Institute of Health
Informatics

UCL

Integrating LLMs into NHS

Case Study -> Automated Discharge Summaries

Presenter: Simon Ellershaw

Supervisors: Kawsar Noor, Anoop Shah, Richard Dobson

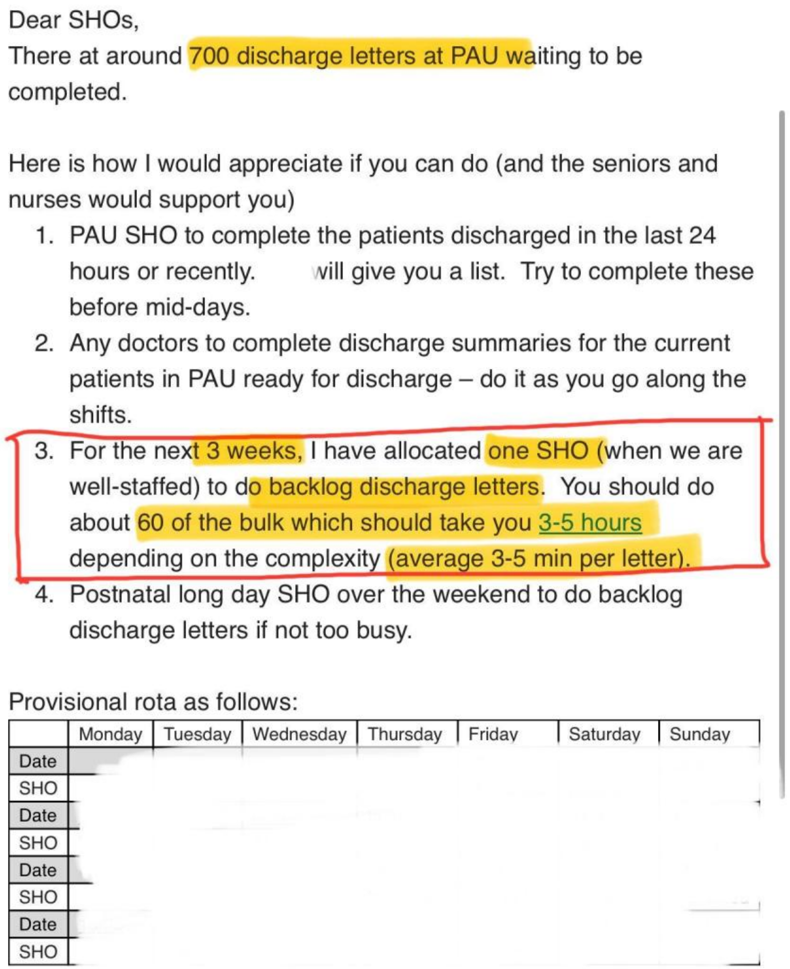
Content

1. Why use an LLM?

2. Easy Proof of Concept

3. Difficult Real World Deployment

Motivation



Dear SHOs,
There at around 700 discharge letters at PAU waiting to be completed.

Here is how I would appreciate if you can do (and the seniors and nurses would support you)

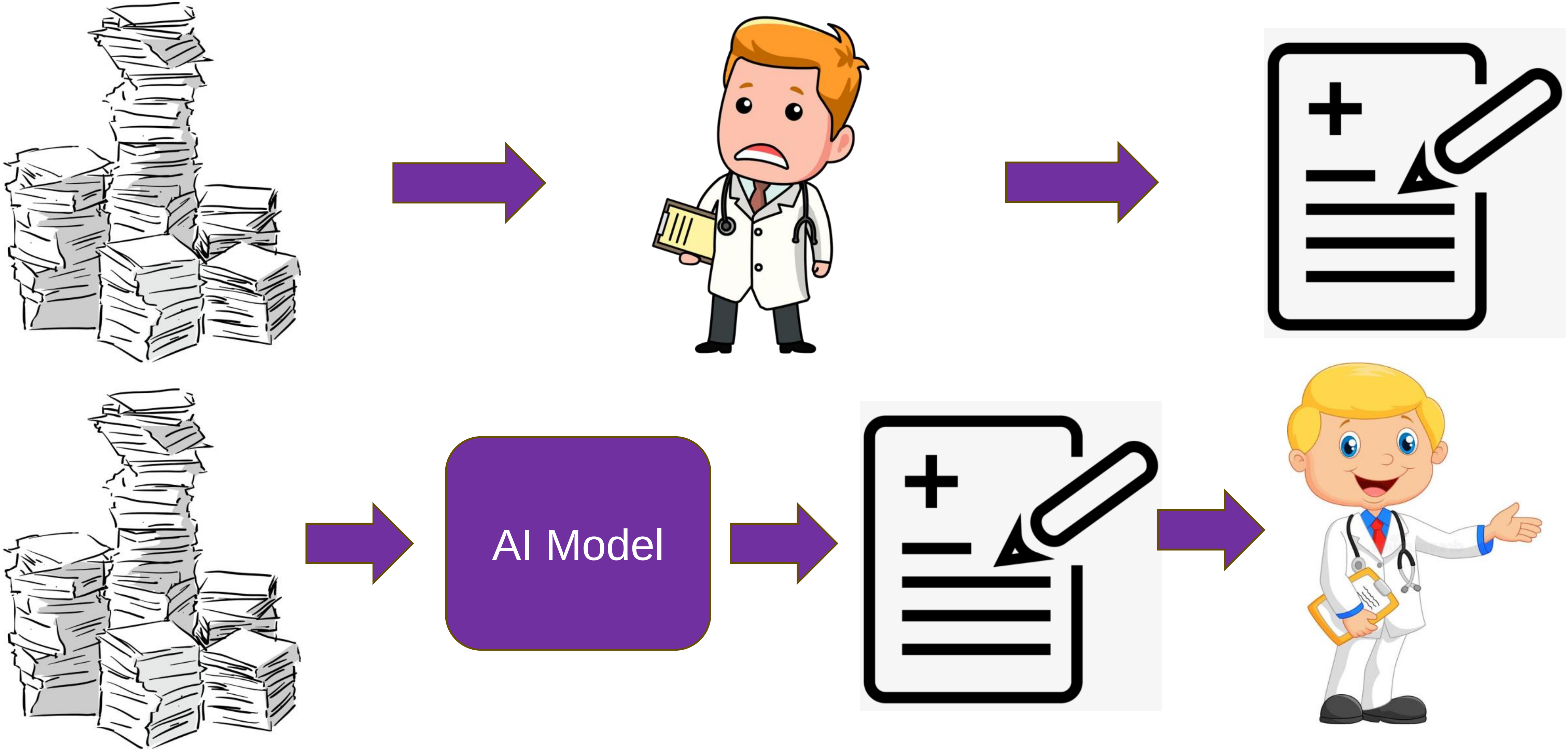
1. PAU SHO to complete the patients discharged in the last 24 hours or recently. will give you a list. Try to complete these before mid-days.
2. Any doctors to complete discharge summaries for the current patients in PAU ready for discharge – do it as you go along the shifts.
3. For the next 3 weeks, I have allocated one SHO (when we are well-staffed) to do backlog discharge letters. You should do about 60 of the bulk which should take you 3-5 hours depending on the complexity (average 3-5 min per letter).
4. Postnatal long day SHO over the weekend to do backlog discharge letters if not too busy.

Provisional rota as follows:

	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday
Date							
SHO							
Date							
SHO							
Date							
SHO							
Date							
SHO							

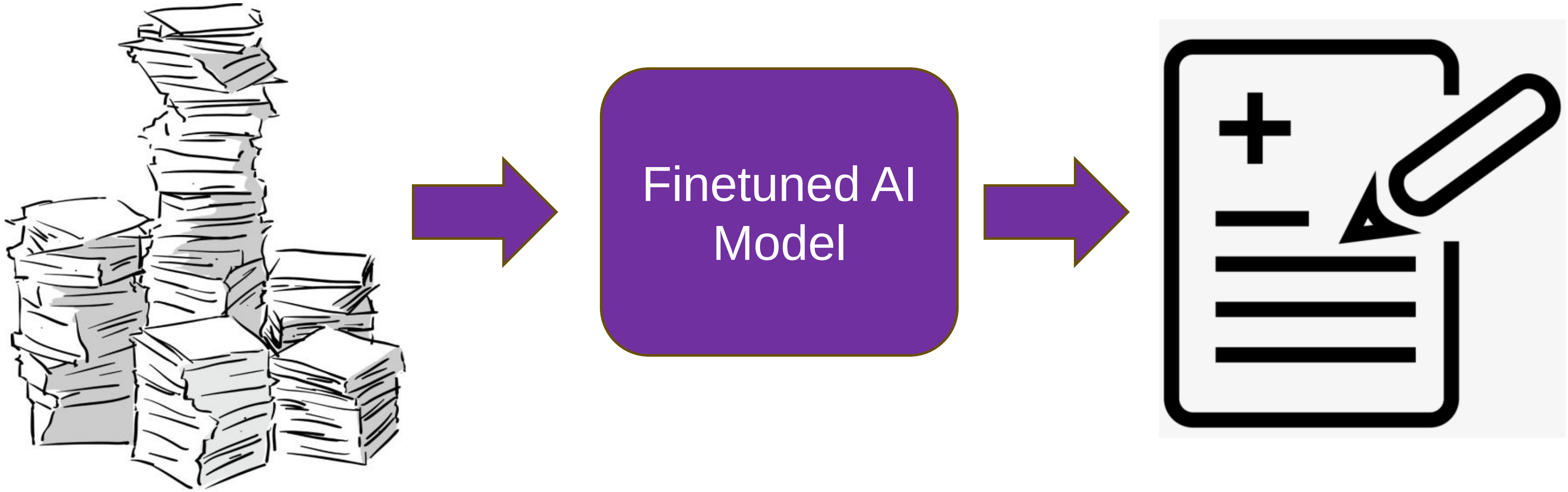
https://x.com/Xeon4f145d96s1/status/1784234951811998123

Motivation



AI Model

Previous Supervised Learning Approaches

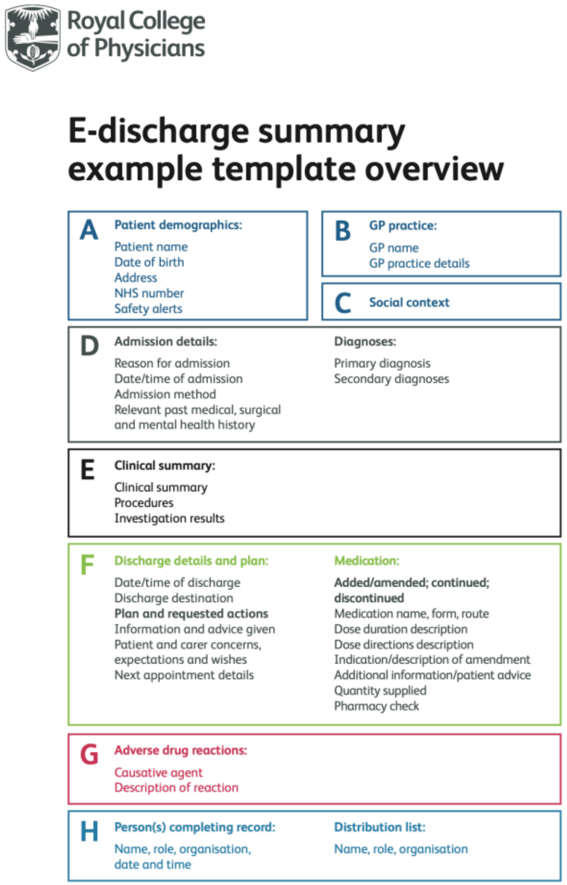


Require notes -> discharge summary dataset

- Real-world discharge summaries “silver standard”
- Generalizability challenge across clinicians, specialties, hospitals, etc...
- Sensitive to input format changes

Searle, T.; Ibrahim, Z.; Teo, J.; and Dobson, R. J. 2023. Discharge summary hospital course summarisation of inpatient Electronic Health Record text with clinical concept guided deep pre-trained Transformer models. Journal of Biomedical Informatics, 141: 104358.

Clinical Guidelines as LLM Prompts



Royal College of Physicians

E-discharge summary example template overview

A **Patient demographics:**
Patient name
Date of birth
Address
NHS number
Safety alerts

B **GP practice:**
GP name
GP practice details

C **Social context**

D **Admission details:**
Reason for admission
Date/time of admission
Admission method
Relevant past medical, surgical and mental health history

Diagnoses:
Primary diagnosis
Secondary diagnoses

E **Clinical summary:**
Clinical summary
Procedures
Investigation results

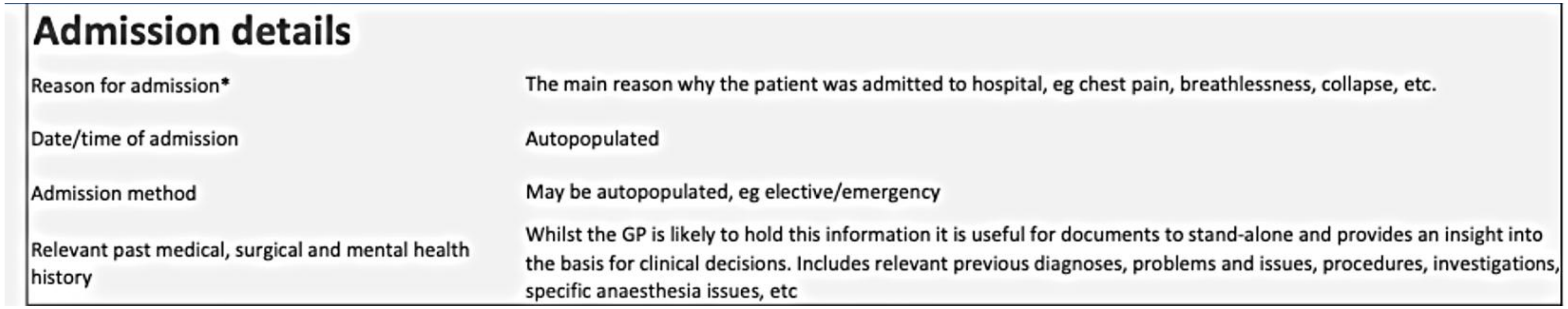
F **Discharge details and plan:**
Date/time of discharge
Discharge destination
Plan and requested actions
Information and advice given
Patient and carer concerns, expectations and wishes
Next appointment details

Medication:
Added/amended; continued; discontinued
Medication name, form, route
Dose duration description
Dose directions description
Indication/description of amendment
Additional information/patient advice
Quantity supplied
Pharmacy check

G **Adverse drug reactions:**
Causative agent
Description of reaction

H **Person(s) completing record:**
Name, role, organisation, date and time

Distribution list:
Name, role, organisation



Admission details

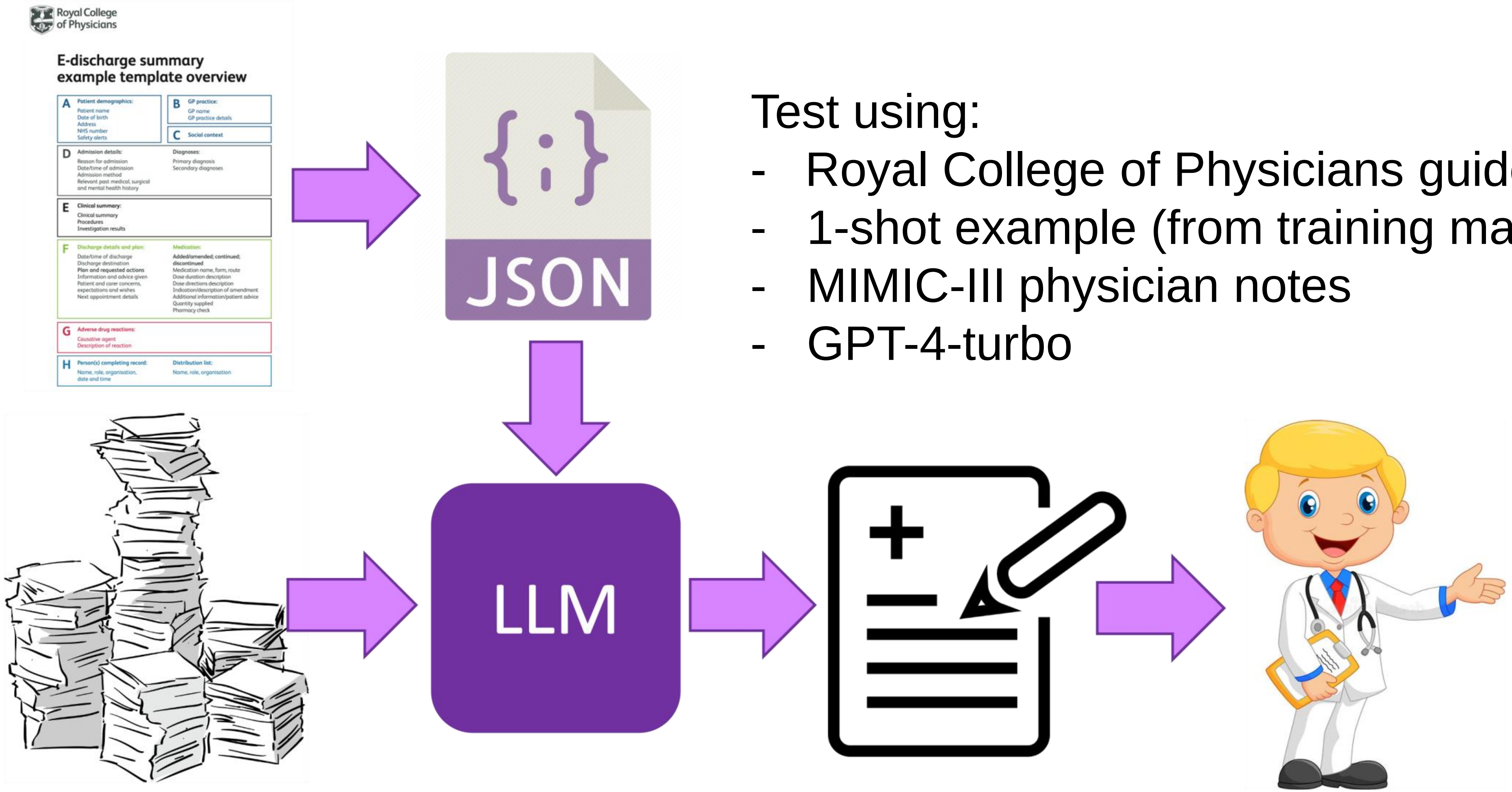
Reason for admission*	The main reason why the patient was admitted to hospital, eg chest pain, breathlessness, collapse, etc.
Date/time of admission	Autopopulated
Admission method	May be autopopulated, eg elective/emergency
Relevant past medical, surgical and mental health history	Whilst the GP is likely to hold this information it is useful for documents to stand-alone and provides an insight into the basis for clinical decisions. Includes relevant previous diagnoses, problems and issues, procedures, investigations, specific anaesthesia issues, etc

https://www.rcplondon.ac.uk/guidelines-policy/improving-discharge-summaries-learning-resource-materials

Method

Test using:

- Royal College of Physicians guidelines
- 1-shot example (from training material)
- MIMIC-III physician notes
- GPT-4-turbo



Royal College of Physicians

E-discharge summary example template overview

A Patient demographics:
Patient name
Date of birth
Address
NHS number
Safety alerts

B GP practice:
GP name
GP practice details

C Social context

D Admission details:
Reason for admission
Date/time of admission
Admission method
Relevant past medical, surgical and mental health history

Diagnoses:
Primary diagnosis
Secondary diagnoses

E Clinical summary:
Clinical summary
Procedures
Investigation results

F Discharge details and plan:
Date/time of discharge
Discharge destination
Plan and requested actions
Information and advice given
Patient and carer concerns, expectations and wishes
Next appointment details

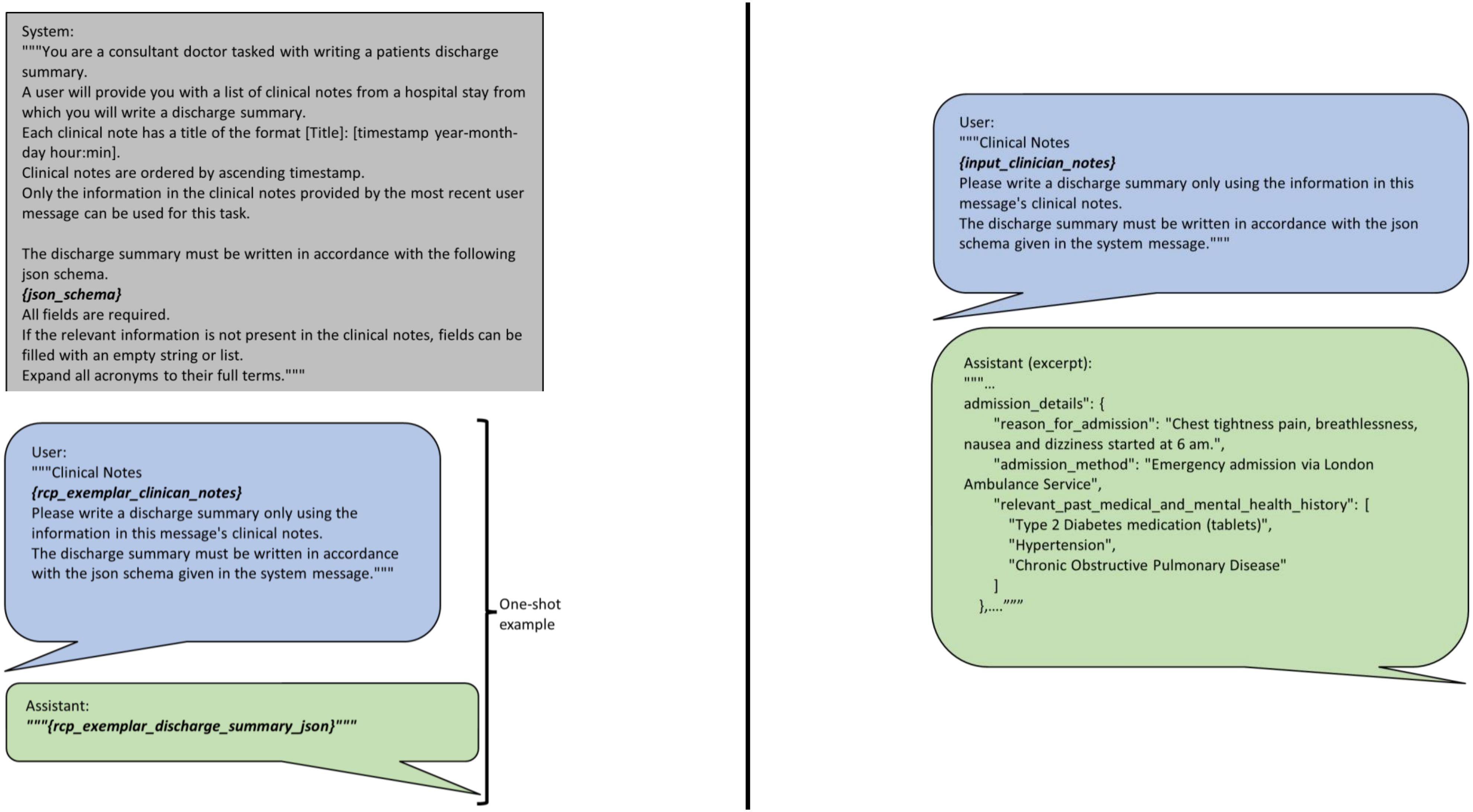
Medication:
Added/amended; continued; discontinued
Medication name, form, route
Dose duration description
Dose directions description
Indication/description of amendment
Additional information/patient advice
Quantity supplied
Pharmacy check

G Adverse drug reactions:
Causative agent
Description of reaction

H Person(s) completing record:
Name, role, organisation, date and time

Distribution list:
Name, role, organisation

Method



System:
"""You are a consultant doctor tasked with writing a patients discharge summary.
A user will provide you with a list of clinical notes from a hospital stay from which you will write a discharge summary.
Each clinical note has a title of the format [Title]: [timestamp year-month-day hour:min].
Clinical notes are ordered by ascending timestamp.
Only the information in the clinical notes provided by the most recent user message can be used for this task.

The discharge summary must be written in accordance with the following json schema.
{json_schema}
All fields are required.
If the relevant information is not present in the clinical notes, fields can be filled with an empty string or list.
Expand all acronyms to their full terms."""

User:
"""Clinical Notes
{rcp_exemplar_clinican_notes}
Please write a discharge summary only using the information in this message's clinical notes.
The discharge summary must be written in accordance with the json schema given in the system message."""

Assistant:
"""{rcp_exemplar_discharge_summary_json}"""

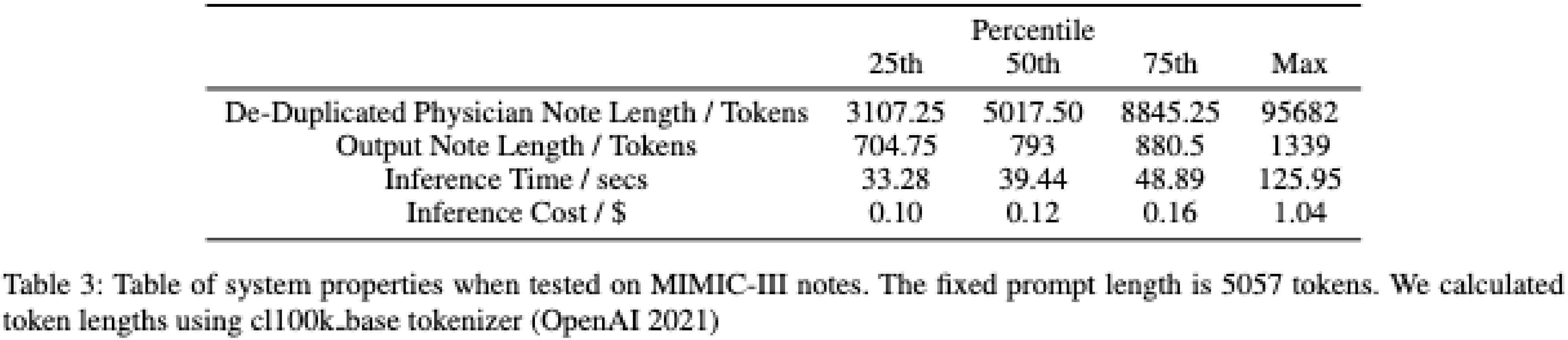
One-shot example

User:
"""Clinical Notes
{input_clinician_notes}
Please write a discharge summary only using the information in this message's clinical notes.
The discharge summary must be written in accordance with the json schema given in the system message."""

Assistant (excerpt):

```
"""...
admission_details": {
      "reason_for_admission": "Chest tightness pain, breathlessness,
nausea and dizziness started at 6 am.",
      "admission_method": "Emergency admission via London
Ambulance Service",
      "relevant_past_medical_and_mental_health_history": [
         "Type 2 Diabetes medication (tablets)",
         "Hypertension",
         "Chronic Obstructive Pulmonary Disease"
      ]
   },....”””
```

Results



	Percentile			
	25th	50th	75th	Max
De-Duplicated Physician Note Length / Tokens	3107.25	5017.50	8845.25	95682
Output Note Length / Tokens	704.75	793	880.5	1339
Inference Time / secs	33.28	39.44	48.89	125.95
Inference Cost / $	0.10	0.12	0.16	1.04

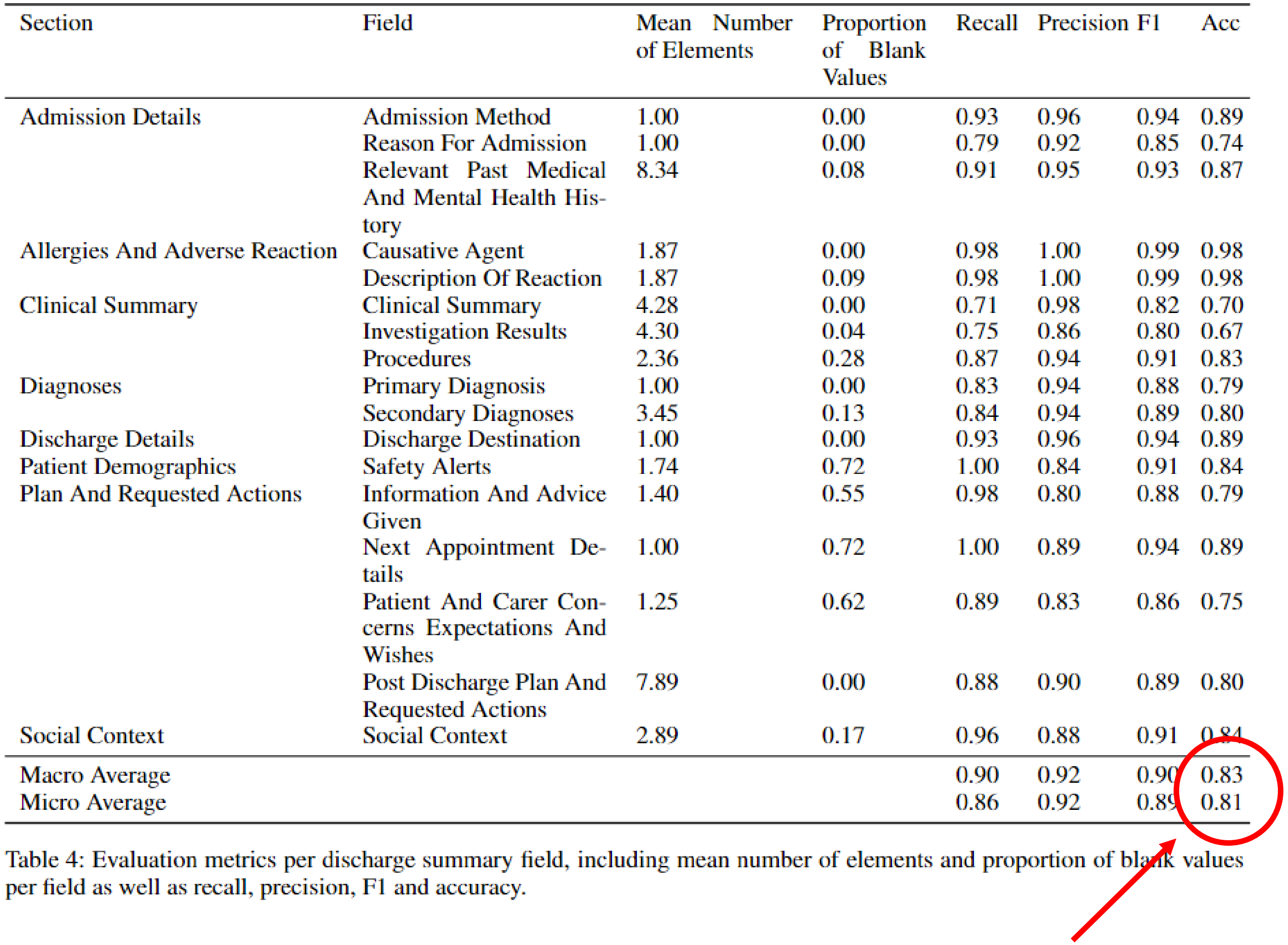
Table 3: Table of system properties when tested on MIMIC-III notes. The fixed prompt length is 5057 tokens. We calculated token lengths using cl100k_base tokenizer (OpenAI 2021)

Results

11 medical professionals evaluated 53 summaries

4 types of error

- Missing (False Negative)
  - Safety Critical
  - Minor
- Additional (False Positive)
  - Hallucination
  - Irrelevant
- Explanation



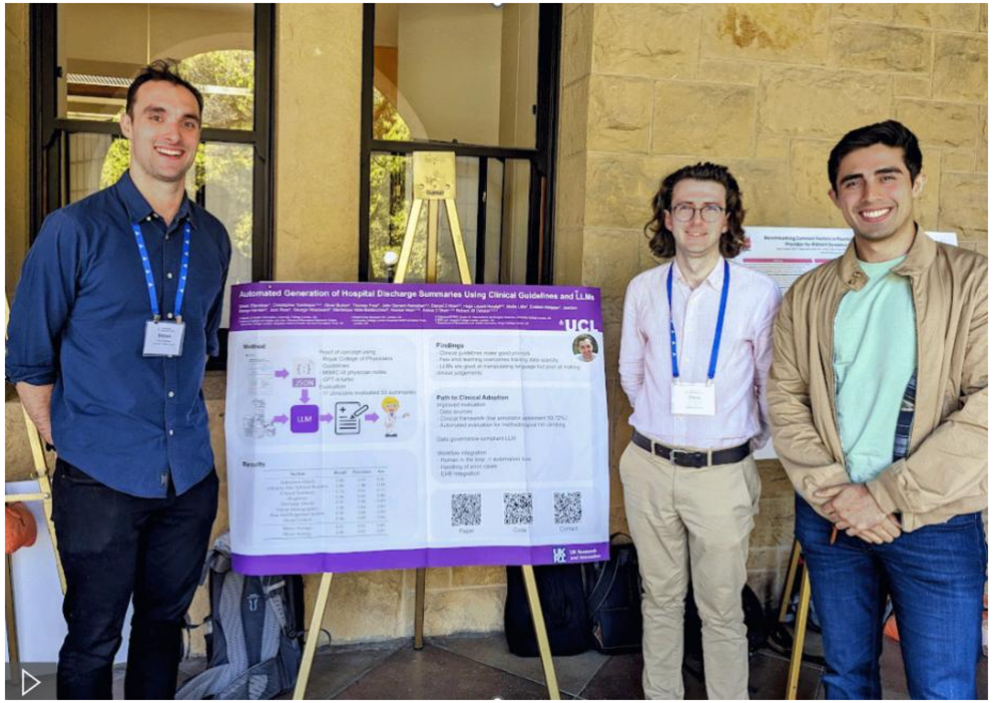
Section	Field	Mean Number of Elements	Proportion of Blank Values	Recall	Precision	F1	Acc
Admission Details	Admission Method	1.00	0.00	0.93	0.96	0.94	0.89
	Reason For Admission	1.00	0.00	0.79	0.92	0.85	0.74
	Relevant Past Medical And Mental Health History	8.34	0.08	0.91	0.95	0.93	0.87
Allergies And Adverse Reaction	Causative Agent	1.87	0.00	0.98	1.00	0.99	0.98
	Description Of Reaction	1.87	0.09	0.98	1.00	0.99	0.98
Clinical Summary	Clinical Summary	4.28	0.00	0.71	0.98	0.82	0.70
	Investigation Results	4.30	0.04	0.75	0.86	0.80	0.67
	Procedures	2.36	0.28	0.87	0.94	0.91	0.83
Diagnoses	Primary Diagnosis	1.00	0.00	0.83	0.94	0.88	0.79
	Secondary Diagnoses	3.45	0.13	0.84	0.94	0.89	0.80
Discharge Details	Discharge Destination	1.00	0.00	0.93	0.96	0.94	0.89
Patient Demographics	Safety Alerts	1.74	0.72	1.00	0.84	0.91	0.84
Plan And Requested Actions	Information And Advice Given	1.40	0.55	0.98	0.80	0.88	0.79
	Next Appointment Details	1.00	0.72	1.00	0.89	0.94	0.89
	Patient And Carer Concerns Expectations And Wishes	1.25	0.62	0.89	0.83	0.86	0.75
	Post Discharge Plan And Requested Actions	7.89	0.00	0.88	0.90	0.89	0.80
Social Context	Social Context	2.89	0.17	0.96	0.88	0.91	0.84
Macro Average				0.90	0.92	0.90	0.83
Micro Average				0.86	0.92	0.89	0.81

Table 4: Evaluation metrics per discharge summary field, including mean number of elements and proportion of blank values per field as well as recall, precision, F1 and accuracy.

TL;DR Good but by no means perfect

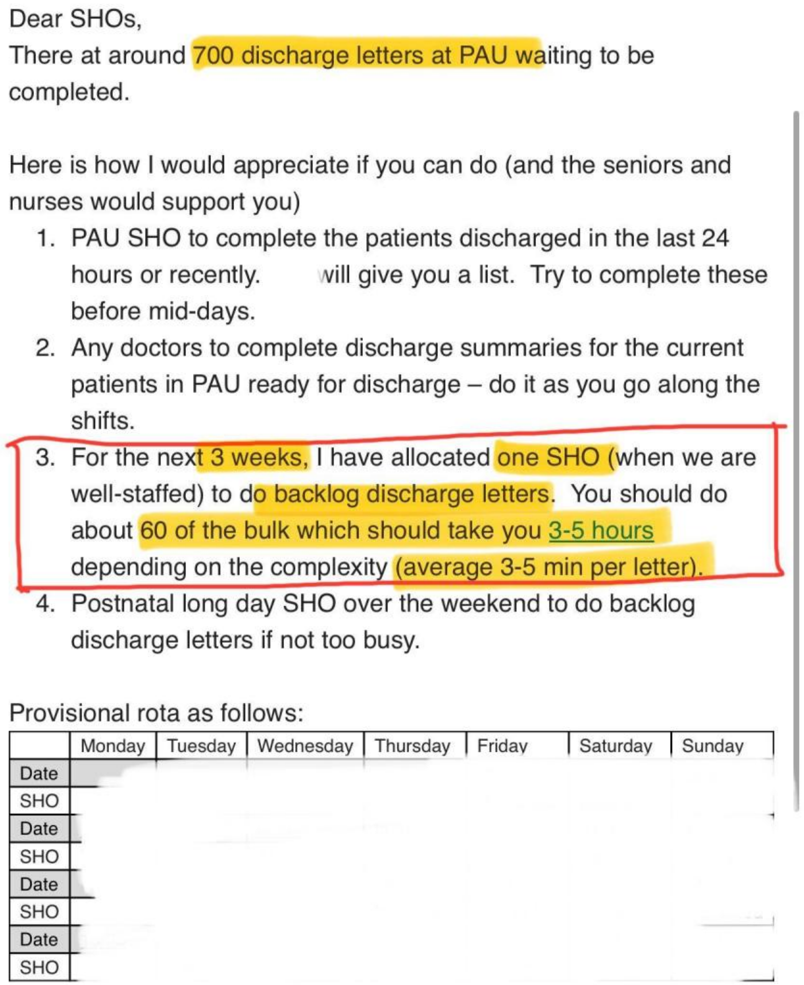
Conclusion

- PoC that LLMs can write valid discharge summaries
- Possible to few shot learn best practice from clinical guidelines


Automated Generation of Hospital Discharge Summaries Using Clinical Guidelines and LLMs
UCL
Method
JSON
LLM
Findings
Path to Clinical Adoption
Results
Paper
Code
Contact

https://openreview.net/forum?id=1kDJJPppRG&trk=public_post_comment-text

That’s nice and all but….



Dear SHOs,
There at around 700 discharge letters at PAU waiting to be completed.

Here is how I would appreciate if you can do (and the seniors and nurses would support you)

1. PAU SHO to complete the patients discharged in the last 24 hours or recently. will give you a list. Try to complete these before mid-days.
2. Any doctors to complete discharge summaries for the current patients in PAU ready for discharge – do it as you go along the shifts.
3. For the next 3 weeks, I have allocated one SHO (when we are well-staffed) to do backlog discharge letters. You should do about 60 of the bulk which should take you 3-5 hours depending on the complexity (average 3-5 min per letter).
4. Postnatal long day SHO over the weekend to do backlog discharge letters if not too busy.

Provisional rota as follows:

	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday
Date							
SHO							
Date							
SHO							
Date							
SHO							
Date							
SHO							

LLM PoC
->
Real World Deployment?

Blockers

- Evaluation
- LLM Deployment
- Regulation

Evaluation- Ideal

Gold standard answer

Reliable

Replicable

Inexpensive

Fast

Evaluation- Ideal

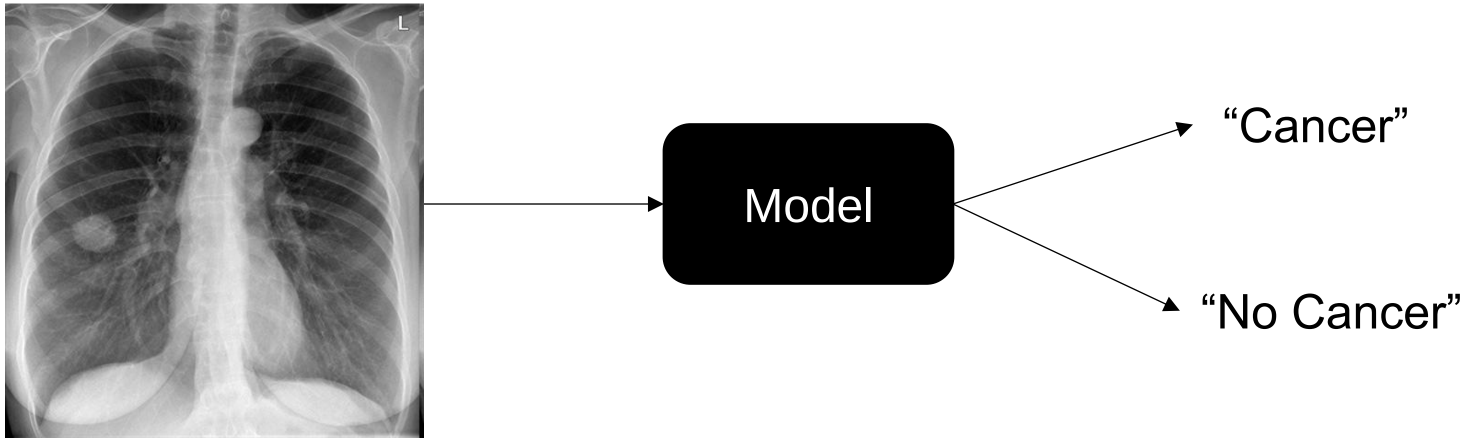
Gold standard answer

Reliable

Replicable

Inexpensive

Fast



Model

“Cancer”

“No Cancer”

Evaluation- Ours

Gold standard answer

- Not in the same format and “silver at best”

Reliable

- 59% inter-annotator agreement

Replicable

- Cannot be replicated without access to same clinicians

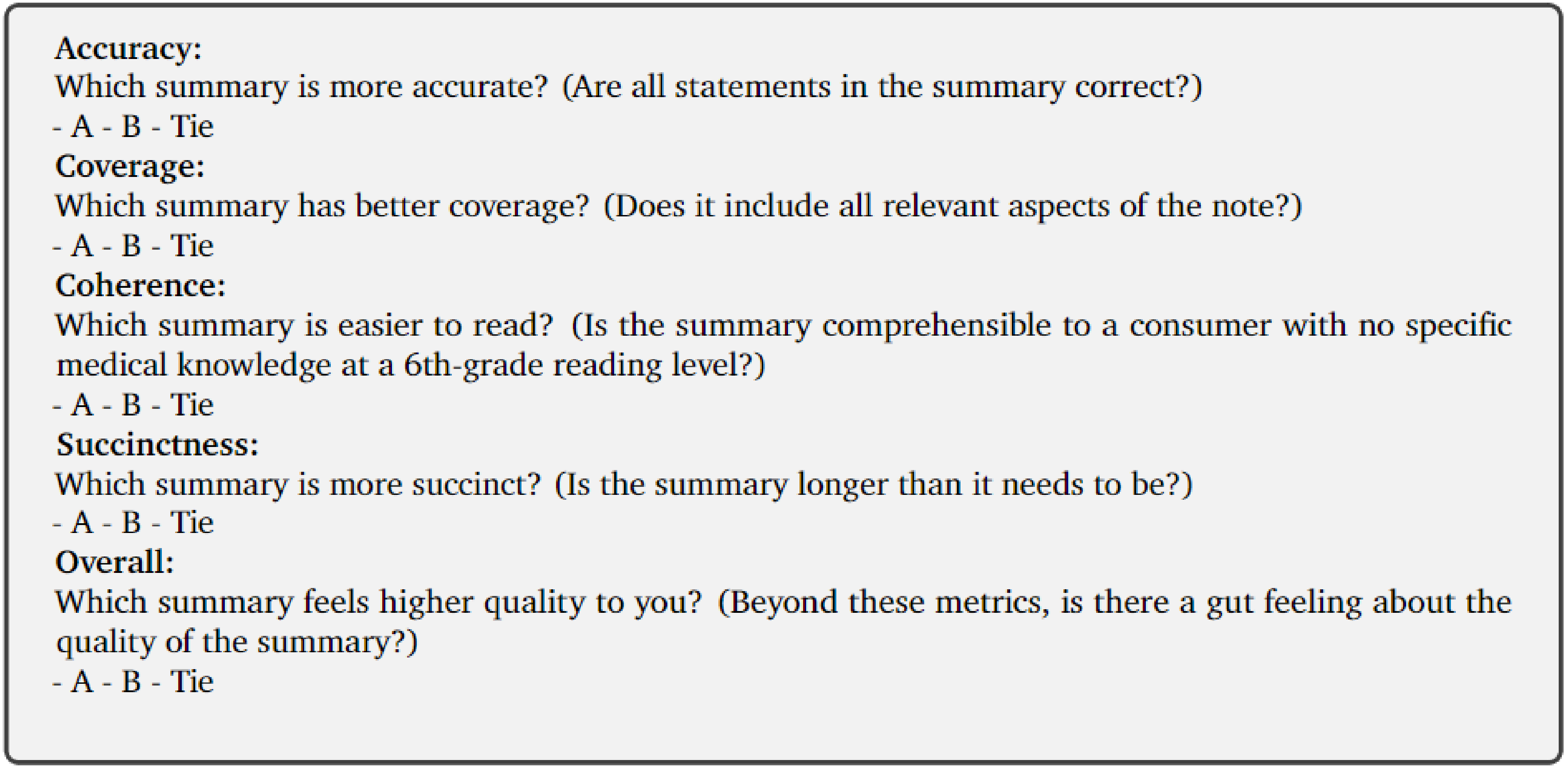
Inexpensive

- Clinician’s our expensive (or want authorship)

Fast

- 1-2 week iteration loop

Evaluation- By Comparison



Accuracy:
Which summary is more accurate? (Are all statements in the summary correct?)
- A - B - Tie

Coverage:
Which summary has better coverage? (Does it include all relevant aspects of the note?)
- A - B - Tie

Coherence:
Which summary is easier to read? (Is the summary comprehensible to a consumer with no specific medical knowledge at a 6th-grade reading level?)
- A - B - Tie

Succinctness:
Which summary is more succinct? (Is the summary longer than it needs to be?)
- A - B - Tie

Overall:
Which summary feels higher quality to you? (Beyond these metrics, is there a gut feeling about the quality of the summary?)
- A - B - Tie

Saab, Khaled, et al. "Capabilities of gemini models in medicine." *arXiv preprint arXiv:2404.18416* (2024).

Evaluation- By Comparison

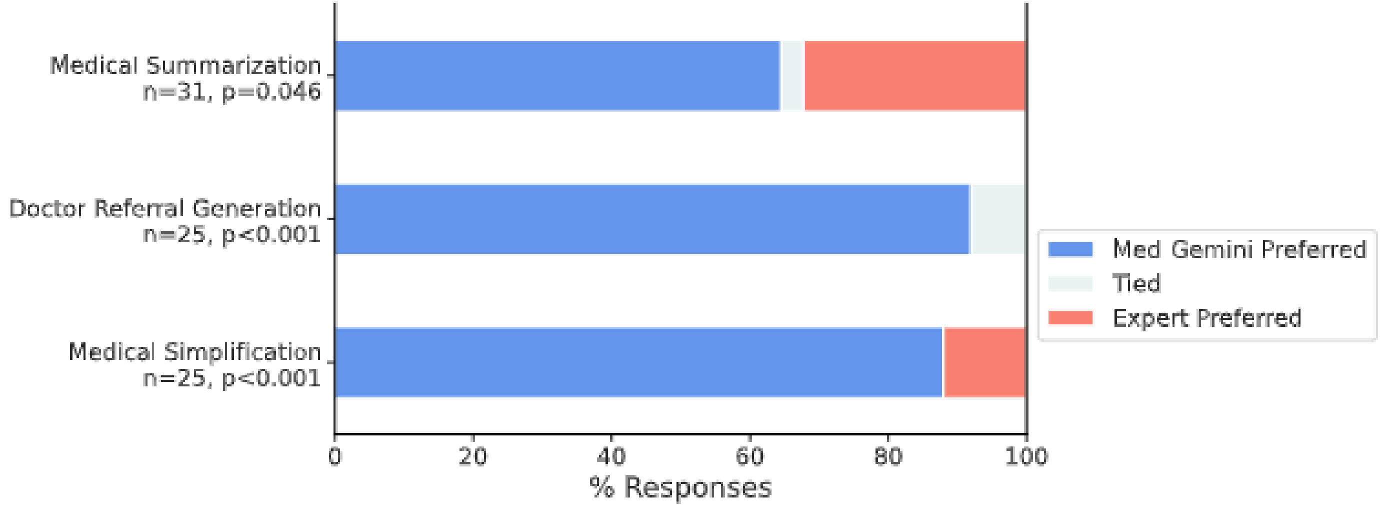
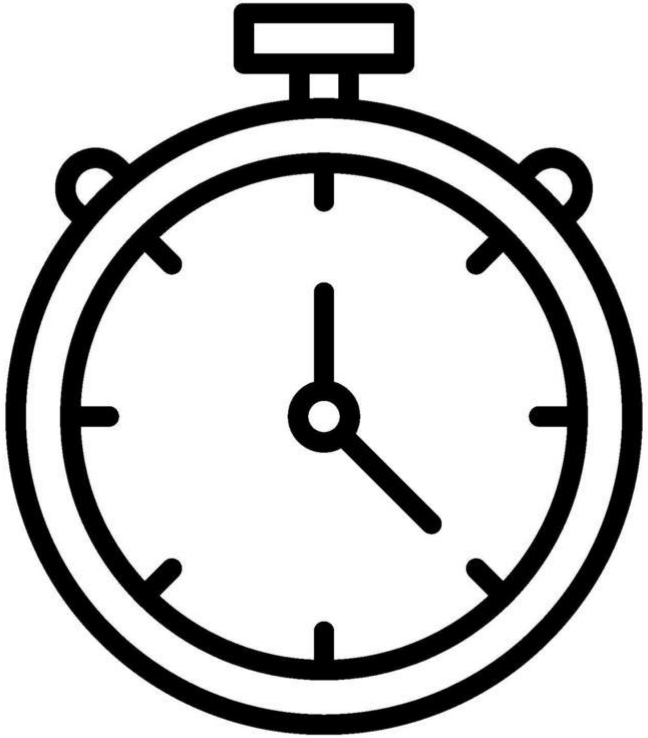


Figure 5 | **Evaluation of Med-Gemini-M 1.0 on long-form text-based tasks via side-by-side comparison with experts.** The tasks considered include generation of after-visit summaries, referral letters and simplified summaries of medical systematic reviews. Evaluation was performed by clinician raters. P-values are used to denote whether the rate at which Med-Gemini-M 1.0 is preferred or tied with experts is 0.5 (two-sided t-test).

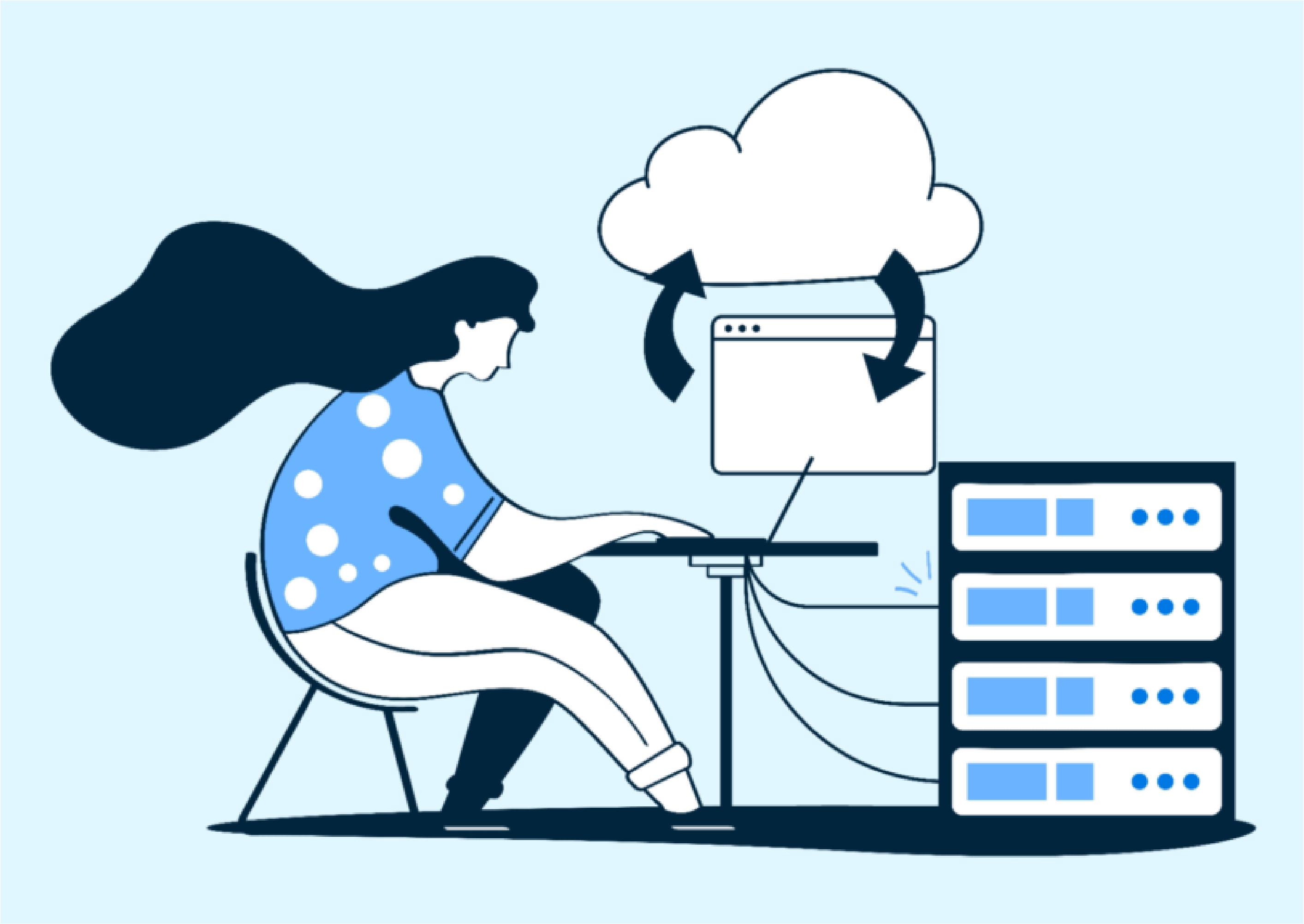
Saab, Khaled, et al. "Capabilities of gemini models in medicine." *arXiv preprint arXiv:2404.18416* (2024).

Evaluation- Automating

Evaluation- Efficiency?



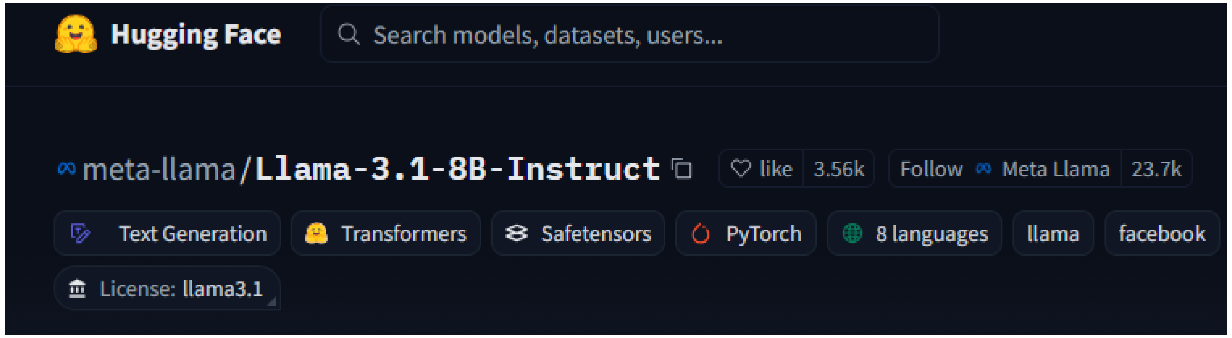
Deploying an LLM on Hospital Infrastructure



1. On-premises
2. On cloud
3. 3^rd party

Deploying an LLM on Hospital Infrastructure

Local LLMs at UCLH



Hugging Face

Search models, datasets, users...

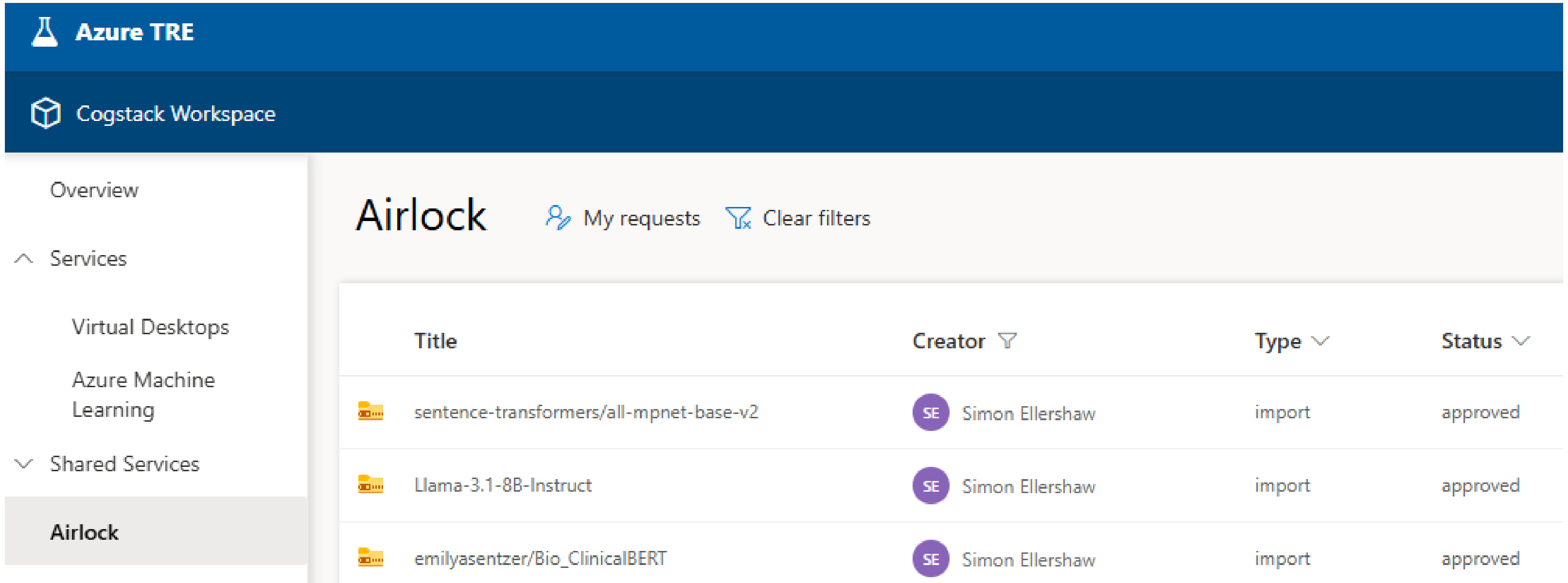
meta-llama/**Llama-3.1-8B-Instruct**

like 3.56k

Follow Meta Llama 23.7k

Text Generation · Transformers · Safetensors · PyTorch · 8 languages · llama · facebook

License: llama3.1



Azure TRE

Cogstack Workspace

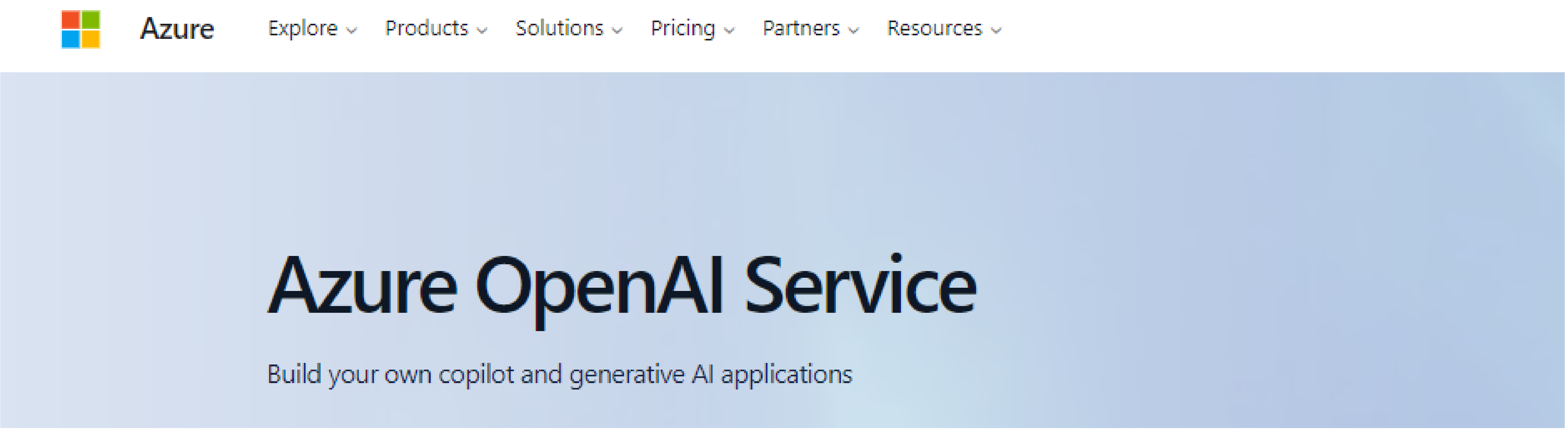
- Overview
- Services
  - Virtual Desktops
  - Azure Machine Learning
- Shared Services
- **Airlock**

Airlock

My requests · Clear filters

Title	Creator	Type	Status
sentence-transformers/all-mpnet-base-v2	SE Simon Ellershaw	import	approved
Llama-3.1-8B-Instruct	SE Simon Ellershaw	import	approved
emilyasentzer/Bio_ClinicalBERT	SE Simon Ellershaw	import	approved

Data governance-compliant 3rd Party LLM

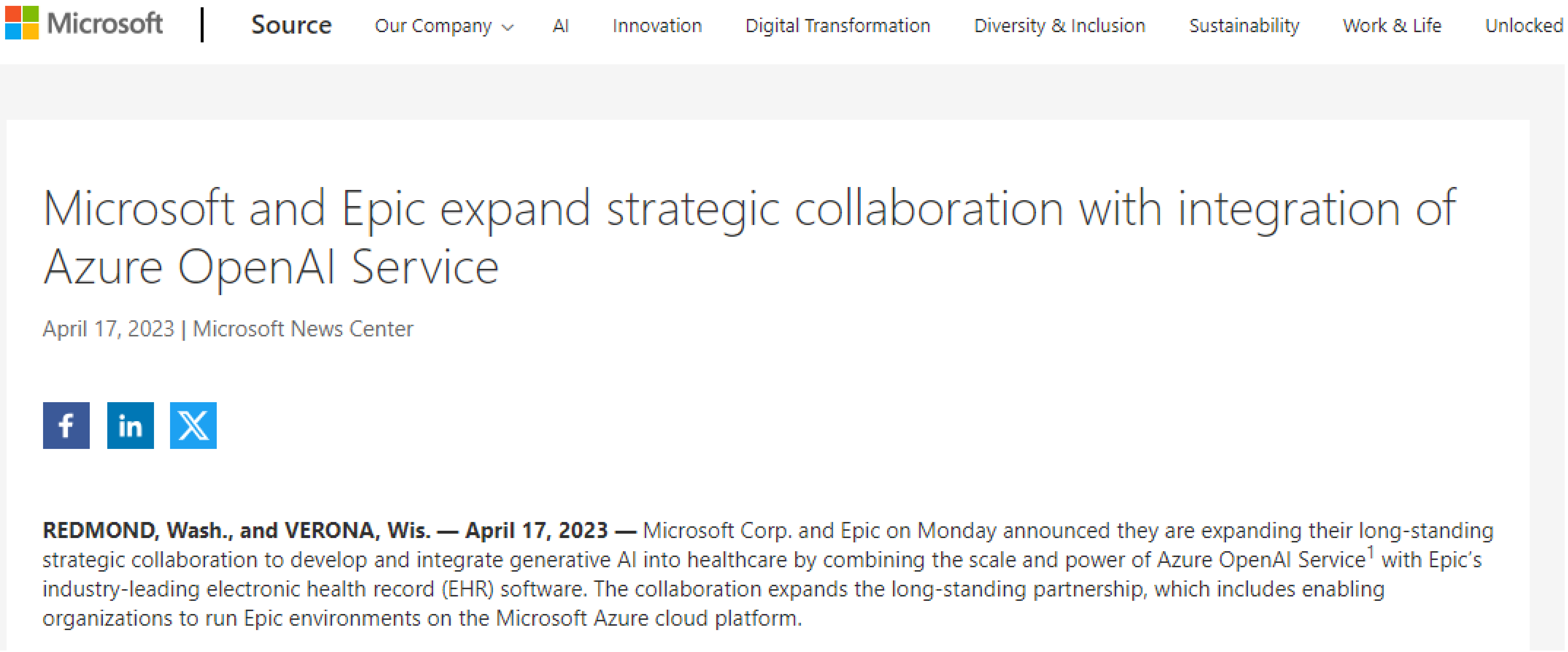


Azure OpenAI Service

Build your own copilot and generative AI applications

https://azure.microsoft.com/en-us/products/ai-services/openai-service

Data governance-compliant 3rd Party LLM



Microsoft and Epic expand strategic collaboration with integration of Azure OpenAI Service

April 17, 2023 | Microsoft News Center

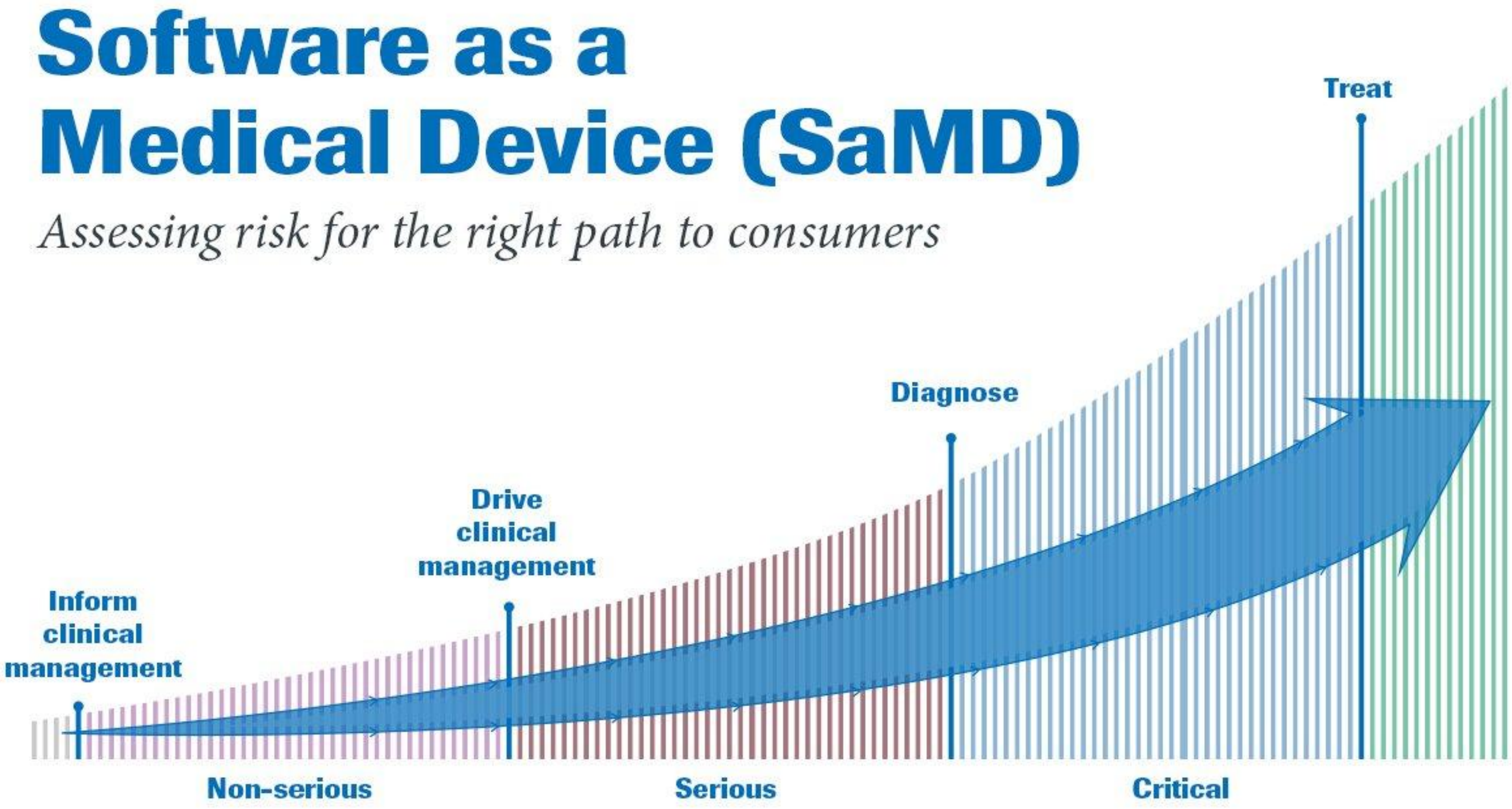
REDMOND, Wash., and VERONA, Wis. — April 17, 2023 — Microsoft Corp. and Epic on Monday announced they are expanding their long-standing strategic collaboration to develop and integrate generative AI into healthcare by combining the scale and power of Azure OpenAI Service[1] with Epic's industry-leading electronic health record (EHR) software. The collaboration expands the long-standing partnership, which includes enabling organizations to run Epic environments on the Microsoft Azure cloud platform.

https://news.microsoft.com/2023/04/17/microsoft-and-epic-expand-strategic-collaboration-with-integration-of-azure-openai-service/

Which one to use?

	Local LLMs (e.g. Llama 3.1 7B)	**3rd Party (e.g. GPT-4.1)**
LM Arena Ranking	70th	3rd
Context Window / tokens	~1000	128,000
Generation speed	Slow	Fast
Throughput	~4000 tokens per min	450,000 tokens per min 2700 request per min
Fixed Model	Yes	No
Virtual Machine Costs / hr	£7.50	£0.07
Inference cost / 1 million tokens	$0	Input- £2.00 Output- £8.00
Available for real time deployment	No	No

All lead to TBC regulation


Software as a
Medical Device (SaMD)
Assessing risk for the right path to consumers
Treat
Diagnose
Drive clinical management
Inform clinical management
Non-serious
Serious
Critical

Fixed model

Provable claims

https://www.roche.com/stories/value-of-digital-health-in-diagnostics

Conclusion

- ~Easy to produce compelling healthcare LLM PoC

But….

- How can you robustly test?
  - Human vs AI comparison
- Which LLM and how to deploy?
  - Open source locally deployed but $$$ and suboptimal performance
  - 3rd Party data governance “pending”
  - No real time access
- Regulation
  - TBC