MedMax: Mixed-Modal Instruction Tuning for Training Biomedical Assistants

Hritik Bansal, Daniel Mingyi Israel*, Siyan Zhao*, Shufan Li, Tung Nguyen, Aditya Grover

NeurIPS 2025

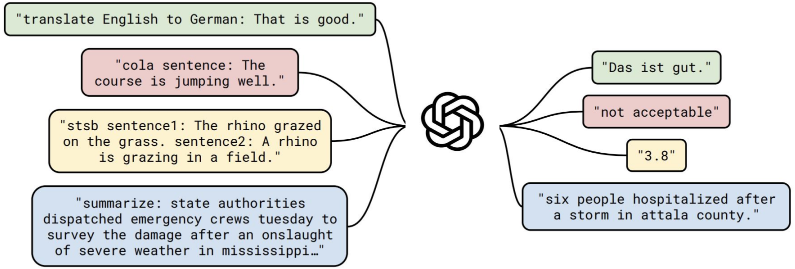
Datasets and Benchmarks Track

UCLA

Foundation Models

ML models that learn from vast amounts of data

Answer questions in text



"translate English to German: That is good."

"cola sentence: The course is jumping well."

"stsb sentence1: The rhino grazed on the grass. sentence2: A rhino is grazing in a field."

"summarize: state authorities dispatched emergency crews tuesday to survey the damage after an onslaught of severe weather in mississippi…"

"Das ist gut."

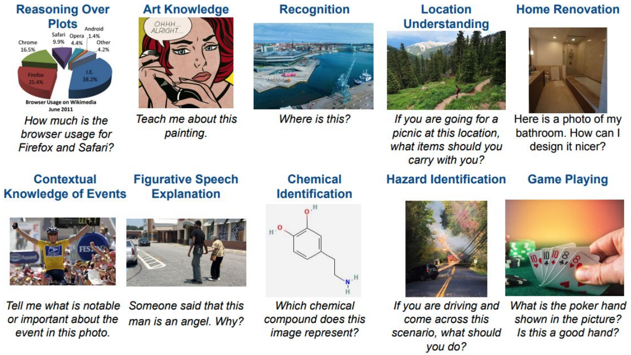
"not acceptable"

"3.8"

"six people hospitalized after a storm in attala county."

Answer questions on images

GPT-4V



How much is the browser usage for Firefox and Safari?

Teach me about this painting.

Where is this?

If you are going for a picnic at this location, what items should you carry with you?

Here is a photo of my bathroom. How can I design it nicer?

Tell me what is notable or important about the event in this photo.

Someone said that this man is an angel. Why?

Which chemical compound does this image represent?

If you are driving and come across this scenario, what should you do?

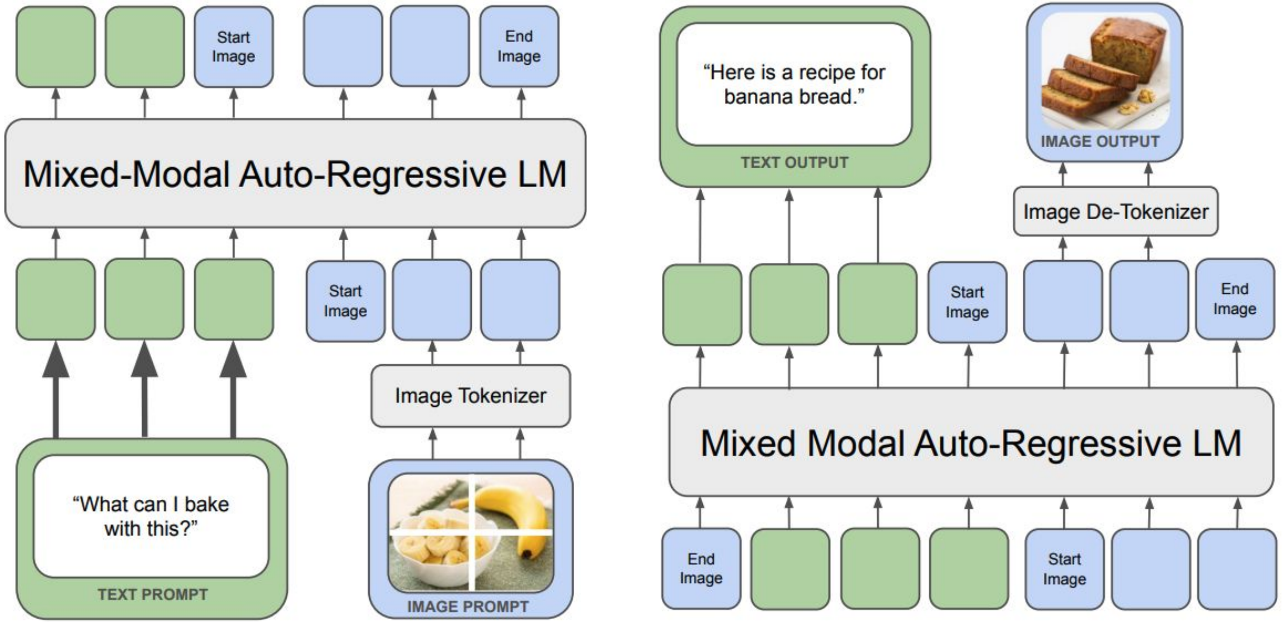
What is the poker hand shown in the picture? Is this a good hand?

Generate images based on descriptions

DALL·E

Mixed Modal Foundation Models: **All in one!**

Understand and generate arbitrary sequences of image and text.

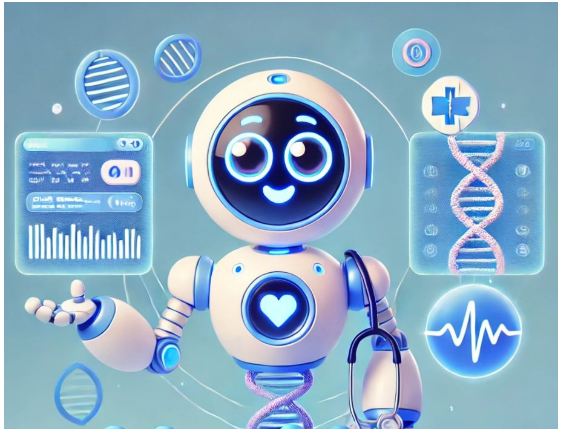


(a) Mixed-Modal Pre-Training

(b) Mixed-Modal Generation

Chameleon: Meta FAIR

Great potential for a unified foundation model for scientific reasoning!

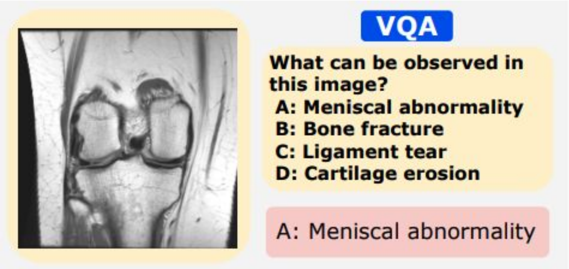


Biomedical Assistant

Diagnosis

Monitoring

Medical Reports



VQA

What can be observed in this image?
A: Meniscal abnormality
B: Bone fracture
C: Ligament tear
D: Cartilage erosion

A: Meniscal abnormality

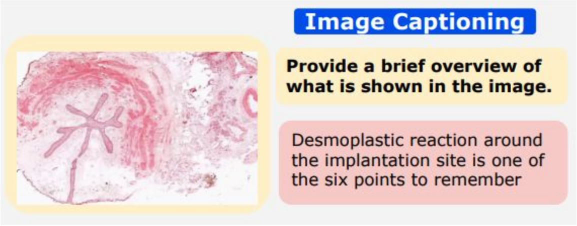
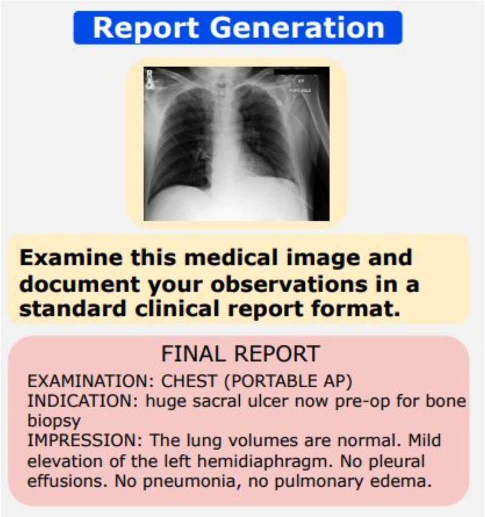


Image Captioning

Provide a brief overview of what is shown in the image.

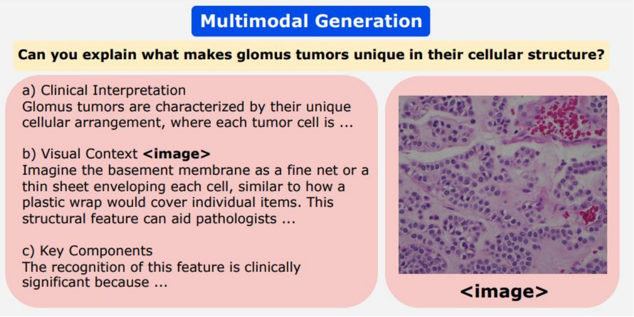
Desmoplastic reaction around the implantation site is one of the six points to remember



Report Generation

Examine this medical image and document your observations in a standard clinical report format.

FINAL REPORT
EXAMINATION: CHEST (PORTABLE AP)
INDICATION: huge sacral ulcer now pre-op for bone biopsy
IMPRESSION: The lung volumes are normal. Mild elevation of the left hemidiaphragm. No pleural effusions. No pneumonia, no pulmonary edema.



Multimodal Generation

Can you explain what makes glomus tumors unique in their cellular structure?

a) Clinical Interpretation
Glomus tumors are characterized by their unique cellular arrangement, where each tumor cell is ...

b) Visual Context **<image>**
Imagine the basement membrane as a fine net or a thin sheet enveloping each cell, similar to how a plastic wrap would cover individual items. This structural feature can aid pathologists ...

c) Key Components
The recognition of this feature is clinically significant because ...

<image>

Lack of large-scale multimodal biomedical datasets to enable complex reasoning across diverse domains!

	✓	✗
VQA datasets (e.g., VQA-RAD)	**Domain experts**	**Limited size** **(~1000s)**
Alignment and chat data (e.g., Llava-Med)	**Scalable**	**Bad quality** **(few biomedical images)**
Curated data (e.g., PubMedVision)	**High quality**	**Limited scope** **(medical research papers)**

MedMax

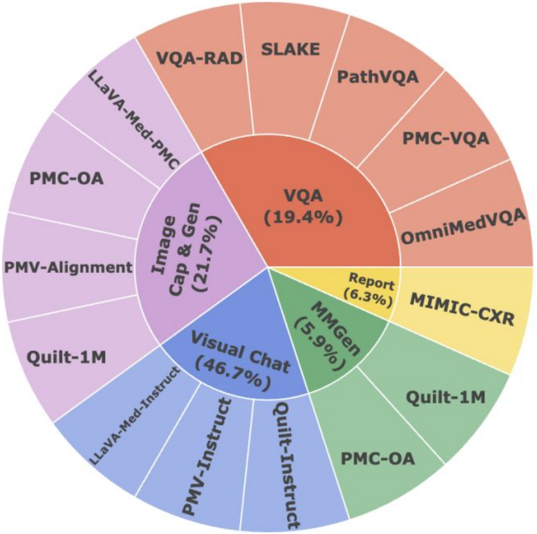
★ **Contains 1.5 million examples spanning many tasks and domains.**

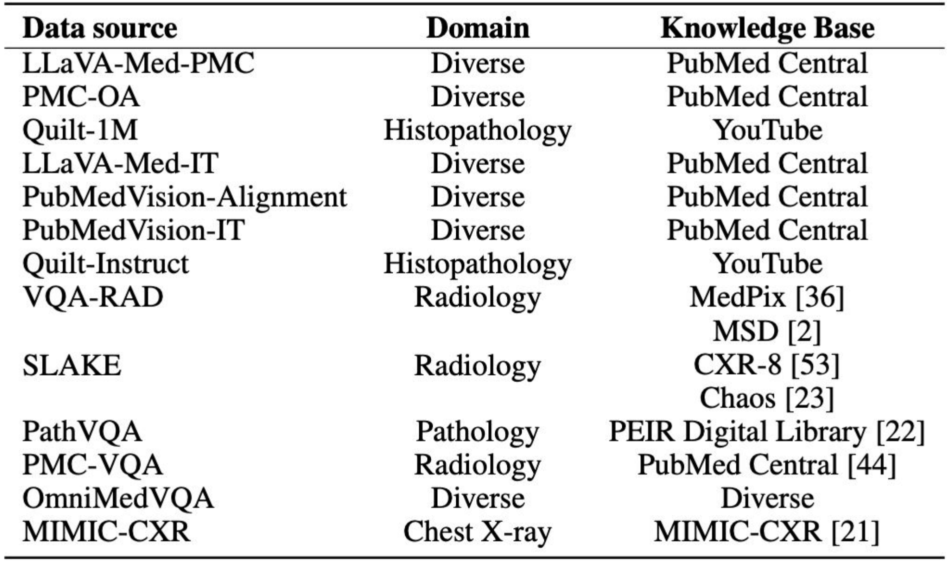
★ **Train mixed-modal foundation model to achieve state-of-the-art performance.**

★ **Added support with a comprehensive and automatic evaluation suite.**

MedMax Data Curation

Data curation across several sources and tasks to enable diverse skills across domains.

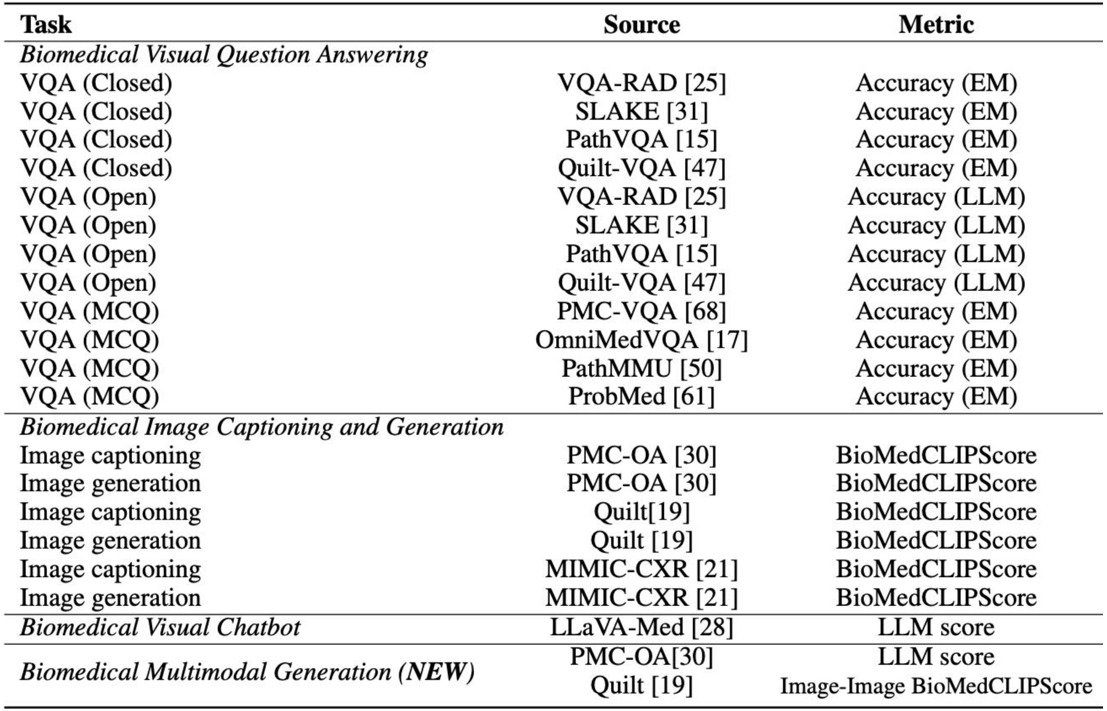




Data source	Domain	Knowledge Base
LLaVA-Med-PMC	Diverse	PubMed Central
PMC-OA	Diverse	PubMed Central
Quilt-1M	Histopathology	YouTube
LLaVA-Med-IT	Diverse	PubMed Central
PubMedVision-Alignment	Diverse	PubMed Central
PubMedVision-IT	Diverse	PubMed Central
Quilt-Instruct	Histopathology	YouTube
VQA-RAD	Radiology	MedPix [36]
SLAKE	Radiology	MSD [2] CXR-8 [53] Chaos [23]
PathVQA	Pathology	PEIR Digital Library [22]
PMC-VQA	Radiology	PubMed Central [44]
OmniMedVQA	Diverse	Diverse
MIMIC-CXR	Chest X-ray	MIMIC-CXR [21]

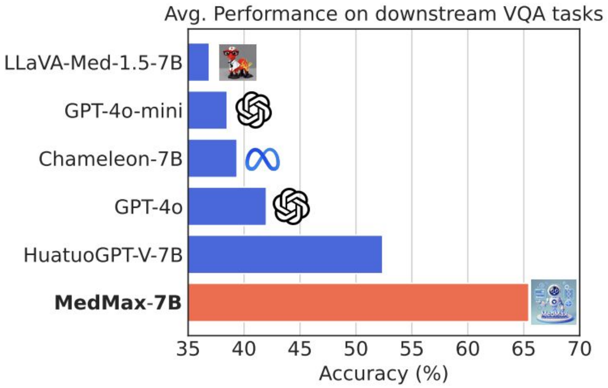
Comprehensive Automatic Evaluation

~10K evaluation examples!

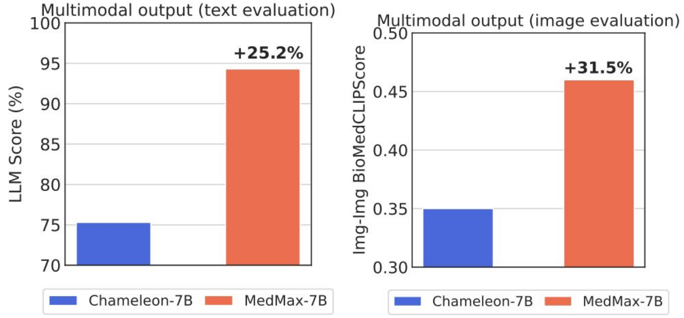


Task	Source	Metric
Biomedical Visual Question Answering		
VQA (Closed)	VQA-RAD [25]	Accuracy (EM)
VQA (Closed)	SLAKE [31]	Accuracy (EM)
VQA (Closed)	PathVQA [15]	Accuracy (EM)
VQA (Closed)	Quilt-VQA [47]	Accuracy (EM)
VQA (Open)	VQA-RAD [25]	Accuracy (LLM)
VQA (Open)	SLAKE [31]	Accuracy (LLM)
VQA (Open)	PathVQA [15]	Accuracy (LLM)
VQA (Open)	Quilt-VQA [47]	Accuracy (LLM)
VQA (MCQ)	PMC-VQA [68]	Accuracy (EM)
VQA (MCQ)	OmniMedVQA [17]	Accuracy (EM)
VQA (MCQ)	PathMMU [50]	Accuracy (EM)
VQA (MCQ)	ProbMed [61]	Accuracy (EM)
Biomedical Image Captioning and Generation		
Image captioning	PMC-OA [30]	BioMedCLIPScore
Image generation	PMC-OA [30]	BioMedCLIPScore
Image captioning	Quilt[19]	BioMedCLIPScore
Image generation	Quilt [19]	BioMedCLIPScore
Image captioning	MIMIC-CXR [21]	BioMedCLIPScore
Image generation	MIMIC-CXR [21]	BioMedCLIPScore
Biomedical Visual Chatbot	LLaVA-Med [28]	LLM score
*Biomedical Multimodal Generation (**NEW**)*	PMC-OA[30]	LLM score
	Quilt [19]	Image-Image BioMedCLIPScore

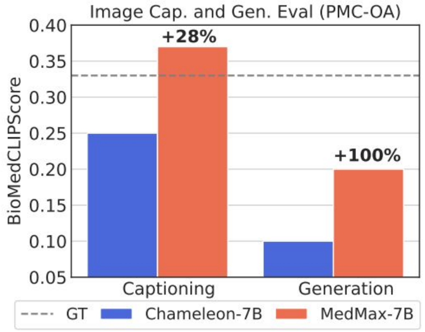
MedMax trains state-of-the-art biomedical assistant across tasks



VQA

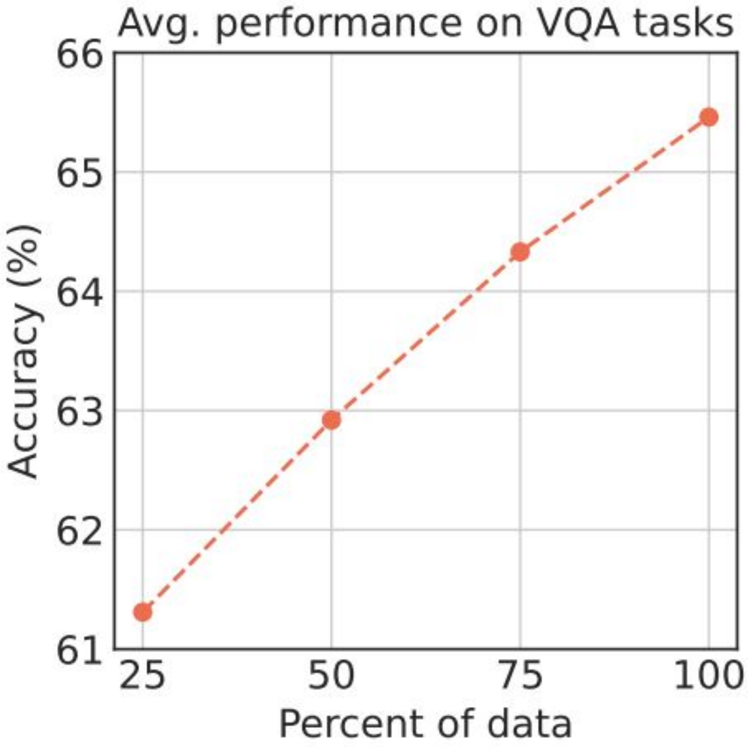


Multimodal Generation



Captioning

Performance scales with data size



Data, Model, Code is publicly available!

https://mint-medmax.github.io/

MedMax:

Mixed-Modal Instruction Tuning for Training Biomedical Assistants

Hritik Bansal, Daniel Israel[†], Siyan Zhao[†], Shufan Li, Tung Nguyen, Aditya Grover

University of California, Los Angeles

[†]Equal Contribution

Thank you!