Large Language Models For Patient Document Summarization

A case study in applying large language models for patient document summarization conducted at Sahlgrenska University Hospital.

Felix Nilsson & Albert Lund

Masters thesis @ Sahlgrenska University Hospital

Supervisor:
Bastiaan Bruinsma

Examiner:
Moa Johansson

Advisors:
Denitsa Saynova - CSE,
Isak Barbopoulos - VGR

CHALMERS
UNIVERSITY OF TECHNOLOGY

The Problem

Reading up on patient backgrounds is a routine part of a clinicians workday,
but it is time consuming.

- Swedish doctors has among the longest meetings with patients from an international context [1].
- 2 out of 3 Swedish doctors see their work as "incredibly stressful" [1].

[1] Vården ur primärvårdsläkarnas perspektiv - International Health Policy Survey (IHP) 2022

Our Approach

Use LLMs + modern machine learning techniques to summarize documents.

Some questions

How well does LLMs transfer to Swedish and specifically our tasks?

What is important? Parameter count, training or/and architecture?

Open source vs proprietary?

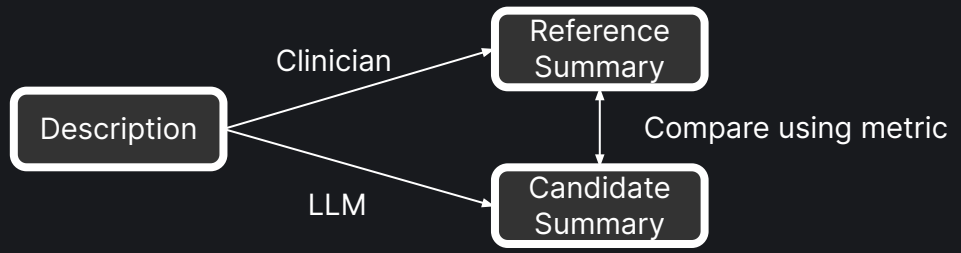
Steps

1. Developed evaluation suite
2. Tested promising models using suite, picked one
3. Prompt engineering + hyperparameter tuning
4. Applied finetuning techniques
5. Performed Expert Evaluation with Doctors

Step 1: Measuring Summarization Quality

Most metrics use a reference to compare to the generated candidate.

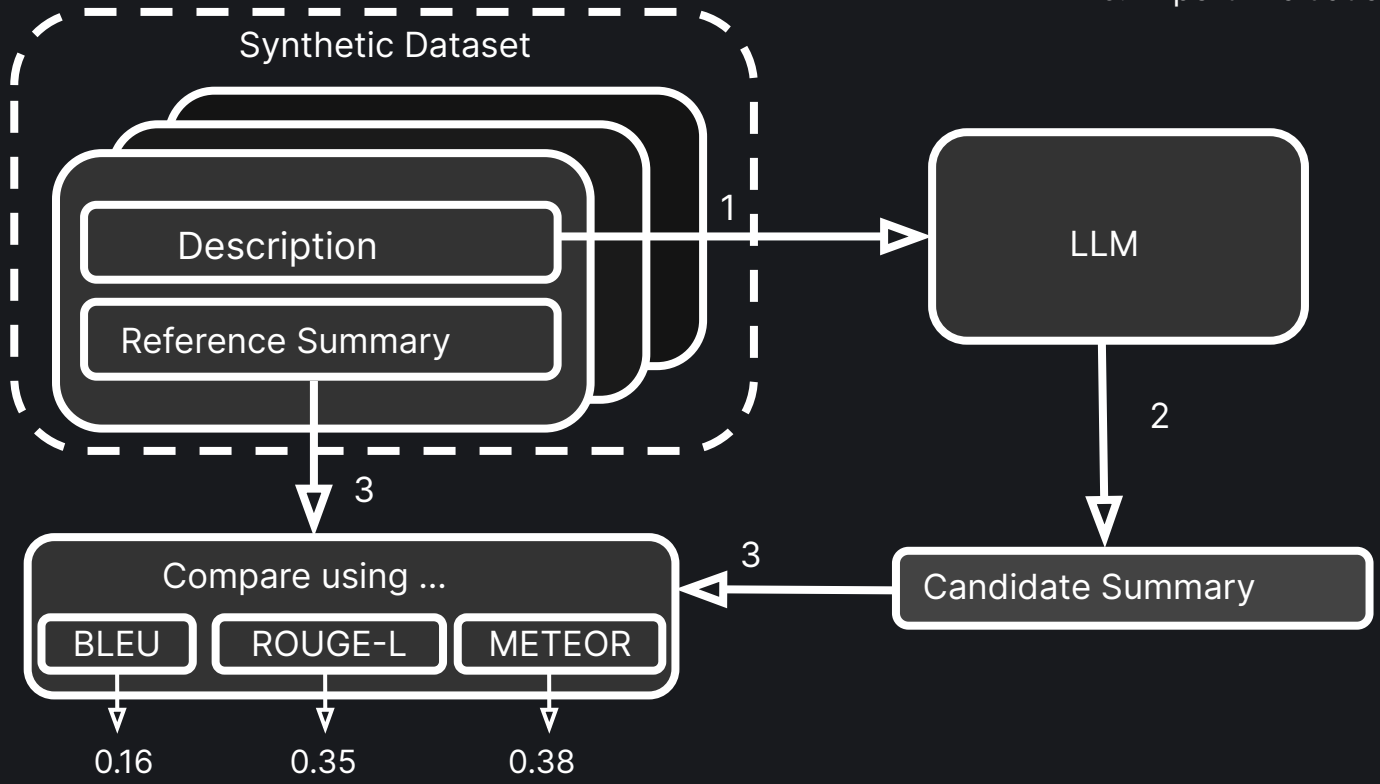
- Rule based:
  - BLEU
  - ROUGE-L
  - METEOR
- ML Model:
  - BERTScore



But what if the reference is bad? Can you exploit the metrics?

1. Automatic Evaluation
2. Model Selection
3. Prompt & Hyperparameters
4. Finetuning
5. Expert Evaluation

Step 1: Automatic Evaluation Suite



Step 1: Datapoint Example

1. Automatic Evaluation
2. Model Selection
3. Prompt & Hyperparameters
4. Finetuning
5. Expert Evaluation

Notes	Operationsberättelse 24-02-03 kl 0900, Läk Dan Frölund: Operationskod: AAF05 VP shunt frontalt höger Strata 1,5 Operation: 5 Sahlgrenska Preoperativ bedömning: 79 årig man med konstaterad NPH, ca 1 års anamnes på tilltagande balansbesvär, bredspårig gång, urininkontinens, närminnespåverkan med MMSE 22 poäng. Bedömd på behandlingskonferens med indikation för shuntoperation. Anestesiform: Generell anestesi och lokalanestesi Operationsberättelse: Semicirkulär incision över Kochers punkt höger. Ett borrhål. Går till buken och tar upp ett växelsnitt ner till peritoneum. Kommer otvetydigt in i fri bukhåla. Tunnelerar därefter upp till ett hjälpsnitt bakom höger öra. Tunnelerar även från skalpsnittet till hjälpsnittet. En förinställd Strataventil (1,5) och distalkateter dras igenom i tvåsteg till buken. Kryssformad durotomi och kortikotomi. För ner en ventrikelkateter (längd [] cm). Klart utbyte. Kopplas på distala systemet. Verifierar distalt utbyte. Matar ner ca 40 cm kateter i fri bukhåla. Syr sedan igen skalpsnittet med inverterad enstaka Vicryl och fortlöpande Etilon. Enstaka etilon över hjälpsnittet. Vicryl i yttre muskelfascian och subkutant på buken, Etilon i huden på buken. Post-operativa ordinationer: Suturtagning 12 dagar post-op. CT kontroll 3 h post-op, remiss skrivs. Ingen mer antibiotika.

Summary	Sjukdomshistoria (Patientens diagnoser, sjukdomshistorik och riskfaktorer (t.ex. sjukdomar i familjen)): o NPH Sökorsaker (patientens symtom och/eller datum för ingrepp): o Balansbesvär o Bredspårig gång o Urininkontinens o Närminnespåverkan o MMSE 22 o Shuntoperation 24-02-03 Åtgärder (planerade undersökningar, behandlingar och åtgärder) o Post-op: o Suturtagning 12 dagar o CT kontroll: 3 h (remiss skriven) o Ingen antibiotika

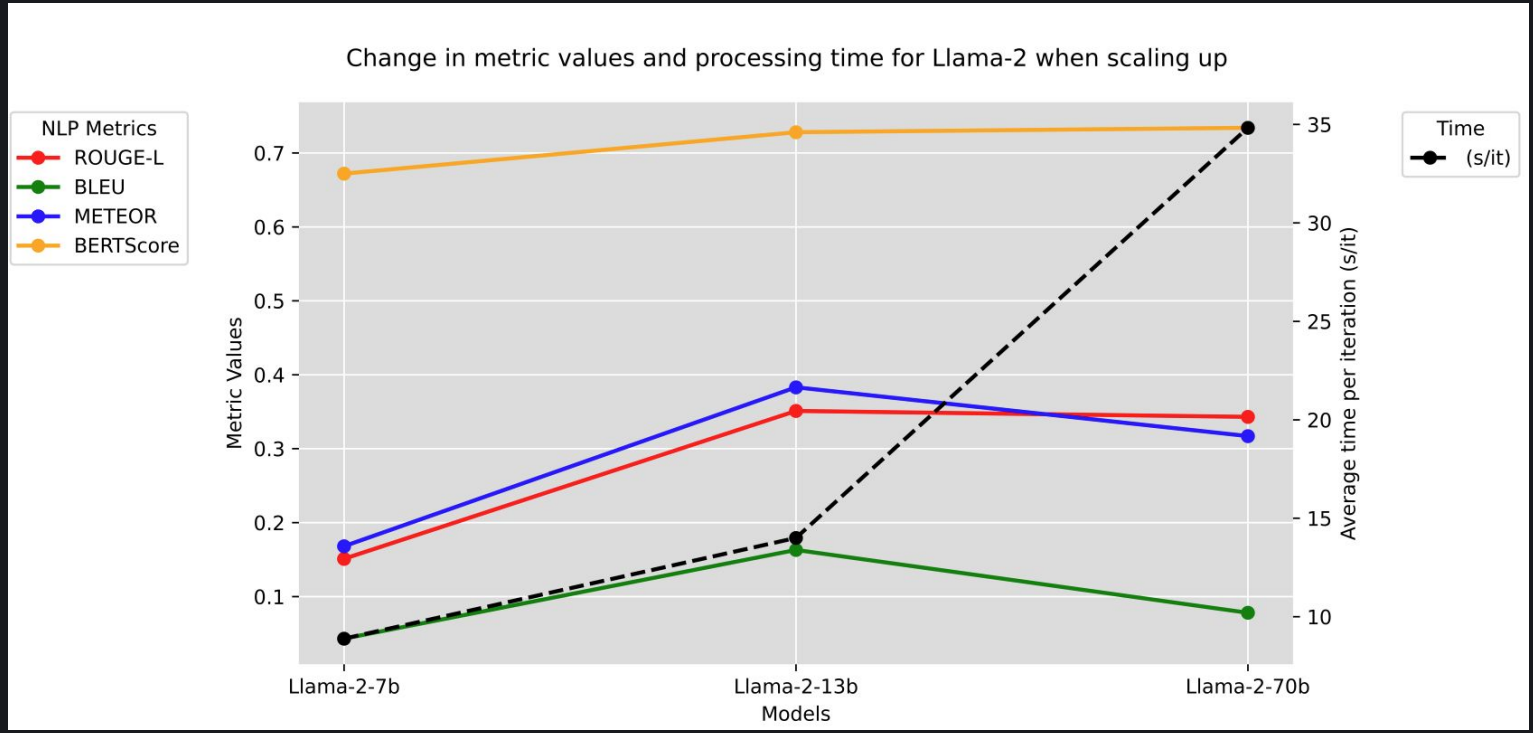
1. Automatic Evaluation
2. Model Selection
3. Prompt & Hyperparameters
4. Finetuning
5. Expert Evaluation

Step 2: Model Selection

Model	ROUGE-L	BLEU	METEOR
Open Source			
`meta-llama/Llama-2-13b-chat-hf`	**0.351**	**0.163**	**0.383**
`meta-llama/Llama-2-70b-chat-hf`	0.343	0.078	0.317
`malhajar/meditron-70b-chat`	0.138	0.032	0.122
`AI-Sweden-Models/gpt-sw3-20b-instruct`	0.301	0.102	0.291
`Falconsai/medical_summarization`	0.104	0.104	0.088
`mistralai/Mixtral-8x7B-Instruct-v0.1`	0.290	0.051	0.230
Proprietary			
`openai/gpt-4-0125-preview`	**0.515**	**0.266**	**0.488**
`openai/gpt-35-turbo-16k`	0.366	0.131	0.310
Rule Based			
`summa-textrank`	**0.171**	**0.058**	**0.130**

Step 2: Llama-2 Scaling Effects

1. Automatic Evaluation
2. Model Selection
3. Prompt & Hyperparameters
4. Finetuning
5. Expert Evaluation



Step 3: Prompt

1. Automatic Evaluation
2. Model Selection
3. Prompt & Hyperparameters
4. Finetuning
5. Expert Evaluation

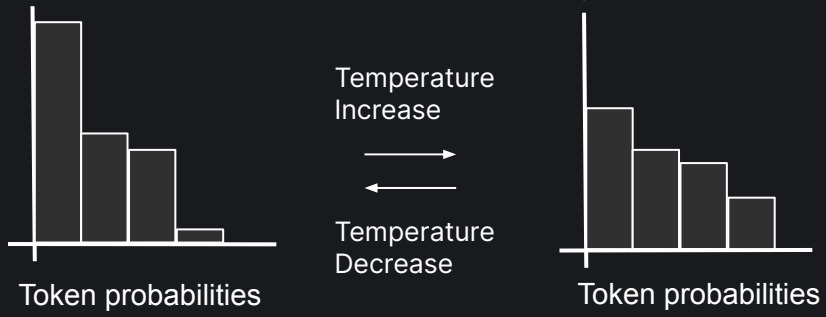
```
[INST] <<SYS>>
You are a helpful medical assistant who helps medical professionals by summarizing information about
patients.
Answer with a bulleted list for each category in the given template.
Answer in Swedish.
<</SYS>>

Below is the history of a patient during one day
<description>
*
</description>

You must select information that fits in the template below. Avoid unnecessary information and only pick out
things that relate to each heading. If relevant information is missing, leave the heading blank. Do not
include any information in multiple headings.

<template>
*Sjukdomshistoria (The patient's diagnoses, medical history and risk factors (e.g. diseases in the family))*
*Sökorsaker (Patient's symptoms and/or date of procedure)*
*Åtgärder (Planned examinations, treatments and measures)*
</template>

[/INST]
```

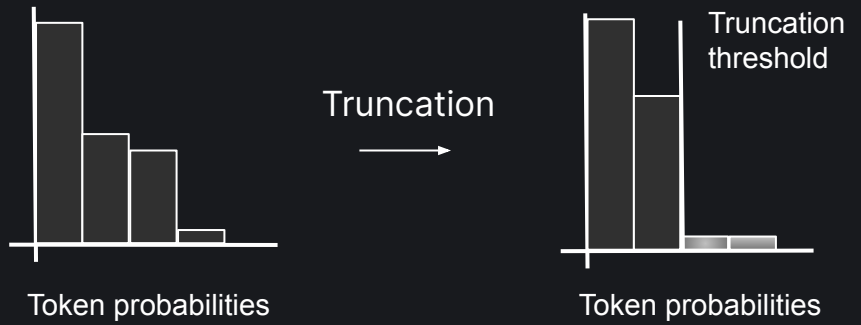
1. Automatic Evaluation
2. Model Selection
3. Prompt & Hyperparameters
4. Finetuning
5. Expert Evaluation

Step 3: Hyperparameters

Explored parameters to:

- Alter the distribution
  - Temperature
- Truncate the distribution
  - Top-k
  - Top-p
  - Local typicality
  - Conditional probability cutoff





1. Automatic Evaluation
2. Model Selection
3. Prompt & Hyperparameters
4. Finetuning
5. Expert Evaluation

Step 4: Training

- Domain adaptation
- Question and answers (QA)
- Finetuning

	Domain adaptation	Question Answering	Finetuning
Model 1	✓	✓	✓
Model 2	✗	✓	✓
Model 3	✗	✗	✓

1. Automatic Evaluation
2. Model Selection
3. Prompt & Hyperparameters
4. Finetuning
5. Expert Evaluation

Step 4: Domain adaptation

Continued pre-training

Deteriorates instruction-tuning

Klinik: En skada innan korsningen (på hjärnstamsnivå eller ovanför) bidrar till kontralaterala symptom medan en skada under korsningen (medulla spinalis) istället bidrar till ipsilaterala symptom. Baserat på vilket neuron som är skadat kan olika skador delas in i supranukleära (övre motorneuron) och nukleära/infranukleära (nedre motorneuron).
Supranukleär skada ger pares, hyperreflexi/klonus, Babinskis tecken, atrofi (sent och diskret) och spasticitet/ökad tonus.
Nukleär/infranukleär skada ger pares hypo- eller areflexi, ingen Babinski, atrofi (tidigt och omfattande) och nedsatt eller normal tonus.

Sensoriska systemet
Fyra modaliteter: Smärta (inkl. temperatur), beröring (inkl. tryck), proprioception och vibration
Klinisk betydande ledningsbanor: Tractus spinothalamicus lateralis (smärta och temperatur), bakströngsbanor/Lemniscus medialis (beröring/proprioception/vibration) och Tractus

1. Automatic Evaluation
2. Model Selection
3. Prompt & Hyperparameters
4. Finetuning
5. Expert Evaluation

Step 4: QA

Domain adaptation without loss of instructions

Difficult to construct from raw text. Used study material for medical doctors.

Question	Hur sker inandning i vila?
Answer	Kontraktion av diafragma och externa intercostalmuskler ökar lungans volym - ökning i lungans volym leder till en trycksäkning och luft kan komma in i lungan - lungan är elastisk (utspänd)

1. Automatic Evaluation
2. Model Selection
3. Prompt & Hyperparameters
4. Finetuning
5. Expert Evaluation

Step 4: Finetuning

Mimics the downstream task

Used synthetic description-summary pairs generated with GPT4

Step 4: Low-Rank Adaptation (LoRA)

1. Automatic Evaluation
2. Model Selection
3. Prompt & Hyperparameters
4. Finetuning
5. Expert Evaluation

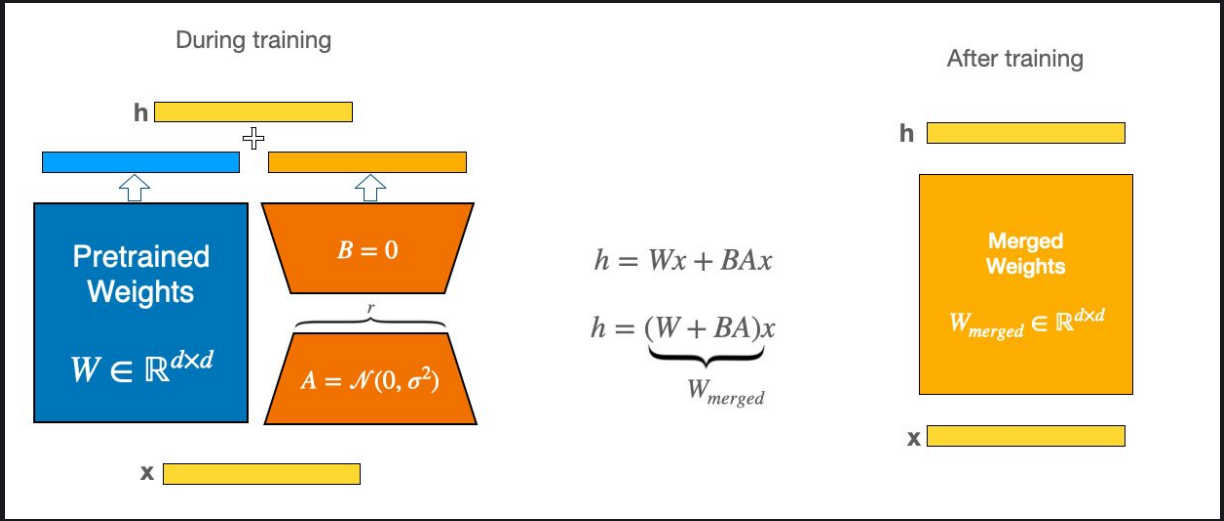


Image source: https://www.linkedin.com/pulse/lora-low-rank-adaptation-efficient-fine-tuning-large-language/

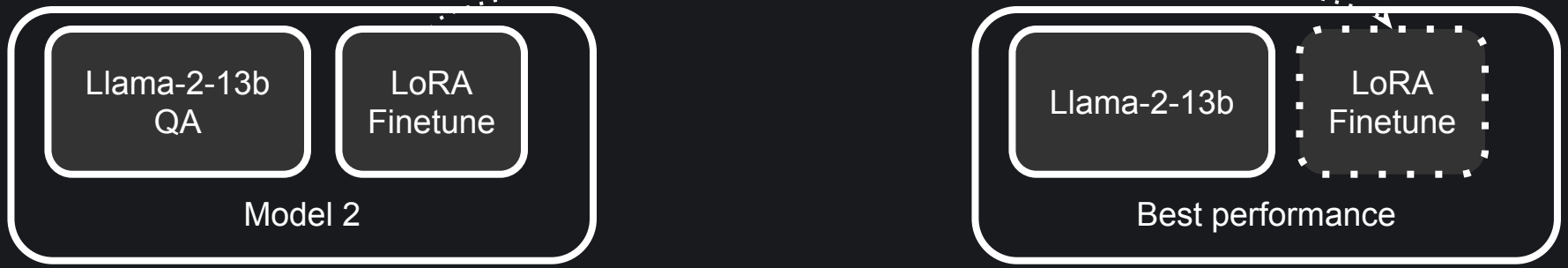
1. Automatic Evaluation
2. Model Selection
3. Prompt & Hyperparameters
4. Finetuning
5. Expert Evaluation

Step 4: Training

Model 3 performed best

Best results came from combination of two strategies:

- Applying LoRA weights from Model 2 to base model



Step 4: Finetuning Performance Gains

1. Automatic Evaluation
2. Model Selection
3. Prompt & Hyperparameters
4. Finetuning
5. Expert Evaluation

Model	ROUGE-L	BLEU	METEOR
`meta-llama/Llama-2-13b-chat-hf`	**0.351**	**0.163**	**0.383**
`Llama-2-13b-chat-hf` + Best Gen-Config.	0.392	0.195	0.414
Model 2 LoRA loaded on base model	0.398	0.206	0.431
Model 2 LoRA loaded on base model + Best config	**0.435**	**0.225**	**0.466**

1. Automatic Evaluation
2. Model Selection
3. Prompt & Hyperparameters
4. Finetuning
5. Expert Evaluation

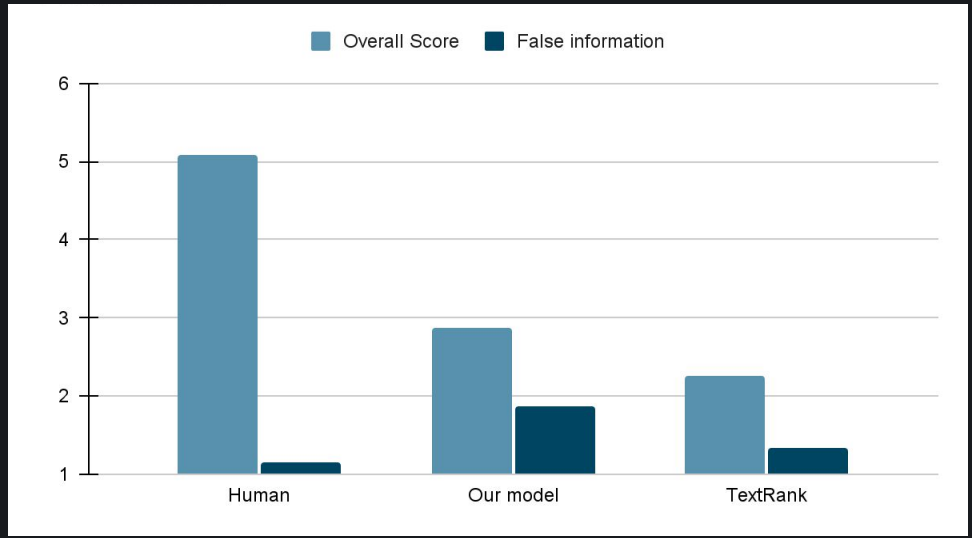
Step 5: Expert Evaluation

Ask clinicians to rate summarization performed by:

- Human
- Our best model
- A non LLM model

According to:

- Overall Score
- Helpfulness
- False information
- Relevancy
- Adherence to medical terminology



Ethics

An LLM may hallucinate when generating an answer.

Can their results be interpreted in a fair way?

Can they be treated like other advanced technologies? E.g. MRI

Conclusion

No need for largest model

Why does LoRA transfer?

LLMs will become even better

Thanks for listening