Clinical Text Summarization from EHR

Presented by
Mirza Mohammad Azwad, 200042121
Dayan Ahmed Khan, 200042105
K.M. Tahlil Mahfuz Faruk, 200042158
H.M. Kawsar Ahamad, 200042108

Content

Content

Motivation & Problem Statement

Motivation

- Clinicians face heavy documentation burden
- Summarizing reports is time-consuming
- LLMs (e.g. GPT) excel at text tasks but clinical summarization is under-explored
- This is a high-impact opportunity to automate summaries and improve patient care

Problem Statement

- Can LLMs generate clinical summaries that match or exceed human expert performance in accuracy and completeness?
- How do adaptation methods like in-context learning and QLoRA fine-tuning affect LLM summarization quality?
- Are open-source LLMs viable alternatives to proprietary models like GPT-4 for clinical text summarization?
- How do LLMs perform across diverse clinical tasks such as radiology, progress notes, and doctor-patient dialogues?
- What are the risks of hallucinations or factual errors in LLM-generated clinical summaries, and how can they be mitigated?

Current State of Art

Adapting LLMs to Clinical Summarization

- Evaluated 8 models (open-source + GPT-3.5/4) on 4 tasks (6 datasets)
- Key goal: Match or surpass human-quality summaries (completeness, correctness, conciseness)

Summarization Tasks & Data

- Four Clinical Task
  - Radiology Reports (findings → impression)
  - Patient Questions (verbose query → concise query)
  - Progress Notes (ICU notes → problem list)
  - Doctor–Patient Dialogues (consult transcripts → assessment & plan)

Models & Adaptation Methods

- LLMs evaluated:
  - Open-source (FLAN-T5, FLAN-UL2, Llama-2, Vicuna, Alpaca-based) vs Proprietary (GPT-3.5, GPT-4)
- Two adaptation strategies:
  - In-Context Learning (ICL): few-shot prompting with example summaries
  - QLoRA: 4-bit quantized LoRA fine-tuning of model adapters

Current State of Art Continued...

Quantitative (NLP) Results

- Adapting with examples dramatically improves performance over zero-shot
- For open models, ICL and QLoRA give similar gains
- GPT-3.5/4 far outperform all others when given full context
- Best setting: GPT-4 (32K token context) with max few-shot examples (FLAN-T5 fine-tuned also strong, but limited by shorter context length

Clinical Reader Study Results

- Physician study (10 doctors) compared GPT-4 vs human expert summaries.
- Completeness/Correctness: GPT-4 was significantly better ($p<0.001$) on all tasks
- In radiology reports, GPT-4 matched/outperformed humans 100% of the time (0 human wins out of 100)
- Overall preference: humans preferred only 19% of summaries; GPT-4 was preferred in 36% and tied/non-inferior in 45%
- Conciseness: GPT-4's summaries were significantly more concise (shorter) on most tasks
- GPT-4 made fewer factual errors (hallucinations) than human summaries

Current State of Art Continued...

Key Findings & Novelty

- Novel result: Machine-generated summaries are clinically non-inferior to experts
- Adapted LLMs outperform humans on completeness, correctness, conciseness
- LLMs hallucinate less than humans (fewer errors)
- This is the first evidence that LLMs can reliably assist clinicians in summarizing notes

Clinical Impact & Integration

- Implication: Adapted LLMs could reduce clinician workload and speed up workflows
- Proposed use: integrate into EHR for auto-summary of notes (doctors review and finalize). This frees more time for patient care, potentially improving outcomes
- Emphasize that LLMs assist rather than replace doctors

Literature Review

Foundational Literature: Clinical Text Summarization: Adapting Large Language Models Can Outperform Human Experts - Van Veen et al. (2023)

- Evaluated 8 large language models (LLMs) including GPT-3.5, GPT-4, FLAN-T5, Llama-2, Vicuna.
- Adaptation via:
  - In-Context Learning (ICL): Using examples in the prompt.
  - QLoRA: Lightweight fine-tuning using low-rank adaptation.
- Assessed across 4 clinical summarization tasks, using 6 datasets.
- NLP Evaluation Metrics:
  - BLEU: Measures word overlap - best for completeness
  - ROUGE-L: Longest matching sequences - good for fluency
  - BERTScore: Compares meaning using BERT - best for correctness
  - MEDCON: Checks medical concept overlap - captures clinical accuracy

Literature Review

Key Findings:

- GPT-4 outperformed human experts in:
    - Completeness
    - Correctness
    - Conciseness
- Reader study with 10 physicians: GPT-4 preferred or non-inferior in >80% of cases.
- NLP metrics showed low correlation with clinical judgments (max ~0.2).

Importance of Adaptation

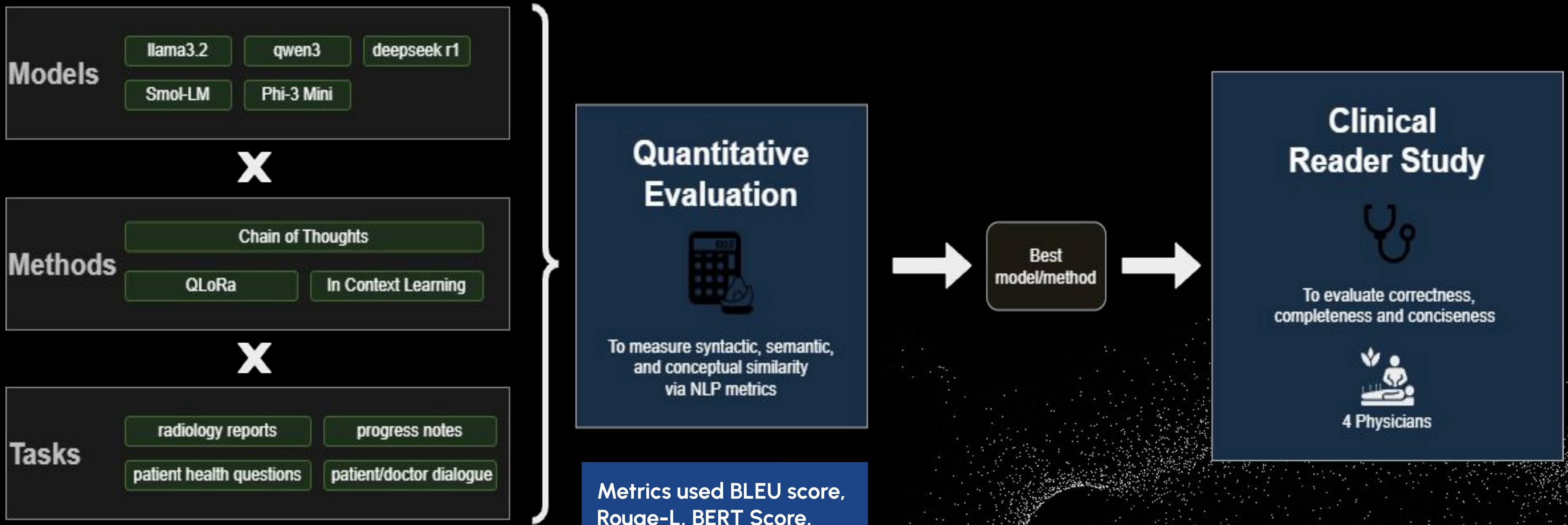
- Zero-shot prompting was inferior to adaptation.
- Task adaptation > domain adaptation
    - E.g., Med-Alpaca (trained on medical Q&A) performed worse than Alpaca in summarization.
- GPT-4 + ICL (with relevant examples) was the best performer across most tasks

Dataset

Task (Dataset)	Task description	Number of samples
Radiol. reports (Open-i)	findings → impression	3.4K
Radiol. reports (MIMIC-CXR)	findings → impression	128K
Radiol. reports (MIMIC-III)	findings → impression	67K
Patient questions (MeQSum)	verbose → short question	1.2K
Progress notes (ProbSum)	notes → problem list	755
Dialogue (ACI-Bench)	dialogue → assessment	126

Figure: Dataset Description [1] by Dave et al.

Proposed Methodology



Metrics used BLEU score, Rouge-L, BERT Score, and Medcon

Toolkits and Technologies

LLMS

Deepseek R1

LLAMA 3

Phi-3

Tools

spaCy

NLTK

LangChain

Chroma

Risks and Challenges

Hallucinations & Errors

- LLMs (e.g., Llama3) can introduce factual inaccuracies or omit key clinical details, especially when lacking relevant context or prior studies.

Over Reliance on AI

- Risk that clinicians depend too much on model outputs, which may occasionally misrepresent or omit critical information, affecting patient care.

Context-Specific Preferences

- Summarization needs vary by specialty and individual clinician; current models may not adequately align with these nuances.

Governance & Privacy

- Use of proprietary models raises issues related to data privacy, regulatory compliance, and transparency around model training and adaptation.

Risks and Challenges

Model Adaptation

- High-quality summaries require both domain and task-specific adaptation; domain adaptation alone is often insufficient.

Metric Limitations

- Standard NLP metrics (BLEU, ROUGE-L, BERTScore, MEDCON) weakly correlate with true clinical quality as judged by physicians.

Prompt Engineering

- Small changes in prompt wording or settings (like temperature) can significantly alter output quality; optimal configurations are task-dependent and require manual tuning.

Long/Complex Input Handling

- Many clinical docs exceed model context windows; open-source models often struggle with lengthy inputs compared to proprietary ones.

Data Diversity

- Performance can degrade with varied or unstructured input formats (e.g., radiology vs. patient questions).

Expected Outcome

Improved Summaries

- Advanced models using in-context learning, generate clinical summaries that are often more complete, accurate, and concise than human-written ones in evaluated cases.

Reduced Documentation Burden

- Decrease clinicians' administrative load, potentially reducing burnout and freeing time for patient care.

Human-AI Collaboration

- LLMs are best used as supportive tools—drafting summaries for clinician review, not as full replacements for expert judgement.

Future Research Directions

- Specialty-specific adaptation, prompt optimization, better long-context handling, hallucination mitigation, and new evaluation frameworks are identified as key research priorities

Comparison with open source models and SLMs

- Using open source models and SLMs can produce comparative results to proprietary models with proper prompt engineering and in context learning

References

1. Clinical Text Summarization: Adapting Large Language Models Can Outperform Human Experts, Dave et al., 2023

2. MIMIC-IV, a freely accessible electronic health record dataset, Alistair et al., 2023

3. MIMIC-III, a freely accessible accessible critical care database, Alistair et al., 2016

Thank You