The University of Texas at Austin
Cockrell School of Engineering

DEC 2023

CLINICAL TEXT SUMMARIZATION

A Benchmark of Domain-Adapted Large Language Models for Generating Brief Hospital Course Summaries

ASAD AALI
MS ECE Student
The University of Texas at Austin

Problem

- Health Care providers at One Medical need to **manually** look through **hundreds** of clinical documents

- Surfacing the most relevant clinical data from these documents can be accomplished with **text summarization**

- This allows for better **health outcomes** as it helps providers:
  - Save valuable **time**
  - Build a **deeper connection** with patients

MIMIC-IV-Notes

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

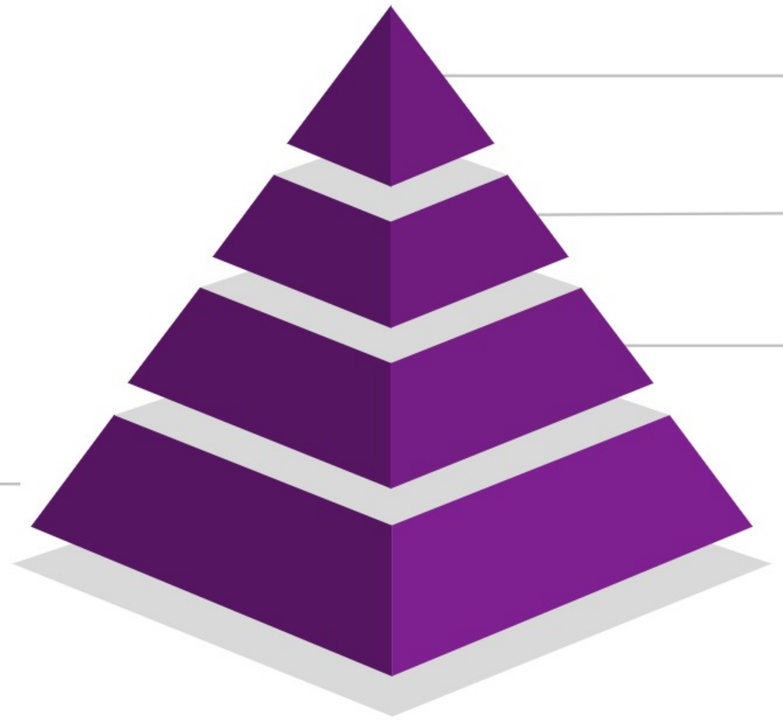
Input	Example
SEX	F
SERVICE	SURGERY
ALLERGIES	No Known Allergies
CHIEF COMPLAINT	Splenic laceration
MAJOR PROCEDURE	NONE
HISTORY OF PRESENT ILLNESS	s/p routine colonoscopy this morning with polypectomy (report not available) ...
PAST MEDICAL HISTORY	Mild asthma, hypothyroid
FAMILY HISTORY	Non-contributory
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 ...

b) Relevant statistics for the pre-processed dataset split across multiple context length ranges for adaptation tasks.

Context Range	Samples	Input Tokens	BHC Tokens
0 - 1,024	2,000	711 ± 199	104 ± 43
1,024 - 2,048	2,000	$1,471 \pm 275$	148 ± 36
2,048 - 4,096	2,000	$2,496 \pm 388$	225 ± 55

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

Evaluation Metrics

Human Evaluation

1. Radiologists
   a. Lost information
   b. Misinterpreted information
   c. Overall Score
2. Domain Experts
   a. Faithfulness
   b. Simplification
   c. Aggregation
   d. Coherence
   e. Relevance
   f. Factual Mistakes

Reference-Based[1]

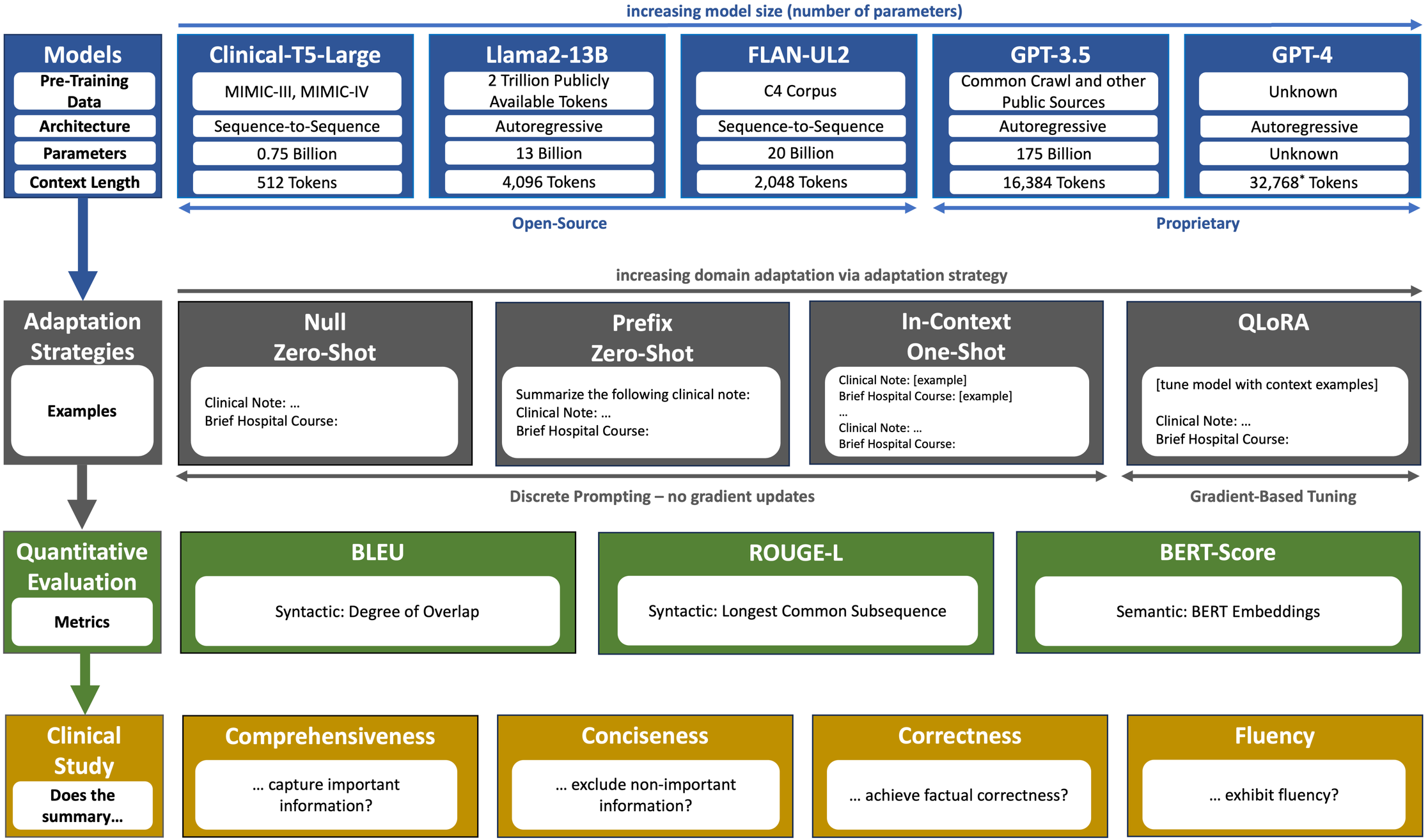
1. Overlap-Based:
   a. ROUGE-L
   b. BLEU
2. Similarity-Based
   a. Jaccard Similarity
   b. TF-IDF Similarity
   c. Sentence-BERT Similarity
   d. BERTScore
   e. Neural CRF Alignment
3. QA-Based:
   a. Exact Match (EM)
   b. F1-RadGraph

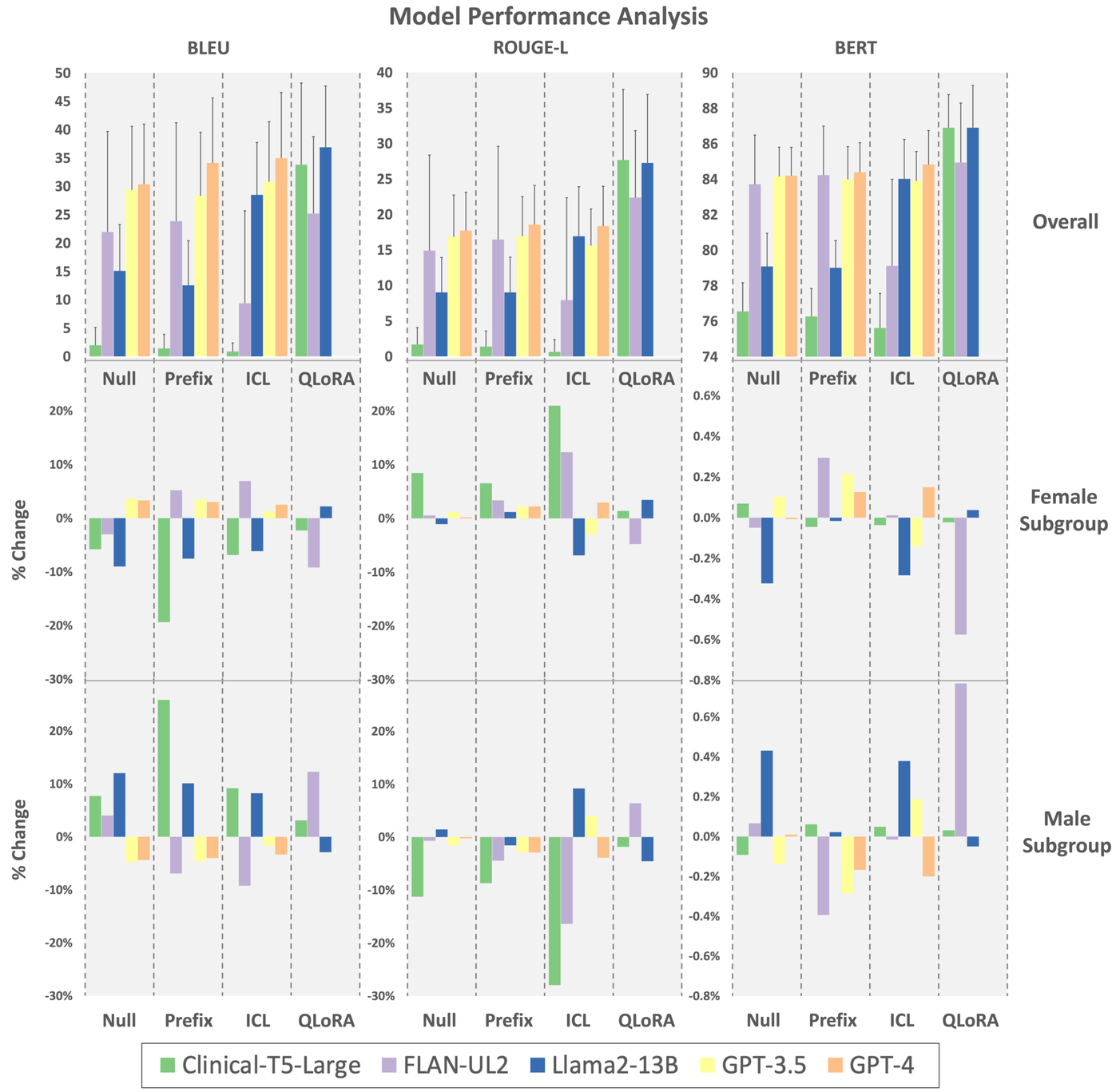
Reference-Free[2]:

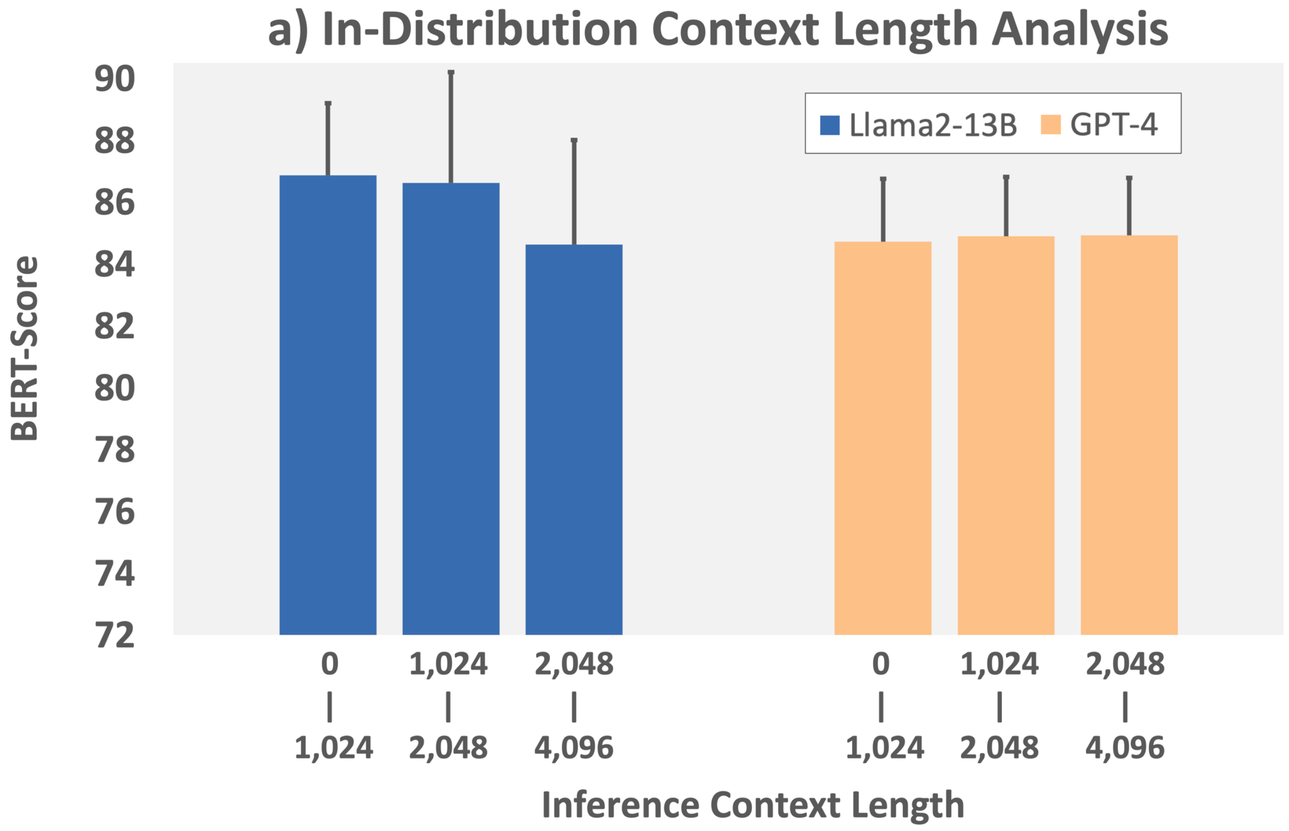
1. Quality Metrics:
   a. SUPERT
   b. BLANC
2. Factuality Metrics:
   a. QuestVal
   b. QAFactEval
   c. FactCC
   d. DAE
   e. SummaC

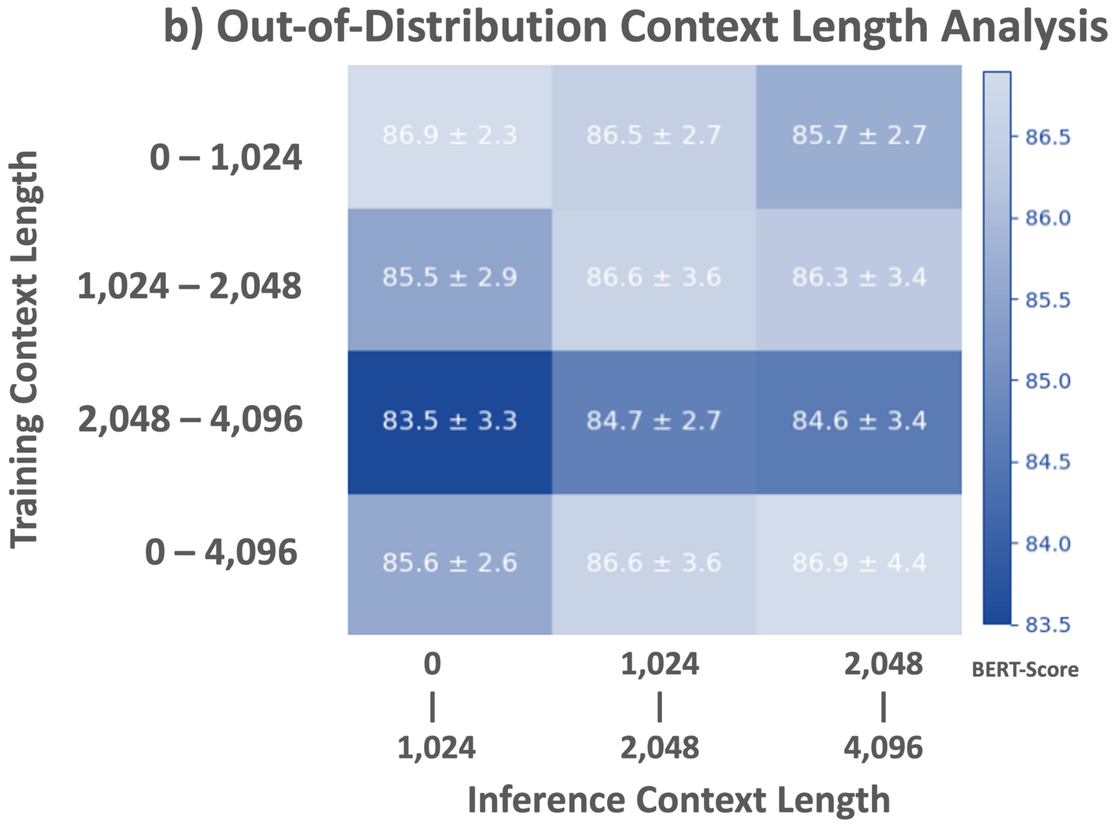
[1] Comparison of LLM generated summary with reference summary
[2] Comparison of LLM generated summary with source material

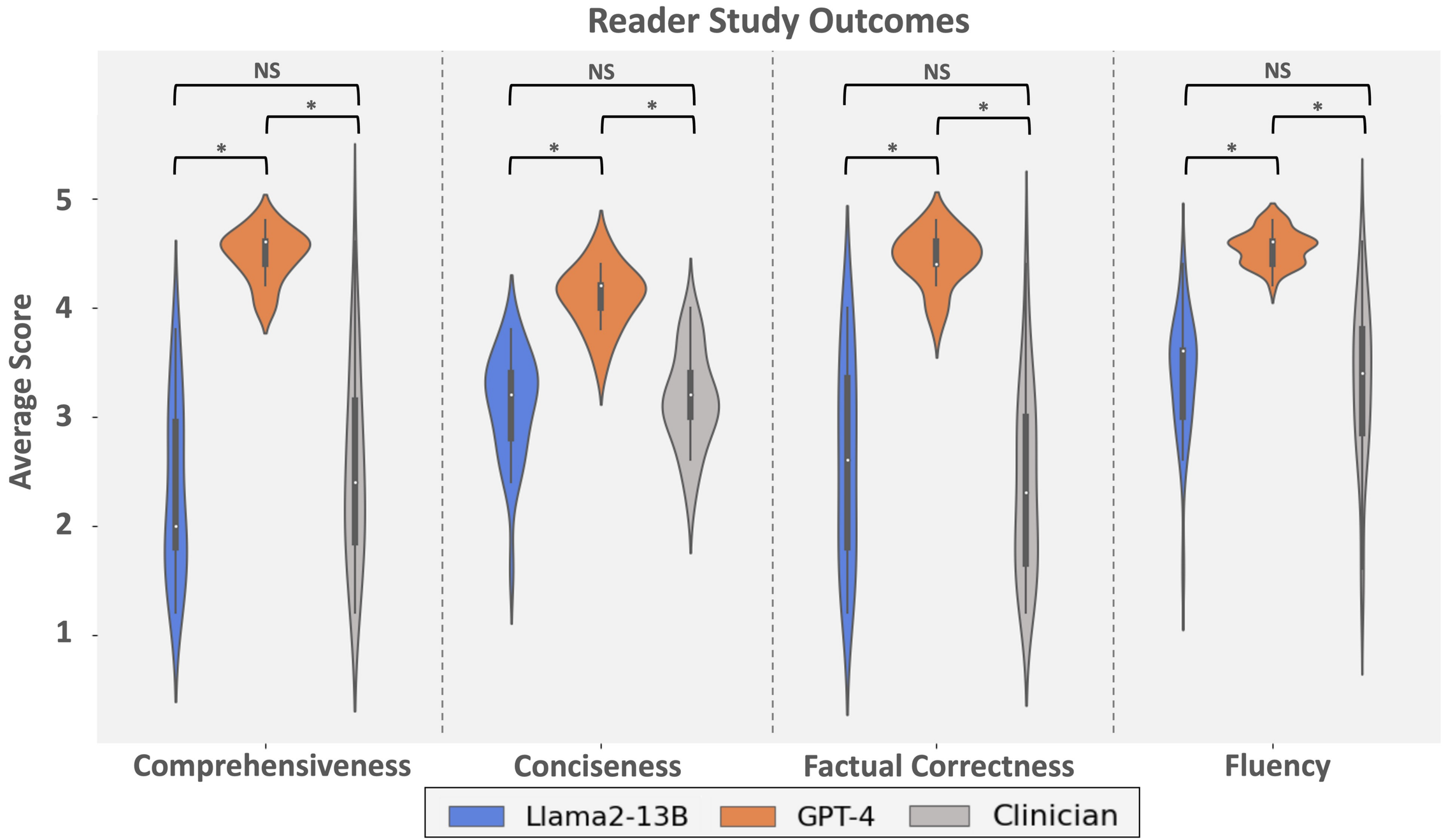
*Reference: News Summarization and Evaluation in the Era of GPT-3


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?


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


a) In-Distribution Context Length Analysis
BERT-Score
90
88
86
84
82
80
78
76
74
72
Llama2-13B
GPT-4
0
1,024
2,048
1,024
2,048
4,096
0
1,024
2,048
1,024
2,048
4,096
Inference Context Length


b) Out-of-Distribution Context Length Analysis
Training Context Length
0 – 1,024
1,024 – 2,048
2,048 – 4,096
0 – 4,096
86.9 ± 2.3
86.5 ± 2.7
85.7 ± 2.7
85.5 ± 2.9
86.6 ± 3.6
86.3 ± 3.4
83.5 ± 3.3
84.7 ± 2.7
84.6 ± 3.4
85.6 ± 2.6
86.6 ± 3.6
86.9 ± 4.4
86.5
86.0
85.5
85.0
84.5
84.0
83.5
BERT-Score
0
1,024
2,048
1,024
2,048
4,096
Inference Context Length


Reader Study Outcomes
Average Score
5
4
3
2
1
NS
*
*
Comprehensiveness
Conciseness
Factual Correctness
Fluency
Llama2-13B
GPT-4
Clinician

The University of Texas at Austin
Cockrell School of Engineering

Innovation starts here