CLIPSyntel: CLIP and LLM Synergy for Multimodal Question Summarization in Healthcare

Akash Ghosh[1*], **Arkadeep Acharya**[1*], **Raghav Jain**[1], **Sriparna Saha**[1], **Aman Chadha**[2,3†], **Setu Sinha**[4]

[1]Department of Computer Science And Engineering, Indian Institute of Technology Patna, India
[2]Stanford University
[3]Amazon AI
[4]Indira Gandhi Institute of Medical Sciences
{akash_2321cs19, arkadeep_2101ai41, sriparna}@iitp.ac.in, raghavjain106@gmail.com, hi@aman.ai, drsinhasetu@gmail.com

Abstract

In the era of modern healthcare, swiftly generating medical question summaries is crucial for informed and timely patient care. Despite the increasing complexity and volume of medical data, existing studies have focused solely on text-based summarization, neglecting the integration of visual information. Recognizing the untapped potential of combining textual queries with visual representations of medical conditions, we introduce the Multimodal Medical Question Summarization (MMQS) Dataset. This dataset, a major contribution of our work, pairs medical queries with visual aids, facilitating a richer and more nuanced understanding of patient needs. We also propose a framework, utilizing the power of Contrastive Language Image Pretraining(CLIP) and Large Language Models(LLMs), consisting of four modules that identify medical disorders, generate relevant context, filter medical concepts, and craft visually aware summaries. Our comprehensive framework harnesses the power of CLIP, a multimodal foundation model, and various general-purpose LLMs, comprising four main modules: the medical disorder identification module, the relevant context generation module, the context filtration module for distilling relevant medical concepts and knowledge, and finally, a general-purpose LLM to generate visually aware medical question summaries. Leveraging our MMQS dataset, we showcase how visual cues from images enhance the generation of medically nuanced summaries. This multimodal approach not only enhances the decision-making process in healthcare but also fosters a more nuanced understanding of patient queries, laying the groundwork for future research in personalized and responsive medical care.

Disclaimer: The article features graphic medical imagery, a result of the subject's inherent requirements.

Introduction

In the midst of the global COVID-19 pandemic, healthcare systems have been inundated with an unprecedented volume of inquiries, concerns, and uncertainties. Individuals worldwide are posing health-related inquiries on online platforms, seeking clarity and guidance on symptoms, prevention, treatments, and vaccinations. These questions often employ everyday language and encompass both pertinent and extraneous details that may not directly pertain to the sought-after solutions. The complexity and urgency of the situation, coupled with a considerable imbalance in the doctor-to-patient ratio across many countries, have made the ability to swiftly and comprehensively comprehend a patient's query paramount (Abacha and Demner-Fushman 2019; Yadav et al. 2021). In this context, medical question summarization has emerged as a vital tool to distill information from consumer health questions, ensuring the provision of accurate and timely responses. However, existing works have overlooked the untapped potential of integrating visual data, particularly images, with textual information. The motivations for focusing on visual aids within medical question summarization (MQS) are manifold. A significant portion of the population lacks familiarity with medical terms needed to accurately describe various symptoms, and some symptoms are inherently challenging to articulate through text alone. Patients may also be confused between closely related symptoms, such as distinguishing between skin dryness and skin rash. The combination of text and images in medical question summarization can offer enhanced accuracy and efficiency, providing a richer context that textual analysis alone may miss. This approach recognizes the complex nature of patient queries, where photographs of symptoms, medical reports, or other visual aids could provide crucial insights. By focusing on the integration of images, researchers and healthcare providers can respond to the evolving challenges of modern healthcare communication.

Large Language Models (LLMs) (Kojima et al. 2023) and Vision Language Models (VLMs) (Zhang et al. 2023) have exhibited remarkable capacities in generating human-like text and multimedia content. This prowess has driven their deployment across the medical domain, predominantly for domain-specific tasks such as chest radiograph summarization (Thawkar et al. 2023) and COVID-19 CT report generation (Liu et al. 2021). Yet, their application in multimodal medical question summarization remains uncharted territory. Leveraging zero-shot and few-shot learning capabilities of these models (Dong et al. 2022) offers a compelling advantage, especially for tasks like multimodal medical question

[*]These authors contributed equally.
[†]Work does not relate to position at Amazon.

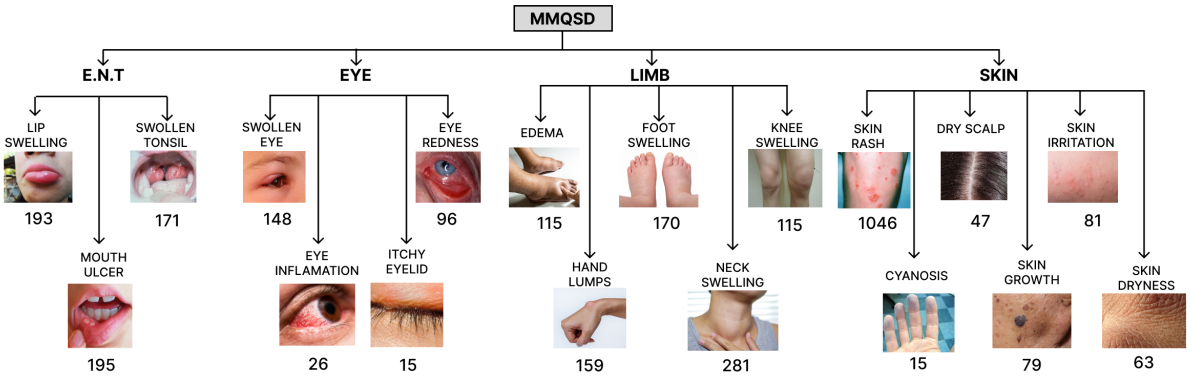
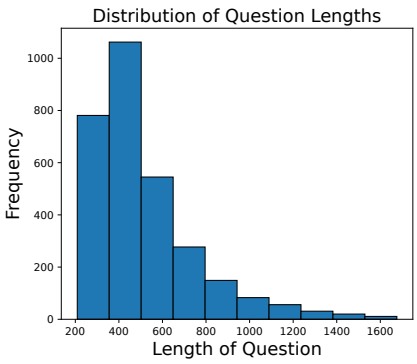
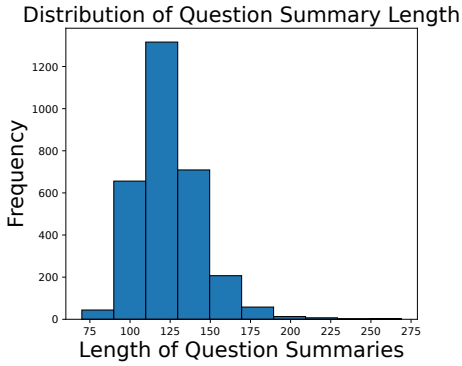


Figure 1: Broad categorization of medical disorders in the MMQS Dataset (MMQSD). The number of data points corresponding to each category has been provided under each category.



(a) Distribution of Question Lengths.



(b) Distribution of Summary Lengths.

Figure 2: Distribution of question lengths and their summaries in the MMQS dataset.

summarization, characterized by inherent data scarcity. However, while the potential of LLMs and VLMs in this domain is undeniable, they aren't without constraints. Predominantly, generic LLMs and VLMs often lack a solid grounding in task-specific knowledge, risking the generation of summaries that might miss intricate details like symptoms, diagnostic tests, and medical complexities. On the visual front, even as VLMs have excelled in typical visual-linguistic tasks, medical imaging presents unique challenges. Efforts like SkinGPT4 (Zhou et al. 2023), which fine-tunes MiniGPT4 (Zhu et al. 2023) on skin disease images and clinical notes, are still notably domain-specific. Medical images are inherently complex, demanding a profound understanding of medical terminology and visual conventions, often necessitating an expert medical practitioner's perspective for accurate interpretation. This complexity, combined with potential gaps in contextual understanding, can result in models producing misleading or irrelevant summaries.

To address the limitations of LLMs and VLMs in multimodal medical question summarization, we've conceived the ***CLIPSyntel*** framework. The first stage, the Medical Disorder Identification Module, combines Large Language Model (LLM) with Contrastive Language Image Pretraining (CLIP) (Radford et al. 2021), forming a novel zero-shot classification approach to identify disorders from visual images, utilizing the unique benefits of zero or few-shot learning. Subsequently, during the Context Generation Phase, LLM adds context to the medical query, focusing on vital components like symptoms, tests, and procedures, mindful of the potential pitfalls of extraneous content. Addressing this challenge, the Context Filtration Module filters the content using a multimodal knowledge selection technique, preserving only the most relevant information. The Summary Generation Phase follows, where an LLM crafts the final medically accurate summaries based on the distilled knowledge. Finally, to validate *CLIPSyntel*, we created the ***MMQS dataset***, rich in visual and textual representations of patients' symptoms and queries[1]. This careful, step-by-step structuring of *CLIPSyntel* allows it to bridge the divide between general-purpose models and the niche requirements of medical question summarization, effectively integrating textual and visual data to create precise and context-aware medical summaries. To summarize we make the following main contributions:

A novel task of Multimodal Medical Question Summarization for generating medically nuanced summaries.

[1]https://github.com/AkashGhosh/CLIPSyntel-AAAI2024

A novel dataset, *MMQS* Dataset, to advance the research in this area.
A novel metric *MMFCM* to quantify how well the model captures the multimodal information in the generated summary.
A novel framework, *ClipSyntel* that harnesses the power of CLIP and LLMs to augment the patient question with an additional rich context from visual symptoms for the generation of final summaries.

Related Works

Medical Question Summarization: The task of Medical Question Summarization (MQS) was initially introduced in 2019 with the development of the MeQSum dataset (Abacha and Demner-Fushman 2019) specifically for this purpose. Early research on MQS employed vanilla seq2seq models and pointer generator networks to generate summaries. In 2021, a contest was held focusing on the generation of medical domain summaries (Abacha et al. 2021). Participants utilized various pre-trained models such as PEGASUS (Zhang et al. 2020), ProphetNet (Qi et al. 2020), and BART (Lewis et al. 2019). Techniques like multi-task learning were employed, utilizing BART to jointly optimize question summarization and entailment tasks (Mrini et al. 2021). Another approach involved reinforcement learning with question-aware semantic rewards derived from two subtasks: question focus recognition (QFR) and question type identification (QTR) (Yadav et al. 2021).

Role of Multimodality: In order to receive accurate treatment and guidance from a medical expert, our responsibility is to effectively and efficiently communicate the medical symptoms which can be done with additional visual cues. Previously, many studies also showed that adding multimodality through visual cues improves the performance of various medical tasks. Tiwari et al. (2022) shows how multimodal information helps in building better Disease Diagnosis Virtual Assistants. Delbrouck, Zhang, and Rubin (2021) also showed how incorporating images helps in better summarization of radiology reports. Gupta, Attal, and Demner-Fushman (2022) also showed the benefit of incorporating videos for medical QA task The current work is motivated by this idea of how integrating the patient's provision of an image of their medical condition alongside the text, improves the generation of more medically rich summaries. We curate a multimodal dataset based on an existing MQS dataset, focusing on a predefined set of symptoms, and propose a novel framework that combines cutting-edge Multimodal Foundation Models like CLIP with Large Language Models (LLMs). To our best understanding, this work is the first to tackle the task of question summarization in the medical domain, particularly within a multimodal setting.

MMQS (Multimodal Medical Question Summarization) Dataset

To the best of our knowledge, a freely available multimodal question summarization dataset that includes both textual questions and corresponding medical images of patients' conditions does not currently exist. To construct such a dataset, we utilized the preexisting HealthCareMagic Dataset, which is derived from the MedDialog data introduced by (Mrini et al. 2021). The initial dataset comprised 226,395 samples, of which 523 were found to be duplicates. To ensure data integrity and fairness, we eliminated these duplicate entries. In order to ascertain the medical symptoms or signs that could effectively be conveyed through visual means, we consulted a medical professional who also happens to be a co-author of this paper. Following a series of brainstorming sessions and carefully analyzing the dataset, we identified 18 symptoms that are hard to specify only through text. These 18 symptoms are broadly divided into four broad groups of multimodal symptoms to incorporate into our dataset. These groups (refer to figure 1) are categorized as ENT (Ear, Nose, and Throat), EYE (Eye-related), LIMB (Limbs-related), and SKIN (Skin-related). From the ENT category, we selected symptoms including lip swelling, mouth ulcers, and swollen tonsils. From the EYE category, we choose swollen eyes, eye redness, and itchy eyelids. For the LIMB category, we included symptoms such as edema, foot swelling, knee swelling, hand lumps, and neck swelling. Lastly, from the SKIN category, we included symptoms of skin rash, skin irritation, and skin growth. To extract images for the corresponding symptoms, Bing Image Search API[2] was used. Then the images extracted are verified by a group of medical students who are led by a medical expert. Therefore, in curating our ultimate multimodal dataset, we selectively included instances from the HealthCareMagic dataset that featured textual mentions of body parts such as skin, eyes, ears, and others, both within the questions and their corresponding summaries. We refrained from directly searching for symptom names, as our intention was to encompass instances from the dataset where uncertainty regarding the medical condition exists. Consequently, patients implicitly allude to symptoms rather than explicitly stating their names. To identify these relevant samples, we employed the Python library FlashText[3] for efficient term matching in the aforementioned dataset. The Python library Textblob4 was used to correct the spelling of many improperly spelled words. To address the misspelling of numerous words, the Textblob4[4] Python library was employed. Following an initial exploration using FlashText, we initially collected approximately 5000 samples. After meticulous manual validation of each sample, our refinement process led us to a final dataset comprising 3015 samples where multimodality could be incorporated for the final dataset creation.

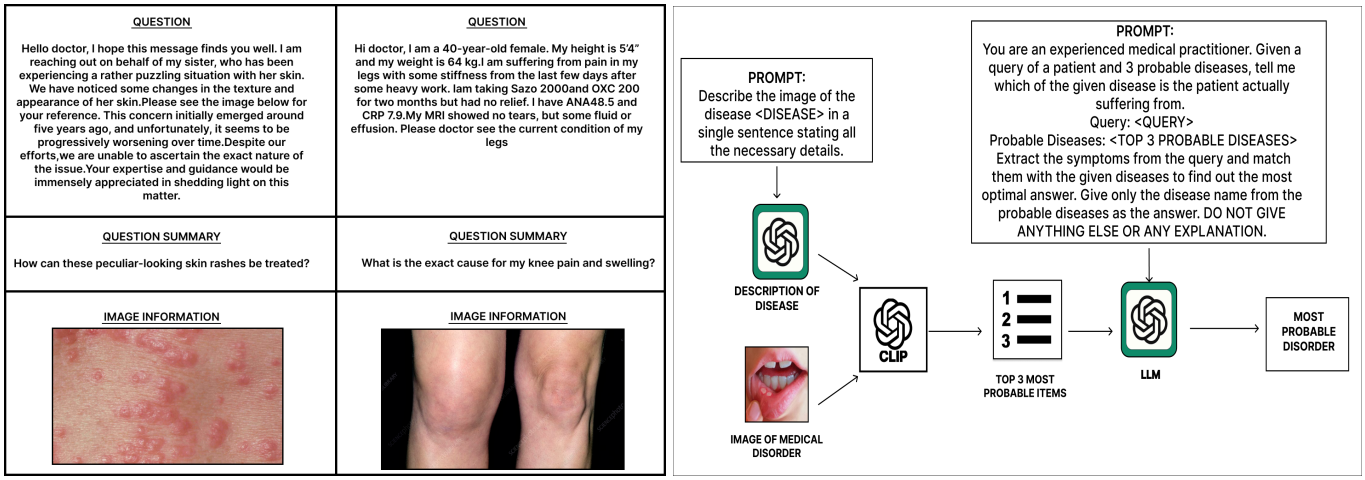
Data Annotation

We extracted 100 random samples from the filtered dataset, each comprising a patient's inquiry alongside its corresponding summary, and gave them to the medical expert for annotation. The methodology through which the medical expert annotated these samples is explained with the given example. For example, suppose the patient is complaining regarding something that has happened near their tonsils, but they are

[2]https://www.microsoft.com/en-us/bing/apis/bing-image-search-api

[3]https://pypi.org/project/flashtext/1.0/

[4]https://pypi.org/project/textblob/0.9.0/



(a) Sample instances from the MMQS-dataset.

(b) Medical Disorder Identification Module

Figure 3: Comprehensive illustration of Sample Instances from the MMQS-dataset and Medical Disorder Identification Module.

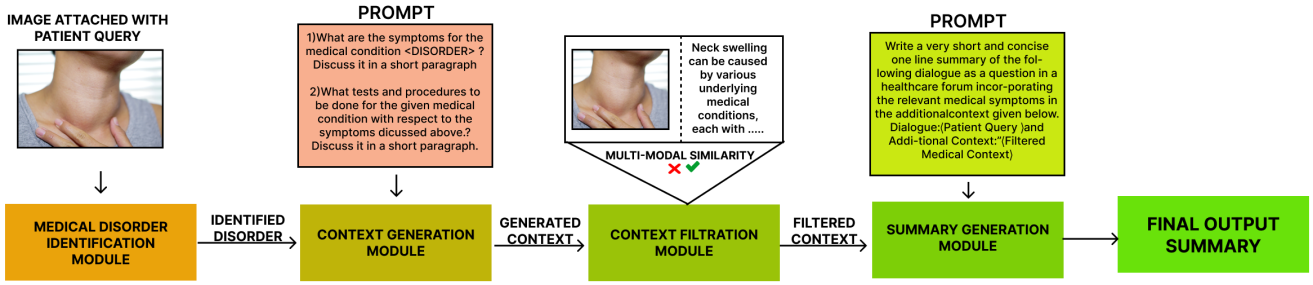


Figure 4: Flowchart depicting the workflow of the proposed model *CLIPSyntel*. The diagram illustrates the step-by-step process of *ClipSyntel*'s functioning.

not able to name the exact medical disorder they are suffering from. But their textual reference gave enough context to understand that the patient is suffering from the disorder named **swollen-tonsils**. After understanding this, the medical expert adds statements like ***Please see what happened to my tonsils in the image below*** in the context and accordingly adds a visual image representation of the disorder. In this way, visual medical signs are incorporated into the textual medical question. Additionally, the medical expert deduced that the conventional golden summaries did not align well with the multimodal queries, prompting the need for corresponding updates. To address these issues, the medical expert himself modified these 100 samples by revising the questions, incorporating multimodal information, and adapting the golden summaries accordingly. Subsequently, we engaged with medical students and graduates to annotate the remaining samples based on the guidelines given by the medical expert[5]. To quantify the level of agreement among annotators, we calculated the kappa coefficient (k), which yielded a value of 0.78. This coefficient indicates a noteworthy level of consistent annotation. Following this rigorous process, the Multimodal Medical Question Summarization (MMQS) dataset comprised a total of 3,015 samples. The median length of the question in the final dataset is 442 words and the median length of the question summary is 121 words, and example data points are also shown in Table-3a. The distributions of question length and question summary length are shown in Figure 2a and Figure 2b, respectively.

[5]The students were compensated through gift vouchers and honorarium.

Methodology

Problem Formulation: Each data point consists of a patient query in textual form T and an image I of the medical disorder that the patient is suffering from or trying to address in his textual query. The final output is a concise summary S of the query which incorporates information from both the modalities. This section presents the novel architecture of ***CLIPSyntel*** shown in Figure-4, a multimodal, knowledge-grounded framework designed to generate medically nuanced summaries.For a comprehensive understanding, we partition our approach into four distinct modules: (i) **Medical Disorder Identification Module**, (ii) **Contextual Information Generation Module**, (iii) **Context Filtration Module**, and (iv) **Summary Generation Module**

Medical Disorder Identification

For the process of generating pertinent medical knowledge it is necessary to accurately identifying the medical disorder that the patient is experiencing. To achieve this, we employed the following structured approach: **(1) Utilization of CLIP:** CLIP, designed to establish connections between images and text, is initially applied. Despite its applicability to various visual classification tasks, CLIP's effectiveness in our Medical Disorder Identification Task using only the names of the disorder is limited due to the complex nature of medical images. **(2) Enhanced Contextualization through Prompts:** To provide better context for a particular medical disorder, we prompted GPT 3.5 using the prompt: **Describe the image of the disease** $\{DISEASE\}$ **in a single sentence stating all the necessary details**. This approach yields improved contextual information for the specific medical condition. **(3) Identification of Probable Disorders:** When presented with an image of a medical disorder, we provided the contextual information for all 18 medical disorders, along with the corresponding image, to the CLIP model. Subsequently, we selected the top 3 most likely medical disorders based on CLIP's analysis. **(4) Final Prediction with GPT-3.5:** These three probable diseases, along with the medical query, is then subsequently passed to LLM(GPT-3.5) using the prompt which is shown in Figure-3b to predict the final medical disorder. Considering only the most probable disorder prediction from CLIP yields an accuracy of 84 %. However, when incorporating the top 3 most probable disorders alongside the context after passing through the LLM, the accuracy is enhanced to 87% in a zero-shot setup. The entire pipeline for this module, displaying its logical flow and connections, is depicted in Figure 3b.

Contextual Medical Knowledge Generation

In cases where patients may lack awareness of their medical condition, and textual inquiries are insufficient, acquiring additional knowledge about their specific ***symptoms*** and the ***necessary tests and procedures*** (as advised by medical experts) proves essential for creating a meaningful summary, especially in a multimodal setup. Formally, after identification of the medical disorder M, we formulate a set of prompts $P = \{p_0, p_1, \ldots, p_n\}$ with the aim of generating diverse contextual information about that disorder M, $KS = \{ks_0, ks_1, \ldots, ks_n\}$ by promting ***GPT-3.5*** as follows. The Prompts used to generate contextual information are as below:

(1) What are the symptoms of the medical condition ⟨**medical disorder** ⟩?

(2) What tests and procedures need to be done for the medical condition ⟨medical disorder ⟩?

Multimodal Medical Knowledge Filtration

While we anticipate that **GPT-3.5** will generate useful and pertinent information based on the given prompt, it occasionally engages in hallucinations and produces irrelevant content in relation to the primary inquiry. This hallucinated text has the potential to divert Large Language Models (LLMs) from generating high-quality summaries. It may also generate content that is not pertinent for doctors to investigate the case swiftly. Particularly in fields such as medicine, one must exercise heightened caution to minimize misinformation, given that errors could have serious consequences for patients. To tackle this issue, we present a filtering strategy called Multimodal Medical Knowledge Filtration, which operates as follows. Formally, for each ks_i, which corresponds to a specific text field in the KS, it is further broken down into a set of k sentences $ks_i = \{s_1, s_2, \ldots, s_m\}$ using sentence tokenization[6]. To achieve this, we employ a pre-trained multimodal model, ImageBind (Girdhar et al. 2023), which serves as an off-the-shelf encoder-based multimodal model $enc(\cdot)$ that maps a token sequence (text) and an image to their respective feature vectors embedded in a unified vector space. Cosine similarity $sim(\cdot, \cdot)$ is then used to measure the relevance of each sentence to the image. Formally, a sentence s_j is retained if $sim(enc(s_j), enc(I)) > Th$, where I is the image associated with the medical disorder, and Th is the predefined similarity threshold. Building upon this, we define the subset of retained knowledge sentences, having undergone the Multimodal Medical Knowledge Filtration module (*MMKFS*), as follows:

$$Ms'_i = \{s_j \mid s_j \in ks_i \text{ and } \text{sim}(\text{enc}(s_j), \text{enc}(I)) > Th\} \quad (1)$$

With this approach, the filtered knowledge sentences $MS' = \{ms'_0, \ldots, ms'_n\}$ not only hold a high degree of visual-textual alignment with the corresponding medical disorder image but are also contextually relevant.

Summary Generation Module

This component is crafted to leverage the refined knowledge sentences, denoted as MS' in order to craft informative and contextually relevant summaries for patients' healthcare inquiries. The Summary Generation Module functions with a pre-trained Large Language Model (LLM), referred to as LM. For summary creation, LM is provided a specially crafted prompt Pin that integrates both the actual patient query X and the generated medical disorder knowledge MS'. The outcome of this process is the generated summary text, denoted as S. In a more formal representation, this process can be articulated as: $S = LM(Pin(X, MS'))$. The prompt used to generate the final summary is presented below:

Write a very short and concise one line summary of the following dialogue as a question in a healthcare forum incorporating the relevant medical symptoms in the additional context given below. Dialogue:⟨Patient Query⟩and Additional Context:⟨Filtered Medical Context⟩

Experimental Results and Discussion

Experimental Setup: We leveraged the following general purpose LLMs for summary generation module: RedPajama[7], FLAN-T5 (Chung et al. 2022), Vicuna (Zheng et al. 2023) and GPT-3.5. We tested these LLMs in different settings: With only the patient's question(text) in the prompt, With the patient's question combined with the knowledge obtained

[6]https://www.nltk.org/api/nltk.tokenize.html

[7]https://huggingface.co/togethercomputer/RedPajama-INCITE-Chat-3B-v1

	Model	ROUGE			BLEU				BERTScore
		R1	R2	RL	B1	B2	B3	B4	
LLAMA-2	LLM(only texual query)	0.188	0.051	0.159	0.144	0.064	0.032	0.016	0.822
	Clip+GPT-3.5	0.189	0.052	0.161	0.156	0.070	0.036	0.018	0.819
	CLIPSyntel	**0.249**	**0.081**	**0.211**	**0.195**	**0.102**	**0.059**	**0.034**	**0.864**
RedPajama	LLM(only texual query)	0.146	0.023	0.125	0.086	0.030	0.011	0.004	0.828
	Clip+GPT-3.5	0.148	0.025	0.132	0.087	0.0314	0.012	0.004	0.831
	CLIPSyntel	**0.157**	**0.0279**	**0.140**	**0.088**	**0.032**	**0.014**	**0.006**	**0.838**
Vicuna-7B	LLM(only textual query)	0.372	0.160	0.318	0.385	0.243	0.161	0.102	0.905
	Clip+GPT-3.5	0.384	0.164	0.321	0.391	0.245	0.167	0.105	0.906
	CLIPSyntel	**0.391**	**0.167**	**0.33**	**0.40**	**0.25**	**0.171**	**0.108**	**0.910**
FLAN-T5	LLM(only textual query)	0.220	0.042	0.205	0.129	0.052	0.026	0.012	0.88
	Clip+ GPT-3.5	0.220	0.042	0.205	0.136	0.056	0.026	0.012	0.88
	CLIPSyntel	**0.243**	**0.068**	**0.215**	**0.182**	**0.096**	**0.052**	**0.0256**	**0.891**
GPT-3.5	LLM(only textual query)	0.451	0.227	0.396	0.471	0.330	0.243	0.170	0.920
	Clip + GPT-3.5	0.452	0.226	0.399	0.471	0.331	0.242	0.171	0.920
	CLIPSyntel	**0.463**	**0.241**	**0.409**	**0.478**	**0.342**	**0.257**	**0.186**	**0.921**

Table 1: Performance of various *CLIPSyntel* models and corresponding baselines, evaluated using automatic metrics with different LLMs.

Model(GPT-3.5)	Clinical-EvalScore	Factual Recall	Omission Rate	MMFCM Score
LLM(Patient Query)	3.4	0.748	0.235	0.9
Clip + GPT-3.5	**3.51**	**0.792**	**0.2215**	**1.52**
CLIPSyntel	**3.62**	**0.818**	**0.2217**	**1.6**
Annotated Summary	**4.1**	**0.889**	**0.144**	**1.86**

Table 2: Human evaluation scores of the best *CLIPSyntel* model and their corresponding baselines across different metrics.

from (CLIP+GPT3.5), and then our proposed framework ((CLIPSyntel)). We have set the similarity threshold parameter Th to 0.5 and temperature to 0.5 across all settings based on a thorough investigation. We utilize ROUGE (Lin 2004), BLEU (Papineni et al. 2002), and BERTScore (Zhang et al. 2019) as automatic evaluation metrics. For the purpose of human evaluation, we collaborate with a medical expert and a few medical students. We have identified four distinct and medically nuanced metrics for this evaluation: clinical evaluation score, factual recall (Abacha et al. 2023), omission rate(Abacha et al. 2023), and our newly introduced MMFCM score.

Automated Evaluation: The results presented in Table 1 provide valuable insights into the performance of various models, highlighting the effectiveness of different approaches in the context of the task. Below, we discuss some key observations and trends: **(1) *CLIPSyntel* Performance:** Across all LLMs, *CLIPSyntel* consistently performed the best. This indicates the robustness and versatility of *CLIPSyntel*, demonstrating its ability to outshine other base models. The results suggest the capability of *CLIPSyntel*'s design to effectively leverage both textual and visual information. These results also show that adding contextual information does help in our task. **(2) GPT-3.5 based Models' Superiority:** Among the various models evaluated, the GPT-3.5 based models stood out as the top performers compared to open-source models. The high scores across multiple metrics reflect the sophisticated design and capability of GPT-3.5. **(3) Best Performance Among Open-Source LLMs:** Vicuna was the best performer among open-source LLMs. Its scores were notably higher compared to other open-source counterparts, signaling its potential as an effective alternative for specialized tasks.

Human Evaluation: The human evaluation was done by a team of medical students led by a doctor. The team was given 10 % of the dataset (selected at random) for evaluation purpose and was asked to rate the summaries generated, which takes only patient question (text) as context, patient query in addition to the medical disorder context (CLIP + GPT-3.5) and finally the summary generated by our proposed pipeline (*CLIPSyntel*). The following metrics are used for the evaluation: **(1) Clinical Evaluation Score:** The doctor and his team were asked to rate the summaries between 1 (poor) and 5 (good) based on their overall relevance, consistency, fluency, and coherence. **(2) Multi-modal fact capturing metric (MMFCM):** We propose a new metric to evaluate how well a model incorporates relevant medical facts and identifies the correct disorder in a multimodal setup. MMFCM is calculated by considering both the facts extracted from the medical query and image and assessing whether they are correctly incorporated in the generated summary. The metric accounts for (1) The ratio of correct facts in the summary to the total number of relevant medical facts. (2) Additional scores based on the accuracy of the disorder's detection, with values ranging from +2 for fully correct to -1 for incorrect identification. See Algorithm 1 for the detailed algorithm.

(3) Medical Fact Based Metrics: We employed the Factual Recall and Omission Recall metrics (Abacha et al. 2023) to assess the extent to which the generated summary captures medical facts compared to the gold standard annotated summary.

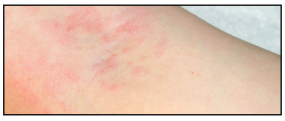
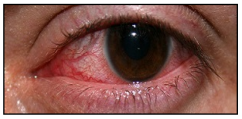
SUPPORTING IMAGE			
MEDICAL QUESTION		We have a 16 month old baby. She has been suffering from serious allergy for last 3/4 months. Her skin gets red as rash.Please doctor see the current condition below . It itches heavily. She feels discomfort in sleeping. Sleeping does not look like sound.Would you please suggest us what we should do now. It is mentioned here that I, the father of the baby have acute allergy from my age 23.	My 8 year old son suffers from allergies. He gets severe reddness under both eyes and across the bridge of his nose. His skin looks blistery in the red areas.Please doctor see the eyes below . The skin is bright red like the color of a sun burn. He says in only itches now and then. The rest of his face in normal colored. Please advise.
GPT 3.5	Only Questoion	What to do for a 16-month-old baby suffering from a serious allergy to develop red, itchy rashes and experience sleep disturbances?	What could be causing severe redness under both eyes and across the bridge of the nose in an 8-year-old boy with occasional itching?
	ClipSyntel	What should be done for a 16-month-old baby suffering from a serious allergy with red, itchy skin rash and difficulty sleeping?	What could be the cause of severe reddness under both eyes and across the bridge of the nose, with blistery and bright red skin in an 8-year-old who only experiences occasional itching?

Figure 5: Sample Summaries generated by two variations GPT-3.5 model: only question and *CLIPSyntel*.

Table 2 offers a comprehensive comparative analysis, underscoring *ClipSyntel*'s distinct advantages over various baseline methods across multiple human evaluation performance metrics. Notably, our proposed architecture excels the summaries that incorporate only the patient's textual information, highlighting the valuable impact of fusing visual symptom cues with textual patient queries, resulting in more clinically nuanced summaries. Moreover, the augmentation of medical knowledge is affirmed by substantial enhancements in Factual Recall and Omission Rate metrics, further corroborated by our proposed Multimodal fact Capturing Metric(MMFCM) score. Overall, *ClipSyntel* emerges as a clear frontrunner in human evaluation metrics, outperforming competing baselines by a considerable margin.

Qualitative analysis: The Figure 5 provides an insightful comparison of the summaries generated by two variations of the GPT-3.5 model, namely the "Only Question" and "*CLIPSyntel*" configurations, for medical inquiries. The analysis mainly focuses on two aspects: (1) *CLIPSyntel* Captures Medical Disorders Correctly: The *CLIPSyntel* variant demonstrates a noteworthy ability to identify and represent medical conditions with precision. In both cases presented in the table, *CLIPSyntel* has shown an understanding of the underlying medical issues. (2) *CLIPSyntel* Does Less Hallucinations: *CLIPSyntel*'s summaries stick to the information explicitly provided in the medical questions. Unlike the "Only Question" variant that might introduce slight changes or uncertainties, *CLIPSyntel* maintains a strict adherence to the facts. This ability to avoid unnecessary extrapolations or inaccuracies enhances *CLIPSyntel*'s reliability.

Risk Analysis

It's important to note several limitations in our approach. We restricted our study to a zero-shot prompting strategy, not fully examining how prompt variations might affect results. Second, although CLIP performs well in low-resource environments, we must be cautious of potential misclassifications of medical images, as these could lead to serious or even fatal misinformation in a summary generation. Despite the promising performance of *CLIPSyntel* in various scenarios, the risk of misinformation in healthcare is significant. Therefore, involving a medical expert is vital to ensure that our AI model serves as an aid, not a replacement, for medical professionals.

Algorithm 1: MMFCM Method

Require: $F_m = \{\text{fact}_{m,1}, \text{fact}_{m,2}, \ldots, \text{fact}_{m,n-1}, \text{fact}_{m,n}\}$
{Relevant Medical facts from query 'm'.}
Require: $Sf_m = \{Summfact_{m,1}, \ldots, Summfact_{m,n}\}$
{ Relevant Medical facts of summary of query 'm'.}
$\#CorrectFacts_m = |F_m \cap Sf_m|$
{Number of correct medical facts in each summary}
if (Correct Medical Disorder phrase $\in \{F_m \cap Sf_m\}$) **then**
  $\#CorrectFacts_m += 2$
else if (Partially correct disorder phrase $\in \{F_m \cap Sf_m\}$) **then**
  $\#CorrectFacts_m += 1$
else if (Incorrect disorder phrase $\in \{F_m \cap Sf_m\}$) **then**
  $\#CorrectFacts_m += -1$
else
  $\#CorrectFacts_m += 0$
end if
return MMFCM = $\frac{\#\text{CorrectFacts}_m}{|F_m \cap Sf_m|}$

Conclusion and Future Work

In this study, we delve into the impact of incorporating multimodal cues, particularly visual information, on question summarization within the realm of healthcare. We present the MMQS dataset, comprising 3015 multimodal medical queries with golden summaries that merge visual and textual data. This novel collection fosters new assessment techniques in healthcare question summarization. We also introduce the *CLIPSyntel* framework, leveraging LLMs and the CLIP model, to enhance summaries with visual symptom details. CLIP enables symptom classification, and an ImageBind-filtering module mitigates content hallucination. In our future endeavors, we aspire to develop a Vision-Language model capable of extracting the intensity and duration details of symptoms and integrating them into the patient query's final summary generation. Furthermore, our expansion plans encompass incorporating medical videos and addressing scenarios involving code-mixed patient queries.

Ethical Statement

Summarization in healthcare necessitates strong ethical considerations, particularly regarding safety, privacy, and potential bias. To address these concerns in our project with the MMQS dataset, we implemented several proactive measures. We collaborated closely with medical professionals and also obtained IRB approval to ensure ethical rigor and patient privacy. We rigorously followed legal and ethical guidelines[8]

[8]https://www.wma.net/what-we-do/medical-ethics/declaration-of-helsinki/

during dataset validation, integration of images, and annotation of summaries. Medical experts were engaged throughout the process, providing validation and correction of the dataset and also validating the outputs of the models. The proposed dataset is based on original HealthCareMagic Dataset; the medical questions/samples are taken from this dataset. The incorporation of multimodality into the task is done under the full supervision of a medical professional. Additionally, we ensured user privacy by not disclosing identities.

Acknowledgements

Akash Ghosh and Sriparna Saha express their heartfelt gratitude to the SERB (Science and Engineering Research Board) POWER scheme (SPG/2021/003801) of the Department of Science and Engineering, Govt. of India, for providing the funding for carrying out this research.

References

Abacha, A. B.; and Demner-Fushman, D. 2019. On the summarization of consumer health questions. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2228–2234.

Abacha, A. B.; M'rabet, Y.; Zhang, Y.; Shivade, C.; Langlotz, C.; and Demner-Fushman, D. 2021. Overview of the MEDIQA 2021 shared task on summarization in the medical domain. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, 74–85.

Abacha, A. B.; Yim, W.-w.; Michalopoulos, G.; and Lin, T. 2023. An Investigation of Evaluation Metrics for Automated Medical Note Generation. *arXiv preprint arXiv:2305.17364*.

Chung, H. W.; Hou, L.; Longpre, S.; Zoph, B.; Tay, Y.; Fedus, W.; Li, E.; Wang, X.; Dehghani, M.; Brahma, S.; et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Delbrouck, J.-B.; Zhang, C.; and Rubin, D. 2021. QIAI at MEDIQA 2021: Multimodal Radiology Report Summarization. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, 285–290. Online: Association for Computational Linguistics.

Dong, Q.; Li, L.; Dai, D.; Zheng, C.; Wu, Z.; Chang, B.; Sun, X.; Xu, J.; and Sui, Z. 2022. A survey for in-context learning. *arXiv preprint arXiv:2301.00234*.

Girdhar, R.; El-Nouby, A.; Liu, Z.; Singh, M.; Alwala, K. V.; Joulin, A.; and Misra, I. 2023. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15180–15190.

Gupta, D.; Attal, K.; and Demner-Fushman, D. 2022. A Dataset for Medical Instructional Video Classification and Question Answering. arXiv:2201.12888.

Kojima, T.; Gu, S.; Reid, M.; Matsuo, Y.; and Iwasawa, Y. 2023. Large language models are zero-shot reasoners. arXiv.

Lewis, M.; Liu, Y.; Goyal, N.; Ghazvininejad, M.; Mohamed, A.; Levy, O.; Stoyanov, V.; and Zettlemoyer, L. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Lin, C.-Y. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, 74–81.

Liu, G.; Liao, Y.; Wang, F.; Zhang, B.; Zhang, L.; Liang, X.; Wan, X.; Li, S.; Li, Z.; Zhang, S.; et al. 2021. Medical-vlbert: Medical visual language bert for covid-19 ct report generation with alternate learning. *IEEE Transactions on Neural Networks and Learning Systems*, 32(9): 3786–3797.

Mrini, K.; Dernoncourt, F.; Chang, W.; Farcas, E.; and Nakashole, N. 2021. Joint summarization-entailment optimization for consumer health question understanding. In *Proceedings of the Second Workshop on Natural Language Processing for Medical Conversations*, 58–65.

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 of the Association for Computational Linguistics*, 311–318.

Qi, W.; Yan, Y.; Gong, Y.; Liu, D.; Duan, N.; Chen, J.; Zhang, R.; and Zhou, M. 2020. Prophetnet: Predicting future n-gram for sequence-to-sequence pre-training. *arXiv preprint arXiv:2001.04063*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Thawkar, O.; Shaker, A.; Mullappilly, S. S.; Cholakkal, H.; Anwer, R. M.; Khan, S.; Laaksonen, J.; and Khan, F. S. 2023. Xraygpt: Chest radiographs summarization using medical vision-language models. *arXiv preprint arXiv:2306.07971*.

Tiwari, A.; Manthena, M.; Saha, S.; Bhattacharyya, P.; Dhar, M.; and Tiwari, S. 2022. Dr. can see: towards a multi-modal disease diagnosis virtual assistant. In *Proceedings of the 31st ACM international conference on information & knowledge management*, 1935–1944.

Yadav, S.; Gupta, D.; Abacha, A. B.; and Demner-Fushman, D. 2021. Reinforcement learning for abstractive question summarization with question-aware semantic rewards. *arXiv preprint arXiv:2107.00176*.

Zhang, J.; Huang, J.; Jin, S.; and Lu, S. 2023. Vision-language models for vision tasks: A survey. *arXiv preprint arXiv:2304.00685*.

Zhang, J.; Zhao, Y.; Saleh, M.; and Liu, P. 2020. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International Conference on Machine Learning*, 11328–11339. PMLR.

Zhang, T.; Kishore, V.; Wu, F.; Weinberger, K. Q.; and Artzi, Y. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Zheng, L.; Chiang, W.-L.; Sheng, Y.; Zhuang, S.; Wu, Z.; Zhuang, Y.; Lin, Z.; Li, Z.; Li, D.; Xing, E.; et al. 2023. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. *arXiv preprint arXiv:2306.05685*.

Zhou, J.; He, X.; Sun, L.; Xu, J.; Chen, X.; Chu, Y.; Zhou, L.; Liao, X.; Zhang, B.; and Gao, X. 2023. SkinGPT-4: An Interactive Dermatology Diagnostic System with Visual Large Language Model.

Zhu, D.; Chen, J.; Shen, X.; Li, X.; and Elhoseiny, M. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.