

Building trustworthy local AI for Healthcare at the Edge

Dr. Shahnawaz Ahmed
Deep Learning Researcher
shahnawaz@embedl.com

Empowering innovative and intelligent solutions at the edge

Outline

- AI in healthcare - envisioning a future
- Challenges: data protection, privacy, trust
- Efficient deep learning at the edge
- Talking to your personal AI health asst.
- Conclusion



SOMEONE SHOULD KEEP AN EYE ON HER.

NO, THAT'S JUST WHAT SHE WANTS!



XKCD

Randall Munroe

Chat GPT

OpenAI

Envisioning some use cases for AI in healthcare

AI in healthcare

GPT 4o Voice,
OpenAI

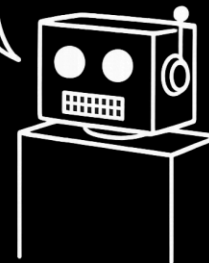
PATIENT

I NEED TO
RESCHEDULE MY
APPOINTMENT



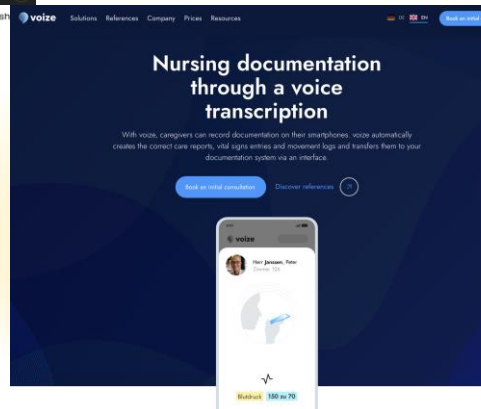
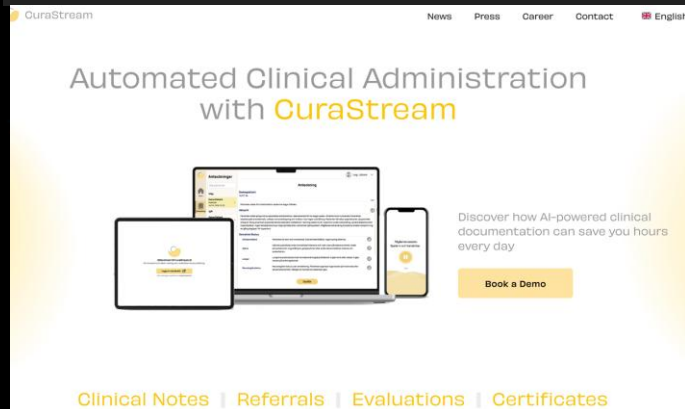
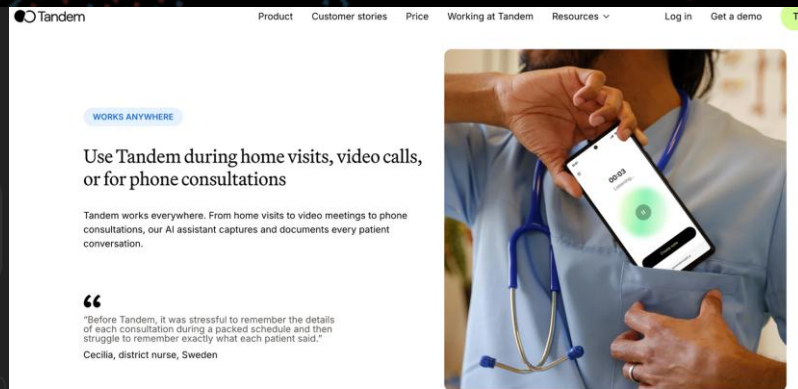
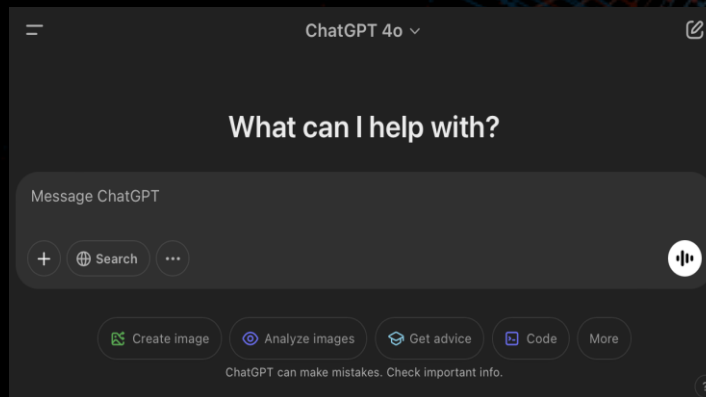
AI

CERTAINLY. WHAT DATE
AND TIME WORK FOR
YOU?



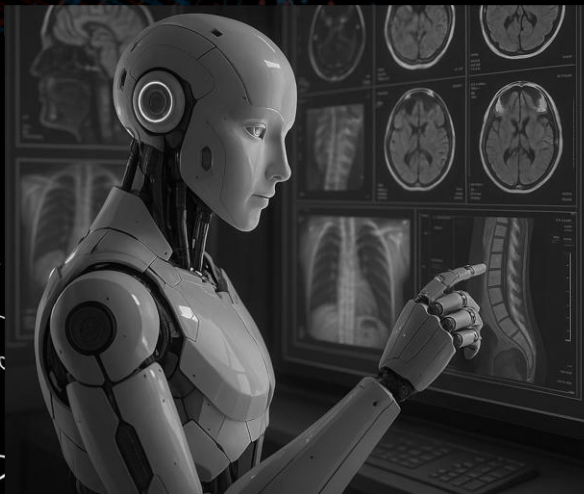
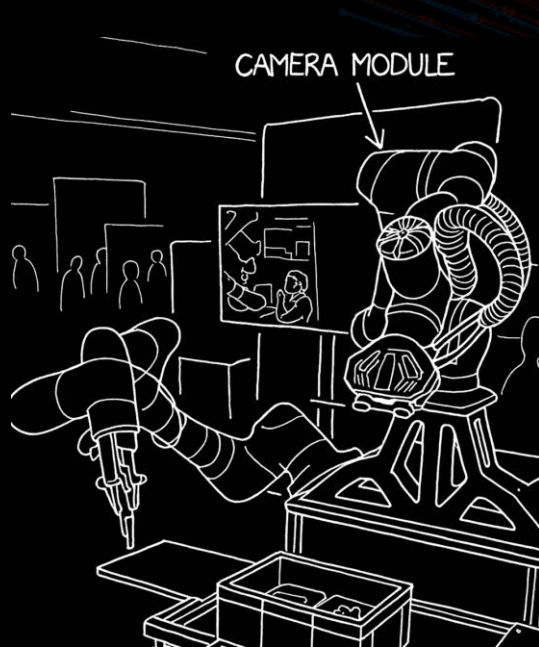
Question-answering, Transcription, Translation, Health records, planning

Innovative healthcare solutions

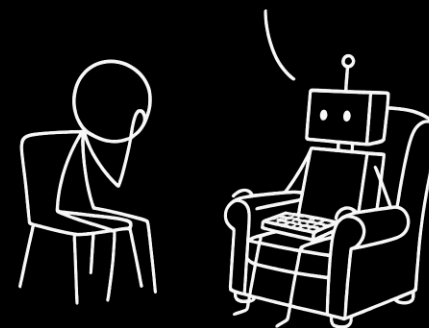


Robotics, diagnostics and image processing

Innovative healthcare solutions



AND HOW DOES THAT
MAKE YOU FEEL?



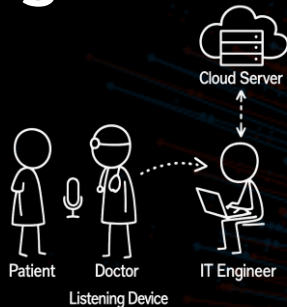
Online Dermatology Consultation by Board-Certified Dermatologists

- Fast and anonymous skin evaluations by board-certified dermatologists.

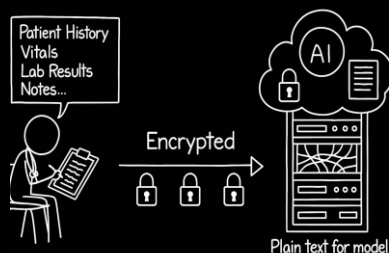
First Derm



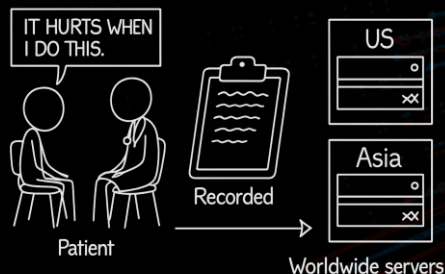
Challenges



✗ data are processed by or under the responsibility of a professional subject to the obligation of professional secrecy (Article 9, GDPR)



✗ data minimization, anonymization and other protections are by design - not an afterthought (Article 25, GDPR)



✗ Transfer of data to third country or international organizations (Article 44, GDPR)

Mistakes can cost lives - healthcare AI necessitates the highest quality

Challenges - trusting AI models in healthcare

- Mistakes in healthcare can cost lives
- Models need to be continuously monitored, tested, improved and re-deployed with medical professionals in the loop
- Sensitive healthcare data necessary for training, cannot be shared - custom model development has to happen locally



"Our patients are in danger, stop Millenium."
Source: RadioSweden

Private, local and trustworthy

Edge AI for healthcare

Privacy by design



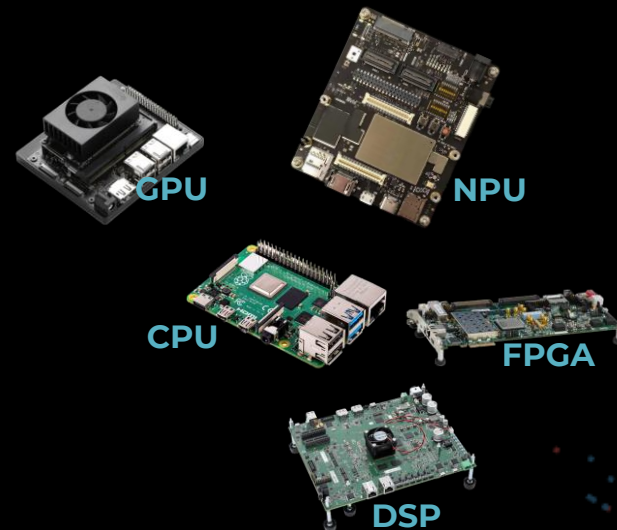
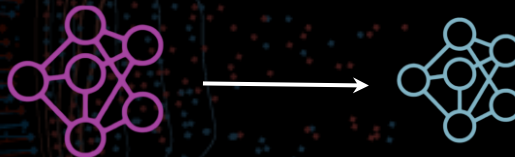
Trustworthy and accurate



Local and secure



embedl



You are in full control of your model, data and applications

Local AI enables healthcare applications

- Do we need trillion parameter black-box models for everything ?
- Easy adaptations, model development and sensitive data processing needs AI on the edge
- Local AI is possible and affordable
- Embedl - right size your model for the right hardware and use



GPT (4o)
\$25 / month / user
?B parameter model?



Nvidia

Orin Nano Super
\$249 + 25 W (electricity)
3-7B parameter models

Building AI for healthcare at the edge is possible

Talking to a personal AI health assistant



```
embedl@orin-nano-sunnyvale:~$ python run.py \  
--prompt "What does a +2 leukocyte count mean?" "" \  
--model "llama-3b"
```

Building AI for healthcare at the edge is possible

Talking to a personal AI health assistant



```
embedl@orin-nano-sunnyvale:~$ python run.py \
--prompt "What does a +2 leukocyte count mean?" "" \
--model "llama-3b"
A leukocyte count of 2+ typically indicates a very mild elevation of white blood cells in the blood.
```

In most cases, a normal leukocyte count ranges from 4,000 to 11,000 cells per microliter (μL) of blood.

A count of 2+ would be considered within the normal range or very slightly elevated, which might not necessarily be cause for concern. However, it's essential to consider the context and any other symptoms or test results that may be available.

To better understand the significance of a 2+ leukocyte count, I would need more information about the patient's medical history, symptoms, and other test results. If you have any additional context or details, I'd be happy to try and provide a more informed response.

```
embedl@orin-nano-sunnyvale:~$
```


Building AI for healthcare at the edge is possible

Talking to a personal AI health assistant



```
embedl@orin-nano-sunnyvale:~$ python run.py \  
--prompt "What does a +2 leukocyte count mean?" "" \  
--model "qwen-7b"
```

Building AI for healthcare at the edge is possible

Talking to a personal AI health assistant



```
embedl@orin-nano-sunnyvale:~$ python run.py \  
--prompt "What does a +2 leukocyte count mean?" "" \  
--model "qwen-7b"
```

A "2+" leukocyte count in a urinalysis typically indicates a moderate presence of white blood cells (leukocytes) in the urine.

This result suggests that there might be some level of urinary tract infection (UTI), inflammation, or irritation in the urinary system. In a standard dipstick test used to screen for urinary abnormalities, the "2+" symbol usually represents a moderately positive result for leukocytes. The exact interpretation can vary slightly depending on the manufacturer of the test strip, but generally:

A "+" indicates a trace amount.
A "2+" indicates a moderate amount.
A "3+" indicates a large amount.

It's important to note that while a "2+" leukocyte count may indicate an issue, it is not a definitive diagnosis. Further testing, such as a microscopic examination of the urine sediment or a culture, may be necessary to confirm the presence of an infection or other conditions. If you have this result, it's advisable to consult with a healthcare provider for further evaluation and appropriate treatment if needed.

```
embedl@orin-nano-sunnyvale:~$ █
```



Optimize with
Context

Automate
Bottleneck Search

Profile, Don't
Assume!

Use-Case
Anchored Design

Optimize for the silicon, not just the server -
Hardware-Aware Neural Architecture Search,
Pruning, Mixed Precision Quantization

Automate analysis to **uncover latency and
accuracy bottlenecks**

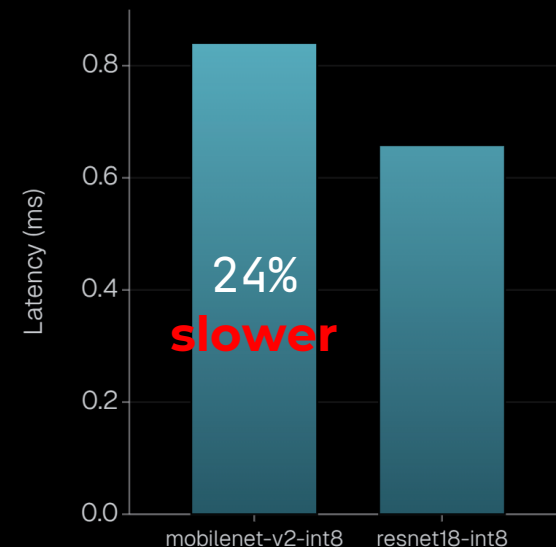
Measure on real hardware with real data early
on!
Ensure **compatibility** - what runs on a GPU may
fail on a DSP

Design backwards from real-world constraints -
latency, power, and task define the architecture

ResNet18 vs MobileNetV2

	FLOPs (params)	Architecture Implications
ResNet18	1.8G (11.7M)	Dense convolutions → Large parallelizable matrix multiplications
MobileNetV2	300M (3.4M) smaller!	Depthwise separable convs → Many small sequential computations

Table 1: Comparison of ResNet18 and MobileNetV2.

NVIDIA Jetson Orin Nano
(GPU)

Hardware-Aware Optimization

embedl + ∞ Meta

LLAMA 3
8B, 70B

LLAMA 3.1
8B, 70B, 405B

Compressed LLAMA 3.2

1B, 3B

Qualcomm friendly!



Qualcomm

Snapdragon

8
Gen 3



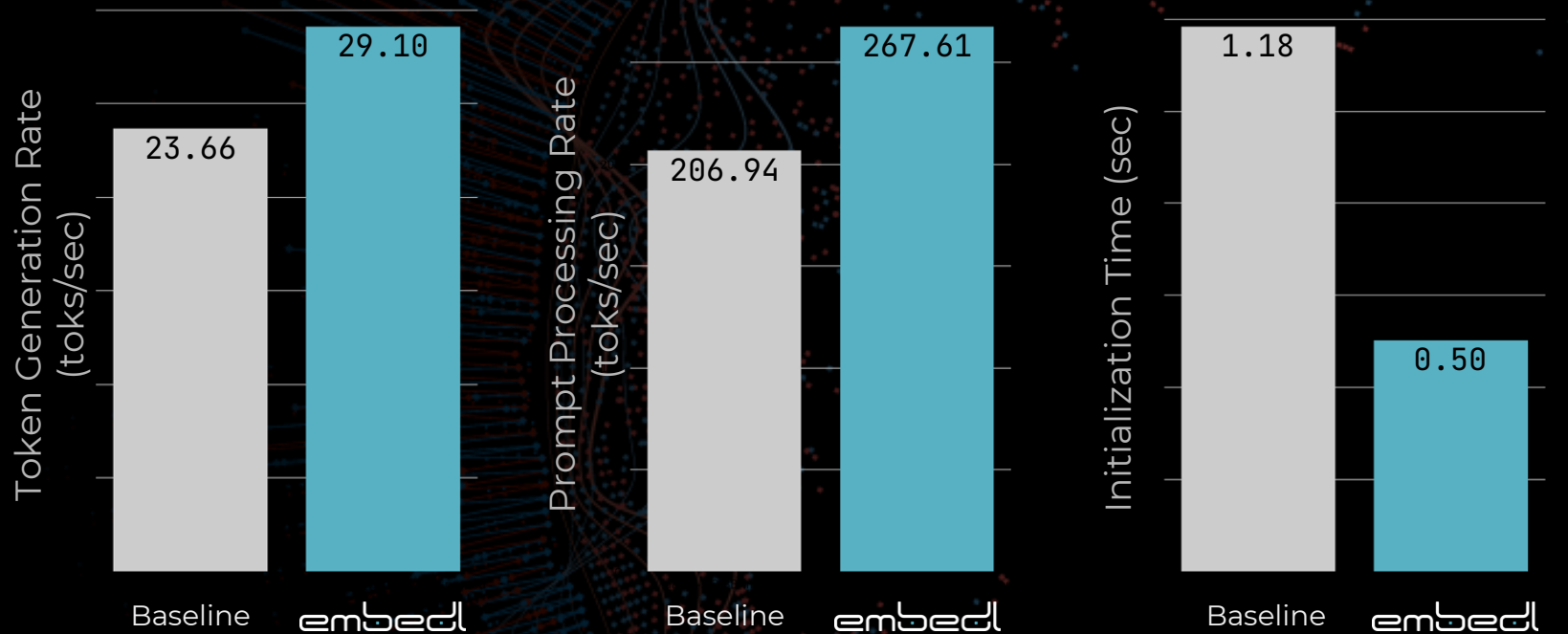
SM8650 Mobile



Method: **Hardware-Aware Pruning**, no fine-tuning or retraining after compression

Method: width pruning, no fine-tuning or retraining after compression

Hardware-Aware Optimization of Llama 3.2 for Qualcomm



Upto*

31% faster token generation

25% faster prompt processing

126% faster initialization

*within ~1%-point accuracy drop on Language model evaluation tasks

Conclusion

- AI can reduce workload for clinicians and lead to new innovative solutions
- Sensitive data must stay local - it is possible to compress models and finetune them for Healthcare applications at the Edge
- Local AI enables private, secure training and adaptations
- Needs open, robust tools for continuous improvement and use in practice



Dr. Shahnawaz Ahmed

Deep Learning Researcher
shahnawaz@embedl.com



About

embed

EFFICIENT DEEP LEARNING



25 people at HQ in Gothenburg



R&D team of 17 engineers and researchers, and growing



Strong connection to Chalmers University of Technology