Optimizing clinical workflows using language models

Guest Lecture
Austin Community College (ACC)

03 Apr 2025

Asad Aali
Research Scientist
Stanford University
asadaali@stanford.edu

Stanford University

About Me

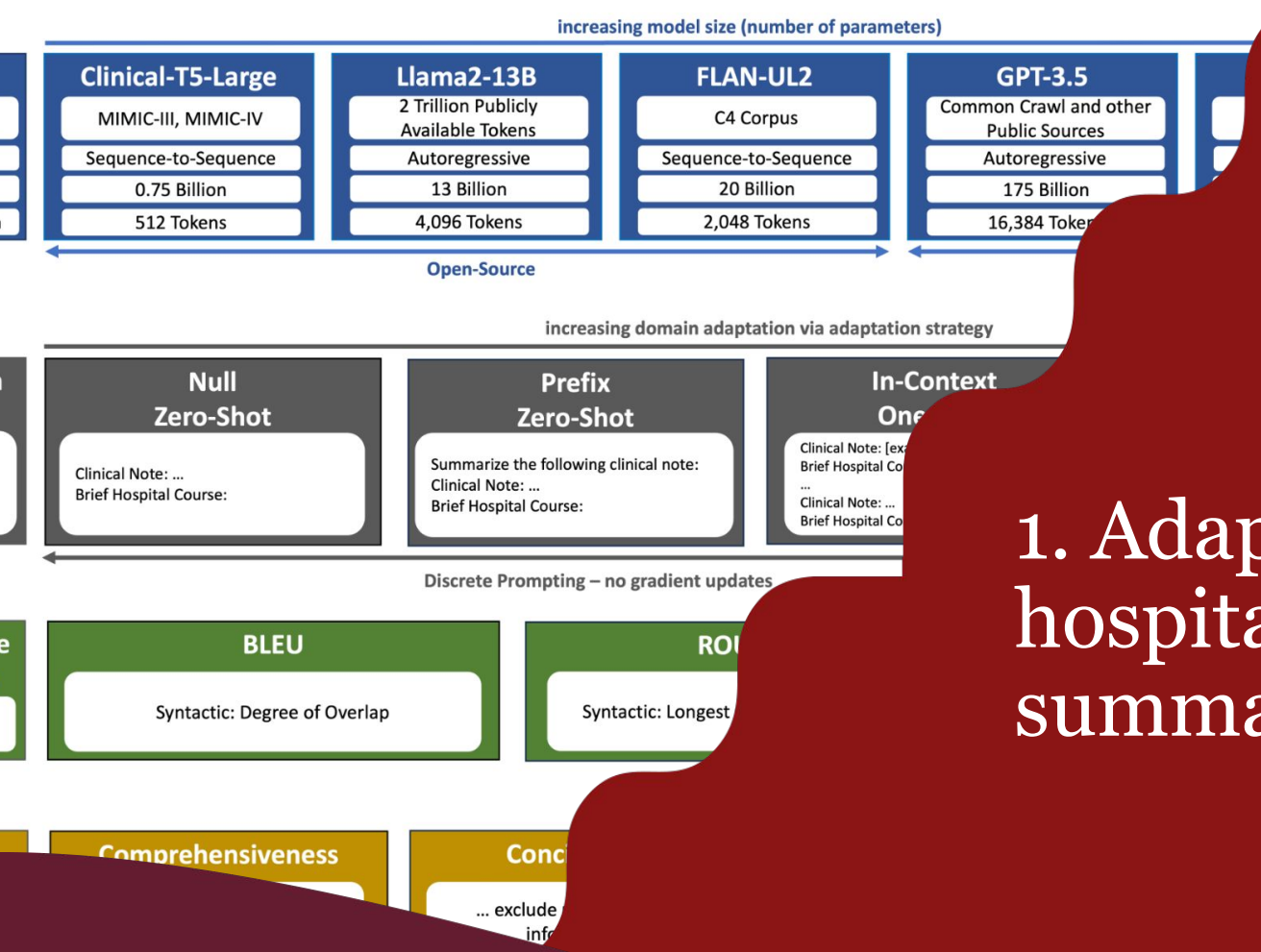
- Research Scientist at Stanford University
  - Lab: Machine Intelligence for Medical Imaging (MIMI)
  - Advisor: Akshay Chaudhari
- Passionate about developing machine learning algorithms for healthcare applications
- Research Interests:
  - Machine Learning
  - Foundation Models
  - Healthcare

Plan for Today

1. Adapting language models (LMs) for hospital discharge summarization

2. Scalable optimization of LMs for healthcare tasks

Stanford University



1. Adapting LMs for hospital discharge summarization

one medical
amazon

Motivation

1. Health Care providers at One Medical need to manually look through hundreds of clinical documents

2. Surfacing the most relevant clinical data can be accomplished with text summarization

3. This can allow for better **health outcomes** as it helps providers:
   a. Save valuable **time**
   b. Build a **deeper connection** with patients

Published in JAMIA

MIMIC-IV-BHC - Sample

Table 1. a) A sample of our novel pre-processed clinical notes dataset, extracted from raw MIMIC-IV notes.

Input	Example
SEX	F
SERVICE	SURGERY
ALLERGIES	No Known Allergies
CHIEF COMPLAINT	Splenic laceration
MAJOR PROCEDURE	NONE
HISTORY OF PRESENT ILLNESS	s/p routine colonoscopy this morning with polypectomy (report not available) ...
PAST MEDICAL HISTORY	Mild asthma, hypothyroid
FAMILY HISTORY	Non-contributory
PHYSICAL EXAM	Gen: Awake and alert CV: RRR Lungs: CTAB Abd: Soft, nontender, nondistended
PERTINENT RESULTS	03:45 PM BLOOD WBC-5.5 RBC-3.95 Hgb-14.1 ...
MEDICATIONS ON ADMISSION	1. Levothyroxine Sodium 100 mcg PO DAILY 2. Flovent HFA (fluticasone) ...
DISCHARGE DISPOSITION	Home
DISCHARGE DIAGNOSIS	Splenic laceration
DISCHARGE CONDITION	Mental Status: Clear and coherent. Level of Consciousness: Alert and interactive ...
DISCHARGE INSTRUCTIONS	You were admitted to ... in the intensive care unit for monitoring after a ...

MIMIC-IV-BHC - Sample

Table 1. a) A sample of our novel pre-processed clinical notes dataset, extracted from raw MIMIC-IV notes.

Input	**Example**
SEX	F
SERVICE	SURGERY
ALLERGIES	No Known Allergies
CHIEF COMPLAINT	Splenic laceration
MAJOR PROCEDURE	NONE
HISTORY OF PRESENT ILLNESS	s/p routine colonoscopy this morning with polypectomy (report not available) ...
PAST MEDICAL HISTORY	Mild asthma, hypothyroid
FAMILY HISTORY	Non-contributory
PHYSICAL EXAM	Gen: Awake and alert CV: RRR Lungs: CTAB Abd: Soft, nontender, nondistended
PERTINENT RESULTS	03:45 PM BLOOD WBC-5.5 RBC-3.95 Hgb-14.1 ...
MEDICATIONS ON ADMISSION	1. Levothyroxine Sodium 100 mcg PO DAILY 2. Flovent HFA (fluticasone) ...
DISCHARGE DISPOSITION	Home
DISCHARGE DIAGNOSIS	Splenic laceration
DISCHARGE CONDITION	Mental Status: Clear and coherent. Level of Consciousness: Alert and interactive ...
DISCHARGE INSTRUCTIONS	You were admitted to ... in the intensive care unit for monitoring after a ...

Output	**Example**
BRIEF HOSPITAL COURSE	Ms. ... was admitted to ... on .... After getting a colonoscopy and polypectomy, she ...

MIMIC-IV-Ext-BHC: Labeled Clinical Notes Dataset for Hospital Course Summarization

Asad Aali, Dave Van Veen, Yamin Arefeen, Jason Hom, Christian Bluethgen, Eduardo Pontes Reis, Sergios Gatidis, Namuun Clifford, Joseph Daws, Arash Tehrani, Jangwon Kim, Akshay Chaudhari

1. A curated collection of **preprocessed and labeled clinical notes** derived from the MIMIC-IV-Note database.

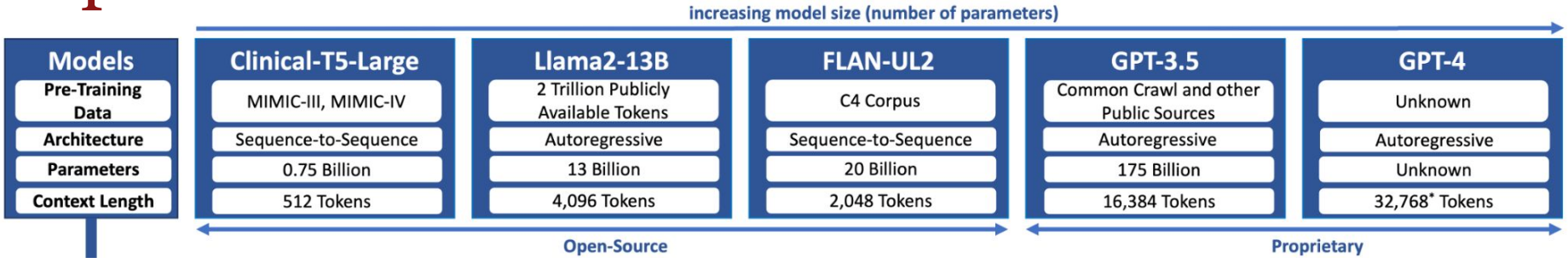
2. To facilitate development and **training of machine learning** models focused on summarizing brief hospital courses (BHC)

3. **270,033 meticulously cleaned and standardized** clinical notes containing an average token length of 2,267

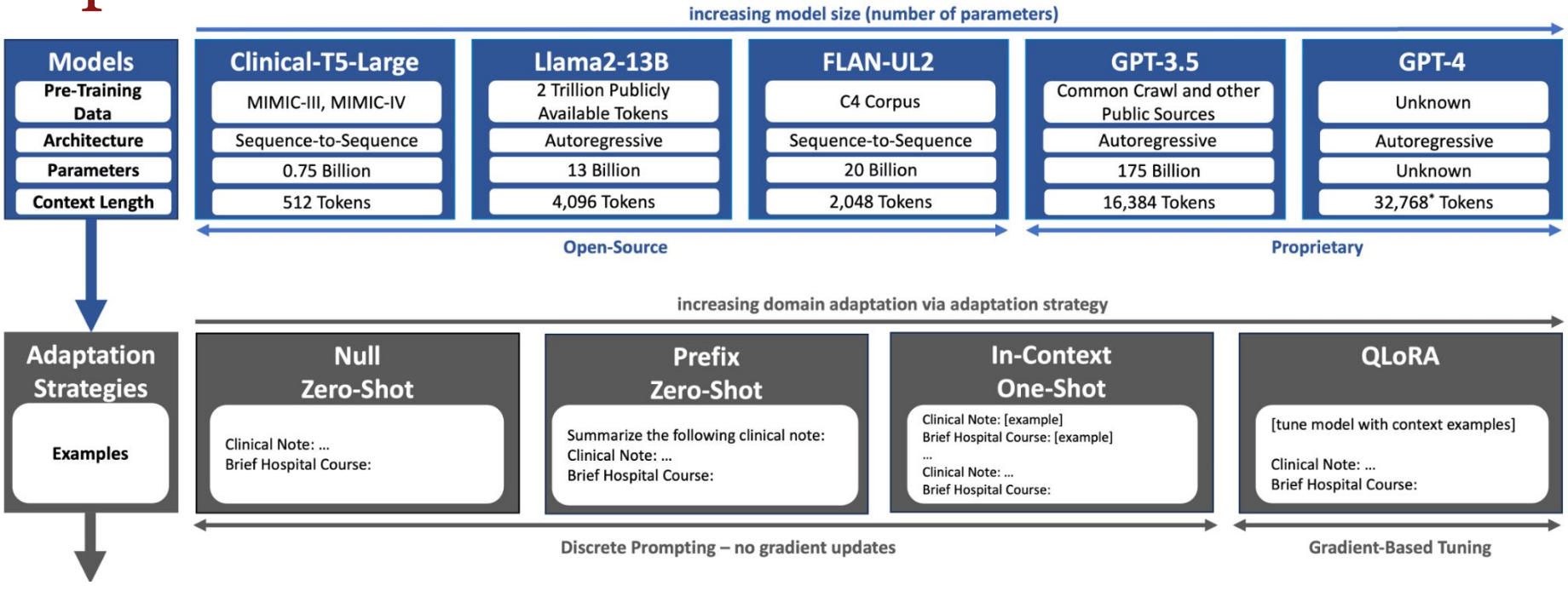
4. Preprocessing pipeline employed uses **regular expressions** to address common issues in the raw clinical text

Published on PhysioNet

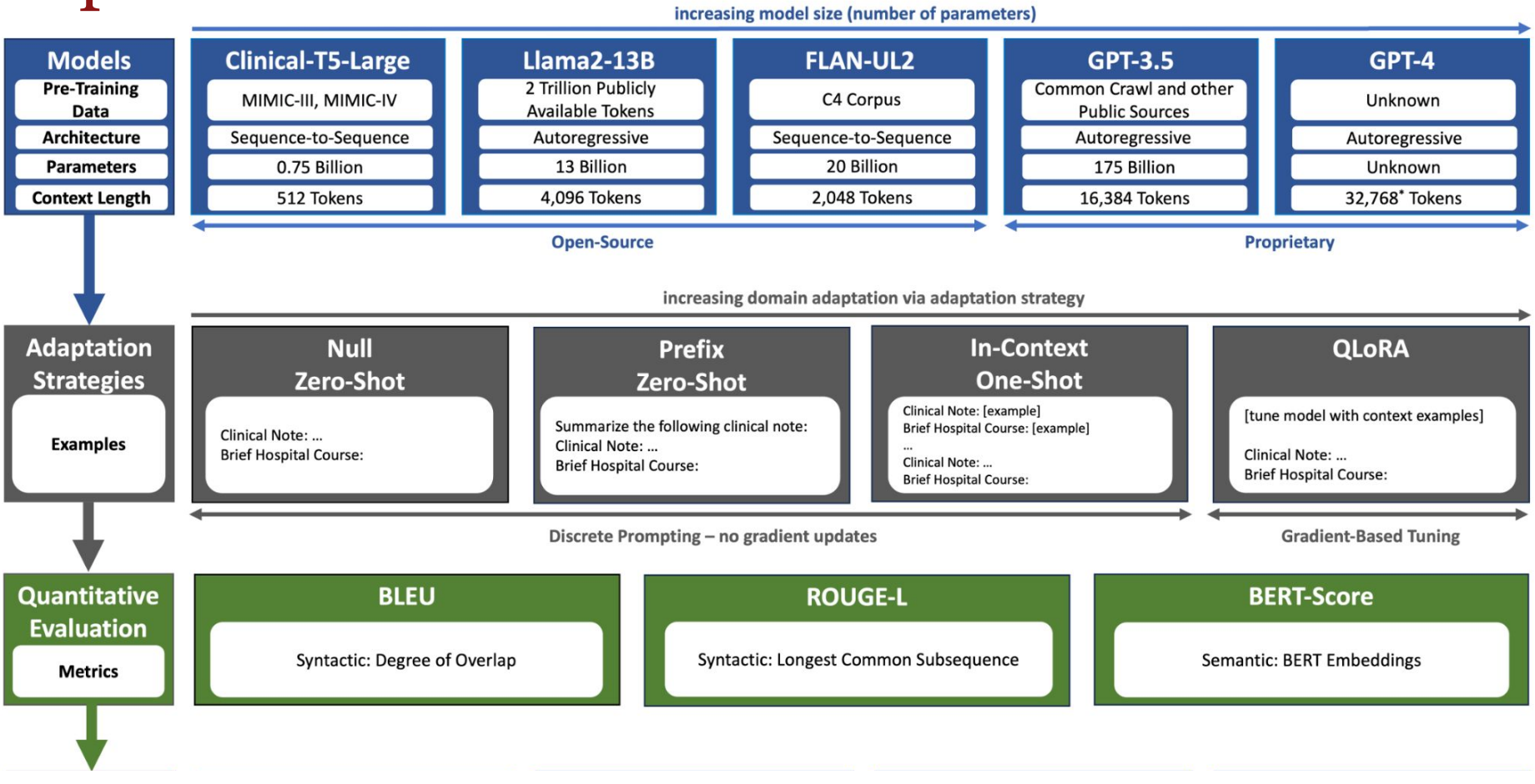
Pipeline


increasing model size (number of parameters)
Models
Pre-Training Data
Architecture
Parameters
Context Length
Clinical-T5-Large
MIMIC-III, MIMIC-IV
Sequence-to-Sequence
0.75 Billion
512 Tokens
Llama2-13B
2 Trillion Publicly Available Tokens
Autoregressive
13 Billion
4,096 Tokens
FLAN-UL2
C4 Corpus
Sequence-to-Sequence
20 Billion
2,048 Tokens
GPT-3.5
Common Crawl and other Public Sources
Autoregressive
175 Billion
16,384 Tokens
GPT-4
Unknown
Autoregressive
Unknown
32,768* Tokens
Open-Source
Proprietary

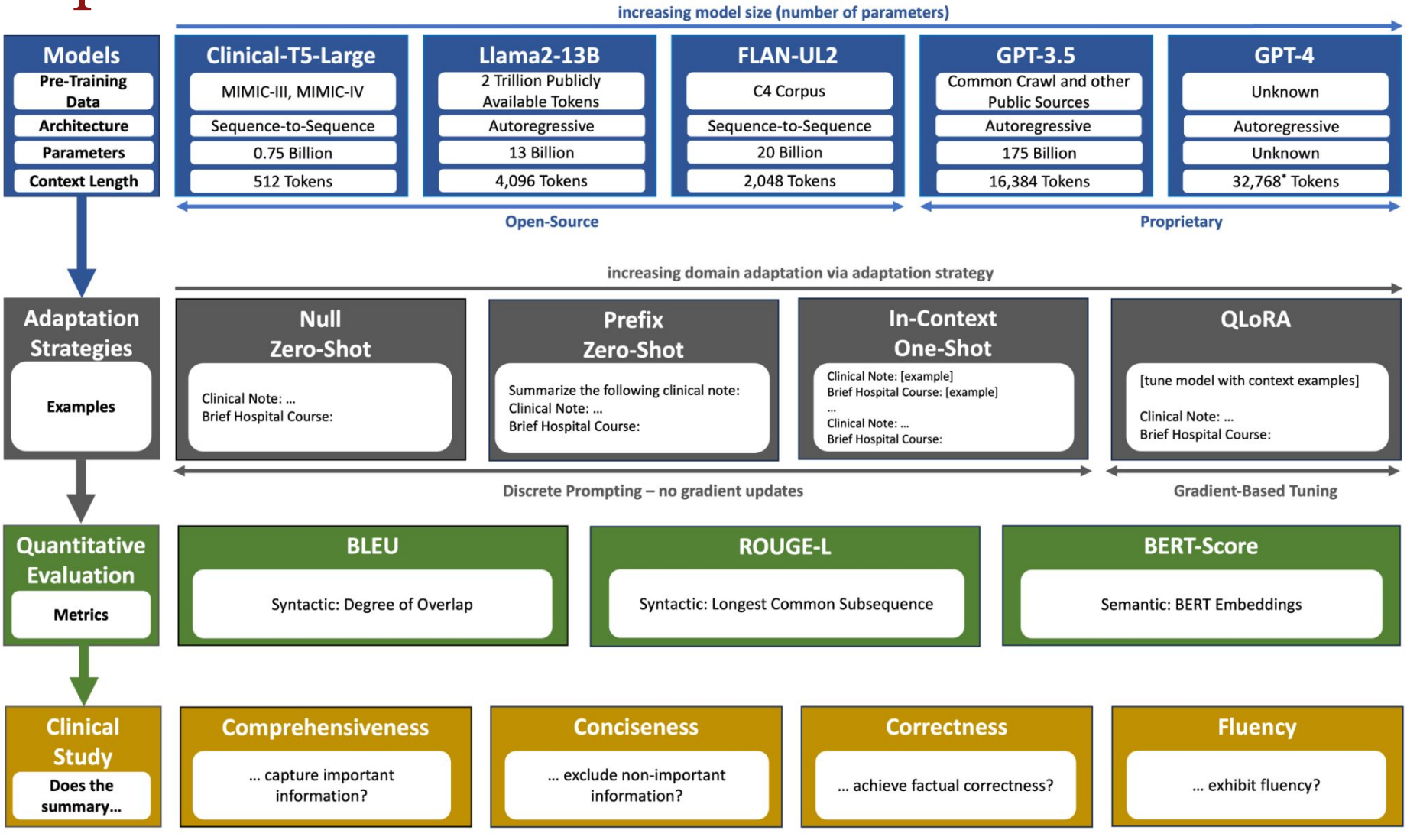
Pipeline


increasing model size (number of parameters)
Models
Pre-Training Data
Architecture
Parameters
Context Length
Clinical-T5-Large
MIMIC-III, MIMIC-IV
Sequence-to-Sequence
0.75 Billion
512 Tokens
Llama2-13B
2 Trillion Publicly Available Tokens
Autoregressive
13 Billion
4,096 Tokens
FLAN-UL2
C4 Corpus
Sequence-to-Sequence
20 Billion
2,048 Tokens
GPT-3.5
Common Crawl and other Public Sources
Autoregressive
175 Billion
16,384 Tokens
GPT-4
Unknown
Autoregressive
Unknown
32,768* Tokens
Open-Source
Proprietary
increasing domain adaptation via adaptation strategy
Adaptation Strategies
Examples
Null Zero-Shot
Clinical Note: ...
Brief Hospital Course:
Prefix Zero-Shot
Summarize the following clinical note:
Clinical Note: ...
Brief Hospital Course:
In-Context One-Shot
Clinical Note: [example]
Brief Hospital Course: [example]
...
Clinical Note: ...
Brief Hospital Course:
QLoRA
[tune model with context examples]
Clinical Note: ...
Brief Hospital Course:
Discrete Prompting – no gradient updates
Gradient-Based Tuning

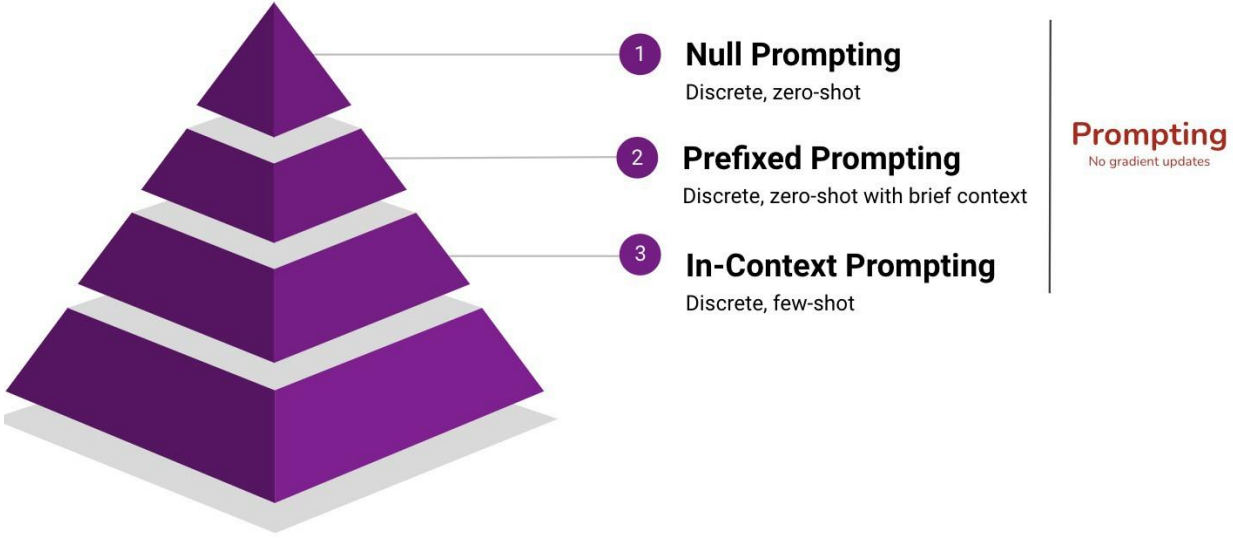
Pipeline


increasing model size (number of parameters)
Models
Pre-Training Data
Architecture
Parameters
Context Length
Clinical-T5-Large
MIMIC-III, MIMIC-IV
Sequence-to-Sequence
0.75 Billion
512 Tokens
Llama2-13B
2 Trillion Publicly Available Tokens
Autoregressive
13 Billion
4,096 Tokens
FLAN-UL2
C4 Corpus
Sequence-to-Sequence
20 Billion
2,048 Tokens
GPT-3.5
Common Crawl and other Public Sources
Autoregressive
175 Billion
16,384 Tokens
GPT-4
Unknown
Autoregressive
Unknown
32,768* Tokens
Open-Source
Proprietary
increasing domain adaptation via adaptation strategy
Adaptation Strategies
Examples
Null Zero-Shot
Clinical Note: ...
Brief Hospital Course:
Prefix Zero-Shot
Summarize the following clinical note:
Clinical Note: ...
Brief Hospital Course:
In-Context One-Shot
Clinical Note: [example]
Brief Hospital Course: [example]
...
Clinical Note: ...
Brief Hospital Course:
QLoRA
[tune model with context examples]
Clinical Note: ...
Brief Hospital Course:
Discrete Prompting – no gradient updates
Gradient-Based Tuning
Quantitative Evaluation
Metrics
BLEU
Syntactic: Degree of Overlap
ROUGE-L
Syntactic: Longest Common Subsequence
BERT-Score
Semantic: BERT Embeddings

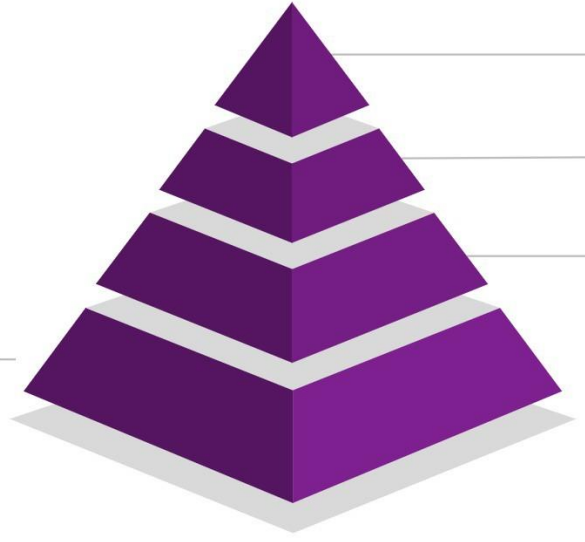
Pipeline


increasing model size (number of parameters)
Models
Pre-Training Data
Architecture
Parameters
Context Length
Clinical-T5-Large
MIMIC-III, MIMIC-IV
Sequence-to-Sequence
0.75 Billion
512 Tokens
Llama2-13B
2 Trillion Publicly Available Tokens
Autoregressive
13 Billion
4,096 Tokens
FLAN-UL2
C4 Corpus
Sequence-to-Sequence
20 Billion
2,048 Tokens
GPT-3.5
Common Crawl and other Public Sources
Autoregressive
175 Billion
16,384 Tokens
GPT-4
Unknown
Autoregressive
Unknown
32,768* Tokens
Open-Source
Proprietary
increasing domain adaptation via adaptation strategy
Adaptation Strategies
Examples
Null Zero-Shot
Clinical Note: ...
Brief Hospital Course:
Prefix Zero-Shot
Summarize the following clinical note:
Clinical Note: ...
Brief Hospital Course:
In-Context One-Shot
Clinical Note: [example]
Brief Hospital Course: [example]
...
Clinical Note: ...
Brief Hospital Course:
QLoRA
[tune model with context examples]
Clinical Note: ...
Brief Hospital Course:
Discrete Prompting – no gradient updates
Gradient-Based Tuning
Quantitative Evaluation
Metrics
BLEU
Syntactic: Degree of Overlap
ROUGE-L
Syntactic: Longest Common Subsequence
BERT-Score
Semantic: BERT Embeddings
Clinical Study
Does the summary...
Comprehensiveness
... capture important information?
Conciseness
... exclude non-important information?
Correctness
... achieve factual correctness?
Fluency
... exhibit fluency?

Overview of Adaptation Methods



Overview of Adaptation Methods



1 **Null Prompting**
Discrete, zero-shot

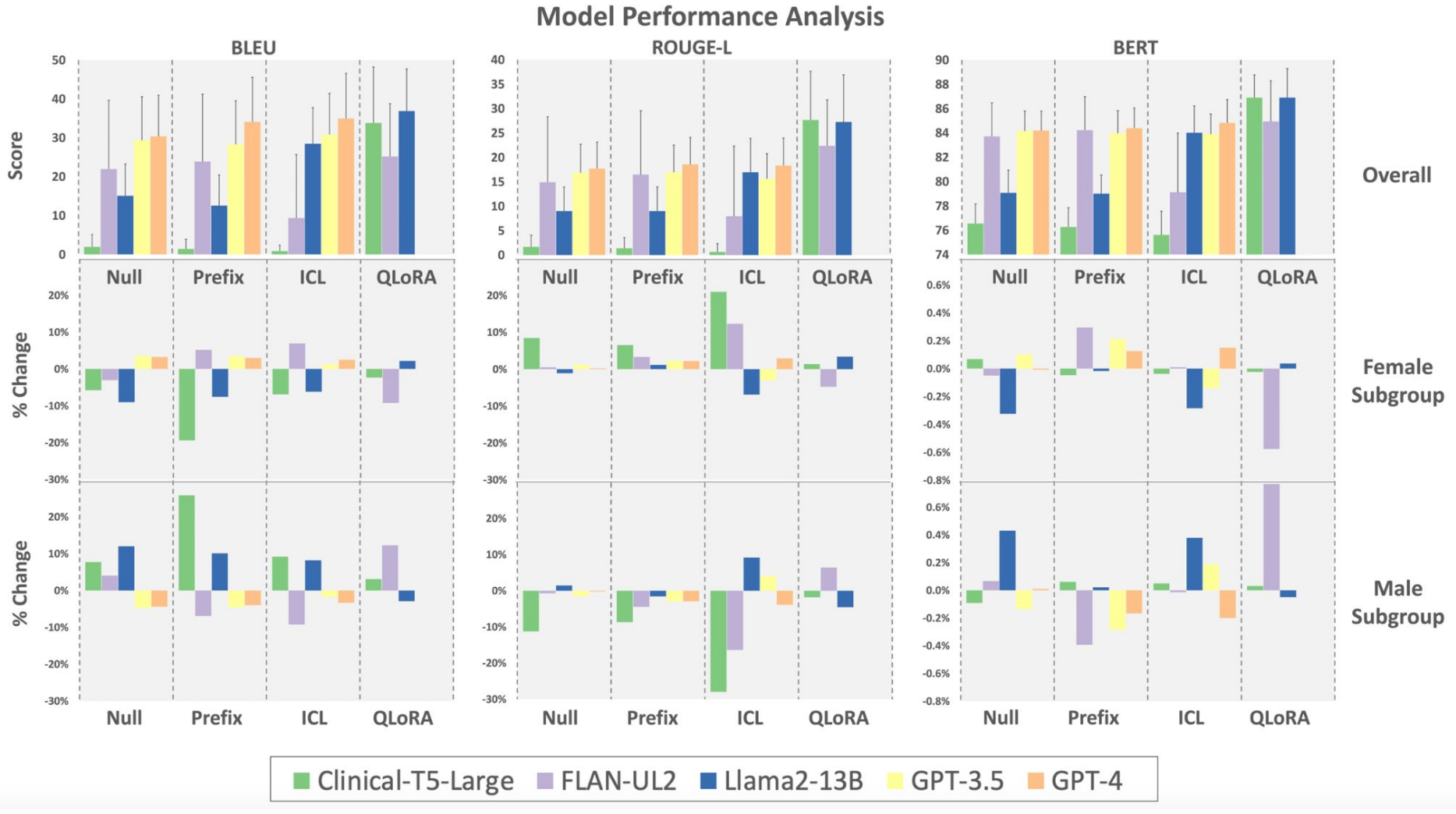
2 **Prefixed Prompting**
Discrete, zero-shot with brief context

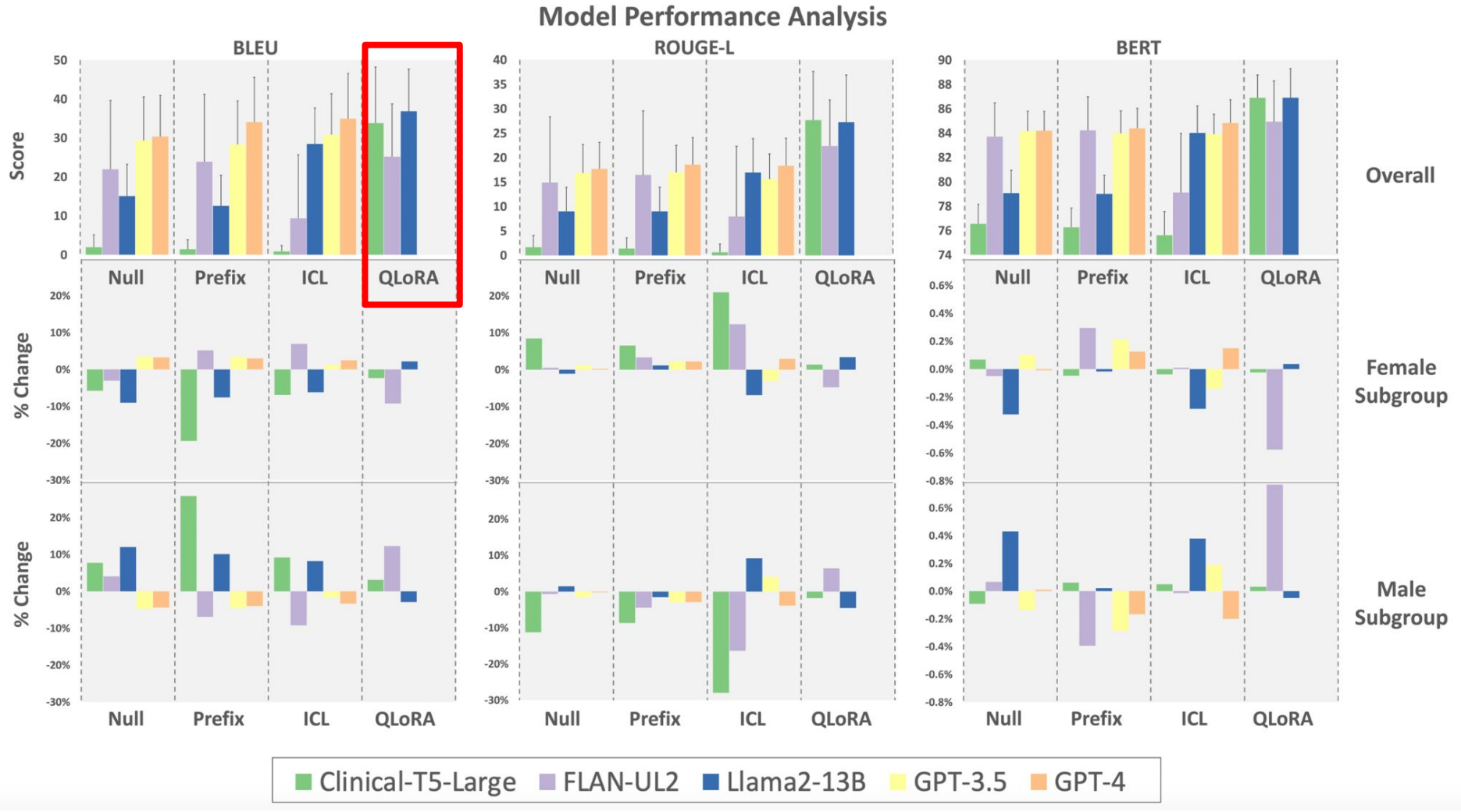
3 **In-Context Prompting**
Discrete, few-shot

Prompting
No gradient updates

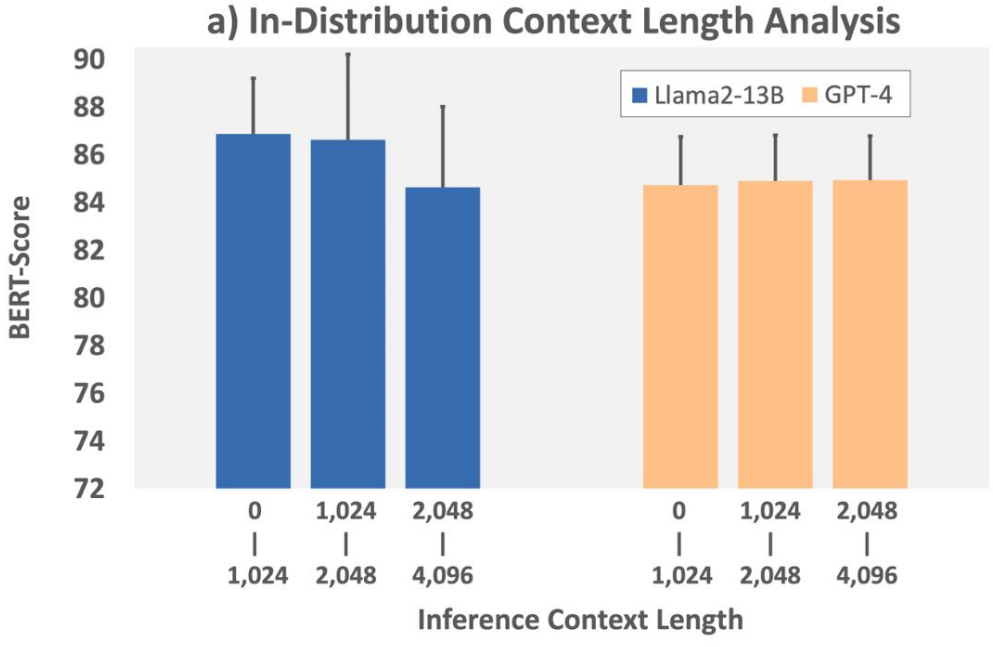
4 **LoRA**
Low-rank adaptation, approximates fine-tuning process by injecting rank decomposition matrices

Tuning
Inexpensive gradient-based updates

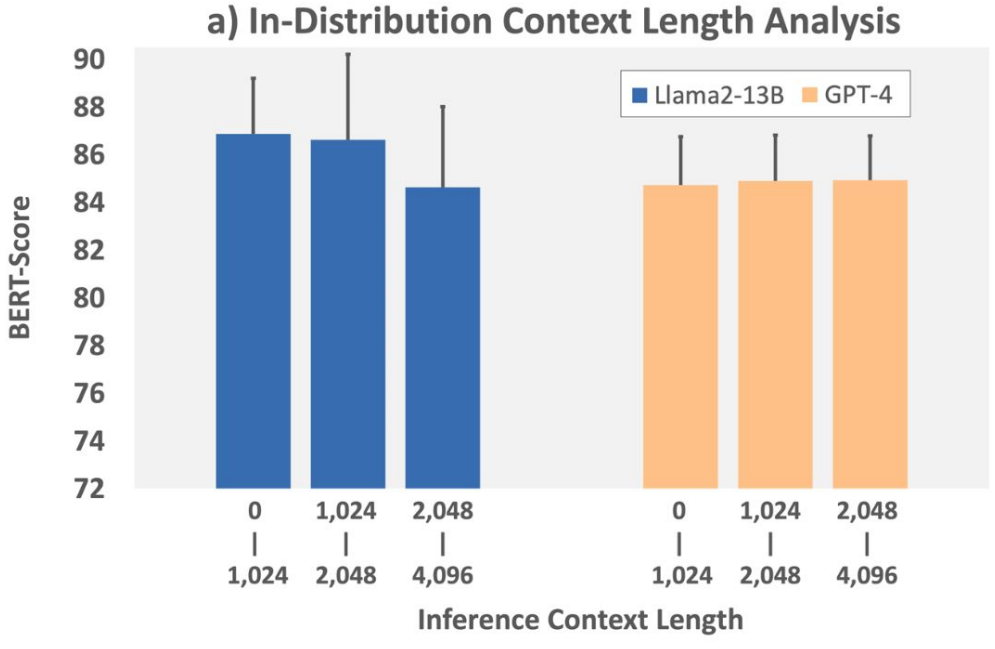

Model Performance Analysis
BLEU
ROUGE-L
BERT
Score
% Change
% Change
Overall
Female Subgroup
Male Subgroup
Null
Prefix
ICL
QLoRA
Clinical-T5-Large
FLAN-UL2
Llama2-13B
GPT-3.5
GPT-4

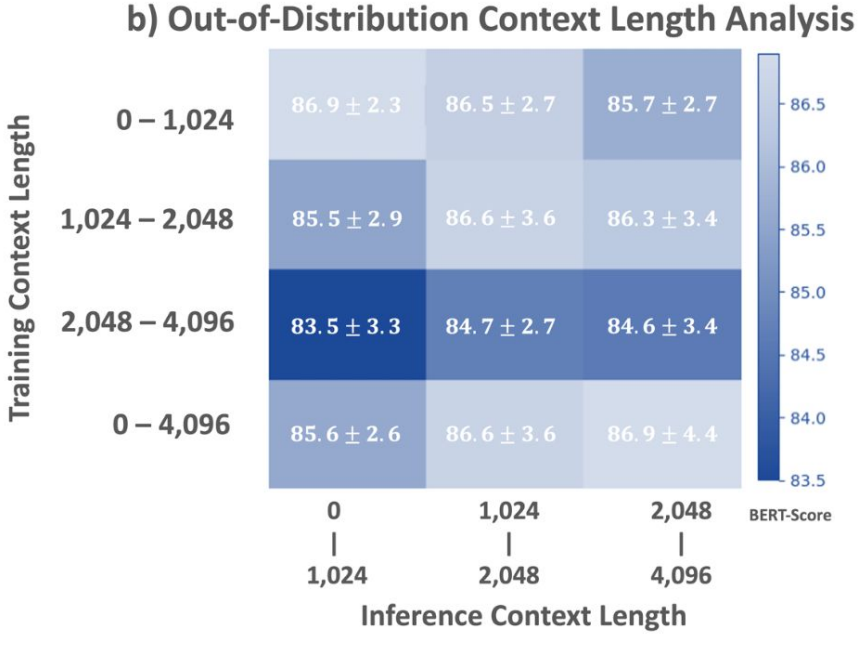

Model Performance Analysis
BLEU
ROUGE-L
BERT
Score
% Change
% Change
Overall
Female Subgroup
Male Subgroup
Null
Prefix
ICL
QLoRA
Clinical-T5-Large
FLAN-UL2
Llama2-13B
GPT-3.5
GPT-4

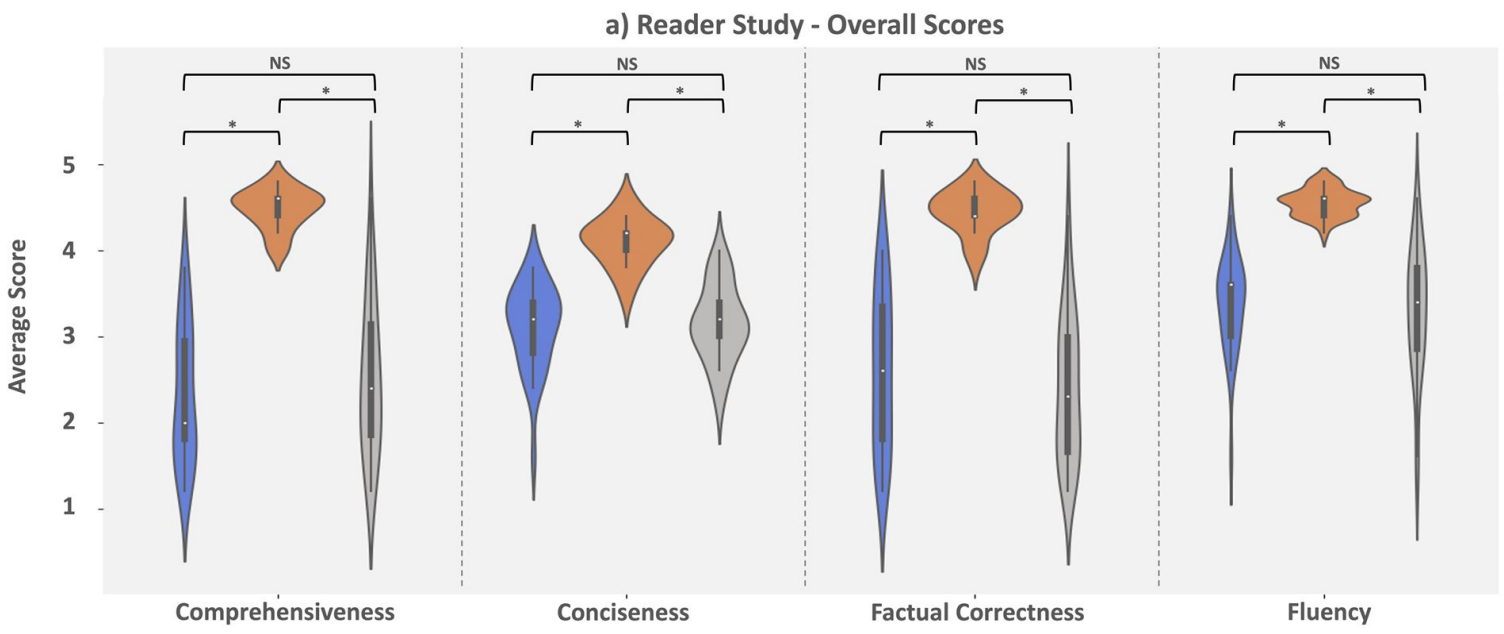
Context Length Analysis

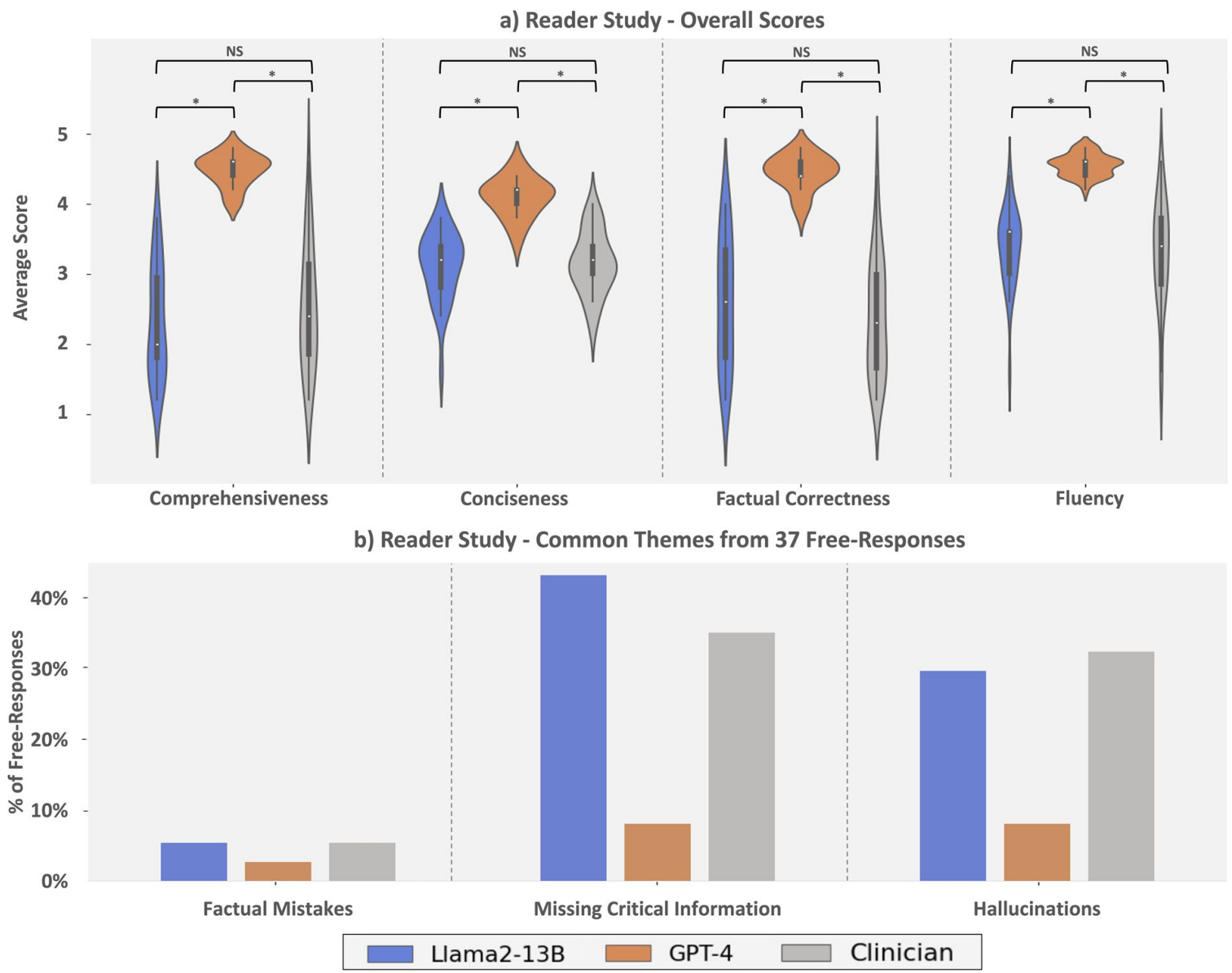


Context Length Analysis






a) Reader Study - Overall Scores
Average Score
5
4
3
2
1
NS
*
*
NS
*
*
NS
*
*
NS
*
*
Comprehensiveness
Conciseness
Factual Correctness
Fluency


a) Reader Study - Overall Scores
Average Score
5
4
3
2
1
NS
*
*
Comprehensiveness
Conciseness
Factual Correctness
Fluency
b) Reader Study - Common Themes from 37 Free-Responses
% of Free-Responses
40%
30%
20%
10%
0%
Factual Mistakes
Missing Critical Information
Hallucinations
Llama2-13B
GPT-4
Clinician

Summarization Example

Expertise	*You are an expert medical professional*
Instruction	*Summarize the clinical note into a brief hospital course*

Summarization Example

Expertise	*You are an expert medical professional*
Instruction	*Summarize the clinical note into a brief hospital course*
In-Context Example	*Use the examples to guide word choice* *input: {example clinical note}* *summary: {example bhc}*

Summarization Example

Expertise	*You are an expert medical professional*
Instruction	*Summarize the clinical note into a brief hospital course*
In-Context Example	*Use the examples to guide word choice* *input: {example clinical note}* *summary: {example bhc}*
Clinical Note Input	***SEX:*** *F* ***SERVICE:*** *OBSTETRICS/GYNECOLOGY* ***ALLERGIES:*** *No Known Allergies / Adverse Drug Reactions* ***ATTENDING:*** *___.* ***CHIEF COMPLAINT:*** *bleeding in pregnancy* ***MAJOR SURGICAL OR INVASIVE PROCEDURE:*** *None* ***HISTORY OF PRESENT ILLNESS:*** *___ G4P0 (h/o) TAB x 3 @ ___ admitted with vaginal bleeding that started 4 days prior.* ***PAST MEDICAL HISTORY:*** *abnormal pap smears anxiety depression warts colposcopy, LEEP ___ TAB x 3 marginal cord insert fibroadenoma of the breast* ***SOCIAL HISTORY:*** *___* ***FAMILY HISTORY:*** *noncontributory* ***PHYSICAL EXAM:*** *VS: 98.3, 109/69, 75, 20, O2 97% Gen: NAD Resp: No evidence of respiratory distress Abd: Soft, non-tender Ext: No lower extremity edema Date: ___ Time: 09: 00 FHT: 130s/mod variability/+accels/-decels (? quick deep variable x 1, assoc w/ loss of pickup); otherwise reactive Toco: rare ctx* ***PERTINENT RESULTS:*** *___ 05: 10PM WBC-9.3 RBC-4.24 HGB-13.7 HCT-39.3 MCV-93 MCH-32.3* MCHC-34.9 RDW-13.0 RDWSD-43.7 ___ 05: 10PM PLT COUNT-229 ___ 05: 10PM ___ PTT-28.6 ___ ___ 05: 10PM ___* ***MEDICATIONS ON ADMISSION:*** *PNV* ***DISCHARGE MEDICATIONS:*** *1. Citalopram 20 mg PO QHS 2. Prenatal Vitamins 1 TAB PO DAILY* ***DISCHARGE DISPOSITION:*** *Home* ***DISCHARGE DIAGNOSIS:*** *Marginal cord insertion Vaginal bleeding in pregnancy* ***DISCHARGE CONDITION:*** *Mental Status: Clear and coherent. Level of Consciousness: Alert and interactive. Activity Status: Ambulatory - Independent.* ***FOLLOWUP INSTRUCTIONS:*** *___* ***DISCHARGE INSTRUCTIONS:*** *Please continue pelvic rest. Avoid heavy lifting or strenuous activity. Otherwise normal activity.*

Summarization Example

Expertise	*You are an expert medical professional*
Instruction	*Summarize the clinical note into a brief hospital course*
In-Context Example	*Use the examples to guide word choice* *input: {example clinical note}* *summary: {example bhc}*
Clinical Note Input	***SEX:*** *F* ***SERVICE:*** *OBSTETRICS/GYNECOLOGY* ***ALLERGIES:*** *No Known Allergies / Adverse Drug Reactions* ***ATTENDING:*** *___.* ***CHIEF COMPLAINT:*** *bleeding in pregnancy* ***MAJOR SURGICAL OR INVASIVE PROCEDURE:*** *None* ***HISTORY OF PRESENT ILLNESS:*** *___ G4P0 (h/o) TAB x 3 @ ___ admitted with vaginal bleeding that started 4 days prior.* ***PAST MEDICAL HISTORY:*** *abnormal pap smears anxiety depression warts colposcopy, LEEP ___ TAB x 3 marginal cord insert fibroadenoma of the breast* ***SOCIAL HISTORY:*** *___* ***FAMILY HISTORY:*** *noncontributory* ***PHYSICAL EXAM:*** *VS: 98.3, 109/69, 75, 20, O2 97% Gen: NAD Resp: No evidence of respiratory distress Abd: Soft, non-tender Ext: No lower extremity edema Date: ___ Time: 09: 00 FHT: 130s/mod variability/+accels/-decels (? quick deep variable x 1, assoc w/ loss of pickup); otherwise reactive Toco: rare ctx* ***PERTINENT RESULTS:*** *___ 05: 10PM WBC-9.3 RBC-4.24 HGB-13.7 HCT-39.3 MCV-93 MCH-32.3* MCHC-34.9 RDW-13.0 RDWSD-43.7 ___ 05: 10PM PLT COUNT-229 ___ 05: 10PM ___ PTT-28.6 ___ ___ 05: 10PM ___* ***MEDICATIONS ON ADMISSION:*** *PNV* ***DISCHARGE MEDICATIONS:*** *1. Citalopram 20 mg PO QHS 2. Prenatal Vitamins 1 TAB PO DAILY* ***DISCHARGE DISPOSITION:*** *Home* ***DISCHARGE DIAGNOSIS:*** *Marginal cord insertion Vaginal bleeding in pregnancy* ***DISCHARGE CONDITION:*** *Mental Status: Clear and coherent. Level of Consciousness: Alert and interactive. Activity Status: Ambulatory - Independent.* ***FOLLOWUP INSTRUCTIONS:*** *___* ***DISCHARGE INSTRUCTIONS:*** *Please continue pelvic rest. Avoid heavy lifting or strenuous activity. Otherwise normal activity.*
Summary 1 Clinician	*Pt was Rh positive and admitted for vaginal bleeding. She had negative cultures (GBS and yeast, GC/CT/trich). She had spotting from admission until ___. She was monitored until ___. She was discharged home in a stable continue without any evidence of vaginal bleeding or spotting.*
Summary 2 GPT-4	*Ms. ___ was admitted under the care of Dr. ___ due to vaginal bleeding during her pregnancy. She has a history of abnormal pap smears, anxiety, depression, and fibroadenoma of the breast. No surgical procedures were performed during her stay. Her condition was monitored and managed with prenatal vitamins and Citalopram for her anxiety. She was discharged in a clear and coherent mental state, able to move independently. She was instructed to continue pelvic rest and avoid heavy lifting or strenuous activity. She will continue her care at home.*

Summarization Example

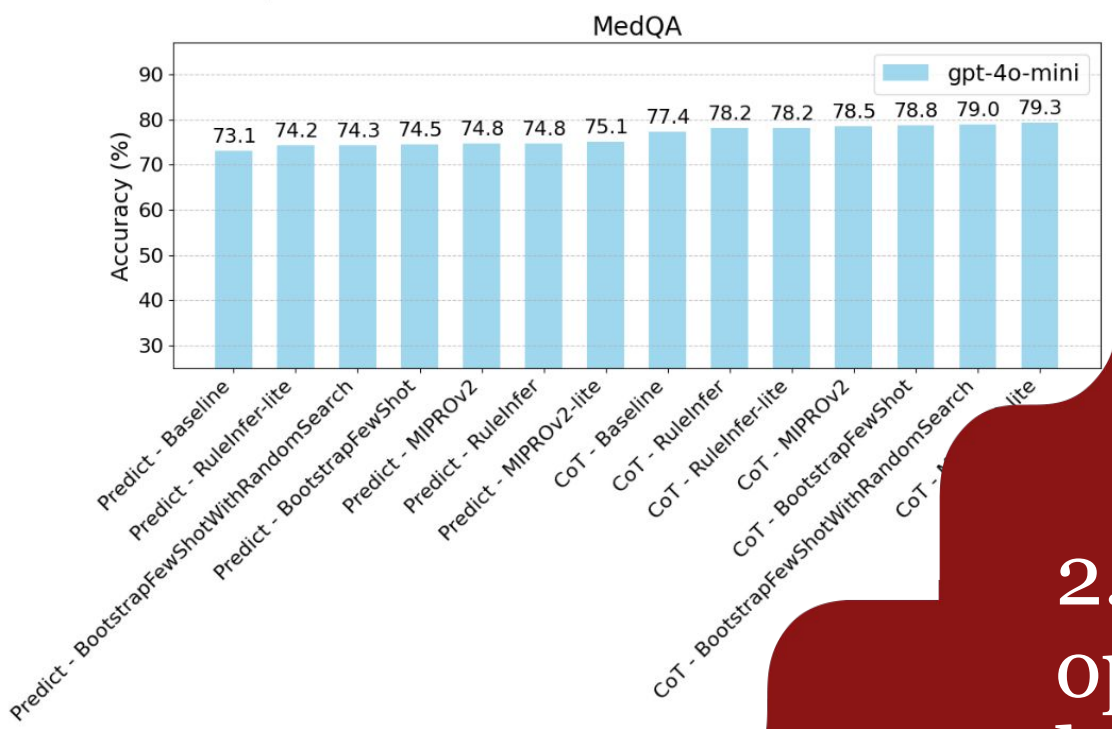
Expertise	*You are an expert medical professional*
Instruction	*Summarize the clinical note into a brief hospital course*
In-Context Example	*Use the examples to guide word choice* *input: {example clinical note}* *summary: {example bhc}*
Clinical Note Input	***SEX:*** *F* ***SERVICE:*** *OBSTETRICS/GYNECOLOGY* ***ALLERGIES:*** *No Known Allergies / Adverse Drug Reactions* ***ATTENDING:*** *___.* ***CHIEF COMPLAINT:*** *bleeding in pregnancy* ***MAJOR SURGICAL OR INVASIVE PROCEDURE:*** *None* ***HISTORY OF PRESENT ILLNESS:*** *___ G4P0 (h/o) TAB x 3 @ ___ admitted with vaginal bleeding that started 4 days prior.* ***PAST MEDICAL HISTORY:*** *abnormal pap smears anxiety depression warts colposcopy, LEEP ___ TAB x 3 marginal cord insert fibroadenoma of the breast* ***SOCIAL HISTORY:*** *___* ***FAMILY HISTORY:*** *noncontributory* ***PHYSICAL EXAM:*** *VS: 98.3, 109/69, 75, 20, O2 97% Gen: NAD Resp: No evidence of respiratory distress Abd: Soft, non-tender Ext: No lower extremity edema Date: ___ Time: 09: 00 FHT: 130s/mod variability/+accels/-decels (? quick deep variable x 1, assoc w/ loss of pickup); otherwise reactive Toco: rare ctx* ***PERTINENT RESULTS:*** *___ 05: 10PM WBC-9.3 RBC-4.24 HGB-13.7 HCT-39.3 MCV-93 MCH-32.3* MCHC-34.9 RDW-13.0 RDWSD-43.7 ___ 05: 10PM PLT COUNT-229 ___ 05: 10PM ___ PTT-28.6 ___ ___ 05: 10PM ___* ***MEDICATIONS ON ADMISSION:*** *PNV* ***DISCHARGE MEDICATIONS:*** *1. Citalopram 20 mg PO QHS 2. Prenatal Vitamins 1 TAB PO DAILY* ***DISCHARGE DISPOSITION:*** *Home* ***DISCHARGE DIAGNOSIS:*** *Marginal cord insertion Vaginal bleeding in pregnancy* ***DISCHARGE CONDITION:*** *Mental Status: Clear and coherent. Level of Consciousness: Alert and interactive. Activity Status: Ambulatory - Independent.* ***FOLLOWUP INSTRUCTIONS:*** *___* ***DISCHARGE INSTRUCTIONS:*** *Please continue pelvic rest. Avoid heavy lifting or strenuous activity. Otherwise normal activity.*
Summary 1 Clinician	*Pt was Rh positive and admitted for vaginal bleeding. She had negative cultures (GBS and yeast, GC/CT/trich). She had spotting from admission until ___. She was monitored until ___. She was discharged home in a stable continue without any evidence of vaginal bleeding or spotting.*
Summary 2 GPT-4	*Ms. ___ was admitted under the care of Dr. ___ due to vaginal bleeding during her pregnancy. She has a history of abnormal pap smears, anxiety, depression, and fibroadenoma of the breast. No surgical procedures were performed during her stay. Her condition was monitored and managed with prenatal vitamins and Citalopram for her anxiety. She was discharged in a clear and coherent mental state, able to move independently. She was instructed to continue pelvic rest and avoid heavy lifting or strenuous activity. She will continue her care at home.*
Reader Feedback	*Summary A (Clinician) contains multiple factual mistakes (serial ultrasounds, no evidence of vaginal bleeding, closed cervix, negative culture). It seems to contain information not at all present in the actual clinical note. Summary B (GPT-4) failed to mention a summary of the patients labs or vital signs, but otherwise looks great.*

Conclusions

1. Adapted **open-source models can match** the quality of clinician-written summaries

2. Adapted **proprietary models can outperform** the quality of clinician-written summaries

3. Adapted LLMs for summarization have the potential to:
   a. streamline documentation
   b. reduce errors
   c. enhance clinical workflows
   d. improve patient safety



2. Scalable optimization of language models for healthcare tasks

Motivation

- Language models (LMs) have significantly advanced **natural language processing capabilities**
- Their deployment is often **constrained** by the necessity for **extensive fine-tuning**
- Goal: **Enhance LM performance** through **refined prompt engineering** without necessitating model fine-tuning
- Plan: **Systematically evaluate** prompt engineering techniques across:
  - a variety of **healthcare tasks**
  - **open-source** and **closed-source LMs**

Introducing DSPy

- **D**eclarative **S**elf-improving **Py**thon

  https://dspy.ai/

- A framework for algorithmically optimizing prompts and LM weights

- **Developed by the Stanford NLP Group**

DSPy Optimizers

1. "give examples within prompt": BootstrapFewShot

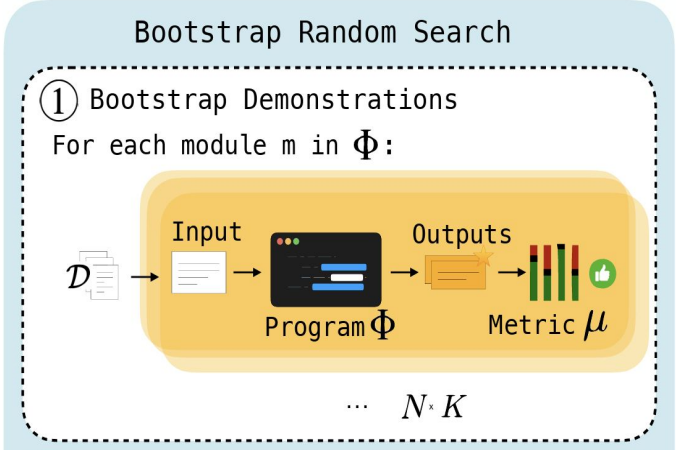
2. "tune the instructions": OPRO

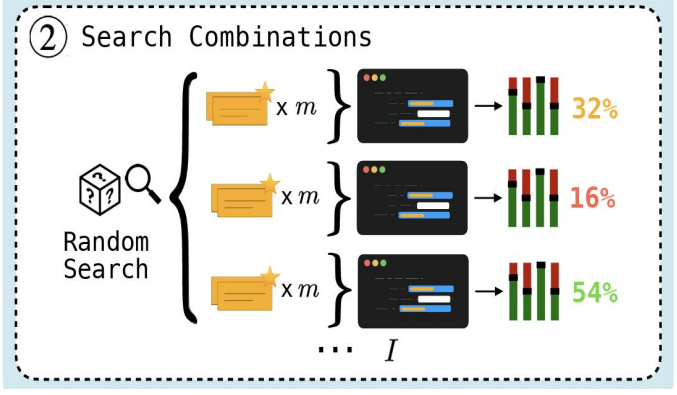
3. "tune the instructions + examples in prompt": MIPROv2

References:
1. Opsahl-Ong, Krista, et al. "Optimizing instructions and demonstrations for multi-stage language model programs." arXiv preprint arXiv:2406.11695 (2024).
2. Khattab, Omar, et al. "Dspy: Compiling declarative language model calls into self-improving pipelines." arXiv preprint arXiv:2310.03714 (2023).

"give examples within prompt": BootstrapFewShot

1. **Randomly sample** examples from your training set
2. If the output from the program is **correct** for this example, it is **kept** as a valid few-shot
3. Try more examples until curated **a specified amount** of few-shot example candidates.

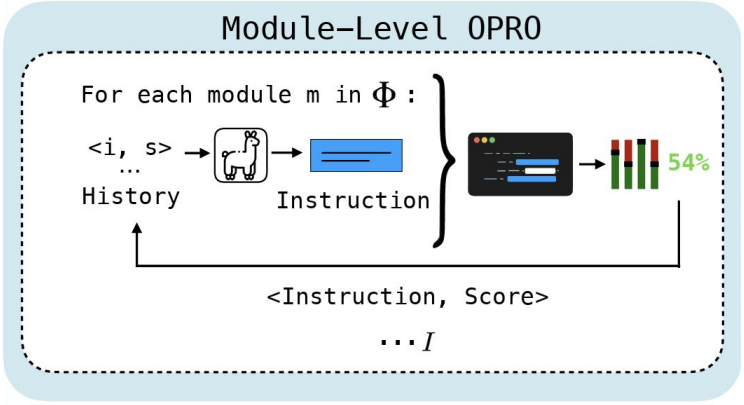




References:
1. Opsahl-Ong, Krista, et al. "Optimizing instructions and demonstrations for multi-stage language model programs." arXiv preprint arXiv:2406.11695 (2024).
2. Khattab, Omar, et al. "Dspy: Compiling declarative language model calls into self-improving pipelines." arXiv preprint arXiv:2310.03714 (2023).

"tune the instructions": OPRO

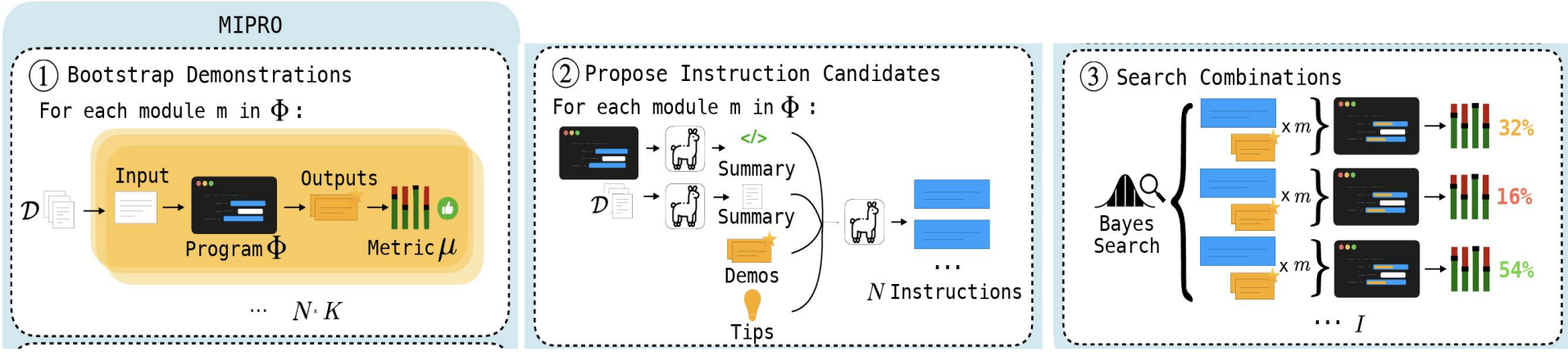
1. Propose **instruction candidates** for each predictor in the program.
   a. It bootstraps & summarizes relevant **information** about the task to propose instructions
2. The instruction proposer includes
   a. a **generated summary of properties** of the training dataset
   b. a **generated summary of your LM** program's code and the specific predictor
   c. previously **bootstrapped few-shot examples** to show reference inputs / outputs
   d. a randomly **sampled tip** for generation (i.e. "be creative", "be concise", etc.)



References:
1. Opsahl-Ong, Krista, et al. "Optimizing instructions and demonstrations for multi-stage language model programs." arXiv preprint arXiv:2406.11695 (2024).
2. Khattab, Omar, et al. "Dspy: Compiling declarative language model calls into self-improving pipelines." arXiv preprint arXiv:2310.03714 (2023).

"tune the instructions + examples in prompt": MIPROv2

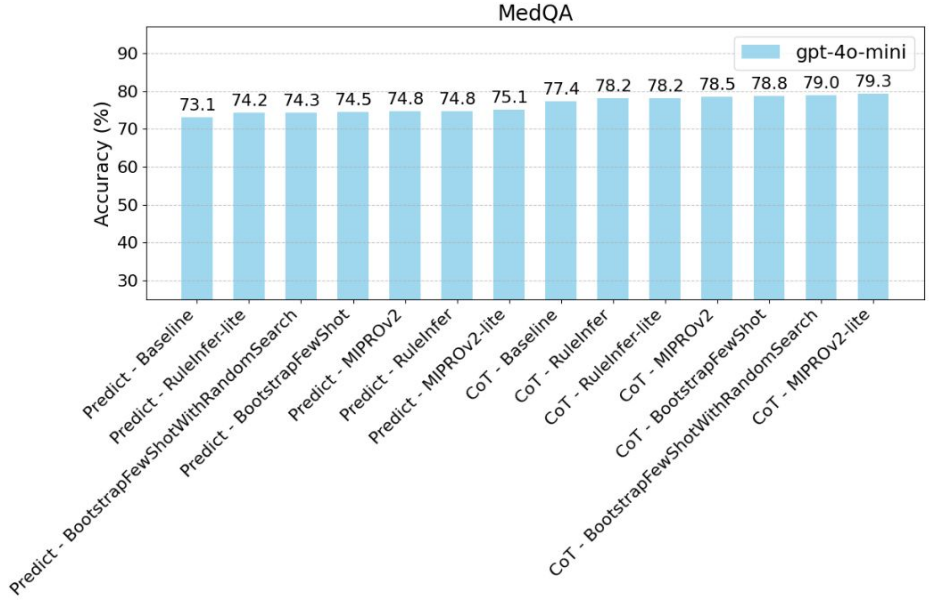
1. Bootstrap Few-Shot Examples: ***BootstrapFewShot***
2. Propose Instruction Candidates: ***OPRO***
3. Find an Optimized Combination of Few-Shot Examples & Instructions

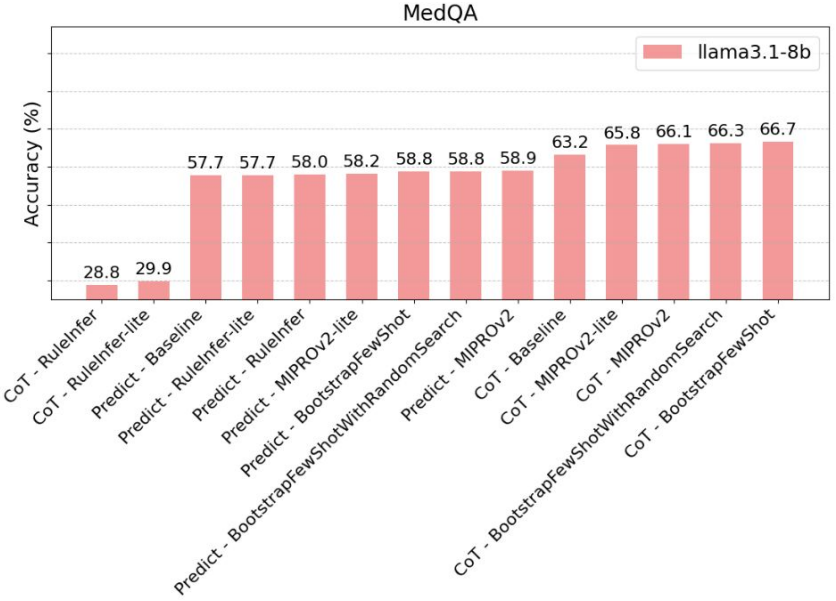


References:
1. Opsahl-Ong, Krista, et al. "Optimizing instructions and demonstrations for multi-stage language model programs." arXiv preprint arXiv:2406.11695 (2024).
2. Khattab, Omar, et al. "Dspy: Compiling declarative language model calls into self-improving pipelines." arXiv preprint arXiv:2310.03714 (2023).

Preliminary Results

- With DSPy optimizers, we were able to **exceed accuracy performance** on the **MedQA** public leaderboard:
  - gpt-4o-mini **72.4% -> 79.3%**
  - llama3.1-8b **62.6% -> 66.7%**





References:

1. MedQA Public Leaderboard

Thank You