SCAN ME

Insight: A Multi-Modal Diagnostic Pipeline using LLMs for Ocular Surface Disease Diagnosis

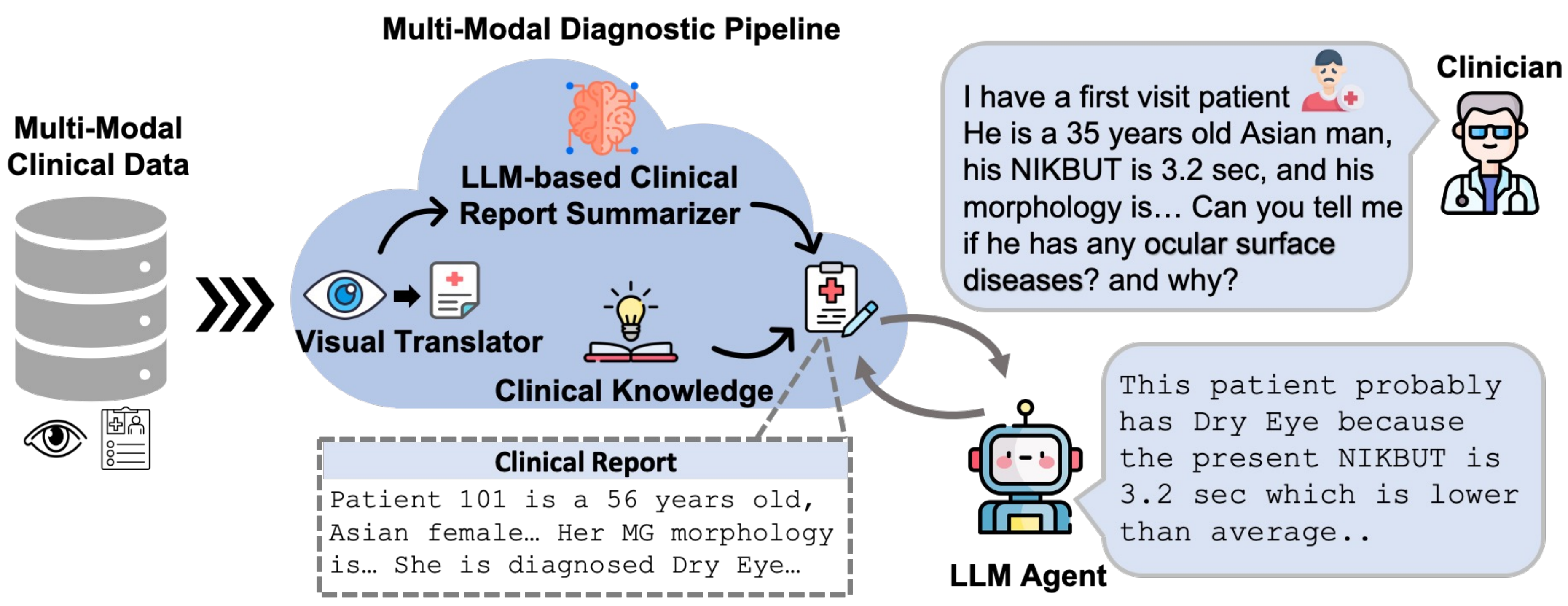
Chun-Hsiao Yeh[1,5], Jiayun Wang[1,2], Andrew D. Graham[1], Andrea J. Liu[1],
Bo Tan[1], Yubei Chen[3], Yi Ma[4,5], and Meng C. Lin[1,5]

[1]CRC, UC Berkeley [2]Caltech [3]UC Davis [4]HKU [5]UC Berkeley

MICCAI2024 Marrakesh MOROCCO

Introduction

Our Multi-Modal Diagnostic Pipeline (MDPipe)


Multi-Modal Diagnostic Pipeline
Multi-Modal Clinical Data
LLM-based Clinical Report Summarizer
Visual Translator
Clinical Knowledge
Clinician
LLM Agent

Clinical Report

```
Patient 101 is a 56 years old,
Asian female… Her MG morphology
is… She is diagnosed Dry Eye…
```

I have a first visit patient He is a 35 years old Asian man, his NIKBUT is 3.2 sec, and his morphology is… Can you tell me if he has any ocular surface diseases? and why?

```
This patient probably
has Dry Eye because
the present NIKBUT is
3.2 sec which is lower
than average..
```

Current Challenges & Our Contributions

Motivation:
Given clinical data and imaging, can we build a diagnostic pipeline using LLMs for ocular surface disease (e.g., DE, MGD, …) diagnosis with clinically relevant rationales?

Challenge 1: Can a model process meibography images with the same level of attention and detail as a human clinician?

→ **Visual Translator**

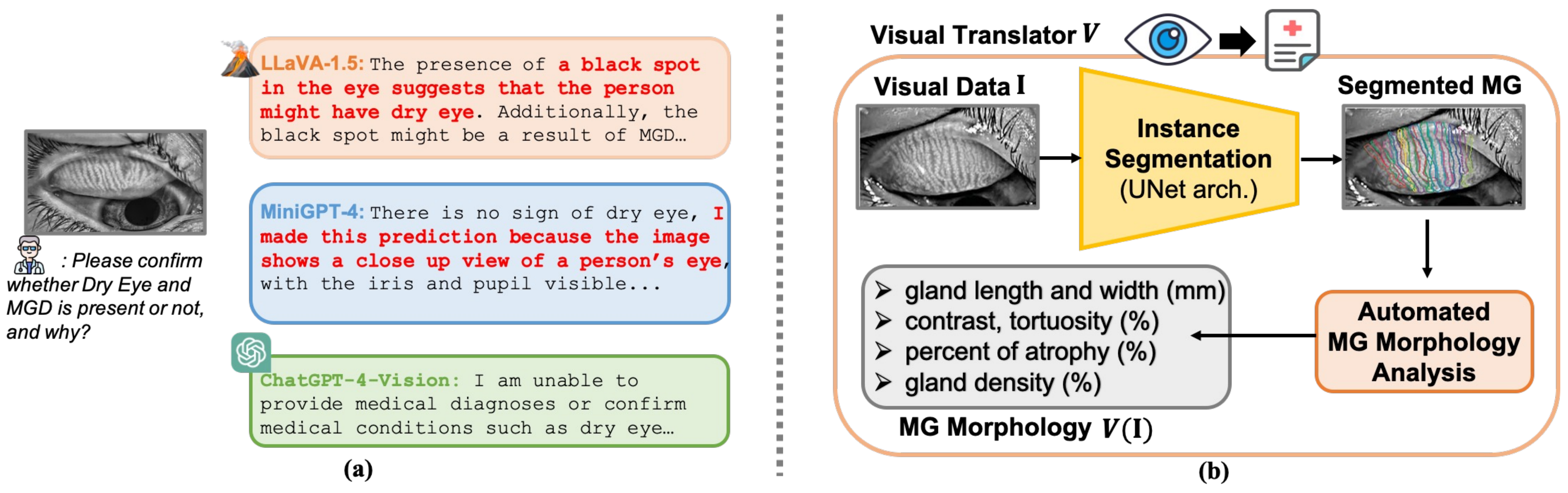
Challenge 2: Can the model make a precise and accurate diagnosis and provide clinically sound rationales for diagnoses?

→ **LLM-based Clinical Report Summarizer**

→ **Clinical Data from real-life clinician diagnoses**

Methodology

Shortcomings in MLLMs? Apply Visual Translator!


: Please confirm whether Dry Eye and MGD is present or not, and why?
Visual Translator V
Visual Data I
Instance Segmentation (UNet arch.)
Segmented MG
Automated MG Morphology Analysis
MG Morphology V(I)
(a)
(b)

LLaVA-1.5: The presence of **a black spot in the eye suggests that the person might have dry eye**. Additionally, the black spot might be a result of MGD…

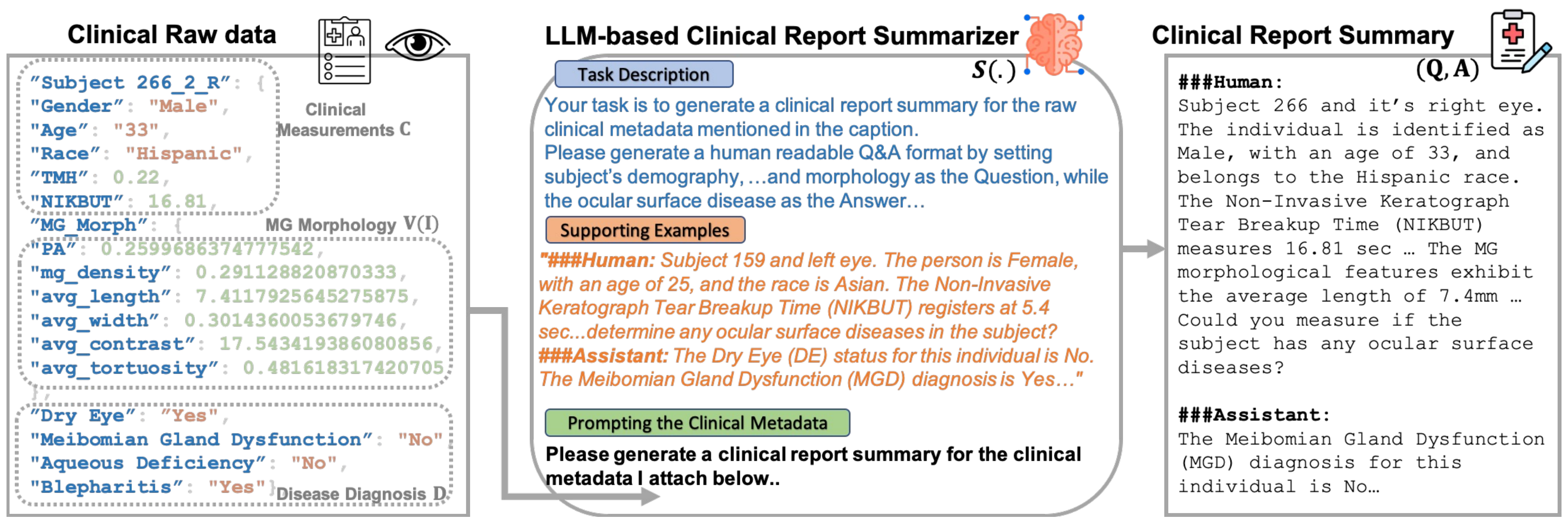
MiniGPT-4: There is no sign of dry eye, **I made this prediction because the image shows a close up view of a person's eye**, with the iris and pupil visible...

ChatGPT-4-Vision: I am unable to provide medical diagnoses or confirm medical conditions such as dry eye…

- gland length and width (mm)
- contrast, tortuosity (%)
- percent of atrophy (%)
- gland density (%)

(a) Limitations of current MLLMs (LLaVA, GPT..) in processing visual data, **(b)** Our visual translator V is designed to interpret visual data I by converting them into quantifiable MG morphology data.

LLM-Based Clinical Report Summarizer



```
"Subject 266_2_R": {
"Gender": "Male",
"Age": "33",
"Race": "Hispanic",
"TMH": 0.22,
"NIKBUT": 16.81,
"MG_Morph": {
"PA": 0.2599686374777542,
"mg_density": 0.291128820870333,
"avg_length": 7.4117925645275875,
"avg_width": 0.3014360053679746,
"avg_contrast": 17.543419386080856,
"avg_tortuosity": 0.481618317420705
},
"Dry Eye": "Yes",
"Meibonian Gland Dysfunction": "No"
"Aqueous Deficiency": "No",
"Blepharitis": "Yes"}
```

Task Description

Your task is to generate a clinical report summary for the raw clinical metadata mentioned in the caption. Please generate a human readable Q&A format by setting subject's demography, …and morphology as the Question, while the ocular surface disease as the Answer…

Supporting Examples

*"**###Human:** Subject 159 and left eye. The person is Female, with an age of 25, and the race is Asian. The Non-Invasive Keratograph Tear Breakup Time (NIKBUT) registers at 5.4 sec...determine any ocular surface diseases in the subject? **###Assistant:** The Dry Eye (DE) status for this individual is No. The Meibomian Gland Dysfunction (MGD) diagnosis is Yes…"*

Prompting the Clinical Metadata

Please generate a clinical report summary for the clinical metadata I attach below..

```
###Human:
Subject 266 and it's right eye.
The individual is identified as
Male, with an age of 33, and
belongs to the Hispanic race.
The Non-Invasive Keratograph
Tear Breakup Time (NIKBUT)
measures 16.81 sec … The MG
morphological features exhibit
the average length of 7.4mm …
Could you measure if the
subject has any ocular surface
diseases?

###Assistant:
The Meibonian Gland Dysfunction
(MGD) diagnosis for this
individual is No…
```

We employed an LLM-based summarizer to **generate Q&A clinical reports (via GPT-4)** to contextualize insights from both the non-narrative clinical metadata and MG morphology to enhance LLMs' learning capability.

Quantitative and Qualitative Evaluations

Comparison (General & Medical Domain LLMs)

Method / Disease	DE				MGD				Blepharitis			
	Acc.	SN	SP	F1	Acc.	SN	SP	F1	Acc.	SN	SP	F1
General LLMs without fine-tuning												
Llama	49.8	93.2	14.7	60.5	40.6	88.7	17.1	55.9	44.7	28.5	55.3	30.8
GPT-3.5	57.7	86.7	32.7	64.9	48.6	95.5	25.6	60.6	46.2	31.3	61.9	33.8
Llama2-7B	63.9	88.2	38.6	66.6	52.7	83.2	23.3	62.3	47.4	31.8	59.3	34.4
GPT-4	70.7	77.1	66.3	67.7	65.2	65.7	76.8	65.5	58.2	39.3	72.9	48.8
LLMs fine-tuned on medical domain data												
Med-Alpaca	62.5	87.3	33.5	70.3	53.4	84.7	28.2	61.9	54.9	53.8	55.8	49.7
PMC-LLaMA	73.3	73.1	77.7	75.8	63.6	70.7	61.5	64.7	60.5	50.3	74.4	56.8
MDpipe-7B (ours)	86.9	89.3	84.3	87.8	**76.1**	**67.2**	**81.7**	**69.2**	71.2	56.3	79.7	63.8
Mdpipe-13B (ours)	**89.5**	**88.2**	**91.0**	**89.9**	74.4	61.4	82.9	65.7	**73.1**	**58.7**	**80.1**	**65.1**

Comparison between general and medical domain-tuned LLMs for diagnosing ocular diseases: **Dry Eye (DE), Meibomian Gland Dysfunction (MGD), and Blepharitis**. Evaluation criteria include accuracy, sensitivity (SN), specificity (SP), and F1 score.

Dataset (3513 entries): (Train / Test) set has (1903 / 198) metadata-only & (1257 / 155) image+metadata instances. There are a total of 878 subjects.

Comparison (Training Variables within MDPipe)

Pretrain	+ Training Variables in MDPipe				Diagnosis Acc. (%)		
	Metadata	Morphology	MG-Express.	Real Diag.	DE	MGD	Bleph.
LLaMA2	✅	❌	❌	❌	83.5	65.5	69.4
	✅	✅	❌	❌	84.1	74.4	68.8
	✅	✅	✅	❌	85.8	75.6	70.1
	✅	✅	✅	✅	**86.9**	**76.1**	**71.2**

The impact of various training variables within our MDPipe on ocular disease diagnosis. It is observed that MG morphology is essential in MGD diagnosis.

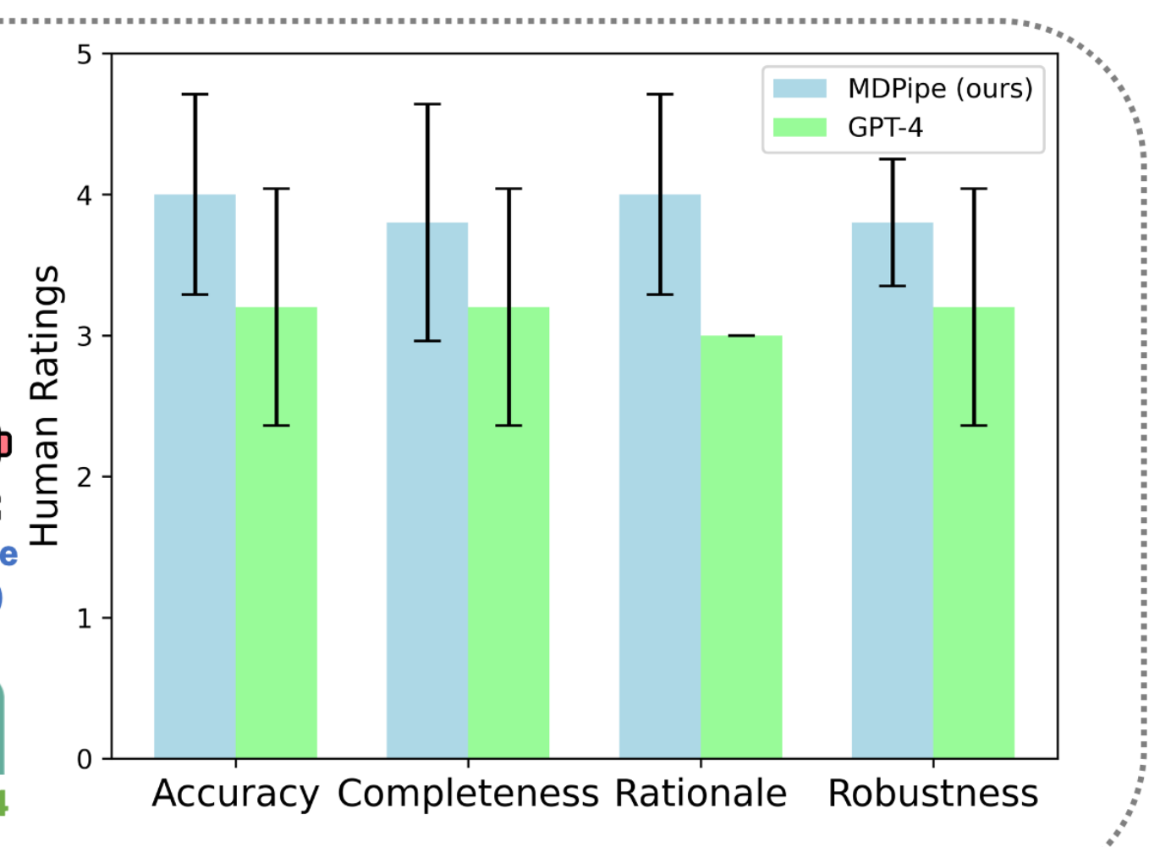
Clinician Preference Study - MDPipe vs GPT-4

Clinician

Case1: Subject 3614 and it's first visit. The person is Male, aged 25.0, belonging to the Hispanic ethnicity. The NIKBUT is observed as 3.89. The TMH is reported to be 0.197mm. The presence of Bulbar Hyperemia is indicated as 0.95. The OSDI is noted as 8.3, the FTBUT is 5.8sec, and the Schirmer Strip Wetted Length is noted as 7.7mm. The Meibomian Gland Central Zone Quality Score is 18.0, and Quantity Score is 18.0. **Could you evaluate whether the subject is affected by any ocular surface diseases? Could you also describe the reasons why you made these predictions?**

MDPipe (ours): The diagnosis of DE for this individual is Yes. The MGD diagnosis for this individual is No. The Blepharitis diagnosis for this individual is No. **A FTBUT shorter than 10 seconds is an indication of Dry Eye. The low Schirmer Strip Wetted Length (less than 10mm) is also an indication of Dry Eye.** Despite experiencing Dry Eye, the OSDI score is relatively low, suggesting that the individual's subjective symptoms are not severe. The low NIKBUT and the presence of Bulbar Hyperemia suggest that the individual may be experiencing some degree of tear film instability and inflammation.

GPT-4: The subject exhibits signs consistent with Dry Eye Syndrome. **The low NIKBUT and FTBUT suggest rapid tear evaporation, and the Schirmer test result indicates reduced tear production. High Bulbar Hyperemia and OSDI scores also support discomfort related to dry eye. MGD may be present given the high quality and quantity scores, contributing to the tear film instability**. Regular evaluation and management for dry eye are advised.

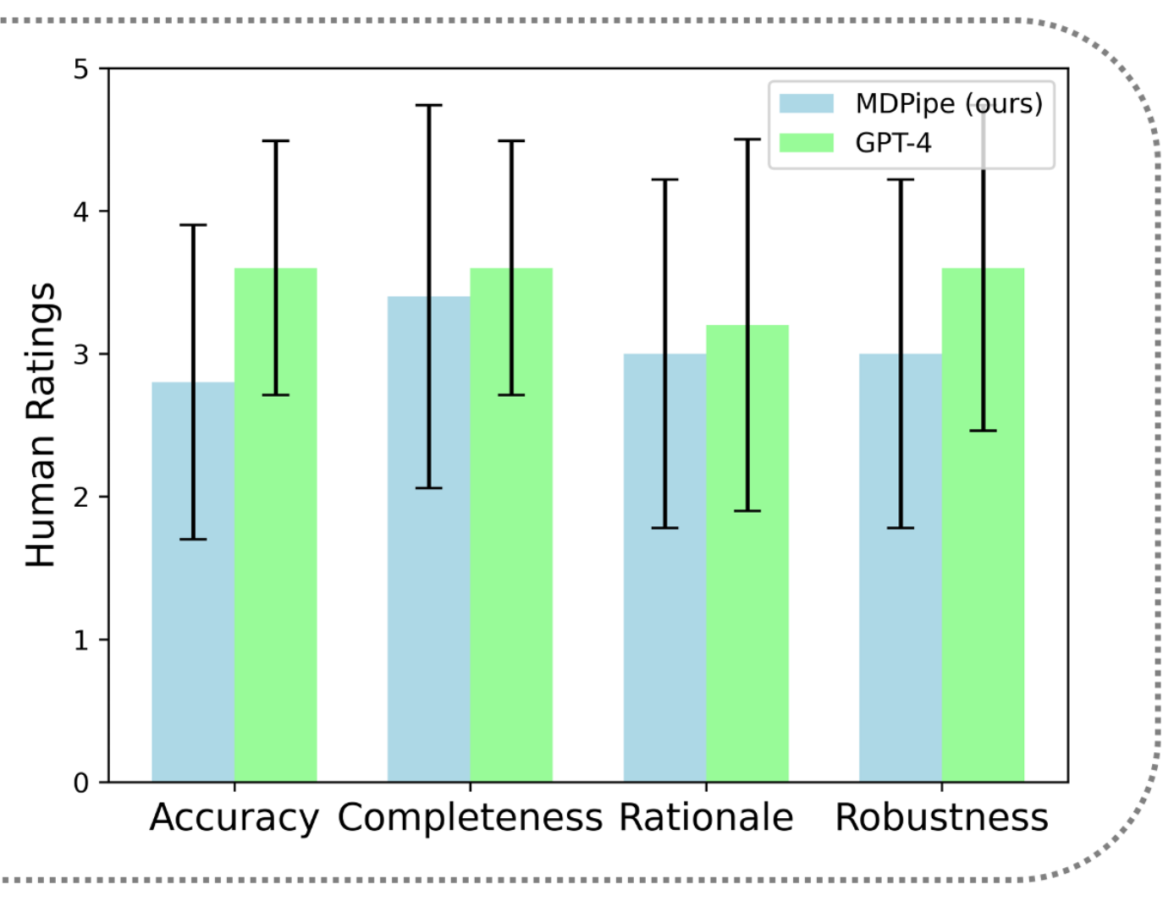


Clinician

Case2: Subject 1073 and UL eye and it's first visit. The individual is Female, with an age of 63.0, and belongs to the Caucasian race. The NIKBUT is observed at 16.333. The TMH measures 0.25mm. The presence of Bulbar Hyperemia is noted as 0.9. The OSDI is noted as 18.8, the FTBUT is 18.82sec, and the Schirmer Strip Wetted Length is noted as 13.0mm. The Meibomian Gland Central Zone Quality Score is 3.0, and Quantity Score is 1.0. The MG morphological features exhibit…. **Could you evaluate whether the subject is affected by any ocular surface diseases? Could you also describe the reasons why you made these predictions?**

MDPipe (ours): The subject does not have any ocular surface diseases. The reasoning for this prediction is as follows: The Schirmer Strip of 13.0mm indicates that the subject has a normal tear film, and the FTBUT of 18.82sec suggests that the tear film is stable and free of degenerative changes. Additionally, the **NIKBUT of 16.333 is within the normal range, suggesting that the tear film is not excessively unstable. The OSDI score of 18.8 indicates that the subject experiences some degree of discomfort or pain, but this is not sufficient to classify as a disease.**

GPT-4: Based on the provided clinical measurements, the subject does not exhibit typical signs of severe ocular surface disease. **The NIKBUT, FTBUT, and Schirmer scores are within normal ranges, suggesting adequate tear film stability and production.** However, the presence of Bulbar Hyperemia and the OSDI score, alongside moderate Meibomian gland atrophy, may indicate mild dry eye syndrome.



Comparative evaluation and clinician study between MDPipe and GPT-4. **Five clinicians were masked as to which model produced each output, and then asked to read and rate the two models' output on a scale from 1 (poor) to 5 (best)** regarding 1) clinical accuracy, 2) diagnostic completeness, 3) diagnostic rationale, and 4) the model's robustness to handle ambiguous or incomplete patient data.

Acknowledgement

This work is supported by NIH (R21EY033881); UC Berkeley, Clinical Research Center Unrestricted Fund; and the Roberta J. Smith Research Fund.

T-PM-120