RadiGist

– LLM Based Impression Prediction Solution

Group member:

Junru (Aeris) Li, Yue Zhang, Kairan Zhong, Xiran Li

Team members introduction

Junru (Aeris) Li

- **Studies:** ADSP in the University of Chicago
- **Work experience:** Data Analytics & Strategy

Yue Zhang

- **Studies:** ADSP in the University of Chicago
- **Work experience:** Data Science & Analytics

Kairan Zhong

- **Studies:** ADSP in the University of Chicago
- **Work experience:** Data Analytics

Xiran Li

- **Studies:** ADSP in the University of Chicago
- **Work experience:** Retail & Finance

Agenda

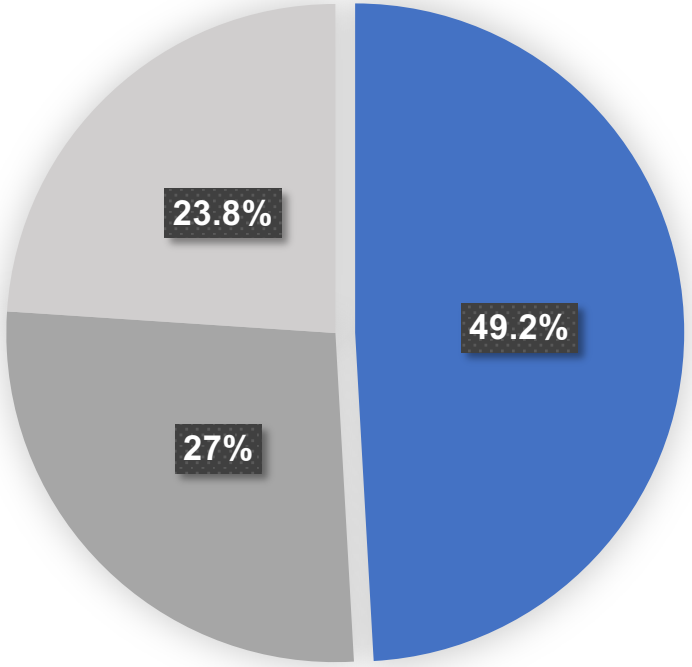
Introduction

- Business Background
- Business Values
- Problem Statement

Physicians spend more time on report summaries than patient treatment

The existing healthcare technology infrastructure struggles with processing unstructured text, leading to several critical challenges:

- **Physician burnout**
- **Lower-quality or even errors in care delivery**
- **Inefficiencies in healthcare workflows**



Source: American Medical Association

RadiGist provides benefits from several aspects

Business Values

Reduction of Physician Burnout

Mitigate physician mental fatigue from repetitive tasks, reducing their burnout

High-quality Patient Care

Minimize patient misdiagnosis caused by errors in medical report interpretation

Precision and Efficiency

Generate accurate and concise impressions in nearly real time

Readability and Usability

Create a user-friendly interface

Inference Analytics advances healthcare with AI-driven medical language processing support

01

Company Background

- A Chicago-based healthcare Artificial Intelligence (AI) startup that focuses on medical language processing

02

Core Business

- AI platform providing real-time clinical decision support to radiologists

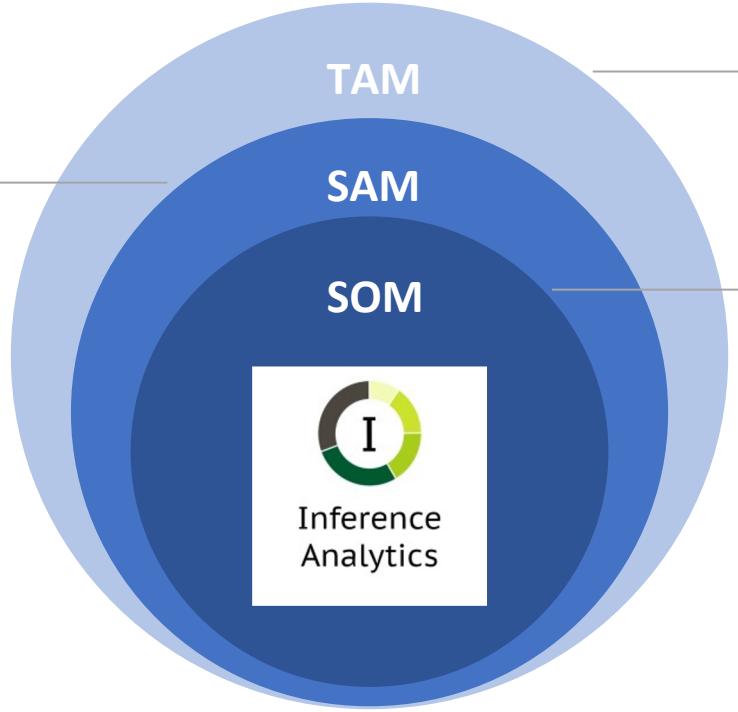
03

Latest Project

- Developing an AI-based solution to generate "impressions" within the radiology report
- Our team will use advanced predictive models, LLM, to create a deployable solution that generates real-time impressions from clinical information and findings

Our value proposition empowers the client to capture a larger market share

Serviceable Available Market: LLMs
- 2023 Market size of $12.75 billion; CAGR[1] at 21.4%



Total Addressable Market: GenAI
- 2023 Market size of $56.8 billion; CAGR at 42%

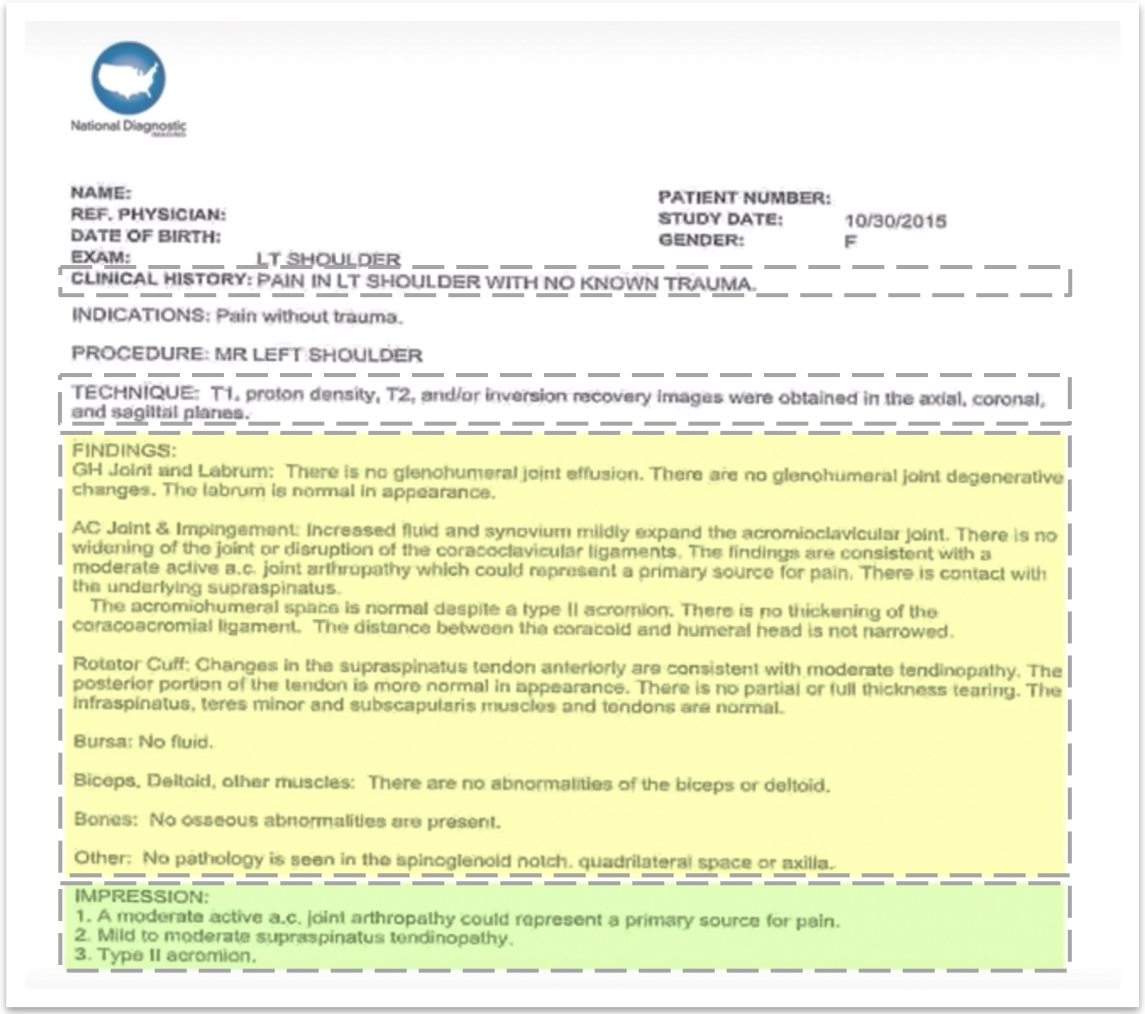
Serviceable Obtainable Market: LLMs in healthcare industry
- 2023 Market size of $0.5 billion

Source: Bloomberg, Allied Market Research

With the empowerment of our RadiGist product,
Inference Analytics can enhance its competitive edge and meet market demands more effectively

[1]: Compound annual growth rate

RadiGist is focused on generating impressions from unstructured narratives within radiology reports



National Diagnostic

NAME:
REF. PHYSICIAN:
DATE OF BIRTH:
EXAM: LT SHOULDER

PATIENT NUMBER:
STUDY DATE: 10/30/2015
GENDER: F

CLINICAL HISTORY: PAIN IN LT SHOULDER WITH NO KNOWN TRAUMA.

INDICATIONS: Pain without trauma.

PROCEDURE: MR LEFT SHOULDER

TECHNIQUE: T1, proton density, T2, and/or inversion recovery images were obtained in the axial, coronal, and sagittal planes.

FINDINGS:
GH Joint and Labrum: There is no glenohumeral joint effusion. There are no glenohumeral joint degenerative changes. The labrum is normal in appearance.

AC Joint & Impingement: Increased fluid and synovium mildly expand the acromioclavicular joint. There is no widening of the joint or disruption of the coracoclavicular ligaments. The findings are consistent with a moderate active a.c. joint arthropathy which could represent a primary source for pain. There is contact with the underlying supraspinatus.
The acromiohumeral space is normal despite a type II acromion. There is no thickening of the coracoacromial ligament. The distance between the coracoid and humeral head is not narrowed.

Rotator Cuff: Changes in the supraspinatus tendon anteriorly are consistent with moderate tendinopathy. The posterior portion of the tendon is more normal in appearance. There is no partial or full thickness tearing. The infraspinatus, teres minor and subscapularis muscles and tendons are normal.

Bursa: No fluid.

Biceps, Deltoid, other muscles: There are no abnormalities of the biceps or deltoid.

Bones: No osseous abnormalities are present.

Other: No pathology is seen in the spinoglenoid notch, quadrilateral space or axilla.

IMPRESSION:
1. A moderate active a.c. joint arthropathy could represent a primary source for pain.
2. Mild to moderate supraspinatus tendinopathy.
3. Type II acromion.

Clinical information + Findings => Impressions

Clinical information: The patient's clinical history

Technique: Describes how the exam was done

Findings: Descriptive observations from the radiologist

Impressions: Interpretation and conclusion about the exam

Technical Background

Understanding these technical concepts is key to grasping our product

- **Text Summarization: Our project is an abstractive summarization task**

Abstractive Summarization
Generate new sentences while keeping the essence of the original text intact.

VS.

Extractive Summarization
Select and combine existing sentences from a text to create a summary.

- **Next Token Prediction**

Definition:
Predicts the most probable subsequent token based on the sequence of tokens that have already been processed.

How it works:
1. Sample a token from ~ p(next token | previous tokens)
2. Append the token to the input
3. Run the new input through the transformer ...

- **Large Language Models**

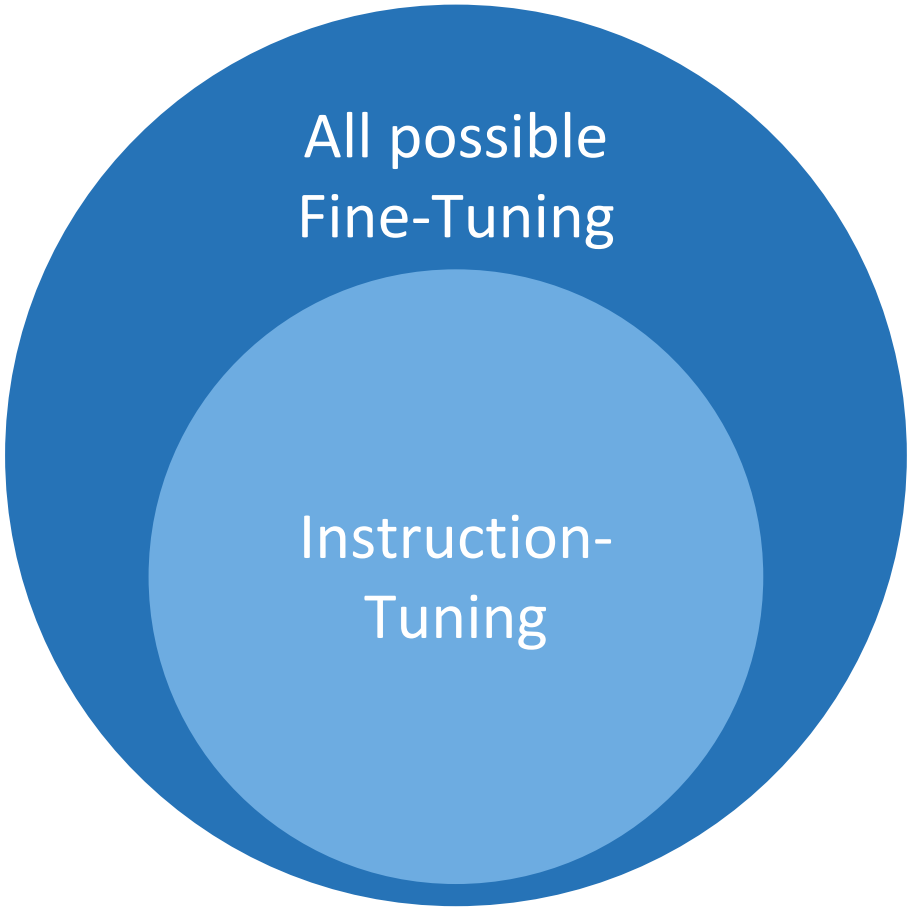
A massive number of parameters which are learned from vast datasets of text during training.

By employing next-token prediction, Large Language Models (LLMs) can generate coherent and contextually relevant text, one token at a time, based on the probabilities learned during their training phase.

Instruction Tuning is a subset of Fine-Tuning

- Focused on **alignment** with humans
- Concerned with **following** “instructions” like a human would
- Augments input-output examples with instructions, which enables instruction-tuned models to **generalize** more easily to new tasks

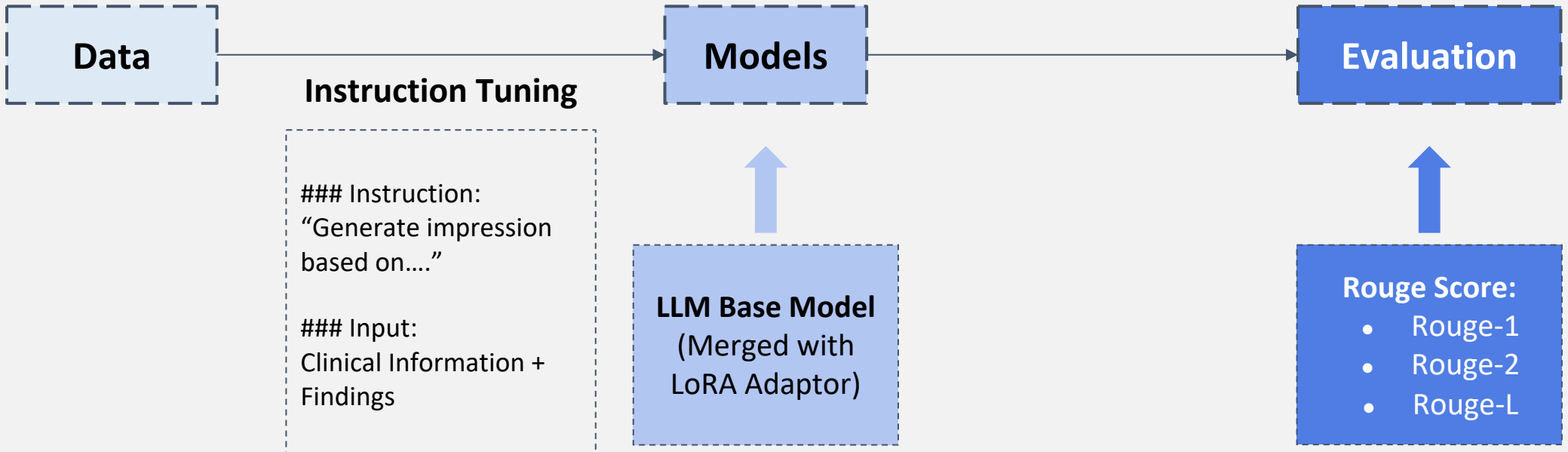
Example
Instruction: ### Input: ### Output:



Framework

- Workflow
- Data
- Models

What is the workflow of RadiGist



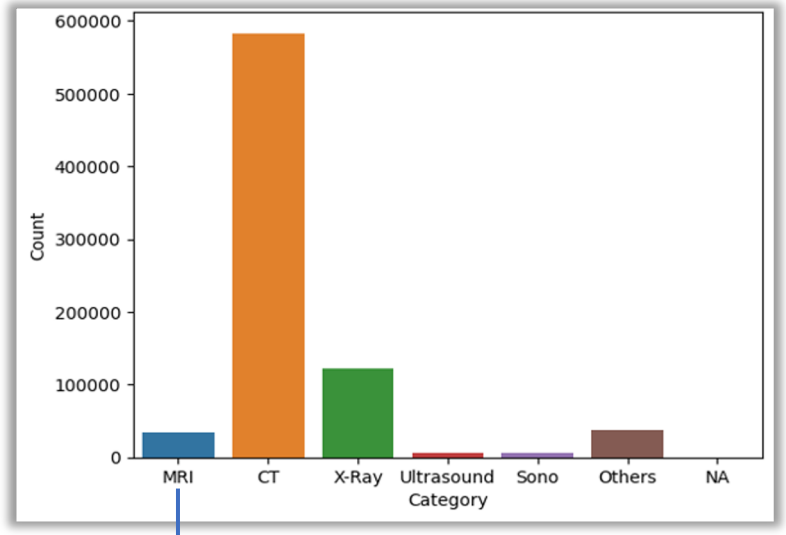
Our data is textual, well-organized, and covers multiple modalities

Data Source: 1M **de-identified** radiology reports from University of Chicago Medicine

Data Overview:

	Unnamed: 0	clinical_information	technique	findings	comparison	impression	report_id
0	0	34 year old female with history of sickle cell...	2 views of the right shoulder at 6:41 on 7/12/12	The right total shoulder arthroplasty componen...	XR shoulder 7/11/12	Right total shoulder arthroplasty components i...	RAD_0
1	1	34 year old female with history of sickle cell...	One portable view of the right shoulder at 17:...	The right total shoulder arthroplasty componen...	XR shoulder 2/1/12	Right total shoulder arthroplasty components i...	RAD_1
2	2	84-year-old female with low back pain	Four views of the lumbar spine	Posterior stabilization rods with transpedicul...	2/13/06	Posterior fixation of L4 and L5, appearing sim...	RAD_2
3	3	NaN	Informed consent was obtained. The patient was...	The colon is adequately cleansed and distended...	NaN	No significant colonic polyps or masses identi...	RAD_3
4	4	Preoperative planning for brain tumor. History...	MRI BRAIN STEALTH W/WO CONTRAST. A total of 17...	There is a heterogeneous left supratentorial a...	Brain MRI dated 11/17/14.	Presurgical planning MRI shows a complex mass ...	RAD_4

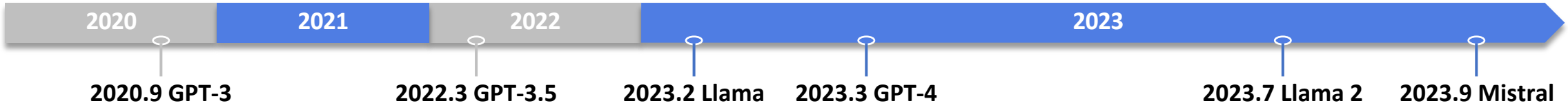
Categorize data into different modalities



Choose **MRI** for modeling considering the data size

Our model selection process is based on several criteria

Selection Criteria	Meta AI		Open AI		Mistral AI
	Llama	**Llama 2**	**GPT3**	**GPT4**	**Mistral**
Open-Source Availability	**Open-source,** accessible through API	**Open source**, available for free and for commercial use	**Not open-source**	**Not open-source**	**Open-source and community-focused**
Manageable Model Size	**Manageable**; there are versions with 7B and 13B parameters, as well as other fine-tuned versions	**Manageable**; there are versions with 7B and 13B parameters, as well as other fine-tuned versions	**Impractical,** with more than 175B parameters.	**Impractical,** with more parameters than GPT3 (precise size not revealed)	**Manageable** parameter sizes (7.3 billion parameters)
Proven Performance	LLaMA-13B outperforms GPT-3 (175B) on most benchmark	Strong improvements over prior LLMs across diverse benchmarks; Demonstrates factual accuracy on par with GPT4 and superior to GPT3.5 when summarizing text.[1]	Performs well in a variety of language tasks like abstract summarization	Better performance compare with the GPT previous version	Beats Llama 2 13B on all benchmarks and Llama 1 34B on many benchmarks [2]
Use Case	Typically include text summarization, language translation, and basic question-answering	More enhanced capabilities, more suitable for applications that demand a higher level of language comprehension	More specialized in creating human-like text and even performing programming tasks, known for its generation capability	Excels in more complex and nuanced language tasks like intricate content creation, complex problem-solving, and advanced conversational AI applications with greater accuracy and nuance	Supports a variety of use cases, such as text summarization, classification, text completion, and code completion



[1] https://promptengineering.org/how-does-llama-2-compare-to-gpt-and-other-ai-language-models/
[2] https://mistral.ai/news/announcing-mistral-7b/

Five LLMs are selected as our base models

Selection of models is based on:

- Open-Source Availability
- Manageable Model Size
- Proven Performance
- Use Case

Base Model

Large Language Models

Meta AI	Llama	Fine-Tuned Versions →	• Alpaca-7b • Medalpaca-7b (specifically fine-tuned on medical data)
	Llama 2.0		• Llama2-7b • Llama2-13b
Mistral AI	Mistral		• Mistral 7b

*** Alpaca-7b:** Alpaca 7B is a fine-tuned version of the seven billion-parameter LLaMA language model from Meta.
*** Medalpaca-7b:** MedAlpaca expands upon both Stanford Alpaca and AlpacaLoRA to offer an advanced suite of large language models specifically fine-tuned f**or medical question-answering** and dialogue applications.

Findings

- ROUGE Metrics Introduction
- ROUGE Metrics Comparison
- Demo

We use ROUGE scores as our evaluation metrics

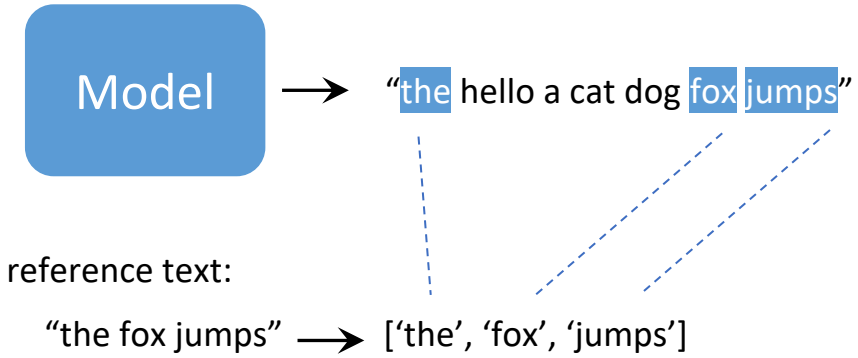
ROUGE (Recall-Oriented Understudy for Gisting Evaluation)

ROUGE-N: measures the number of matching "**n-grams**" between our model-generated text and a "reference"

ROUGE-L: measures the longest common subsequence (LCS) between our model output and reference

n-grams	Original: "the fox jumps" ↓ Unigrams: ['the', 'fox', 'jumps'] Bigrams: ['the fox', 'fox jumps'] Trigrams: ['the fox jumps']

Taking ROUGE-1 as an example:



$$\text{Recall} = \frac{\text{count}_{\text{match}}(\text{gram}_n)}{\text{count}(\text{gram}_n)\text{ in reference}} = \frac{3}{3} = 100\%$$

$$\text{Precision} = \frac{\text{count}_{\text{match}}(\text{gram}_n)}{\text{count}(\text{gram}_n)\text{ in model generated output}} = \frac{3}{7} = 43\%$$

$$\text{F-measure} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} = \frac{2 * 0.43 * 1.0}{0.43 + 1.0} = 60\%$$

We compare our models to the baseline result

Specification

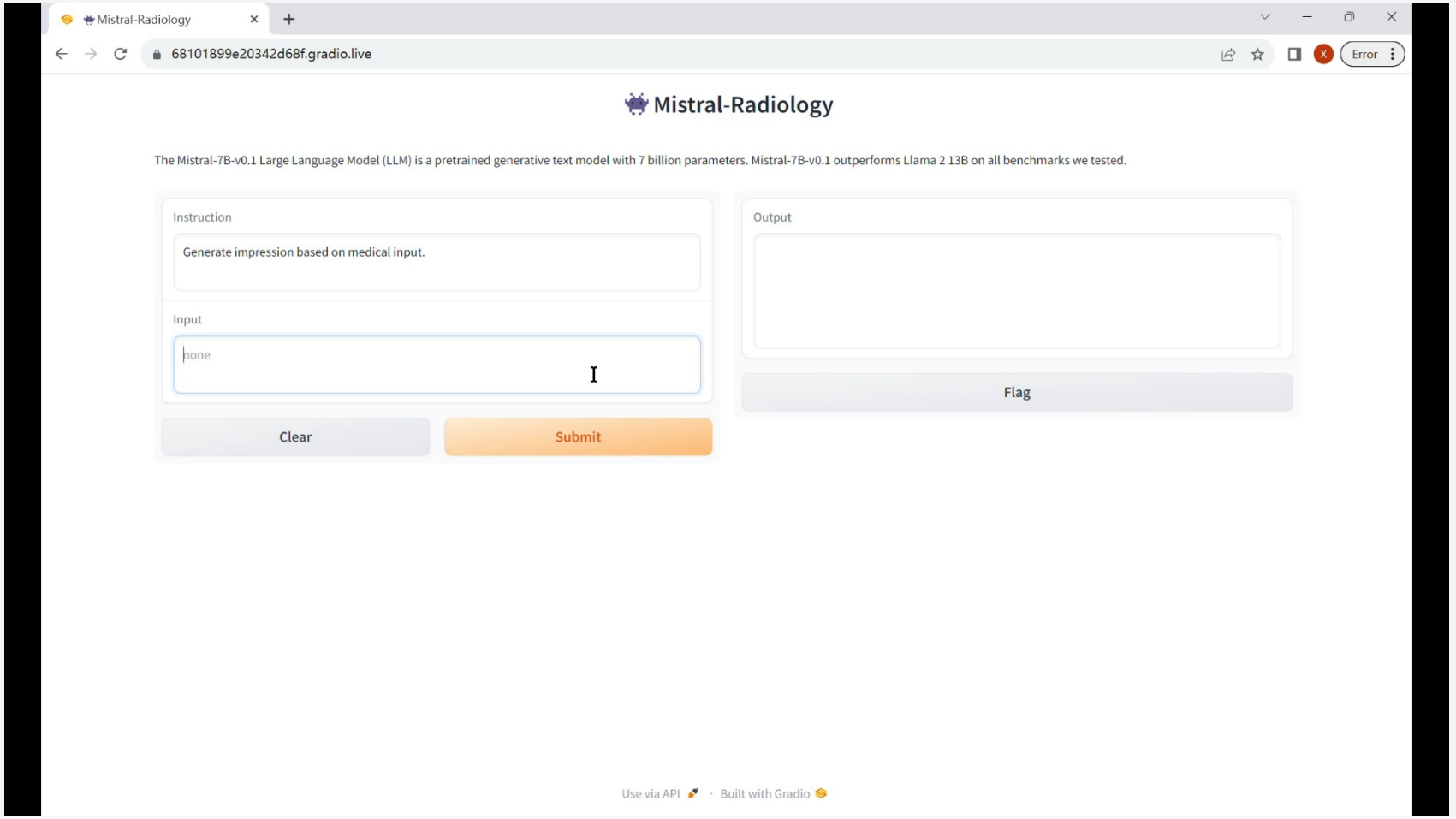
- Modality: MRI
- Data Size: 33,894
- Training : Test = 0.8 : 0.2

System	ROUGE-1	ROUGE-2	ROUGE-L
PG ⊕ BERT (Baseline Result)	47.17	32.89	45.91
Alpaca-7b	45.63	27.34	38.08
Medalpaca-7b	50.47	31.02	42.10
Llama-7b	46.75	28.02	39.07
Llama-13b	44.98	26.54	38.53
Mistral-7b	56.42	40.11	48.66

The fine-tuned **Mistral-7b** model emerged as the **top performer**, surpassing other models and baseline results

Our final deliverable is a user-friendly interface

Successfully deployed our models using Gradio to generate real-time impressions



Future Improvements

Future improvements could focus on refining evaluation methods and expanding generalizability

Evaluation Method

Strategy: Include judgements from Subject Matter Experts (SMEs) in radiology

Generalizability

Strategy: Test on other data (e.g., Indiana University data) to enhance generalization

Acknowledgement

Client

INFERENCE ANALYTICS

Utku Pamuksuz

- Chief Data Scientist

Benan Akca

- Senior Data Scientist

University

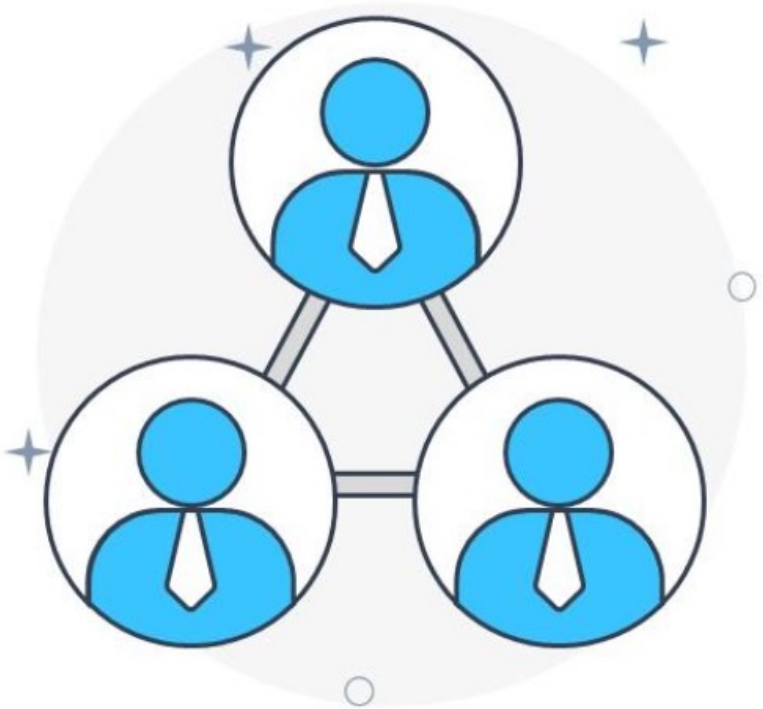
THE UNIVERSITY OF CHICAGO | MS-APPLIED DATA SCIENCE

Utku Pamuksuz

- Capstone Advisor
- Capstone Instructor

Roger Moore

- Capstone Instructor



Thanks!

Junru (Aeris) Li, Yue Zhang, Kairan Zhong, Xiran Li

11/28/2023

References

[1] Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M. A., Lacroix, T., ... & Lample, G. (2023). Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971.

[2] Gundogdu, B., Pamuksuz, U., Chung, J. H., Telleria, J. M., Liu, P., Khan, F., & Chang, P. J. (2021). Customized impression prediction from radiology reports using bert and lstms. IEEE Transactions on Artificial Intelligence.

[3] Lee, M., Srivastava, M., Hardy, A., Thickstun, J., Durmus, E., Paranjape, A., ... & Liang, P. (2022). Evaluating Human-Language Model Interaction. arXiv preprint arXiv:2212.09746..

[4] Zhang, A., Xing, L., Zou, J. et al. Shifting machine learning for healthcare from development to deployment and from models to data. Nat. Biomed. Eng 6, 1330–1345 (2022). https://doi.org/10.1038/s41551-022-00898-y

[5] Sezgin E, Sirrianni J, Linwood S. Operationalizing and Implementing Pretrained, Large Artificial Intelligence Linguistic Models in the US Health Care System: Outlook of Generative Pretrained Transformer 3 (GPT-3) as a Service Model. JMIR Med Inform 2022;10(2):e32875 https://medinform.jmir.org/2022/2/e32875

[6] Edjinedja, L., Omar, E., Barakat, O., Desmettre, T., & Marx, T. (2023). LEVERAGE CHATGPT AS TOOL FOR DATA AUGMENTATION FOR SUPERVISED MACHINE LEARNING TASKS IN HEALTHCARE NATURAL LANGUAGE PROCESSING.

[7] Cascella, M., Montomoli, J., Bellini, V., & Bignami, E. (2023). Evaluating the feasibility of ChatGPT in healthcare: an analysis of multiple clinical and research scenarios. Journal of Medical Systems, 47(1), 1-5.

References (cont.)

[8] Liu, Z., Yu, X., Zhang, L., Wu, Z., Cao, C., Dai, H., ... & Li, X. (2023). Deid-gpt: Zero-shot medical text de-identification by gpt-4. arXiv preprint arXiv:2303.11032.

[9] M Lafortune, G Breton, and JL Baudouin, “The radiological report: what is useful for the referring physician?,” Canadian Association of Radiologists journal= Journal l’Association canadienne des radiologistes, vol. 39, no. 2, pp. 140, 1988.

[10] Korngiebel, D.M., Mooney, S.D. Considering the possibilities and pitfalls of Generative Pre-trained Transformer 3 (GPT-3) in healthcare delivery. npj Digit. Med. 4, 93 (2021). https://doi.org/10.1038/s41746-021-00464-x

[11] Puspitaningrum, D. (2022). A Survey of Recent Abstract Summarization Techniques. In: Yang, XS., Sherratt, S., Dey, N., Joshi, A. (eds) Proceedings of Sixth International Congress on Information and Communication Technology. Lecture Notes in Networks and Systems, vol 217. Springer, Singapore. https://doi.org/10.1007/978-981-16-2102-4_71

[12] Zhaopu Teng. Abstractive summarization of COVID-19 with transfer text-to-text transformer. ACE (2023) Vol. 2: 232-238. DOI: 10.54254/2755-2721/2/20220520.

[13] Vo, T. A novel semantic-enhanced generative adversarial network for abstractive text summarization. Soft Comput (2023). https://doi.org/10.1007/s00500-023-07890-x

[14] Sinsky C, Colligan L, Li L, Prgomet M, Reynolds S, Goeders L, Westbrook J, Tutty M, Blike G. Allocation of Physician Time in Ambulatory Practice: A Time and Motion Study in 4 Specialties. Ann Intern Med. 2016 Dec 6;165(11):753-760. doi: 10.7326/M16-0961. Epub 2016 Sep 6. PMID: 27595430.

Appendix 1: Project timeline

TIME			Fall									
Phase	PROJECT WEEK:		1	2	3	4	5	6	7	8	9	10
1	Research Design	- Initial Proposal										
		- Business Problem										
2	Implementation	- final Proposal										
		- Research Proposal Draft										
		- Literature Review & Methodology										
3	Implementation (cont.)	- Research Paper 1st Draft	Research paper 1st draft									
		- Research Paper 2nd Draft				Research paper 2nd draft						
		- Research Paper							Final research paper			
4	Communication	- Showcase Draft					Showcase draft					
		- Video									Video	
		- Showcase										Showcase
5	Client Meeting (iterate)	- Client Meeting		meeting	meeting	meeting	meeting	meeting	meeting	meeting	meeting	

Project End

Appendix 2: Literature review summary

For background:	• The primary qualities that makes a radiological report useful for referring physicians are clarity, brevity, and clinical correlation. [9] • AI NLP applications can improve work satisfaction for providers and reduce time spent interacting with computer systems, which is a well-documented concern. This could be achieved by routinizing tedious work, such as navigating complex electronic health record (EHR) systems, automating documentation with human review, preparing orders, etc. [10]
For methodology:	• Large Language Model (BERT and LSTMs) [2] • Generative Pretrained Models (GPT, T-5, etc.) [7][8]
For model evaluation:	• non-interactive (Measuring similarity: e.g., ROUGE metrics; BLEU metrics;...etc.) [2] • interactive (Human-AI Language-based Interaction Evaluation (HALIE)) [3]
For model improvement:	• Use more recent transformer models for handling larger datasets and enhancing the modeling of clinical text. [4] • Data augmentation : 1. Deep generative models and federated learning [4] 2. Text generation using the ChatGPT API. [6]

Appendix 3: Popular LLMs existing in the market

	Meta AI		Open AI		Mistral AI
	Llama	**Llama 2**	**GPT3**	**GPT4**	**Mistral**
Introduction	• An auto-regressive language model that uses an optimized transformer architecture • Comes in a range of parameter sizes (**7B, 13B, and 70B …**)	• **Fine-tuned with 40% more data** than Llama • **Double the context** length of Llama • Involves **Reinforcement Learning from Human Deedback (RLHF)*** during training • Comes in a range of parameter sizes (**7B, 13B, and 70B …**)	• Uses a transformer architecture relies on self-attention mechanisms • **More than 175 billion parameters** (Significantly larger than previous LLM)	• **More parameters** than previous versions • **Includes more data** • **Longer memory** • Able to handle **multimodal** inputs (mixed text or image data)	• The only version being published is **Mistral-7b** (a pre-trained generative text model with 7.3 billion parameters) • **Different from** the competition: **open and community-focused** instead of proprietary
Use Case	• Typically include text summarization, language translation, and basic question-answering • Effective for applications where a balance between performance and computational efficiency is needed	• More enhanced capabilities, particularly in tasks requiring a deeper understanding of context and complex language structures • More suitable for applications that demand a higher level of language comprehension, such as more sophisticated content creation or complex language interpretation	• Versatile, with a wide range of applications from creating human-like text to coding: generating creative writing, business communication, technical writing, and even programming tasks. • - Ability to generate contextually relevant and coherent text	• Excels in more complex and nuanced language tasks like intricate content creation, complex problem-solving, and advanced conversational AI applications with greater accuracy and nuance • Valuable in scenarios where deep understanding and sophisticated responses are required	• Designed for specialized applications, often in specific industries like healthcare, finance, or legal • It can provide detailed analysis and generate industry-specific content

***Reinforcement learning from human feedback (RLHF):** incorporating human feedback into the learning process, allowing them to better align their outputs with human expectations and preferences.

Appendix 4: ROUGE Score

ROUGE	ROUGE (Recall-Oriented Understudy for Gisting Evaluation) is a commonly used evaluation metric in NLP to assess the quality of generated text. It compares a generated summary to one or more reference summaries.
Different Types	• ROUGE-1 evaluates the overlap of unigrams between the generated impressions and reference. • ROUGE-2 extends this assessment to bigrams, shedding light on the fluency and coherence of the generated text. • ROUGE-L quantifies the longest common subsequence, which is crucial in capturing the essence and contextual accuracy of the reports. • ROUGE-Lsum applies the ROUGE-L calculation method at the sentence level and then aggregates all the results for the final score.
Precision, Recall and F1-score in ROUGE Score	Taking ROUGE-1 as an example: • ROUGE-1 recall = Num word matches / Num words in reference • ROUGE-1 precision = Num word matches / Num words in generated summary • ROUGE-1 F1-score = 2 * (precision * recall) / (precision + recall)

Appendix 5: Findings

Performance Comparison

Alpaca-7b

Type	Low			Mid			High		
	Precision	Recall	F-measure	Precision	Recall	F-measure	Precision	Recall	F-measure
rouge1	0.1148	0.3644	0.1543	0.1358	0.4563	0.1744	0.1621	0.4885	0.1961
rouge2	0.0478	0.3843	0.0705	0.0603	0.2734	0.0877	0.0779	0.3229	0.1087
rougeL	0.0925	0.3083	0.1278	0.1076	0.3808	0.1441	0.1276	0.4371	0.1673
rougeLsum	0.0916	0.3057	0.1281	0.1076	0.3837	0.1447	0.1284	0.4378	0.1675

Medalpaca7b

Type	Low			Mid			High		
	Precision	Recall	F-measure	Precision	Recall	F-measure	Precision	Recall	F-measure
rouge1	0.3543	0.4426	0.3628	0.4198	0.5047	0.4154	0.4829	0.5648	0.4611
rouge2	0.1995	0.2450	0.2008	0.2491	0.3102	0.2530	0.3000	0.3781	0.3046
rougeL	0.2814	0.3595	0.2929	0.3343	0.4210	0.3391	0.3845	0.4850	0.3846
rougeLsum	0.2868	0.3602	0.2942	0.3332	0.4217	0.3377	0.3829	0.4842	0.3859

Appendix 5: Findings (cont.)

Performance Comparison

Llama-7b

Type	Low			Mid			High		
	Precision	Recall	F-measure	Precision	Recall	F-measure	Precision	Recall	F-measure
rouge1	0.4716	0.4319	0.4239	0.5084	0.4675	0.4537	0.5455	0.5071	0.4858
rouge2	0.2615	0.2442	0.2358	0.2948	0.2802	0.2665	0.3313	0.3197	0.3022
rougeL	0.3823	0.3585	0.3455	0.4132	0.3907	0.3734	0.4448	0.4270	0.4026
rougeLsum	0.3803	0.3563	0.3437	0.4125	0.3916	0.3739	0.4489	0.4275	0.4051

Llama-13b

Type	Low			Mid			High		
	Precision	Recall	F-measure	Precision	Recall	F-measure	Precision	Recall	F-measure
rouge1	0.4239	0.3890	0.3816	0.4913	0.4498	0.4345	0.5589	0.5210	0.4961
rouge2	0.2252	0.2095	0.2031	0.2832	0.2654	0.2558	0.3576	0.3372	0.3228
rougeL	0.3460	0.3238	0.3098	0.4064	0.3853	0.3659	0.4697	0.4495	0.4259
rougeLsum	0.3412	0.3224	0.3098	0.4068	0.3847	0.3656	0.4711	0.4556	0.4286

Appendix 5: Findings (cont.)

Performance Comparison

Mistral 7b

Type	Low			Mid			High		
	Precision	Recall	F-measure	Precision	Recall	F-measure	Precision	Recall	F-measure
rouge1	0.6239	0.4949	0.5238	0.6616	0.5642	0.5593	0.6980	0.5746	0.5950
rouge2	0.4192	0.3404	0.3569	0.4626	0.4011	0.3986	0.5098	0.4308	0.4467
rougeL	0.5301	0.4227	0.4453	0.5695	0.4866	0.4841	0.6101	0.5109	0.5315
rougeLsum	0.5303	0.4245	0.4461	0.5700	0.4622	0.4837	0.6092	0.5031	0.5249

Appendix 6: Factual Correctness

Factual Correctness refers to the ability of these models to generate responses or information that is accurate and true to reality.

Background: radiographic examination of the chest. clinical history: 80 years of age, male ... **Findings**: frontal radiograph of the chest demonstrates repositioning of the right atrial lead possibly into the ivc. ... a right apical pneumothorax can be seen from the image. moderate right and small left pleural effusions continue. no pulmonary edema is observed. heart size is upper limits of normal.
Human Summary: pneumothorax is seen. bilateral pleural effusions continue.
Summary A (ROUGE-L = 0.77): no pneumothorax is observed. bilateral pleural effusions continue.
Summary B (ROUGE-L = 0.44): pneumothorax is observed on radiograph. bilateral pleural effusions continue to be seen.

Compared to the human summary:

- Summary A has high textual overlap (i.e., ROUGE-L) but makes a factual error
- Summary B has a lower ROUGE-L score but is factually correct

The factual correctness result of our final Mistral-7b model:

- The percentage of negative items is: 13.4%
- The percentage of neutral items is: 57.0%
- The percentage of positive items is: 27.6%

Appendix 7: Challenges & methods to overcome

 Generalizability	1.Train on University of Chicago Medicine data and test on Indiana University data
 Quantitative and Qualitative Validation	2.Validation with both statistical language metrics and physician logs
 High Precision	3.Fine-tune for some physician personas and predictions adjusted with patients' clinical history
 Personalization	4.Develop Proprietary embeddings for different institutions and second attention mechanism for medical language processing
 Seamless Deployment	5.Adaptation to various deployment options including cloud and on-prem

Appendix 8: Data bias and ethics discussion

Bias	Definition	In our project:
Historical Bias	Arises even if data is perfectly measured and sampled, reflecting existing societal stereotypes in datasets	The risk of perpetuating historical stereotypes or inaccuracies present in medical data.
Representation Bias	Occurs when the development sample under-represents part of the population, leading to failures in generalization for certain subsets	Occurs when the dataset does not adequately represent the diversity of the patient population
Aggregation Bias	A "one-size-fits-all" approach that overlooks underlying groups or types in the data, potentially leading to a model that is suboptimal for any group or biased towards the dominant population	The challenge of creating models that are applicable to varied medical scenarios and patient groups without losing specificity
Evaluation Bias	Concerns the quality of the learned function and how it is measured. Addressing this may involve redefining the function that computes evaluation metrics and adjusting the data on which metrics are computed	Risk that the metrics used do not accurately reflect model's efficacy in real-world medical settings. (e.g. Rouge Score vs.. Factual Correctness)

Solution

- Ensure diversity and completeness in data
- Implement model validation on different datasets
- Engage with healthcare professionals and potentially patients
- Continuously monitor and update the model