WSIs) requires a nuanced understanding beyond examining isolated patches taken within the WSI. Pathologists typically analyze these WSIs holistically, compounding

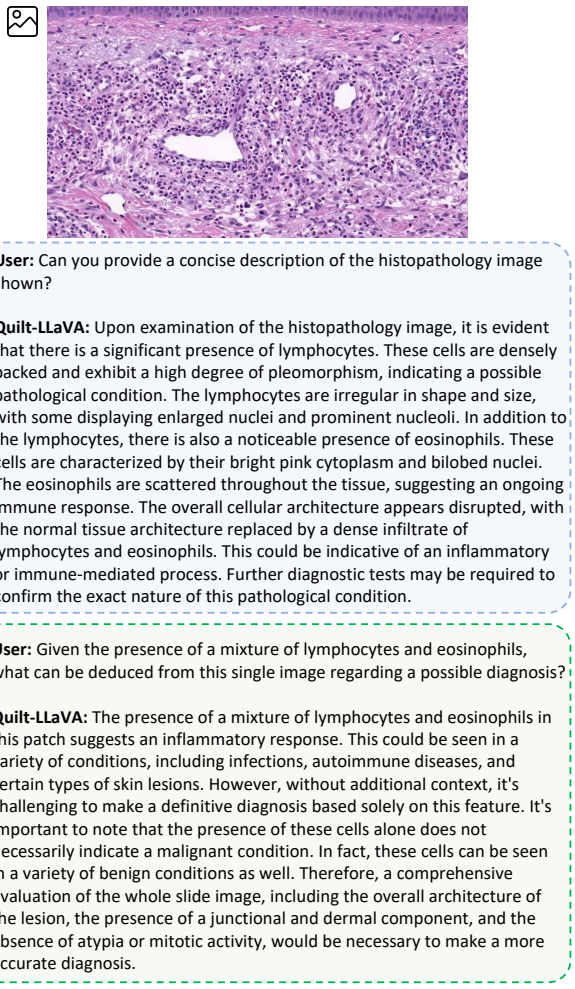

Figure 1. QUILT-LLAVA is capable of describing the prominent medical regions within a histopathology patch. Additionally, it can be utilized to reason towards a diagnosis based on the current observations. Note: The image includes eosinophils and lymphocytes, and is sampled from a WSI showing rare benign dermatitis.

*msaygin@cs.washington.edu; ♠ Denotes equal contribution.
[1]Our code, data, and model is publicly accessible at quilt-llava.github.io.

evidence from various regions to form a diagnosis. While navigating the WSI regions, pathologists engage in a critical reasoning process over a multi-turn conversation, determining what morphological concepts within the regions confirm or rule out specific diagnoses, making the process dynamic and context-sensitive. For example, in liver tissue WSI, isolated steatosis (fat accumulation) might not be indicative of significant pathology, but if accompanied by ballooning of hepatocytes and lobular inflammation, it strongly suggests non-alcoholic steatohepatitis, a more serious condition that could lead to cirrhosis or liver cancer [5]. This approach involves understanding the diagnostic relevance of histological visual features, spatial relationships, and aggregating insights across the WSI [1]. Although current histopathology multi-modal models can analyze isolated image patches effectively, they often lack the capability to reason beyond that patch to determine next steps, thereby limiting their utility.

Meanwhile in natural images, the success of the Large Language and Vision Assistant (LLaVA) [14] demonstrates the possibility of multi-modal models as chat partners. LLaVA is trained by extracting instruction-tuning data with Large Language Models (LLMs), such as GPT-4 [19]. Moreover, photos usually include multiple objects which can be referred to using bounding boxes, which can be used to create spatially grounded instruction-tuning data [14]. Recently, multi-modal models have also begun employing video content, allowing for more complex reasoning beyond single images [13]. However, this progress has not translated to histopathology yet, where existing multi-modal models often rely on PubMed articles for image-caption pairs [14, 18]. These articles lack two critical elements: **1)** visually grounded captions necessary for effective visual grounding of histopathology concepts, and **2)** the broader contextual information from the WSI, essential for reasoning beyond single image patches, limiting their effectiveness in detailed histopathological analysis.

To fill this void, we present QUILT-INSTRUCT, an instruction-tuning dataset of $107,131$ histopathology-specific question/answer pairs. Similar to QUILT [9], QUILT-INSTRUCT capitalizes on educational histopathology videos from YouTube, featuring pathologists narrating their examination of WSIs. **1)** To alleviate the limited spatial awareness, we extract narrators' mouse cursors from videos, grounding histopathology concepts within each patch using spatio-temporal clustering. **2)** To enable reasoning, we propose novel instruction-tuning QA prompting techniques for histopathology: Complex reasoning, and iterative abductive reasoning, which incorporates the global WSI diagnosis and its supporting facts with image captions to ground factual information, preventing hallucinations (see Fig. 18).

Using QUILT-INSTRUCT, we train QUILT-LLAVA, a multi-modal model for histopathology, with its capabilities illustrated in Fig 1. QUILT-LLAVA undergoes a two-stage training process (see Fig 3). First, it is aligned with the histopathology domain using 723K image-text pairs from QUILT [9], and then it is further instruction-tuned with QUILT-INSTRUCT. QUILT-LLAVA analyzes given images in detail, spatially localizes medical concepts, and reasons beyond the given image patch by guiding users on what further evidence is needed to validate or rule out certain diagnoses, and can even be used as an educational tool that, instead of directly revealing a diagnosis, subtly hints at it, aiding pathology students in their training (Appendix Section 2.1.)

Finally, educational videos are valuable for their interactivity, as narrators often engage viewers by asking and answering questions throughout their presentations. Leveraging this, we propose QUILT-VQA, an organic evaluation dataset extracted from naturally occurring questions and answers from QUILT videos with the help of GPT-4 and manual verification, to evaluate QUILT-LLAVA's reasoning capabilities. We evaluate QUILT-LLAVA on QUILT-VQA and two public histology VQA test sets on both open and closed questions. Using red circle ◯ [22] marking the area of interest in the image, we can prompt QUILT-LLAVA to focus on specific regions of the image patch. We outperform the SOTA by 4% and 9% on open and closed set VQA tasks. To further evaluate the reasoning capabilities of QUILT-LLAVA, we utilize GPT-4 to score the model's generated response against two other SOTA multi-modal models: LLAVA [16] and LLAVA-MED [14]. QUILT-LLAVA outperforms LLAVA and LLAVA-MED by over 16% and 7%, respectively, and upon increasing the size of instruction-tuning data and only pre-training for three epochs, we achieve even better results of 10.8% over LLAVA-MED.

2. Related work

We built our work from the expanding body of literature in visual instruction-tuning and video-based dataset generation, with a specific focus on their application in the areas of medical image analysis and histopathology.

Visual instruction-tuning in natural images. Thanks to the open-source availability of LLMs [3, 11, 23, 24], studies in the general vision-language domain have advanced the training of multi-modal models by harnessing implicit visual understanding through generating an instruction-tuning dataset from image captions. Prior work [4, 17, 32] demonstrated significant capabilities, with LLaVA-1.5 matching GPT-4's performance in certain multi-modal tasks [16].

Visual instruction-tuning in medical images. Visual Med-Alpaca [6] created 54K question-answer pairs for instruction-tuning using GPT-3.5. PMC-VQA [31] curated a larger multiple choice answer based dataset from general medical domains using PubMed, yet its coverage of histopathology is limited. For instance, LLaVA-Med [14] employs image captions from PubMed articles and, in cases of brief captions, supplements them with sentences from the article, which

may not directly pertain to the referenced figure. Within LLaVA-Med, a subset of 17k images relates to histology, yielding 49K question-answer pairs. Furthermore, these works rely on isolated image-caption pairs for constructing instruction-tuning datasets, which limits GPT-4's capacity to reason beyond its context or, if attempted, raises its likelihood of hallucination.

Video-based image-text datasets. Multiple works have attempted to curate datasets from videos in the natural domain [13, 25–28]. MIMIC-IT [13] uses video data to create various prompts, including some reasoning-based ones, to create instruction-tuning datasets. Video Localized Narratives [25] involve annotators verbally describing an image while concurrently moving their mouse cursor over the regions they refer to, yielding dense visual groundings. Acquiring a dataset in the same way is expensive in histopathology. PathNarratives [29] employed eight pathologists to manually annotate spatially grounded pathology data; however, it is not available for open access. More recently, QUILT [9] was proposed, which curates large-scale data from educational histopathology content from YouTube to create image-caption pairs, however, it lacks spatially grounded captions. We expand upon QUILT's video content to generate grounded histopathology data. To the best of our knowledge, our work is the first to use videos for creating spatially grounded instruction-tuning datasets in the medical domain.

3. Curating QUILT-INSTRUCT

To construct a comprehensive histopathology instruction-tuning dataset with visual groundings in WSIs, we harness the rich narrative content of educational YouTube videos.

3.1. Data preparation

We make the key observation that, in educational videos, narrators often pause while exploring large-scale WSIs before indicating salient areas with their cursor [10, 12, 20, 25]. Our process uses three steps to convert unstructured videos into usable visually-grounded instruction data: First, we **localize narrators' cursors** in videos. Then, we perform **spatio-temporal clustering of cursor location** to visually ground histopathology concepts within images. These two steps are illustrated in Fig. 2. Finally, using the extracted grounded captions, we use an LLM to **generate our instruction-tuning dataset - QUILT-INSTRUCT**. This process involves prompting techniques, from independent prompts generating diverse Q/A pairs for each image patch to reasoning-based prompts combining information across patches in WSIs, creating Q/A pairs that reason towards diagnoses.

Data source. We construct QUILT-INSTRUCT from 4149 educational YouTube videos totaling over 1,000 hours of content. These videos are part of a recent histopathology dataset, QUILT [9]. The videos feature a *narrative style*, where pathologists articulate histopathological concepts while interacting with WSIs. QUILT provides rich image-text data but does not offer spatial annotations to connect text descriptions with specific regions in the images—a key aspect often absent in most medical datasets. Fortunately, narrators' mouse cursors highlight morphological features, offering visual grounding to bridge this gap. We next detail the process of extracting these cursors from raw video footage.

Localizing Cursors. Extracting the cursor location from histopathology images poses a significant challenge due to the homogeneity in color and texture, alongside the presence of other dynamic elements in videos, such as minor pixel variations and variations in the narrators' cursor movement speed and style. Despite these obstacles, we have made a critical observation: Narrators typically pause their exploration of a WSI before gesturing with their cursor, guiding the audience's attention. Therefore, if the background is stable for several consecutive frames, it is highly likely that the narrator may be using their cursor to mark medically relevant regions. Leveraging this observation, we employ a relatively straightforward method to detect mouse cursors. First, we isolate segments in each video where the background is mostly static, termed stable chunks shown in the leftmost section of Fig. 2. To detect the stable chunks, we utilize a simple algorithm (Algorithm 1 in the Appendix) that detects stable chunks of frames by employing a frame-differencing approach to detect chunks with minimal background movement. Our algorithm computes the absolute difference between consecutive frames and then applies a Gaussian filter for adaptive thresholding to pinpoint frames with minor changes. Given the typically uniform coloration of histopathology backgrounds due to histopathology staining, simple pixel-wise difference yields a high rate of false positives, namely, chunks that had a change occur in them are identified as stable. To alleviate this, we incorporate a perceptual metric, Structural Similarity Index Measure (SSIM). By evaluating SSIM on randomly sampled patches, we further verify whether a change has occurred between two frames. For each stable chunk where the background remains consistent, we calculate a median frame in the pixel domain to represent the background. This median frame is then subtracted from the frames within the stable chunk, and a thresholding is applied to mitigate noise or other artifacts, effectively isolating the cursor as illustrated in the middle section of Fig. 2. Subsequently, we capture the cursor by identifying the coordinates of the maximum pixel value. However, this approach assumes the absence of other slight movements with respect to pixel changes, which is not always the case, as there may be movements like the narrator's facial expressions within the scene. We address this by applying a face detection model [21] to mask such distractions, ensuring our analysis concentrates solely on cursor movement. This algorithm offers a surprisingly generalizable way

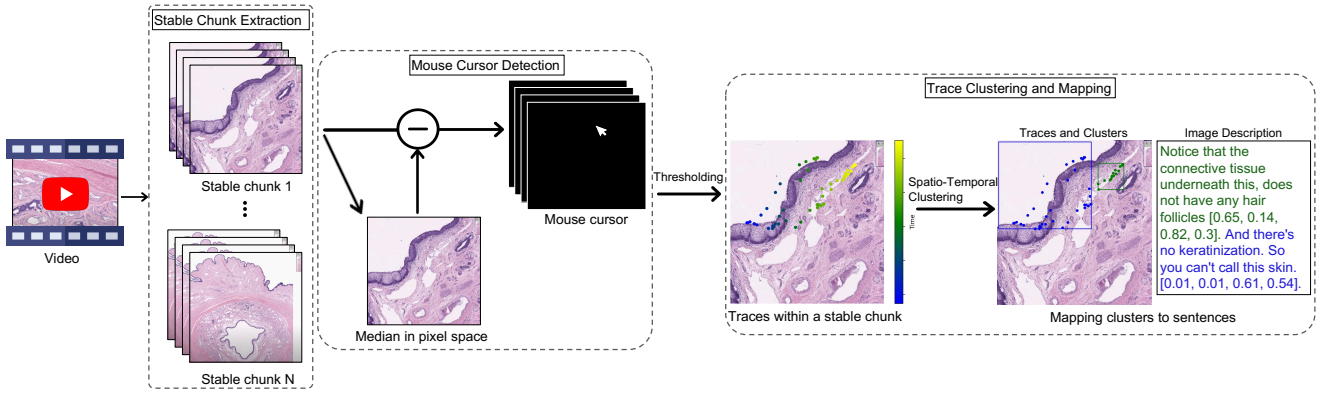

Figure 2. To create QUILT-INSTRUCT, we first identify stable chunks within the video. For each chunk, we compute a median frame in the pixel domain and subtract it from every frame within the chunk. We then apply a threshold to reduce noise and take the maximum value to capture the mouse cursor points. These cursor points are then clustered to localize medical content in image captions. Please note that color encodes time in the "Trace Clustering and Mapping" part of the figure.

to collect cursor traces from any educational videos with similar *narrative styling* at a meager computational cost.

After pinpointing these candidate stable chunks, we retain those with a minimum duration of three seconds. Subsequently, we utilize a histology classifier as described in [9] to eliminate non-histological segments using each stable chunk's background frame. This methodology generates 132K image-caption pairs, represented as $\langle I_j^v, c_j^v \rangle$, where I_j^v is the median frame of the j^{th} stable chunk within the v^{th} video, and c_j^v is the corresponding caption for I_j^v, which is obtained by converting the narrator's speech to text similar to the strategy outlined in [9]. Each pair spans an average of 24 seconds, corresponding to 55 words in the caption. From these 132K image caption pairs, 44,163 exhibits active cursor movement where we preserved cursor position as (x_j^t, y_j^t), where $x \in [0, W]$ and $y \in [0, H]$, with W and H representing the image width and height, respectively, t spans from 0 up to the total duration of the j^{th} stable chunk. Next, we demonstrate our approach for clustering extracted mouse cursors, which serves to visually ground our captions.

Spatio-temporal clustering of the mouse cursor. Next we proceed to cluster the tracer points (x_j^t, y_j^t) for each stable chunk to spatially ground narrators' words. For spatio-temporal clustering, we transformed our inputs into (x_j^t, y_j^t, t, w_j^t), where t is time and w_j^t denotes the cumulative word count at each cursor position, to encode the pace of speech better, and we have normalized the inputs to ensure each input is scaled within the standardized range of 0 to 1. Given this input format, to achieve a balance between spatial and temporal aspects — specifically, to determine the relative importance of spatial proximity and temporal proximity in cluster formation — we apply an exponential decay to the spatial coordinates, $e^{-\lambda t}$, where λ is set at 0.05, ensuring that points closer in time are preferentially clustered together. The algorithm dynamically selects the number of clusters based on the number of words in c_j^v. To refine the process of mapping words to clusters, we first determine the temporal center of each cluster. This is achieved by calculating the mean timestamp of the tracer points within a cluster, which gives us a 'temporal midpoint.' Subsequently, we assign words to clusters based on their temporal proximity to these midpoints, ensuring that each word matches the cluster whose average time is closest to the word's occurrence time. Finally, the resulting clusters are represented by bounding boxes, denoted in the format [x1, y1, x2, y2], where each coordinate is a floating-point number normalized between 0 and 1. These coordinates specify the boxes' top-left (x1, y1) and bottom-right (x2, y2) corners. A sample is shown in Fig. 2. Next, we show how we used our grounded captions to generate our instruction-tuning data QUILT-INSTRUCT.

3.2. Generating QUILT-INSTRUCT

We generate two sets of question-answer types. First, we employ Conversations and Detailed Descriptions prompts, akin to those in [17], that take as input single patch-level grounded-text to generate Q/A pairs constrained by the independent input sample; hence we call these prompts – Independent prompts. Secondly, we leverage the contextual continuity inherent in our dataset – that is, for a video reviewing a single WSI, we leverage the sequential unraveling of concepts/clues toward a final diagnosis, by introducing novel Reasoning-based Prompts which receive as input patch-level grounded-text and global WSI-level information. These enable an LLM (GPT-4) to extend its reasoning beyond its immediatecontext while still being anchored by the factual insights derived from the entire video, reducing its hallucinative behavior.

3.2.1 Independent Prompts

Following the approach of [17], we generate Q/A pairs from each image caption c_j^v, creating an array of questions that aim to explain the context presented in the image. However, these prompts are designed to elicit answers based solely on the information within the caption and do not attempt to extrapolate beyond it.

Conversation. The objective of the Conversation prompt is to construct a dialogue between an AI assistant and an individual inquiring about a histopathology image. We generate a diverse set of Q/A pairs based on the image caption c_j^v, which includes bounding boxes of medical concepts. The assistant, equipped with descriptions and bounding box coordinates, responds as if directly observing the image, discussing elements like cell types, activities, and relative positions of medical concepts within the image. The dialogue includes around 3-4 Q/A pairs, maintaining a conversational tone while focusing on the visual content of the images.

Detailed Descriptions. Similar to [17], we instruct GPT-4 to use c_j^v to generate a detailed description for a given image. Subsequently, questions are randomly selected from a pre-compiled list given in the Appendix Fig.19.

3.2.2 Reasoning-based Prompts

Most image-caption pairs used to create instruction-tuning datasets in other works, such as PubMed or COCO [15], exhibit contextual isolation, meaning $< \mathrm{I}_i, \mathrm{c}_i >$ and $< \mathrm{I}_j, \mathrm{c}_j >$ where $i \neq j$ do not share context with each other. Consequently, Q/A pairs generated by GPT-4 will be constrained to the context of a single image. Given the gigapixel nature of histopathology images, deriving a comprehensive diagnosis based on a single image patch is often inadequate. The model should ideally reason beyond the given image, guiding the human user on subsequent steps or areas to scrutinize. This may be attempted using GPT-4's inherent medical knowledge. However, coaxing GPT-4 to extrapolate beyond a single caption without an interconnected context could lead to hallucinations, compromising data quality.

In contrast, our approach capitalizes on video-extracted image-caption pairs derived from a single patient's WSI. To that end, we manually reviewed the entire video content of QUILT identifying 2066 videos that feature only a single WSI for our reasoning-based prompts. By utilizing the entire text from these single whole-slide videos, we initially used GPT-4 (see Fig. 10 in Appendix) to infer the final diagnosis d_j^v and the supporting facts f_j^v. Utilizing f_j^v and d_j^v enables us to craft reasoning-based prompts that implicitly direct GPT-4 towards abductive reasoning. This approach fosters extrapolations more anchored in context, reducing hallucination tendency. We developed two prompt types, **Complex Medical Reasoning** and **Iterative Abductive Reasoning**, for reasoning-based instruction-tuning dataset generation.

Complex Medical Reasoning. Given a caption c_j^v, along with a diagnosis d_j^v and contributory facts f_j^v, we prompt GPT-4 in a diagnostic reasoning task designed to extrapolate beyond the immediate context of c_j^v. More broadly, we instruct GPT-4 to utilize its inherent medical knowledge to interpret the contents of a single image caption c_j^v, while subconsciously incorporating the diagnosis and supporting facts extracted from the entire video. If the observations from c_j^v suffice for making a diagnosis based on GPT-4's general medical knowledge, it proceeds to provide a diagnosis along with the facts leading up to it. If not, it extrapolates using f_j^v to suggest what to look for in the WSI to validate or rule out certain diagnoses. We term this approach "Subconscious Knowledge Injection," as it allows GPT-4 to leverage its medical knowledge while being subtly guided by the provided d_j^v and f_j^v, effectively constraining its context and ensuring focused reasoning and fewer hallucinations.

Iterative Abductive Reasoning: We simulate a conversation between two GPT-4 agents, mimicking a scenario where a professional pathologist uses our model to ask longer medically intricate questions about an image. This contrasts with the shorter questions typically found in other prompts that we use. The first agent, termed Human-GPT, is provided with an image patch caption c_j^v and is tasked with abductively reasoning about the possible diagnoses and the facts used to arrive at these conclusions. This is presented in the format: *User: [{Abduction: xxx}, {Facts Used: xxx}]*. The second agent, referred to as the AI Assistant GPT, is privy to the diagnosis d_j^v and facts f_j^v, simulating someone who has viewed the WSI of this particular patient. The AI Assistant evaluates the accuracy of the abduction derived by Human-GPT and provides comments or hints at potentially overlooked details using its inherent medical knowledge while utilizing d_j^v and f_j^v. This is communicated in the format: *GPT: [{Comments: xxx}, {Hint: xxx}]*. If Human-GPT has exhausted all information from c_j^v and a diagnosis cannot be made, the AI Assistant suggests exploring other patches for additional evidence, such as "Consider looking for evidence of X in other patches to validate your diagnosis.". Furthermore, leveraging Iterative Abductive Reasoning-based instruction-tuning data, QUILT-LLAVA can serve as an educational tool, aiding human users in brainstorming by providing hints about the next steps, without fully revealing the answer. (See Fig. 16 in Appendix). The conversation between the AI Assistant and Human-GPT proceeds with an upper limit of iterations, randomly selected between two to four interchanges per agent, where after each exchange, the start of a new exchange incorporates the post-exchange history, allowing GPT to retain the memory of past conversations. The dialogue may terminate before reaching this limit if Human-GPT arrives at a conclusive diagnosis, or if the AI Assistant determines that Human-GPT has fully exhausted all relevant information from c_j^v, and then guides it on what to do next using f_j^v.

Dataset Statistics. We extracted 162,566 image-caption pairs from QUILT. To refine this data, we filtered out captions with fewer than 20 words and those with more than 150 words. This process resulted in a dataset of 114,343 pairs, with an average caption length of 55 words. From this, we created QUILT-INSTRUCT, comprising 107,131 question/answer pairs where, on average, we have questions with 16.5 words and answers with 101 words. For reasoning-based prompts, we manually reviewed 4,149 videos and selected 2,066 that focused on a single WSI from a single patient.

4. Training QUILT-LLAVA & evaluating with QUILT-VQA

In this section, we detail using QUILT-INSTRUCT to train QUILT-LLAVA. Next, we curate QUILT-VQA independently of QUILT-INSTRUCT, to evaluate QUILT-LLAVA. Finally, we generate an Instruction Following Test Set from QUILT-VQA to assess QUILT-LLAVA's ability to follow instructions.

4.1. Training QUILT-LLAVA

We embrace the LLAVA autoregressive model architecture for its simple yet efficient design. Additionally, our selection of the LLAVA architecture aimed to ensure consistency in evaluation against our baselines LLAVA [17] and LLaVA-MED [14], both of which utilize a curriculum-learning strategy [14] on instruction tuning sets. Overall, LLAVA integrates a vision module, an LLM, and an MLP connector, allowing the LLM to process visual information. Initially, the MLP —serving as a projector— is trained until it converges. During this stage, both LLM and the vision module kept frozen. Subsequently, both the MLP and the LLM are fine-tuned with instruction-following data to align the model with human pathologists. Our overall architecture is shown in Fig. 3. LLAVA typically uses a pre-trained CLIP image encoder; for our domain, we use the pre-trained CLIP model trained using public histopathology datasets such as QUILT-NET [9] and PLIP [8]. We also run ablations with various image encoders, training strategies, and visual prompts.

Histopathology domain alignment. First, we align our vision and language models within the histopathology domain. To that end, we extract 723K image-text pairs from QUILT and convert the captions into the Q/A format. To do that, we randomly select an instruction (Question) and prepend it to the caption (Answer) to create Q/A pairs. The instructions, drawn from a predefined list (See Appendix Fig. 18), are designed to variably describe the visual information present in the images. At this stage, we freeze vision and language models and only train the MLP layer, which is tasked to project the embeddings from the image encoder to enable the language model predict the images' caption given the question. This stage aligns the histology image embeddings to their corresponding text embeddings.

Histopathology instruction-tuning. Finally, we fine-tune our model with QUILT-INSTRUCT. At this stage, we keep the visual encoder weights frozen and continue to train the MLP layer and the language module.

4.2. Evaluation Data Generation: QUILT-VQA

In the field of histopathology, researchers rely on evaluation datasets like PathVQA [7] and PMC-VQA [31] to assess the performance of their models. However, these datasets exhibit notable shortcomings, including significant repetitiveness due to paraphrasing the same question. Even worse, oftentimes, there are contradictory answers to the same question (See Appendix Section 3.4). In contrast, educational video content provides a valuable resource: the interactive element introduced by narrators who often pose questions during their presentations and subsequently provide the answers themselves. For instance, a narrator says "*Do you know what kind of organ we're dealing with?*" and then proceeds to elaborate by "*Yes, this is a colon.*". This Q/A format within the videos offers a rich organic Q/A dataset to be extracted and repurposed for evaluation.

To harness this potential, we map the "?"s in the video's transcript to our stable chunks. If a "?" appears within a 45-second time frame of a stable chunk, we expand the stable chunk's caption to encompass the complete sentence that includes the question mark. This method ensures that the questions are related to the visual content displayed. Following the data pre-processing and the strategic mapping of question marks, we prompt GPT-4 to extract the question and answer pairs directly given in a text. We input GPT-4 the stable chunk's text, as well as any sentences from the text that end in a question mark, indicating a posed question. Fig. 11 in the Appendix demonstrates our GPT-4 prompt to generate QUILT-VQA. After the initial extraction by GPT-4, we perform a manual verification, ensuring that each question-answer pair in our dataset is not only medically relevant but also correctly corresponds to the content provided in the stable chunk's text. See table 7 in the Appendix for statistics of QUILT-VQA. Additionally, we categorize questions into two groups: Image-dependent (1055 Q/A pairs), referencing the narrator's specific image, and general-knowledge (228 Q/A pairs), rooted in broader medical understanding. This division lets us craft evaluation sets that comprehensively test the model's image analysis and medical knowledge.

4.3. Evaluation data generation: Instruction Following Test Set

In addition to QUILT-VQA, which is focused on assessing QUILT-LLAVA's medical knowledge, we also aimed to evaluate the model's capability in following instructions during multi-modal conversations. To that end, we constructed a set of 326 questions, including 256 conversational and 70 detailed description questions, all derived from image-text

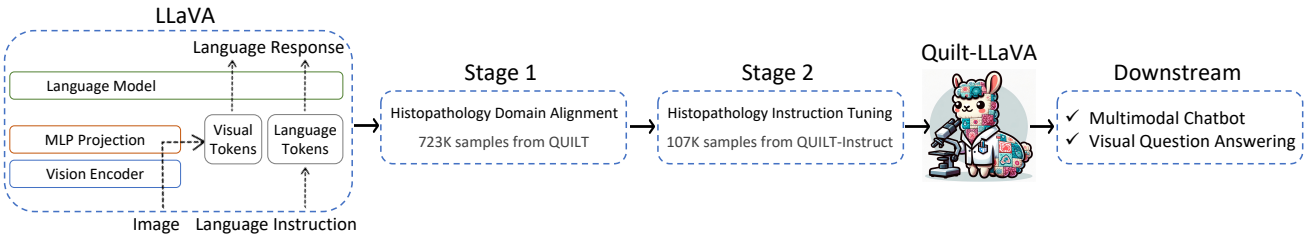

Figure 3. QUILT-LLAVA was initialized with the general-domain LLaVA and trained for two stages: Histopathology Domain Alignment on QUILT and instruction-tuning on QUILT-INSTRUCT. We evaluated QUILT-LLAVA on visual conversation and question answering tasks.

pairs extracted from unseen videos within QUILT-VQA. To generate this evaluation set, we followed the same Conversation and Detailed Description based prompts that we used when generating QUILT-INSTRUCT.

5. Experiments

This section covers the performance of QUILT-LLAVA against existing SOTA multi-modal models on histopathology VQA benchmarks using a variety of metrics. First, using GPT-4 we conducted an alignment of output generations with the ground truth answer. Second, we run open and closed set VQA tasks. Finally, we ablate with visual promptings as well as different trained models.

Oracle (GPT-4) Alignment Evaluation. To assess the effectiveness of QUILT-LLAVA in multi-modal conversations, using the Instruction Following test-set (see section 4.3), we leverage the GPT-4 (language-only model) to evaluate the *helpfulness, relevance, accuracy, and level of details* of the responses from the two assistants (candidate model and GPT-4), and give an overall score on a scale of 1 to 10 of the answers generated from the two assistants, where a higher score indicates better overall performance, while also providing comprehensive explanation of its evaluation, for us to better understand the models. We then compute the relative score using GPT-4 reference score for normalization [14, 17], as seen in Table 1a across 14 sub-pathologies and 2 QA types. See Fig. 12 in the Appendix for the exact prompt used. All QUILT-LLAVA models outperform baselines, with QUILT-LLAVA trained on single epoch stage-1 and a balanced subset of 40K instruction-tuning pairs for stage-2 fine-tuning outperforming LLAVA and LLAVA-MED by over 16% and 7% respectively, and upon increasing the size of instruction-tuning data and pre-training for three epochs we achieve even better results of 10.8% over LLAVA-MED.

Visual Question Answering. We evaluated QUILT-LLAVA on three histopathology VQA datasets, the details of which are provided in Table 7 in the Appendix. These datasets feature a mix of open-ended and close-ended Q/A pairs. For closed-set questions, accuracy is used as the metric to determine the proportion of correct answers given by the model. In contrast, for open-set questions, we focus on recall to assess how frequently the model's responses include ground-truth tokens [14]. We compare all versions of QUILT-LLAVA against the medical domain LLAVA-MED, and general domain LLAVA in Table 1b. All variants of QUILT-LLAVA outperform LLAVA. We see the difference in text encoder initialization from Vicuna [3] in the first set of experiments paired with pre-trained open-sourced QUILTNET models doing better on open-set questions with an average of 4% over SOTA and the text encoder initialization from LLAVA doing better on closed-set questions with an average of 9% over SOTA. Performance of QUILT-LLAVA on closed-set questions showcases our model's instruction following capacity (in both binary *yes/no* and true multi-choice *{A/B/C/D}* QA unlike LLAVA-MED which underperforms on the latter due to lack of following such multi-choice instructions), whilst outperforming both baselines on PathVQA and QUILT-VQA, albeit performance is limited on PMC-VQA-Subset, we believe it stems from the PMC-VQA's issues as discussed in Section 3.4 of the Appendix. QUILT-LLAVA outperforms both baselines on Open-set evaluation on PathVQA and QUILT-VQA, with significant margins. We also run ablations with multiple image encoders and training durations in Stages 1 and 2 summarized in Table 2a and qualitative examples in Tables 3-6 in the Appendix.

VQA with Visual Prompting. We adopt the visual prompting methodology from [22] for evaluating our model. This involves utilizing the subset of QUILT-VQA with bounding boxes to create ellipses that encapsulate the concepts highlighted by these boxes (See Appendix Fig. 17). Our QUILT-LLAVA surpasses the two baselines in both open and closed-set visually prompted VQA, achieving over 4% in open-set and 10% in closed-set accuracy. We attribute this enhanced performance to our grounded instruction-tuning dataset, underscoring QUILT-LLAVA's capability to reason within a region of interest, a critical aspect in histopathology.

Instruction-tuning Ablation Studies. To study the impact of QUILT-INSTRUCT on performance, we train several models: ***I*** – Only balanced independent prompts data 3.2.1 model, ***R*** – Only balanced reasoning prompts data 3.2.2, and ***IR*** –

	Question Types		Sub-Domains														Overall
	Conv	Desc	Bone	Breast	Cyto	Derm	Endo	Gastro	Bone	Geni	Gyne	H&N	Neuro	Pulm	Renal	Soft	
(Question Count)	(256)	(70)	(25)	(23)	(23)	(21)	(23)	(23)	(23)	(22)	(24)	(22)	(24)	(25)	(23)	(25)	(326)
LLaVA [16]	61.4	36.5	54.5	62.0	49.2	48.0	60.1	49.5	62.5	62.2	61.9	49.7	59.7	44.8	53.9	62.7	55.7
LLaVA-MED [14]	70.1	46.9	62.1	69.3	54.1	64.0	61.0	60.7	71.2	68.1	70.3	66.9	66.0	58.9	62.7	73.4	64.8
QUILT-LLAVA @ 40K	76.3	58.7	83.4	73.3	69.2	66.7	71.7	67.2	84.5	81.1	78.4	63.2	68.9	55.2	63.5	87.7	72.3
QUILT-LLAVA @ 107K	78.4	66.0	82.5	84.4	75.0	79.0	76.2	72.8	75.3	82.1	79.1	69.1	68.7	58.1	67.8	89.0	75.6

(a) Performance comparison of multi-modal chat instruction-following abilities, measured by the relative score via language GPT-4 evaluation. Our best model QUILT-LLAVA with ViT-B-32 Vision Encoder [9], 7B Language Model (trained for Stage1: 3 epochs, Stage2:1 epoch) outperforms the baselines.

QUILT-LLAVA Model Variants			PathVQA		PMC-VQA-Subset	QUILT-VQA		QUILT-VQA ○		Average	
Instruct	Stage 1	Stage 2	Open	Closed	Closed	Open	Closed	Open (w/o ● w/)	Closed (w/o ● w/)	Open	Closed
QUILTNET ViT-B-32 Vision Encoder [9], 7B Language Model											
107K	1	0	14.34	53.78	27.05	47.69	56.56	49.62 ● 54.13	55.56 ● 58.33	41.45	50.26
107K	1	1	14.24	58.42	19.63	59.82	64.43	58.81 ● 61.08	68.52 ● 70.37	48.49	56.27
107K	1	3	12.79	56.30	17.21	57.62	63.55	56.21 ● 58.32	65.74 ● 69.44	46.24	54.45
107K	3	1	15.30	54.93	16.01	60.97	60.64	59.24 ● 64.06	56.48 ● 59.26	49.89	49.46
LLAVA [16] checkpoint, 7B Language Model											
107K	1	0	11.65	54.03	33.91	55.80	58.02	54.77 ● 59.56	51.85 ● 60.19	45.45	51.60
107K	1	1	15.06	58.68	28.56	55.39	68.81	54.24 ● 59.83	71.30 ● 75.00	46.13	60.47
Baselines											
LLaVA-Med [14] 7B	0	0	11.97	56.15	1.34	54.81	61.22	52.58 ● 53.97	69.44 ● 64.81	43.33	50.59
LLaVA [16] 7B	0	0	11.65	54.02	33.91	55.81	57.73	54.74 ● 59.96	51.85 ● 60.19	45.54	51.54

(b) Results with varying training epochs at different stages and models alongside baselines. 107K indicates the size of instruct data used in Stage-2.

QUILT-LLAVA Instruction Data				PathVQA		PMC-VQA-Subset	QUILT-VQA		QUILT-VQA ○	
Conv	Detail	Complex	Abductive	Open	Closed	Closed	Open	Closed	Open (w/o ● w/)	Closed (w/o ● w/)
10k	10k	0	0	12.56	50.99	12.04	57.52	64.14	55.81 ● 55.29	68.52 ● 59.26
0	0	10k	10k	17.53	44.50	30.46	73.95	39.65	74.37 ● 69.64	39.81 ● 41.67
5k	5k	5k	5k	14.04	57.12	22.13	61.75	65.60	60.15 ● 62.04	59.82 ● 67.59
10k	10k	10k	10k	14.10	57.33	22.17	61.42	65.60	59.67 ● 58.50	66.67 ● 67.59

(c) Instruction-tuning Ablation: finetuning QUILT-LLAVA on the instruction-tuning data subsets with varying number of samples in each Q/A type.

Table 1. Quantitative results on histopathology VQA datasets. For open-set questions, we report recall for our free-form text generation method in column *Open*. For closed-set questions, we report the accuracy in column *Closed*. Red indicates the best-performing model.

balanced on all prompt types. We find that ***I*** model performs better on closed-set questions, whereas ***R*** model outperforms on open-set questions. We argue this is due to the contextual distribution of the questions and answers. ***I*** lends itself well to closed-set questions directly asking option-constrained questions and ***R*** lends itself well to open-set questions because these questions often require reasoning about the diagnosis. ***IR*** improves performance on closed-set questions; however, we see a drop in open-set VQA performance.

6. Conclusion and Limitations

Limitations. Our data, derived from raw video footage, inherently contains noise. This can manifest in many ways: clusters may occasionally be generated incorrectly, or erroneous mouse cursor detection. Additionally, despite our efforts to keep its context limited, GPT-4 still hallucinates, leading to instances where QUILT-LLAVA also hallucinates. Also, despite being explicitly instructed not to, GPT-4 at times reads from the caption rather than extracting information from the image. We discuss the limitations of QUILT-LLAVA and QUILT-INSTRUCT in Section 2 of the Appendix.

Conclusion. We presented: **1)** QUILT-INSTRUCT, in which we automatically extracted human narrators' mouse movements to spatially ground our concepts and leverage WSI to introduce novel reasoning-based prompts. **2)** Utilizing QUILT-INSTRUCT, we trained QUILT-LLAVA, a multimodal model that outperforms existing baselines in both open-ended and close-ended histopathology question answering. **3)** To further test QUILT-LLAVA's reasoning capabilities, we curated an evaluation dataset, QUILT-VQA, comprised of human-generated question-answer pairs from pedagogical videos. In the future, we plan to collaborate with pathologists to appraise our model and we aim to broaden our focus from histopathology to the wider medical field.

Acknowledgements. We thank Microsoft for OpenAI credits, Department of Defense W81XWH-20-1-0798, and National Cancer Institute U01 CA231782, and R01 CA200690.

References

[1] Tad T Brunyé, Agnes Balla, Trafton Drew, Joann G Elmore, Kathleen F Kerr, Hannah Shucard, and Donald L Weaver. From image to diagnosis: Characterizing sources of error in histopathologic interpretation. *Modern Pathology*, 36(7): 100162, 2023. 2

[2] Hyaline Change. Cellular responses to stress and toxic insults: Adaptation, injury, and death. *Robbins and Cotran Pathologic Basis of Disease, Professional Edition E-Book*, page 1, 2009. 2

[3] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2023. 2, 7

[4] Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He, Xiangyu Yue, et al. Llama-adapter v2: Parameter-efficient visual instruction model. *arXiv preprint arXiv:2304.15010*, 2023. 2

[5] Shah Giashuddin and Mouyed Alawad. Histopathological diagnosis of nonalcoholic steatohepatitis (nash). In *Non-Alcoholic Steatohepatitis: Methods and Protocols*, pages 1–18. Springer, 2022. 2

[6] Tianyu Han, Lisa C Adams, Jens-Michalis Papaioannou, Paul Grundmann, Tom Oberhauser, Alexander Löser, Daniel Truhn, and Keno K Bressem. Medalpaca–an open-source collection of medical conversational ai models and training data. *arXiv preprint arXiv:2304.08247*, 2023. 2

[7] Xuehai He, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. Pathvqa: 30000+ questions for medical visual question answering. *arXiv preprint arXiv:2003.10286*, 2020. 6, 2, 3

[8] Zhi Huang, Federico Bianchi, Mert Yuksekgonul, Thomas J Montine, and James Zou. A visual–language foundation model for pathology image analysis using medical twitter. *Nature medicine*, 29(9):2307–2316, 2023. 6

[9] Wisdom Oluchi Ikezogwo, Mehmet Saygin Seyfioglu, Fatemeh Ghezloo, Dylan Stefan Chan Geva, Fatwir Sheikh Mohammed, Pavan Kumar Anand, Ranjay Krishna, and Linda Shapiro. Quilt-1m: One million image-text pairs for histopathology. *arXiv preprint arXiv:2306.11207*, 2023. 2, 3, 4, 6, 8

[10] Julie A Jacko. Human computer interaction handbook: Fundamentals, evolving technologies, and emerging applications. 2012. 3

[11] Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023. 2

[12] Daniel Kahneman. *Attention and effort*. Citeseer, 1973. 3

[13] Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Fanyi Pu, Jingkang Yang, Chunyuan Li, and Ziwei Liu. Mimic-it: Multi-modal in-context instruction tuning. *arXiv preprint arXiv:2306.05425*, 2023. 2, 3

[14] Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *arXiv preprint arXiv:2306.00890*, 2023. 2, 6, 7, 8, 4, 18

[15] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 5

[16] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*, 2023. 2, 8, 4

[17] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023. 2, 4, 5, 6, 7

[18] Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Cyril Zakka, Yash Dalmia, Eduardo Pontes Reis, Pranav Rajpurkar, and Jure Leskovec. Med-flamingo: a multimodal medical few-shot learner. *arXiv preprint arXiv:2307.15189*, 2023. 2

[19] R OpenAI. Gpt-4 technical report. *arXiv*, pages 2303–08774, 2023. 2

[20] Jordi Pont-Tuset, Jasper Uijlings, Soravit Changpinyo, Radu Soricut, and Vittorio Ferrari. Connecting vision and language with localized narratives. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*, pages 647–664. Springer, 2020. 3

[21] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015. 3

[22] Aleksandar Shtedritski, Christian Rupprecht, and Andrea Vedaldi. What does clip know about a red circle? visual prompt engineering for vlms, 2023. 2, 7

[23] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. Alpaca: A strong, replicable instruction-following model. *Stanford Center for Research on Foundation Models. https://crfm. stanford. edu/2023/03/13/alpaca. html*, 3(6):7, 2023. 2

[24] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 2

[25] Paul Voigtlaender, Soravit Changpinyo, Jordi Pont-Tuset, Radu Soricut, and Vittorio Ferrari. Connecting vision and language with video localized narratives. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2461–2471, 2023. 3

[26] Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinyuan Chen, Yaohui Wang, Ping Luo, Ziwei Liu, et al. Internvid: A large-scale video-text dataset for multimodal understanding and generation. *arXiv preprint arXiv:2307.06942*, 2023.

[27] Rowan Zellers, Ximing Lu, Jack Hessel, Youngjae Yu, Jae Sung Park, Jize Cao, Ali Farhadi, and Yejin Choi. Merlot: Multimodal neural script knowledge models. *Advances in Neural Information Processing Systems*, 34:23634–23651, 2021.

[28] Rowan Zellers, Jiasen Lu, Ximing Lu, Youngjae Yu, Yanpeng Zhao, Mohammadreza Salehi, Aditya Kusupati, Jack Hessel, Ali Farhadi, and Yejin Choi. Merlot reserve: Neural script knowledge through vision and language and sound. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16375–16387, 2022. 3

[29] Heyu Zhang, Yan He, Xiaomin Wu, Peixiang Huang, Wenkang Qin, Fan Wang, Juxiang Ye, Xirui Huang, Yanfang Liao, Hang Chen, et al. Pathnarratives: Data annotation for pathological human-ai collaborative diagnosis. *Frontiers in Medicine*, 9:1070072, 2023. 3

[30] Sheng Zhang, Yanbo Xu, Naoto Usuyama, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, Cliff Wong, et al. Large-scale domain-specific pretraining for biomedical vision-language processing. *arXiv preprint arXiv:2303.00915*, 2023. 4

[31] Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Weixiong Lin, Ya Zhang, Yanfeng Wang, and Weidi Xie. Pmc-vqa: Visual instruction tuning for medical visual question answering. *arXiv preprint arXiv:2305.10415*, 2023. 2, 6, 3

[32] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023. 2